AIGR Gene Hypothesis Deep Research — Final Report

Target: *Aquarana catesbeiana* (American bullfrog) protein A0A2G9RZF1

Hypothesis: ProtNLM2 prediction of ovochymase-type role in single fertilization (GO:0007338)

Summary

Verdict: REFUTED / OVER-ANNOTATED.

The ProtNLM2 prediction that bullfrog A0A2G9RZF1 functions in *single fertilization* (GO:0007338) as an ovochymase-type egg-envelope serine protease is **not supported** by the protein's actual sequence and domain architecture. The core of the ProtNLM2 hypothesis is that the protein carries the defining feature of an ovochymase — a chymotrypsin-like (S1/trypsin-family) serine protease domain with an intact His-Asp-Ser catalytic triad, combined with one or more CUB domains. Independent inspection shows that A0A2G9RZF1 is a **short, 156-amino-acid, CUB-domain-only protein with no serine protease domain whatsoever and no catalytic triad.**

Three lines of evidence converge on this conclusion. First, the entire protein is occupied by a single CUB domain (residues ~31–147), annotated concordantly across Pfam (PF00431), SMART (SM00042), InterPro (IPR000859/IPR035914), PROSITE (PS01180) and Gene3D (2.60.120.290); the four canonical CUB cysteines are present. Second, a chymotrypsin-like S1 protease domain requires ~220–260 residues, which cannot be accommodated in a 156-aa protein that is already almost entirely CUB. Third, a direct motif scan of the sequence found **none** of the diagnostic S1-

protease signatures — the His-loop motif TAAHC (His57), the nucleophile motif GDSGGP (Ser195), and the DSGG motif were all absent — meaning the His-Asp-Ser catalytic triad that defines a competent serine protease is not present.

By contrast, every bona fide ovochymase characterized to date — including the founding *Xenopus laevis* enzyme (a 30-kDa chymotrypsin-like protease, PMID: 7875375) and orthologs across mammals and multiple frogs — is a large (564–1575 aa) multidomain protein built from one to three Peptidase S1 domains interleaved with CUB domains. A0A2G9RZF1 has neither the size, the domain complement, nor the catalytic machinery of an ovochymase. The prediction is best explained as **paralog/family over-annotation**: the single CUB module places the protein in a broad "ovochymase-related" or CUB-containing family by homology, and the ProtNLM2 label propagated the family's protease/fertilization function to a protein that lacks the catalytic domain entirely. The most important caveat is that the UniProt entry is a genome-derived predicted ORF (proteinExistence "Predicted", TrEMBL); a partial gene model cannot be fully excluded from sequence data alone, though the entry is not flagged as a fragment and the AlphaFold model spans the full 1–156 as a single compact domain.

Key Findings

Finding 1 — A0A2G9RZF1 is a CUB-domain-only protein with no serine protease domain or catalytic triad

The UniProt entry A0A2G9RZF1 (organism code AQUCT) encodes a protein of only **156 amino acids** (predicted molecular weight ~17.7 kDa). Domain annotations from five independent resources all map to a **single CUB domain**: Pfam PF00431 (CUB), SMART SM00042, InterPro IPR000859 (CUB domain) / IPR035914 (Spermadhesin CUB-like superfamily), PROSITE PS01180, and Gene3D 2.60.120.290 (the CUB β-sandwich fold). This CUB domain spans approximately residues 31–147 — that is, essentially the entire mature protein after the N-terminal region. All four canonical CUB cysteines that form the two conserved disulfide bonds are present (C31, C61, C88, C110), confirming a structurally intact CUB module.

Crucially, a chymotrypsin-like S1 serine protease domain is typically ~220–260 residues long and cannot physically fit within a 156-aa protein whose residues are already accounted for by the CUB domain. A **direct motif scan of the protein sequence** confirmed the absence of the S1-protease catalytic apparatus: the His-loop signature **TAAHC** (surrounding the catalytic His57 in chymotrypsin numbering) was **absent**; the nucleophile-elbow signature **GDSGGP**

(surrounding the catalytic Ser195) was **absent**; and the **DSGG** motif was likewise **absent**. In other words, no His-Asp-Ser catalytic triad — the sine qua non of a functional serine protease — is encoded in this sequence.

This directly contradicts the mechanistic premise of the ProtNLM2 prediction. Authentic ovochymase, as biochemically defined in *Xenopus laevis*, is a chymotrypsin-like serine protease of the S1 (serine protease I) family with a characteristic N-terminal sequence **VVGGQQAAPR** (PMID: 7875375). A0A2G9RZF1 possesses **neither** the protease domain **nor** that diagnostic N-terminal signature. Meanwhile, the CUB domain on its own is a generic ~110-residue β-sandwich protein–protein interaction module found in a wide variety of extracellular proteins (PMID: 21954942); its mere presence carries no implication of protease activity or a fertilization role.

Supporting citation snippets: - PMID: 7875375: *"The N-terminal amino acid sequence of SDS-PAGE-isolated ovochymase was determined to be VVGGQQAAPR. This conserved amino acid sequence, plus active site specific inhibition and substrate specificity studies, places ovochymase in the serine protease I family of enzymes."* — defines authentic ovochymase as an S1-family chymotrypsin-like serine protease; A0A2G9RZF1 lacks both the protease domain and this N-terminal signature. - PMID: 21954942: *"CUB domains are 110-residue protein motifs exhibiting a β-sandwich fold and mediating protein-protein interactions in various extracellular proteins."* — establishes that a CUB domain alone is a generic protein–protein interaction module that does not confer protease or fertilization function.

Finding 2 — Every bona fide ovochymase is a large multidomain S1-protease; A0A2G9RZF1 lacks the defining protease domain

A survey of UniProt entries for characterized ovochymases and their orthologs shows a highly conserved architecture: **one-to-three Peptidase S1 (chymotrypsin-like serine protease) domains interspersed with multiple CUB domains**, in proteins ranging from 564 to 1575 amino acids. The table below summarizes the comparison and makes the mismatch immediate.

Protein	UniProt	Organism	Length (aa)	Domain architecture	S1 protease domain?
OVCH1	Q7RTY7	*Homo sapiens*	1134	S1 + CUB + CUB + S1 + CUB	Yes (×2)
OVCH2	Q7RTZ1	*Homo sapiens*	564	S1 + CUB + CUB	Yes
Ovochymase	P79953	*Xenopus laevis*	1004	S1 + CUB + CUB + S1	Yes (×2)
Ovochymase-like	Q90WD8	*Bufo japonicus*	974	multidomain S1 + CUB	Yes
Ovochymase-like	Q66TN7	*Rhinella arenarum*	980	multidomain S1 + CUB	Yes
Ovochymase-like	A0AAV2ZYZ9	*Pyxicephalus adspersus*	820	SEA + CUB + CUB + S1	Yes
Query	**A0A2G9RZF1**	***Aquarana catesbeiana***	**156**	**single CUB only**	**No**

The contrast is stark: genuine ovochymases are 5- to 10-fold longer than the query protein and always retain at least one intact Peptidase S1 domain, including in three independent frog lineages (*Xenopus*, *Bufo*, *Rhinella*, plus *Pyxicephalus*). A0A2G9RZF1, at 156 aa with a single CUB domain and no Peptidase S1 domain, does not fit this family template.

Provenance/context: UniProt does not annotate A0A2G9RZF1 as a fragment. It is a genome-derived predicted ORF (locus AB205_0007200; EMBL genomic entry KV928989; proteinExistence "Predicted"; TrEMBL/unreviewed). The AlphaFold model AF-A0A2G9RZF1-F1 spans the full 1–156 as a single compact domain, consistent with a genuinely CUB-only protein rather than an obviously truncated protease fragment.

Mechanistic Model / Interpretation

The seed hypothesis rests on a testable structural claim: *does A0A2G9RZF1 carry the two-part ovochymase architecture — CUB module(s) + a chymotrypsin-like S1 protease domain with an intact His-Asp-Ser triad?* The answer, from direct sequence and domain analysis, is unambiguous: it has the CUB half but is entirely missing the protease half.

```
AUTHENTIC OVOCHYMASE (e.g., Xenopus P79953, 1004 aa)
N-[ S1 protease ]--[ CUB ]--[ CUB ]--[ S1 protease ]-C
      |His-Asp-Ser triad          |His-Asp-Ser triad
      +-- cleaves egg-envelope substrates at fertilization

QUERY  A0A2G9RZF1 (156 aa)
N--[ CUB (31-147) ]--C
      | 4 canonical Cys present (C31,C61,C88,C110)
      | NO S1 domain, NO TAAHC / GDSGGP / DSGG motifs
      +-- generic protein-protein interaction module only
```

Mechanistically, ovochymase's role in fertilization is **catalytic**: at egg activation it is released and proteolytically cleaves egg-envelope / vitelline-layer targets, contributing to envelope remodeling and the block to polyspermy. This is exactly the kind of activity documented for related fertilization proteases — for example, the sea urchin cortical granule serine protease CGSP1 cleaves the surface protein p160 to detach the fertilization envelope (PMID: 15242800). Such a function **requires a competent catalytic domain**. Because A0A2G9RZF1 lacks any protease domain, it cannot perform the molecular event (peptide-bond hydrolysis) that underlies the ovochymase contribution to GO:0007338. The prediction therefore assigns a biological process that depends on an enzymatic activity the protein does not possess.

What A0A2G9RZF1 *could* plausibly do, given its architecture, is act as a stand-alone extracellular CUB-based protein–protein interaction module. CUB domains mediate diverse extracellular interactions (PMID: 21954942); CUB-containing surface proteins can scaffold or cluster partners (e.g., LEV-10 clustering acetylcholine receptors, PMID: 15457263) and can serve as protease substrates or membrane-linkers at fertilization (e.g., sea-urchin p160, PMID: 15242800). Notably, in mammalian CUB-protease families such as BMP1/mTLL1, catalytic activity resides in the protease domain and not the CUB domains (PMID: 16507574) — reinforcing that a CUB-only protein cannot inherit the enzyme's function. However, none of these non-catalytic possibilities is specifically evidenced for A0A2G9RZF1, and none justifies the specific catalytic-protease-driven annotation to GO:0007338. The most parsimonious explanation for the ProtNLM2 label is **family/paralog over-annotation**: the single CUB

domain triggered similarity to a broad "ovochymase-related" or CUB-containing family, and the label-transfer model propagated the family's dominant protease/fertilization function to a protein that lacks the enzyme.

Evidence Base / Evidence Matrix

Citation	Evidence type	Supports/ Refutes/ Qualifies	Claim tested	Key finding	Context	Co & lim
PMID: 7875375	Direct biochemical assay	**Refutes** (defines standard the query fails)	What is authentic ovochymase?	30-kDa chymotrypsin-like S1-family serine protease, N-terminal VVGGQQAAPR	*Xenopus laevis* egg exudate	Hig fan def que pro dor N-t sig
PMID: 21954942	Review/ structural	**Qualifies**/ **Refutes**	Does CUB imply protease/ fertilization function?	CUB is a generic ~110-aa β-sandwich protein–protein interaction module	General extracellular proteins	Hig est CU nor
UniProt A0A2G9RZF1 + Pfam/ SMART/ InterPro/ PROSITE/ Gene3D	Computational/ database	**Refutes**	Domain architecture of the query	156 aa; single CUB (31–147); 4 canonical Cys; no S1 domain	*A. catesbeiana* predicted ORF	Hig dor ent "Pr TrE
Direct sequence motif scan (this investigation)	Computational	**Refutes**	Presence of His-Asp-Ser triad	TAAHC, GDSGGP, DSGG all absent → no catalytic triad	Query sequence	Hig bas cor wit cal
UniProt survey (Q7RTY7, Q7RTZ1, P79953, Q90WD8,	Structural/ evolutionary	**Refutes**	Does query match ovochymase family architecture?	All bona fide ovochymases 564–1575 aa with ≥1 S1 domain;	Human + 4 frogs	Hig cor ten acr

Citation	Evidence type	Supports/ Refutes/ Qualifies	Claim tested	Key finding	Context	Co & lim
Q66TN7, A0AAV2ZYZ9)				query 156 aa CUB-only		
PMID: 15242800	Direct assay (comparator)	**Qualifies** (mechanistic context)	How do fertilization proteases act on CUB substrates?	CGSP1 serine protease cleaves CUB-containing p160 to detach fertilization envelope	Sea urchin egg	Illu cat req CU car sub not
PMID: 15457263	Mutant phenotype/ localization (comparator)	**Qualifies** (alternative function)	Can CUB-only extracellular regions have non-protease roles?	LEV-10 CUB domains mediate AChR clustering via extracellular interactions	*C. elegans* NMJ	Sho pro via pro int not
PMID: 16507574	Structure-function (comparator)	**Qualifies** (domain-role context)	Which domain carries catalysis in CUB-protease families?	In BMP1/ mTLL1 the protease domain (not CUB) is catalytic	Mammalian metalloproteinases	Rei cat res the dor que
PMID: 8844694	Biochemical (comparator)	Context only	Regulation of ovochymase	pNiXa serpin partially inhibits ovochymase in egg exudates	*Xenopus* oocytes/ embryos	Cor ovo is a pro tar per que

GO Curation Implications

Lead requiring curator verification: The computational prediction of **single fertilization (GO:0007338)** for A0A2G9RZF1 should **not be accepted** and should be treated as an over-annotation arising from generic CUB-family/paralog similarity.

- **Biological Process (GO:0007338, single fertilization):** Not supported. The annotation depends on an ovochymase-type catalytic protease activity that the protein does not possess (no S1 domain, no His-Asp-Ser triad). Recommend **do not annotate / remove** this prediction from the review, or at minimum down-weight it to an explicitly speculative status pending experimental evidence.
- **Molecular Function:** No evidence supports serine-type endopeptidase activity (e.g., GO:0004252) — the catalytic residues are absent, so any protease MF term should be **rejected**. The only defensible molecular-function characterization from the sequence is that the protein contains a CUB protein-interaction module; however, "protein binding" (GO:0005515) is uninformative and is explicitly discouraged as a final recommendation. No stronger, better-supported MF term is available from current evidence.
- **Cellular Component:** Not directly established here. CUB domains are characteristic of extracellular/secreted proteins, so an extracellular-region CC term could be a *lead* if a signal peptide is confirmed, but this was not tested and should not be asserted.

Net recommendation: Retain the gene as an uncharacterized CUB-domain-containing protein; reject the ProtNLM2 GO:0007338 assignment as a molecular-function-driven annotation. Flag the discrepancy between the ProtNLM2 "ovochymase" label and the actual single-CUB architecture for curator attention.

Mechanistic Scope

The immediate molecular function being tested is **serine-type endopeptidase (chymotrypsin-like S1) activity**, which is what an ovochymase contributes to the biological process of *single fertilization*. That contribution is a **direct catalytic activity** (peptide-bond hydrolysis of egg-envelope/vitelline substrates), not a downstream phenotype. The analysis shows the gene product lacks the catalytic domain required for this direct activity. Everything the seed hypothesis attributes to the protein — a role in egg-envelope processing, in the block to polyspermy, in fertilization — is downstream of an enzymatic step the protein cannot perform.

Any conceivable residual role for A0A2G9RZF1 would be a **non-catalytic, protein–protein-interaction** function of its CUB module, which is distinct from, and far less specific than, the ovochymase protease function the prediction asserts.

Conflicts and Alternatives

1. **Paralog / family over-annotation (most likely explanation).** ProtNLM2 is a label-transfer model; the single CUB module is enough to place the protein in a broad CUB-containing or "ovochymase-related" family whose best-characterized members are S1-protease ovochymases. The dominant family function (protease → fertilization) was transferred despite the query lacking the catalytic domain. This is a textbook frequency-bias/paralog-overannotation failure mode.
2. **Partial or mis-predicted gene model.** The entry is a genome-derived predicted ORF ("Predicted"/TrEMBL, locus AB205_0007200, EMBL KV928989). It is conceivable the true gene is longer and a protease domain exists in an unannotated exon. However, UniProt does not flag it as a fragment, and the AlphaFold model treats residues 1–156 as a complete single domain — so the CUB-only interpretation is the best current reading, and even a truncated protease fragment would not itself be a competent enzyme.
3. **Genuine CUB-only interaction protein.** The protein may be a bona fide short, secreted CUB protein that functions purely through protein–protein interactions (cf. LEV-10, PMID: 15457263) or as a substrate/linker at the egg surface (cf. sea-urchin p160, PMID: 15242800). If so, it could conceivably participate in fertilization-adjacent processes *non-catalytically* — but this is speculative, unproven, and would not be captured by an ovochymase-protease annotation.

Limitations and Knowledge Gaps

- **Predicted, unreviewed entry.** All conclusions rest on an automatically annotated TrEMBL/"Predicted" sequence; no experimental protein has been isolated for *A. catesbeiana* A0A2G9RZF1. *Checked:* UniProt/domain databases. *Why it matters:* a corrected/extended gene model could change the architecture. *Resolution:* RNA-seq/EST evidence and a full-length cDNA to confirm the ORF boundaries.

- **Signal peptide / localization not established.** *Checked:* not computed in this investigation. *Why it matters:* extracellular localization would strengthen a CUB-interaction interpretation and bear on any fertilization-adjacent role. *Resolution:* run SignalP/DeepTMHMM on the sequence and inspect the AlphaFold N-terminus.
- **No functional data.** No expression, interaction, or phenotype data specific to this bullfrog protein exist. *Why it matters:* rules in/out any non-catalytic fertilization role. *Resolution:* tissue expression profiling (ovary/egg), interaction assays, and, if feasible, knockdown phenotyping.
- **Motif scan is signature-based.** The absence of TAAHC/GDSGGP/DSGG is strong but is a motif heuristic; it was corroborated by concordant Pfam/InterPro domain calls and by the protein's length, but a formal HMM search against the full S1/trypsin profile would be the most rigorous confirmation. *Resolution:* run hmmscan against Pfam PF00089 (Trypsin) to confirm no S1 hit.

Discriminating Tests

1. **HMM profile search (fastest, decisive).** Run `hmmscan` of A0A2G9RZF1 against Pfam-A (especially PF00089 Trypsin / S1). Expected result under the refutation: a single significant CUB (PF00431) hit and **no** S1 hit — a rigorous confirmation of the motif-scan conclusion.
2. **Full-length gene-model verification.** Retrieve the genomic locus (KV928989 / AB205_0007200) and adjacent exons; check bullfrog transcriptome/RNA-seq to confirm whether the ORF is truly 156 aa or a truncated model missing a downstream protease-encoding exon.
3. **Catalytic-triad structural check.** Inspect the AlphaFold model AF-A0A2G9RZF1-F1 for any His/Asp/Ser spatial cluster; absence confirms no cryptic triad.
4. **Ortholog/synteny comparison.** Compare A0A2G9RZF1 to the *Xenopus*/*Bufo*/*Rhinella*/*Pyxicephalus* ovochymases and to any full-length bullfrog ovochymase paralog; a separate full-length bullfrog ovochymase would strengthen the case that A0A2G9RZF1 is a distinct, protease-less CUB protein.
5. **Signal-peptide / topology prediction.** SignalP + DeepTMHMM to establish secretion, informing any non-catalytic extracellular role.

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

- **Action change:** Reject the ProtNLM2 GO:0007338 (*single fertilization*) prediction for A0A2G9RZF1; do not carry it into the review as an accepted annotation. Rationale: catalytic protease domain absent; ovochymase family architecture not met.
- **Reject associated MF:** Do not annotate serine-type endopeptidase activity (GO:0004252) or any protease MF — catalytic triad absent.
- **Candidate accurate description:** "CUB domain-containing protein of unknown function; predicted secreted; single CUB module (residues ~31–147); no serine protease domain." Avoid "protein binding" as a terminal annotation.
- **Candidate reference + snippet to verify (family standard the protein fails):** PMID: 7875375 — *"...places ovochymase in the serine protease I family of enzymes."*
- **Candidate reference + snippet to verify (CUB is non-catalytic):** PMID: 21954942 — *"CUB domains are 110-residue protein motifs exhibiting a β-sandwich fold and mediating protein-protein interactions in various extracellular proteins."*
- **Suggested curator question:** Is the 156-aa ORF the complete gene product, or a truncated genome-derived model? Resolve via transcriptome/full-length cDNA before any functional annotation.
- **Suggested experiments:** HMM search vs PF00089; SignalP/DeepTMHMM; AlphaFold triad inspection; ortholog/synteny analysis to locate the true bullfrog ovochymase paralog.

Conclusion

The ProtNLM2 prediction is a **paralog/family over-annotation**. Bullfrog A0A2G9RZF1 is a 156-aa, single-CUB-domain protein that lacks the chymotrypsin-like S1 serine protease domain and the His-Asp-Ser catalytic triad that define an ovochymase. Because the ovochymase contribution to *single fertilization* is intrinsically catalytic, and this protein cannot catalyze peptide-bond hydrolysis, GO:0007338 should not be accepted for A0A2G9RZF1. The only material caveat is that the entry is a predicted, unreviewed ORF whose gene model would benefit from experimental confirmation.