SDPF SPECIFICATION

FoldContract

Protein Structure Prediction for Novel Pathogen Variants

Style 2 — Exploratory Problem-Solving

SDPF Language Specification v1.3.1 Conforming

Problem Owner	Head of Computational Biology
Date	April 2026
Status	**SPEC_LOCKED**

PHASE 0 — Problem Definition

Problem Statement

> *"Novel pathogen variant protein structure prediction requires an average of 18 months of laboratory validation per target, against a target of 0 validation cycles at ≥85% structural confidence, resulting in $240M delayed drug development cost per year."*

Problem Components

Component	Definition
Current State	Novel pathogen variant protein structure prediction requires an average of 18 months of laboratory validation per target.
Desired State	Computational prediction produces candidate structures with ≥85% structural confidence requiring 0 laboratory validation cycles before drug development handoff.
Gap	18 months of laboratory validation per target above the target of 0 validation cycles.
Impact	$240M in delayed drug development cost per year across the active pipeline.

Validation Tests

Test	Name	Pass Condition	Result
T-1	Observable	Months and validation cycles are measurable.	✅
T-2	Bounded	Scoped to novel pathogen variant targets entering drug pipeline.	✅
T-3	Cause-Free	No "because" or "due to" language present.	✅
T-4	Solution-Free	No algorithm, tool, or method named.	✅

PHASE 1 — SDPF Specification

S-01 — Style Declaration

> *Style 2 — Exploratory Problem-Solving. Requirements are unknown and must emerge through exploration. The specification defines the search space and abort criteria. No solution is asserted to exist before validation.*

S-02 — Style Context

Domain: Computational structural biology for pharmaceutical drug discovery.

System boundary: Intake of novel pathogen variant amino acid sequences → output of ranked candidate protein structures with confidence scores.

Out of scope: Wet-lab validation, molecular dynamics simulation beyond initial scoring, clinical assessment.

S-03 — External Contract

Problem statement reference: 18-month laboratory validation per target → target 0 validation cycles at ≥85% structural confidence

Interface: Computational pipeline

Input: Amino acid sequence (FASTA format)

Output: Ranked list of predicted 3D structures (PDB format) with per-structure confidence scores and uncertainty bounds

GUI: Not required at this stage

S-04 — Input Contract

[CRITICAL] Accept a valid amino acid sequence in FASTA format, minimum 50 residues, maximum 2,500 residues.

[CRITICAL] Accept an optional homolog reference database path for template-based scoring calibration.

[REQUIRED] Accept a compute budget parameter expressed in GPU-hours (integer, 1–500).

[REQUIRED] Accept a minimum confidence threshold parameter (float, 0.0–1.0; default 0.85).

[OPTIONAL] Accept a target organism annotation to bias homolog search.

S-05 — Processing Rules

[CRITICAL] For each input sequence, execute all active hypotheses in parallel up to the declared compute budget. Do not assume any single hypothesis will succeed.

[CRITICAL] Score every candidate structure against the declared confidence threshold using template-matched LDDT comparison via biotite where a reference homolog exists, and pLDDT-equivalent scoring where no homolog exists.

[REQUIRED] Rank all candidate structures by confidence score, descending. Include uncertainty bounds per structure.

[REQUIRED] Halt hypothesis execution for any individual hypothesis branch that has consumed 40% of its allocated compute budget without producing a candidate scoring above 0.50 confidence — this is a branch-level abort, not a full pipeline abort.

[OPTIONAL] If a target organism annotation is provided, weight homolog search toward evolutionary neighbours.

S-06 — Output Guarantees

[CRITICAL] Every run produces a ranked output list of at least 1 candidate structure, regardless of confidence achieved. If no candidate meets the threshold, the list is still returned with actual scores and a below-threshold flag.

[CRITICAL] Every output structure includes: predicted 3D coordinates (PDB), per-residue confidence score, global confidence score, uncertainty bound, and the hypothesis branch that produced it.

[REQUIRED] Output is produced within the declared compute budget. The pipeline does not exceed the GPU-hour limit.

[REQUIRED] If the top-ranked structure achieves ≥0.85 confidence, the output record is flagged THRESHOLD_MET. Otherwise it is flagged THRESHOLD_NOT_MET with the highest score achieved recorded.

S-07 — Exception Handling

Condition	Action
Invalid FASTA characters in sequence	Reject immediately: "Invalid sequence — FASTA format required, valid amino acid characters only."
Sequence length < 50 or > 2,500 residues	Reject: "Sequence length out of bounds — accepted range 50–2,500 residues."
Compute budget exhausted before completion	Halt remaining branches. Return all completed results. Flag: COMPUTE_BUDGET_EXHAUSTED.
No homolog and pLDDT scoring unavailable	Return candidate with confidence score 0.00. Flag: CONFIDENCE_UNSCORED. Do not suppress output.
All branches abort below 0.50 before budget exhaustion	Trigger Fallback Strategy (S-13) immediately.

S-08 — Technical Verification Gate (TVG)

> *HALT rule: Any failed TVG entry halts the specification at DRAFT. The specification is corrected — the environment is not patched around the failure.*

Asset	Asserted Value	Verification Command	Pass Condition	HALT Action
Python runtime	3.11.x	python --version	Returns Python 3.11.*	Update environment
BioPython	≥1.81	pip show biopython	Version ≥ 1.81 confirmed	Update package
FASTA parser	Handles up to 2,500 residues	Parse test sequence of 2,500 residues	No exception, correct residue count returned	Fix parser config
GPU availability	≥1 CUDA device	nvidia-smi	At least 1 device listed	Provision GPU
PDB output writer	Valid PDB format	Write test structure, validate with PDB validator	Validation passes with 0 errors	Fix output serialiser
biotite	≥1.6.0	pip show biotite	Version ≥1.6.0 confirmed	pip install biotite

S-09 — Verification Requirements

TEST-ID	REQ-ID	Description	Pass Condition
T-001	R-001	Valid FASTA accepted	200-residue sequence ingested, pipeline initiates
T-002	R-001	Invalid FASTA rejected	Non-FASTA input returns rejection message, no crash
T-003	R-002	Compute budget respected	10 GPU-hour budget → pipeline halts at or before 10 GPU-hours
T-004	R-003	Output always produced	Pipeline run with no homolog available → at least 1 structure returned
T-005	R-004	Confidence threshold flagging — above threshold	Known structure scoring 0.91 → output flagged THRESHOLD_MET
T-006	R-004	Confidence threshold flagging — below threshold	Sequence producing max 0.62 → flagged THRESHOLD_NOT_MET, score recorded
T-007	R-005	Branch-level abort	Branch at 40% budget, 0.48 confidence → halted, others continue
T-008	R-006	Fallback triggered	All branches abort below 0.50 → fallback executes, partial results returned
T-009	R-007	PDB output valid	Every returned structure passes PDB format validation
T-010	R-008	Uncertainty bounds present	Every returned structure includes a numeric uncertainty bound

S-10 — Traceability Matrix

REQ-ID	Requirement Summary	TEST-ID(s)
R-001	Accept valid FASTA, reject invalid	T-001, T-002
R-002	Respect compute budget	T-003
R-003	Always produce output	T-004
R-004	Confidence threshold flagging	T-005, T-006
R-005	Branch-level abort at 40% / 0.50	T-007

R-006	Fallback on all-branch failure	T-008
R-007	Valid PDB output format	T-009
R-008	Uncertainty bounds per structure	T-010

S-11 — Hypothesis List (Style 2 Required Section)

Each hypothesis is an independent prediction approach executed in parallel. No hypothesis is assumed to succeed.

H-ID	Hypothesis	Testable Condition
H-01	Template-based homology modelling against known PDB structures will produce ≥0.85 LDDT score for variants with >30% sequence identity to a known homolog.	LDDT score ≥ 0.85 on held-out validation set of 50 sequences with known structures.
H-02	MSA co-evolutionary signal extraction will produce ≥0.85 confidence for variants with sufficient homologous sequences (≥100 MSA depth).	pLDDT-equivalent score ≥ 0.85 on MSA-depth-qualified sequences in validation set.
H-03	De novo structure prediction without template or MSA reliance will produce ≥0.85 confidence for at least 20% of novel sequences with no known homologs.	≥20% of the no-homolog subset in the validation set scored ≥ 0.85.
H-04	Ensemble consensus across H-01, H-02, and H-03 outputs will improve confidence score by ≥0.05 over the best single-hypothesis result.	Ensemble score exceeds best individual hypothesis score by ≥ 0.05 on ≥60% of validation sequences.

S-12 — Abort Criteria (Style 2 Required Section)

> *Abort criteria are hard stops. They are not suggestions. When triggered, the stated action is taken immediately.*

ID	Criterion	Trigger Condition	Action
AC-01	Branch abort	Any single hypothesis branch consumes 40% of its allocated GPU-hour share without producing a candidate at confidence ≥ 0.50.	Halt that branch. Reallocate remaining budget to active branches. Record branch as ABORTED.
AC-02	Full pipeline abort	All hypothesis branches aborted under AC-01 before 80% of total compute budget is consumed.	Trigger Fallback Strategy (S-13). Do not consume remaining budget on exhausted approaches.
AC-03	Validation failure abort	Fewer than 10% of sequences in the held-out validation set achieve ≥ 0.85 confidence across all hypotheses combined.	Flag run as VALIDATION_FAILED. Return all results produced. Do not suppress output. Report highest confidence achieved.
AC-04	Time limit abort	Wall-clock time exceeds 72 hours regardless of GPU-hour budget status.	Halt all branches. Return all results completed to that point. Flag: TIME_LIMIT_REACHED.

S-13 — Fallback Strategy (Style 2 Required Section)

The fallback activates when AC-02 is triggered (all branches aborted) or when confidence threshold cannot be met within budget.

Step	Action
1	Return ranked partial solutions. All candidate structures produced before abort, regardless of confidence score, are returned in ranked order with actual scores. No output is suppressed.
2	Report the confidence ceiling. The output record includes the highest confidence score achieved across all branches and which hypothesis produced it.
3	Report the search boundary. The output states which hypotheses were attempted, which were aborted, and at what confidence level each was aborted.
4	Recommend next action. If highest confidence achieved is 0.70–0.84: recommend targeted laboratory validation of top-ranked structure only. If below 0.70: recommend sequence re-evaluation or expanded compute budget.
5	Never return empty output. Even if all branches abort at 0.00 confidence, the pipeline returns the best candidate produced with full abort metadata. The drug development team always receives a documented artefact.

Lifecycle — What Happens Next

With the specification locked and TVG verified, the following stage sequence applies:

Stage	Action	Gate Condition
1	Generate Tests	Derive all 10 tests from S-09 before any implementation code exists.
2	Lock Tests	Test suite frozen. No modification permitted after this point.
3	Generate Implementation	AI framing: "Expert algorithm researcher. Implement hypotheses and abort criteria. Do not assert a solution exists before validation."
4	Verification Gate	All 11 structural invariant checks must pass.
5	Export Evidence	Signed audit trail from the $240M problem statement to verified pipeline.

SDPF Guarantee at This Scale

> *SDPF does not solve protein folding. It ensures that no one wastes 18 months discovering they were solving the wrong protein folding problem. The gap is a number. The desired state is testable. Abort criteria prevent infinite compute burn. The fallback ensures the drug development team always receives a documented artefact — even when the problem proves harder than anticipated.*