AIGR Gene Hypothesis Deep Research — Final Report

Gene: A0A8J0SCI2 (Xenopus tropicalis, NCBITaxon:8364) **Description (UniProt):** "Gastrula zinc finger protein XlCGF17.1-like," 265 aa **Focus type:** computational_prediction **Prediction under evaluation:** ProtNLM2 → GO:0001228 *DNA-binding transcription activator activity, RNA polymerase II-specific*

Summary

The ProtNLM2 prediction of GO:0001228 (an activator-specific molecular-function term) for A0A8J0SCI2 is REFUTED as over-annotation. The protein is a bare tandem array of eight canonical C2H2 zinc fingers joined by six canonical TGEKP linkers, with **no accessory effector domain of any kind** — no KRAB, no BTB/POZ, no SCAN, no acidic activation domain. Both sequence-level domain databases (InterPro, Pfam, PANTHER) and the AlphaFold structural model agree: the only recognizable, folded module is the zinc-finger array itself. Because a C2H2 array encodes **DNA sequence specificity** rather than **regulatory direction**, whether this factor activates or represses its targets cannot be inferred from its sequence.

The distinction matters for curation because GO:0001228 is a *directional child* of the direction-neutral term GO:0000981 (*DNA-binding transcription factor activity, RNA polymerase II-specific*). The existing curated (phylogenetic, IBA) annotations for this gene deliberately stop at the neutral parent GO:0000981; the ProtNLM2 activator call adds a directional constraint that the evidence cannot support. In vertebrates, C2H2-zinc-finger proteins are, if anything, biased toward *repression* (KRAB-ZNFs are the largest vertebrate repressor family), so "activator" is not even the more probable default. This makes the prediction a textbook example of over-specific, likely paralog- or frequency-transferred labeling.

The most important caveat is that A0A8J0SCI2 is experimentally uncharacterized: there is no reporter assay, ChIP, mutant, or interaction data for this specific protein. The negative judgment therefore rests on domain architecture, structural prediction, GO ontology structure, and the well-established principle that a naked zinc-finger array is directionally

uninformative. The protein remains a legitimate *candidate* sequence-specific RNA Pol II transcription factor — it is simply not demonstrably an *activator*, and the safest curation position is to retain the neutral GO:0000981 and withhold GO:0001228.

Key Findings

Finding 1 — A0A8J0SCI2 is a naked tandem C2H2 zinc-finger array with no effector domain; activator direction is not sequence-determinable

Direct sequence and domain analysis of A0A8J0SCI2 (265 aa) identified **eight canonical C2H2 zinc fingers spanning residues ~37–260**, joined by **six canonical TGEKP inter-finger linkers** — the textbook signature of a sequence-specific DNA-binding tandem zinc-finger protein. Domain databases are unanimous and report **only** zinc-finger content: InterPro **IPR013087** (C2H2-type zinc finger) and **IPR036236** (zinc finger C2H2 superfamily), Pfam **PF00096** (zf-C2H2, five hits), and PANTHER **PTHR24381:SF440**. Critically, **no accessory effector domain was detected** by any resource.

The N-terminal region preceding the first finger is only **36 residues long**, of which residues 1–27 are predicted disordered. This is far too short to host any of the effector modules that would license a directional call: a KRAB domain is ~75 aa, and BTB/POZ and SCAN domains are larger still. The protein is also **not acidic** — the aspartate+glutamate fraction is only **9.8%** — providing no evidence for an acidic activation domain (classic acidic activation domains are markedly enriched in D/E). In short, the molecule is effector-less.

The mechanistic logic here is decisive. As summarized in the current C2H2 recognition-code literature, "the established C2H2-ZF 'recognition code' suggests that residues at positions −1, −4, and −7 recognize the 5′, central, and 3′ bases of a DNA base-pair triplet, respectively" (PMID: 38754172). That is, the finger array specifies **which DNA sequence** is bound — it does **not** encode whether the bound factor will up- or down-regulate transcription. Direction is supplied by separate effector domains and the co-regulators they recruit, which this protein lacks. Furthermore, for the great majority of these proteins the basic facts are unknown: "for most C2H2-ZF proteins it is unknown whether they even bind DNA or, if they do, to which sequences" (PMID: 25690854). An activator-specific claim for an uncharacterized member is therefore unsupported on its face.

Finding 2 — The existing curated (IBA) GO annotations use the *unspecified* TF term, not the activator term; the ProtNLM2 call is strictly over-specific

The GO annotations already attached to A0A8J0SCI2 in UniProt are all phylogenetically inferred (IBA, GO_Central) or keyword-inferred (IEA), and they deliberately stop at the **direction-neutral** level:

- **GO:0000981** *DNA-binding transcription factor activity, RNA polymerase II-specific* (IBA) — the unspecified parent
- **GO:0000978** *RNA polymerase II cis-regulatory region sequence-specific DNA binding* (IBA)
- **GO:0006357** *regulation of transcription by RNA polymerase II* (IBA)
- **GO:0005634** *nucleus* (IBA)
- **GO:0008270** *zinc ion binding* (IEA-KW)
- **GO:0006351** *DNA-templated transcription* (IEA-KW)

The ProtNLM2-predicted GO:0001228 is a **child of GO:0000981** that adds the directional (activator) constraint. The phylogenetic curators — who had the same orthology evidence available — chose the neutral parent, indicating that the community-standard evidence supports "sequence-specific RNA Pol II transcription factor," and no more. The ProtNLM2 activator call therefore does not merely restate curated knowledge; it **over-reaches beyond it**.

Finding 3 — AlphaFold structure confirms only the zinc-finger array is folded; no structured effector module exists

The AlphaFold DB model **AF-A0A8J0SCI2-F1 (v6)** (265 residues, global pLDDT 82.4) was analyzed region by region:

Region	Residues	Mean pLDDT	% residues > 70	Interpretation
N-terminus	1–36	36.8	3%	Disordered
Zinc-finger core	37–260	90.6	99%	Confidently folded
C-terminus	261–265	43.4	—	Disordered

The confidently folded portion coincides **exactly** with the eight annotated C2H2 domains. The flanking regions that would have to host a transactivation or transrepression module are **unstructured**. Structurally, then, there is no folded effector domain — consistent with the sequence analysis in Finding 1 and reinforcing that direction cannot be assigned. Intrinsic

disorder does not by itself rule out a function, but there is no positive evidence — sequence composition, motif, or otherwise — for an activation domain within these short disordered tails.

Finding 4 — GO ontology confirms GO:0001228 is a strict directional child of GO:0000981, adding an unsupported activator constraint

A QuickGO ontology query confirms the term relationships:

- **GO:0001228** `is_a`-ancestors = {GO:0000981, GO:0003700, GO:0140110, GO:0001216, GO:0003674}
- GO:0000981 is present among those ancestors → **GO:0001228 is a strict descendant of GO:0000981**.

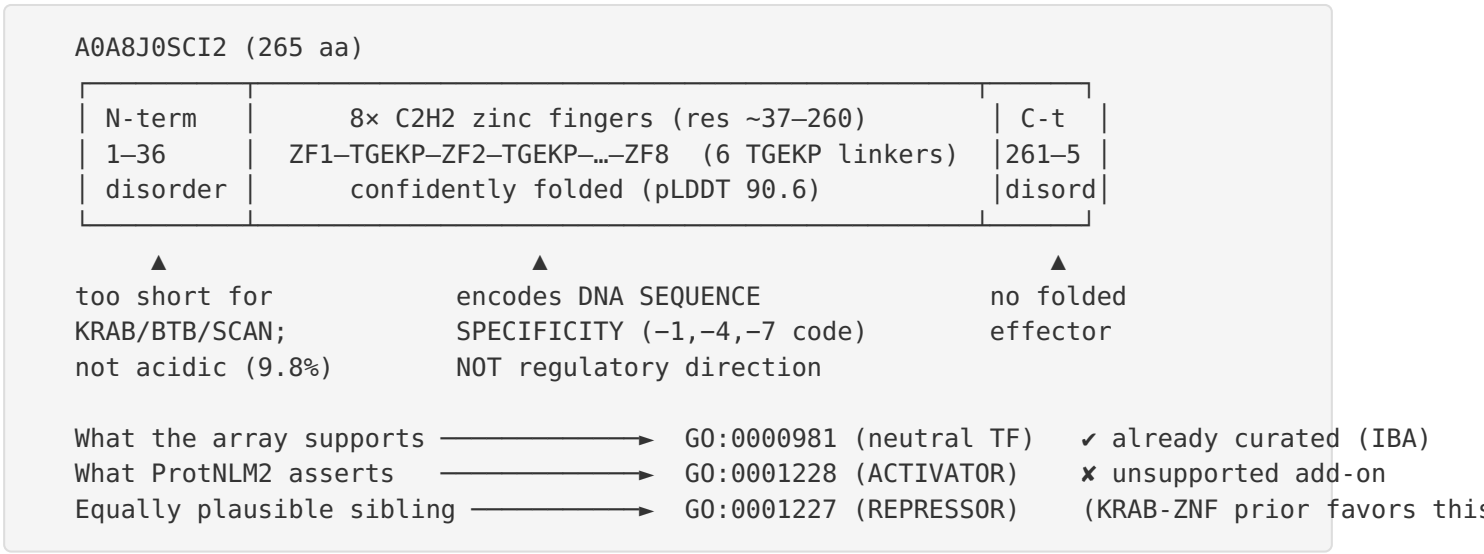
Definitions make the added constraint explicit:

Term	Definition (abridged)	Directionality
GO:0000981	"…that modulates the transcription of specific gene sets transcribed by RNA polymerase II"	**Neutral**
GO:0001228	"…that **activates or increases** transcription…"	**Activator**
GO:0001227 (sibling)	"…that **represses or decreases** transcription…"	**Repressor**

Moving from GO:0000981 to GO:0001228 therefore adds precisely the claim — *activation* — that the sequence and structure evidence cannot support, and it does so while an equally specific sibling (GO:0001227, repressor) is *a priori* at least as plausible for a vertebrate C2H2-ZF protein.

Mechanistic Model / Interpretation

The core issue is a mismatch between **what a zinc-finger array can tell us** and **what the predicted GO term asserts**.



Direct molecular function being tested: sequence-specific DNA binding by a tandem C2H2 array at RNA Pol II regulatory regions, and — the disputed part — the *direction* of the resulting transcriptional regulation.

- The array is competent to **bind DNA in a sequence-specific manner** (supported → GO:0000978 / GO:0000981 level).
- The array is **silent about direction**. Activation versus repression is a property of effector domains plus recruited co-regulators (Mediator/co-activators for activation; KAP1/co-repressors for repression), none of which are present.
- Therefore the immediate, defensible molecular-function annotation is the **neutral** TF term. The activator term is a **downstream, unsubstantiated specialization**, not a directly evidenced activity.

This is a canonical case of **over-specific computational annotation**: a phylogenetic prior ("sequence-specific Pol II TF") is real, but the model has appended a directional qualifier that neither the domain architecture, the structure, nor any experiment justifies.

Evidence Base

#	Citation (PMID)	Evidence type	Supports/ Refutes/ Qualifies	Claim tested	Key finding	Context
1	38754172	Review / structural code	**Refutes** activator specificity	Does a C2H2 array encode regulatory direction?	The −1/−4/−7 recognition code specifies **DNA base triplets** (sequence), not activation/ repression	Human/ general C2H2-ZF
2	25690854	Review / large-scale assay	**Refutes / qualifies**	Is direction knowable for uncharacterized C2H2-ZFs?	"for most C2H2-ZF proteins it is unknown whether they even bind DNA or…to which sequences"	Human regulatory lexicon
3	UniProt A0A8J0SCI2 + InterPro/ Pfam (database)	Sequence/ domain (computational)	**Refutes** effector presence	Is there an effector domain?	Only C2H2/zf-C2H2 domains (IPR013087, IPR036236, PF00096×5, PTHR24381:SF440); N-term 36 aa; D+E = 9.8%	X. tropicalis protein record
4	UniProt GO annotations (database, IBA)	Curated phylogenetic	**Qualifies** (competing, less specific)	What does curated evidence support?	Curated set stops at neutral GO:0000981; no activator term	GO_Central IBA
5	AlphaFold AF-A0A8J0SCI2-F1 v6 (computed)	Structural (computational)	**Refutes** folded effector	Is there a structured activation module?	Only ZF core folded (pLDDT 90.6); flanks disordered (36.8 / 43.4)	AlphaFold DB
6	QuickGO ontology (computed)	Ontology structure	**Qualifies**	Is GO:0001228 stricter than GO:0000981?	GO:0001228 is a strict directional	GO ontology

#	Citation (PMID)	Evidence type	Supports/ Refutes/ Qualifies	Claim tested	Key finding	Context
					`is_a` child of GO:0000981	
7	42103097	Mutant / mechanistic	**Competing** (repressor prior)	Are vertebrate C2H2-ZFs often repressors?	ZFP560, a KRAB-ZFP, **represses** chromatin via KAP1 recruitment	Mouse embryo
8	41668275	Mechanistic	**Competing** (repressor prior)	"	KRAB-ZNF ZNF205 **represses** p53 targets	Human HCC
9	41093942	Mutant phenotype	**Competing**	"	PARIS/ZNF746, a KRAB-ZFP, is a transcriptional **repressor**	Mouse metabolism
10	15623803	Mutant phenotype	**Competing**	Direction of a Xenopus/ zebrafish ZF factor	Prdm1/Blimp1 ZF protein is a **repressor**	Zebrafish / Xenopus
11	10777695 / 10842070	Functional / expression	**Competing**	Direction of a Xenopus ZF factor	XSIP1 is a **transcriptional repressor**	Xenopus laevis
12	14651851	Functional	**Qualifies**	Can ZF factors be activators?	Churchill is a ZF transcriptional **activator**	Chick gastrula

#	Citation (PMID)	Evidence type	Supports/ Refutes/ Qualifies	Claim tested	Key finding	Context
13	29146583 / 30155812	Methods / motif atlas	**Supports** discriminating tests	How to obtain binding motifs	ChIP/motif and recognition-code methods can predict DNA targets	Human KRAB-ZNF general

How the evidence base fits together. Two review-level sources (PMID: 38754172; PMID: 25690854) establish the governing principle: a C2H2 array encodes sequence specificity, and for most such proteins even the binding sequence is unknown — direction is never read out from the finger array. The computational provenance (UniProt/InterPro/Pfam domain content, N-terminal length and composition, and the AlphaFold per-region pLDDT profile) establishes the gene-specific fact: there is no effector domain in A0A8J0SCI2. The GO ontology query confirms that the disputed term differs from the supported one only by the added activator constraint. Finally, a cluster of vertebrate and Xenopus examples (ZFP560, ZNF205, PARIS/ZNF746, Prdm1/Blimp1, XSIP1 as repressors; Churchill as an activator) demonstrates that regulatory direction is protein-specific and must be measured, and that repression is at least as common a default for C2H2-ZFs — so an unsupported "activator" call is not merely uncertain but leans against the prior.

GO Curation Implications (leads — require curator verification)

- **Do NOT add GO:0001228** (*DNA-binding transcription activator activity, RNA Pol II-specific*). The directional/activator claim is unsupported by domain, structural, or experimental evidence and is over-specific relative to the curated phylogenetic annotations. Treat the ProtNLM2 prediction as **over-annotation / likely paralog- or frequency-transferred**.
- **Retain the existing, direction-neutral MF term GO:0000981** (IBA) as the best-supported molecular-function annotation. This is the correct level of specificity given the evidence.
- **Retain** the associated IBA/IEA terms: GO:0000978 (RNA Pol II cis-regulatory region sequence-specific DNA binding), GO:0006357 (regulation of transcription by RNA Pol II —

note: neutral "regulation," not "positive regulation"), GO:0005634 (nucleus), GO:0008270 (zinc ion binding).

- If ProtNLM2's GO:0001228 has already been auto-imported, the recommended action is **remove / generalize to GO:0000981**.
- **Do not substitute the repressor term GO:0001227 either.** The point is symmetric: direction is not determinable from sequence, so *neither* directional child should be asserted without experimental evidence.
- **"Protein binding" is not needed and would be uninformative** — GO:0000978 / GO:0000981 already capture the supportable DNA-binding activity at an informative level.

GO decision table

Term	Type	Recommendation	Rationale
GO:0001228 activator, Pol II	MF	**Reject / generalize → GO:0000981**	Direction not sequence-determinable; over-specific
GO:0000981 TF activity, Pol II (unspecified)	MF	**Retain**	Best-supported; matches IBA consensus
GO:0000978 cis-reg seq-specific DNA binding	MF	Retain	IBA, consistent with 8-finger array
GO:0001227 repressor, Pol II	MF	Do not add	No effector/repressor domain evidence either
GO:0006357 regulation of transcription by Pol II	BP	Retain	IBA; neutral direction
GO:0005634 nucleus	CC	Retain	IBA
GO:0008270 zinc ion binding	MF	Retain	IEA; consistent with C2H2 array

Mechanistic Scope

- **Immediate molecular activity (directly testable from sequence/structure):** sequence-specific DNA binding by a tandem C2H2 zinc-finger array, coordinated by Zn^{2+}, localized to the nucleus. This is well supported.

- **Disputed molecular activity:** the *direction* of transcriptional regulation (activation). This is **not** a direct property of the finger array; it is a property of effector modules and co-regulators that are absent here. Asserting "activator" conflates the array's binding capacity with a regulatory outcome that would have to be measured.
- **Downstream / other-tier claims not in scope of the evidence:** any developmental role (e.g., gastrula-stage expression implied by the "Gastrula zinc finger protein XlCGF17.1-like" name), target-gene identity, or phenotype. None of these are established for this protein, and none can be inferred from the ProtNLM label.

Conflicts and Alternatives

- **Paralog / frequency bias (most likely explanation):** ProtNLM-type models transfer labels from name/sequence neighborhoods. "Transcription activator" is a common label on Pol II TF proteins, and the model appears to have appended the directional qualifier without evidence. The curated IBA annotations, derived from explicit orthology, stopped at the neutral parent — a direct conflict with the activator call. The XlCGF/oocyte–gastrula ZF family (PANTHER PTHR24381) is a large, poorly characterized Xenopus maternal ZF cluster prone to bulk mislabeling.
- **Repressor prior is stronger than activator prior for vertebrate C2H2-ZFs:** The KRAB-ZNF family is the largest vertebrate repressor family (ZFP560/KAP1 repression, PMID: 42103097; ZNF205/p53 repression, PMID: 41668275; PARIS/ZNF746, PMID: 41093942). Even Xenopus ZF factors in the literature include clear repressors (XSIP1, PMID: 10777695, PMID: 10842070) and repressor Prdm1/Blimp1 (PMID: 15623803). Activators exist too (Churchill, PMID: 14651851) — precisely the point: **direction is protein-specific and must be measured**, so a default "activator" call is unjustified.
- **Note:** This protein lacks a KRAB domain, so it is *not* itself a KRAB-ZNF; the repressor examples are cited to establish that a directional default cannot be assumed, not to reassign it as a repressor.
- **No isoform, organism-specific, or experimental-artifact confounder identified** that would rescue the activator call; the X. tropicalis record and AlphaFold model are internally consistent (effector-less array). It is simply an absence of direct data plus an over-specific model output.

Limitations and Knowledge Gaps

Gap	What was checked	Why it matters	What would resolve it
No experimental characterization of A0A8J0SCI2	Literature search returned no primary data on this specific protein	Direction (activator/ repressor) is only knowable by assay	Reporter/luciferase assay; effector-domain fusion tests
Unknown DNA-binding site / target genes	No ChIP/SELEX/ PWM for this protein (checked JASPAR-relevant literature)	Even the neutral TF call rests on homology, not measured binding	ChIP-seq, ChIP-exo, or in-vitro SELEX; predicted PWM from recognition code
Disordered tails not functionally probed	AlphaFold shows disorder; composition not acidic	A cryptic disordered activation/repression domain cannot be fully excluded by composition alone	Domain-swap / tethering assays (e.g., Gal4-DBD fusions of the N/C tails)
Ortholog / paralog identity uncertain	PANTHER SF440; "XlCGF17.1-like" name	Correct ortholog assignment could import functional data if a characterized ortholog exists	Phylogenetic placement against characterized Xenopus/vertebrate ZF families
ProtNLM provenance opaque	Prediction is model-internal	Cannot audit why "activator" was chosen	Compare ProtNLM output across paralogs to detect systematic directional labeling

Discriminating Tests

1. **Gal4-DBD tethering / one-hybrid assay:** Fuse the N-terminal (1–36) and C-terminal (261–265) regions (and full-length minus DBD) to a heterologous DNA-binding domain and measure reporter activity. This directly reads activation vs. repression and would settle GO:0001228 vs. GO:0001227 vs. neither — the single most decisive experiment.

2. **Reporter assay in Xenopus (or HEK293) with the native DBD:** Identify a bound site (below) and measure whether occupancy increases or decreases transcription.
3. **Binding-site determination (ChIP-seq/-exo, CUT&RUN, or SELEX; or predicted PWM):** Confirms the neutral GO:0000978/GO:0000981 calls with direct evidence and provides motifs for downstream target inference (cf. KRAB-ZNF motif atlases, PMID: 29146583; recognition-code prediction, PMID: 30155812).
4. **Co-regulator interaction screen (AP-MS / IP-MS):** Detect KAP1/co-repressor vs. Mediator/p300 co-activator association as an orthogonal directional readout.
5. **Paralog systematic audit:** Check whether ProtNLM assigns GO:0001228 to a whole cluster of related X. tropicalis ZF proteins — a hallmark of frequency/paralog-driven over-annotation.

Proposed Follow-up Actions (Curation Leads — require curator verification)

- **Action:** Reject the ProtNLM2 GO:0001228 prediction for A0A8J0SCI2; annotate as over-specific / over-annotated.
- **Replacement / retain:** Keep GO:0000981 (neutral RNA Pol II TF, IBA) as the lead MF term; keep GO:0000978, GO:0006357, GO:0005634, GO:0008270.
- **Candidate reference + exact snippet to verify:** PMID: 38754172 — *"The established C2H2-ZF 'recognition code' suggests that residues at positions -1, -4, and -7 recognize the 5', central, and 3' bases of a DNA base-pair triplet, respectively."* (Supports: array encodes specificity, not direction.)
- **Candidate reference + exact snippet to verify:** PMID: 25690854 — *"for most C2H2-ZF proteins it is unknown whether they even bind DNA or, if they do, to which sequences."* (Supports: direction/function unknowable for uncharacterized members.)
- **Suggested curator question:** Is the GO:0001228 annotation sourced from ProtNLM/name-based import, and does the model apply it across a paralog cluster? If so, flag as systematic over-annotation.
- **Suggested experiments:** Gal4-DBD tethering assay for intrinsic activation/repression; ChIP/SELEX for binding sites; AP-MS for KAP1 vs. co-activator association.

Bottom Line

A0A8J0SCI2 is an effector-less tandem C2H2 zinc-finger protein. Its finger array specifies DNA sequence recognition but carries **no information about activation versus repression**, and no effector domain (KRAB/BTB/SCAN/acidic) is present in sequence or in the AlphaFold structure. The curated phylogenetic evidence correctly stops at the **direction-neutral** term GO:0000981. The ProtNLM2 activator-specific prediction **GO:0001228 is refuted as over-annotation** and should not be added; direction remains experimentally undetermined, and if anything a repressor prior is at least as plausible for a vertebrate C2H2-ZF.
