

Molecular Systematics & Complexity | Signal Rating: 4/5 Stars | Oct 29, 2026 | 15 min read

2x2 Factorial Calibration of Multiple Sequence Alignment and Substitution Model Selection: Anchor-Segmented Gotoh & Felsenstein Pruning vs. MAFFT v7.526 & jModelTest v2.1.9

A 2x2 Factorial Calibration of Anchor-Segmented Gotoh Alignment and Column-Compressed Felsenstein Pruning Against MAFFT v7.526 (L-INS-i) and jModelTest v2.1.9 (PhyML v3.0) Across Mitochondrial COX1, Bacterial 16S rRNA, and Fungal ITS1-5.8S-ITS2 Loci

Phylogenomic inference relies on two sequential computational stages -- multiple sequence alignment (MSA) and nucleotide substitution model selection -- whose algorithmic interactions are rarely evaluated in a controlled factorial design. Here we execute a 2x2 factorial calibration crossing Native Rust Anchor-Segmented Gotoh Alignment v2 and MAFFT v7.526 (L-INS-i) with Native Rust Column-Compressed Felsenstein ModelTest and jModelTest v2.1.9 (PhyML v3.0) across 20 verified NCBI RefSeq accessions spanning vertebrate mitochondrial COX1, bacterial 16S rRNA, and fungal ITS1-5.8S-ITS2 cistrons. Across all four factorial cells, the pipelines achieve unanimous Bayesian Information Criterion (BIC) base-matrix concordance while revealing how upstream loop-indel consolidation shifts absolute log-likelihood surfaces by 146.0 to 330.8 log-units and delivering a 119.6x to 267.7x end-to-end speedup in the all-native configuration.

1. Joint Aligner x Model-Selection Interactions in Molecular Systematics

In computational phylogenetics and molecular systematics, nucleotide substitution model selection is mathematically downstream of multiple sequence alignment (MSA). Because Felsenstein's tree-pruning algorithm evaluates conditional site likelihoods independently across aligned columns, any upstream shift in gap placement -- such as expanding vs. consolidating indels inside ribosomal expansion loops or internal transcribed spacers -- alters the number of alignment columns (n), the empirical invariant-site fraction (p_{inv}), the number of unique site patterns (u), and the Bayesian Information Criterion penalty term ($k \ln n$). Despite this tight mathematical coupling, alignment engines and substitution model-selection engines are almost always benchmarked in isolation rather than across a crossed factorial matrix.

To quantify how upstream alignment architecture interacts with downstream maximum-likelihood model selection, we executed a full 2×2 factorial calibration crossing two multiple sequence alignment engines with two 24-model nucleotide substitution selection engines:

* Cell 1 -- Native Aligner v2 + Native ModelTest (Default All-Native Stack): Combines Codon-Anchored / 9-mer Anchor-Segmented Gotoh 3-state affine-gap dynamic programming ($G_{open} = -13.0$, $G_{ext} = -1.5$) with in-memory column-compressed Felsenstein pruning, scaling-and-squaring continuous-time Markov chain exponentiation $P(t) = \exp(Qt)$, and Yang discrete Gamma (Γ_4) quadrature

across all 24 candidate models (JC, F81, K80, HKY, SYM, GTR x [Uniform, +I, +G 4 , +I+G 4]).

* Cell 2 -- MAFFT v7.526 (L-INS-i) + Native ModelTest (Hybrid External-MSA Stack): Aligns sequences via iterative local-pair consistency refinement and pipes the resulting alignment directly into the native column-compressed Felsenstein pruning engine.

* Cell 3 -- Native Aligner v2 + jModelTest v2.1.9 (Hybrid External-ModelTest Stack): Aligns sequences via the native Anchor-Segmented Gotoh engine and evaluates the 24 substitution models via external jModelTest v2.1.9 invoking PhyML v3.0 maximum-likelihood optimization across AIC, AICc, and BIC criteria.

* Cell 4 -- MAFFT v7.526 (L-INS-i) + jModelTest v2.1.9 (All-External Reference Stack): Runs iterative local alignment in MAFFT v7.526 followed by 24-process PhyML v3.0 likelihood optimization in jModelTest v2.1.9 .

Signal Rating Assessment (**** 4* Strong Emergent Signal): This synthesis evaluates 12 full factorial pipeline executions across three empirical structural regimes (vertebrate mitochondrial coding COX1 , cross-phylum prokaryotic 16S SSU rRNA , and eukaryotic nuclear ITS1 -5.8S-ITS2 cistrons, plus an unanchored ITS1 negative-control boundary), totaling 288 complete maximum-likelihood tree and substitution-model evaluations over 20 verified NCBI RefSeq accessions. The study earns a **** 4* Strong Emergent Signal rating because it surfaces a reproducible cross-stage interaction: upstream alignment loop compression shifts absolute log-likelihoods by hundreds of log-units ($\Delta \ln L = 146.0$ to 330.8) identically across both model-selection engines while preserving 100% BIC base-matrix concordance when structural anchors remain intact.

2. Master 2x2 Factorial Calibration Matrix Across Structural Regimes

Table 1 summarizes end-to-end execution telemetry, alignment geometry, unique site-pattern compression, top-ranked BIC/AIC substitution models, maximized log-likelihoods ($-\ln L$), BIC scores, and cross-cell Pearson log-likelihood correlations (r) across all three multi-clade structural cohorts.

Cohort & Structural Regime	Factorial Cell	Aligner Engine	ModelTest Engine	MSA Time (ms)	ModelTest Time (ms)	Total Time (ms)	Speedup vs Cell 4 (x)	Align Len (bp)	Gap Frac (%)	Unique Patterns	Invariant Frac (%)	Best BIC Model	Best AIC Model	Best BIC $-\ln L$	Best BIC Score	BIC Weight
----------------------------	----------------	----------------	------------------	---------------	---------------------	-----------------	-----------------------	----------------	--------------	-----------------	--------------------	----------------	----------------	-------------------	----------------	------------

Cohort I: Carnivora COX1 (10 Taxa)	Cell 1 (Default)	Native Gotoh v2	Native Felsenstein	0.44	94.87	95.31	267.74	1545	0.00	162	70.87	HKY+I	HKY+I	4780.6745	9722.8901	0.8567
------------------------------------	------------------	-----------------	--------------------	------	-------	-------	--------	------	------	-----	-------	-------	-------	-----------	-----------	--------

Cohort I: Carnivora COX1 (10 Taxa)	Cell 2 (Hybrid MSA)	MAFFT v7.526 L-INS-i	Native Felsenstein	3375.10	87.51	3462.61	7.37	1545	0.00	162	70.87	HKY+I	HKY+I	4780.6745	9722.8901	0.8567
------------------------------------	---------------------	----------------------	--------------------	---------	-------	---------	------	------	------	-----	-------	-------	-------	-----------	-----------	--------

Cohort I: Carnivora COX1 (10 Taxa)	Cell 3 (Hybrid MT)	Native Gotoh v2	jModelTest v2.1.9	0.44	22070.33	22070.77	1.16	1545	0.00							
------------------------------------	--------------------	-----------------	-------------------	------	----------	----------	------	------	------	--	--	--	--	--	--	--

162 || 70.87 || HKY+I || HKY+I || 4684.3863 || 9537.6565 || 1.0000 |

| Cohort I: Carnivora COX1 (10 Taxa) || Cell 4 (All-External) || MAFFT v7.526

L-INS-i || jModelTest v2.1.9 || 3375.10 || 22143.26 || 25518.36 || 1.00 ||

1545 || 0.00 || 162 || 70.87 || HKY+I || HKY+I || 4684.3915 || 9537.6669 ||

1.0000 |

| Cohort II: Bacterial 16S rRNA (5 Operons) || Cell 1 (Default) || Native Gotoh v2

|| Native Felsenstein || 24.35 || 43.43 || 67.78 || 119.58 || 1868 || 17.32 |

| 104 || 62.31 || HKY || GTR || 5073.8473 || 10230.5535 || 0.9902 |

| Cohort II: Bacterial 16S rRNA (5 Operons) || Cell 2 (Hybrid MSA) || MAFFT v7.526

L-INS-i || Native Felsenstein || 2030.33 || 30.01 || 2060.34 || 3.93 || 1803 |

| 14.34 || 100 || 49.20 || HKY || HKY || 5404.3508 || 10891.1709 || 0.9954 |

| Cohort II: Bacterial 16S rRNA (5 Operons) || Cell 3 (Hybrid MT) || Native Gotoh

v2 || jModelTest v2.1.9 || 24.35 || 5982.57 || 6006.92 || 1.35 || 1868 ||

17.32 || 104 || 62.31 || HKY || GTR || 5068.6458 || 10227.6831 || 0.9247 |

| Cohort II: Bacterial 16S rRNA (5 Operons) || Cell 4 (All-External) || MAFFT

v7.526 L-INS-i || jModelTest v2.1.9 || 2030.33 || 6075.18 || 8105.52 || 1.00 |

| 1803 || 14.34 || 100 || 49.20 || HKY || GTR || 5399.4535 || 10888.8735 ||

0.9528 |

| Cohort III: Fungal ITS1-5.8S-ITS2 (5 Taxa) || Cell 1 (Default) || Native Gotoh

v2 || Native Felsenstein || 19.21 || 50.32 || 69.54 || 170.14 || 766 || 23.21

|| 174 || 67.62 || K80+G || K80+G || 2294.4525 || 4648.6756 || 0.7513 |

| Cohort III: Fungal ITS1-5.8S-ITS2 (5 Taxa) || Cell 2 (Hybrid MSA) || MAFFT

v7.526 L-INS-i || Native Felsenstein || 1639.55 || 70.34 || 1709.89 || 6.92 ||

756 || 22.20 || 187 || 64.15 || K80+I || K80+I || 2440.4441 || 4940.5406 ||

0.8145 |

| Cohort III: Fungal ITS1-5.8S-ITS2 (5 Taxa) || Cell 3 (Hybrid MT) || Native Gotoh

v2 || jModelTest v2.1.9 || 19.21 || 6879.97 || 6899.18 || 1.71 || 766 ||

23.21 || 174 || 67.62 || K80+G || SYM+G || 2284.2192 || 4634.8502 || 0.7997 |

| Cohort III: Fungal ITS1-5.8S-ITS2 (5 Taxa) || Cell 4 (All-External) || MAFFT

v7.526 L-INS-i || jModelTest v2.1.9 || 1639.55 || 10191.31 || 11830.87 || 1.00

|| 756 || 22.20 || 187 || 64.15 || K80+G || SYM+G || 2430.3382 || 4926.9568

|| 0.9173 |

Generated by: Native Rust Column-Compressed Felsenstein ModelTest Engine (

modeltest.rs) * jModelTest v2.1.9 / PhyML v3.0 * Native Rust Gotoh 3-State

Affine-Gap & 9-mer Anchor Multiple Sequence Aligner (phylo.rs) * MAFFT v7.526 (

L-INS-i)

Generated by: Native Rust Column-Compressed Felsenstein ModelTest Engine (

modeltest.rs) * jModelTest v2.1.9 / PhyML v3.0 * Native Rust Gotoh 3-State

Affine-Gap & 9-mer Anchor Multiple Sequence Aligner (phylo.rs) * MAFFT v7.526 (

L-INS-i)

3. Cohort I: Vertebrate Mitochondrial COX1 -- Four-Way Identity on Ungapped Reading Frames

Across the 10 Carnivora reference mitogenomes (Felidae, Canidae, and Ursidae; 1,545 bp full-length COX1 gene), purifying selection against frameshift mutations enforces strict length invariance. Both the Native Codon-Anchored aligner (0.44 ms) and MAFFT v7.526 L-INS-i (3,375.10 ms) produce byte-for-byte identical 1,545 bp alignments with 0.00% gaps, 87.14% sum-of-pairs nucleotide identity, 70.87% invariant columns (1,095 / 1,545), and exactly \$162\$ unique site patterns.

Because the upstream alignments in Cell 1 vs. Cell 2 and Cell 3 vs. Cell 4 are identical, the 2 x 2 design isolates pure model-selection concordance across engines. All four cells unanimously select HKY+I under BIC, AIC, and AICc (w BIC = 0.8567 in Cell 1 and Cell 2; w BIC = 1.0000 in Cell 3 and Cell 4), with a cross-aligner Pearson log-likelihood correlation of $r = 1.0000$ and a cross-model-selector correlation of $r = 0.9933$. Replacing the all-external workflow (Cell 4: 25,518.36 ms) with the default all-native workflow (Cell 1: 95.31 ms) yields a 267.74x end-to-end speedup with zero change in alignment topology or substitution model selection.

BIC Rank	Cell 1 (Native+Native) Model	Cell 1 -lnL	Cell 1 BIC	Cell 2 (MAFFT+Native) Model	Cell 2 -lnL	Cell 2 BIC	Cell 3 (Native+jMT) Model	Cell 3 -lnL	Cell 3 BIC	Cell 4 (MAFFT+jMT) Model	Cell 4 -lnL	Cell 4 BIC
----------	------------------------------	-------------	------------	-----------------------------	-------------	------------	---------------------------	-------------	------------	--------------------------	-------------	------------

1	HKY+I	4780.6745	9722.8901	HKY+I	4780.6745	9722.8901	HKY+I	4684.3863	9537.6565	HKY+I	4684.3915	9537.6669
---	-------	-----------	-----------	-------	-----------	-----------	-------	-----------	-----------	-------	-----------	-----------

2	HKY+G	4782.4907	9726.5225	HKY+G	4782.4907	9726.5225	GTR+I	4680.4864	9559.2278	GTR+I	4680.4785	9559.2120
---	-------	-----------	-----------	-------	-----------	-----------	-------	-----------	-----------	-------	-----------	-----------

3	HKY+I+G	4782.3719	9733.6277	HKY+I+G	4782.3719	9733.6277	HKY+G	4696.9781	9562.8401	HKY+G	4696.9613	9562.8065
---	---------	-----------	-----------	---------	-----------	-----------	-------	-----------	-----------	-------	-----------	-----------

4	K80+I	4817.6652	9774.8432	K80+I	4817.6652	9774.8432	HKY+I+G	4696.9758	9570.1783	HKY+I+G	4696.9758	9570.1783
---	-------	-----------	-----------	-------	-----------	-----------	---------	-----------	-----------	---------	-----------	-----------

5	K80+G	4819.6853	9778.8834	K80+G	4819.6853	9778.8834	GTR+G	4693.8106	9585.8762	GTR+G	4693.8253	9585.9056
---	-------	-----------	-----------	-------	-----------	-----------	-------	-----------	-----------	-------	-----------	-----------

6	GTR+I	4794.7769	9780.4661	GTR+I	4794.7769	9780.4661	GTR+I+G	4693.8262	9593.2502	GTR+I+G	4693.8262	9593.2502
---	-------	-----------	-----------	-------	-----------	-----------	---------	-----------	-----------	---------	-----------	-----------

Generated by: Native Rust Column-Compressed Felsenstein ModelTest Engine (modeltest.rs) * Native Rust Gotoh 3-State Affine-Gap & 9-mer Anchor Multiple Sequence Aligner (phylo.rs) * MAFFT v7.526 (L-INS-i)

4. Cohort II: Prokaryotic 16S rRNA -- How Loop Indel Consolidation Shifts Log-Likelihood Surfaces

Cohort II examines five prokaryotic 16S SSU rRNA operons spanning Gammaproteobacteria (Escherichia coli K-12 rrsA , rrsB , rrsH and Salmonella enterica LT2 rrsA) and Bacillota (Bacillus subtilis 168 rrnO , 1,542-1,554 bp). While the universal decoding stems are invariant across phyla, the hypervariable

expansion loops (V1, V2, V4, and V6) exhibit substantial length polymorphism between Enterobacteria and Firmicutes.

Crossing the two aligners with the two model selectors exposes a striking quantitative phenomenon (Table 3):

* Four-Way Unanimous BIC Selection of HKY: All four factorial cells independently select HKY as the #1 BIC model with overwhelming posterior weight: Cell 1 (w BIC = 0.9902), Cell 2 (w BIC = 0.9954), Cell 3 (w BIC = 0.9247), and Cell 4 (w BIC = 0.9528).

* Parallel 330.8 Log-Unit Likelihood Shift Between Aligners: Native Gotoh v2 preserves slightly wider affine gap blocks around the *B. subtilis** V1/V6 loop transitions (1,868 bp, 17.32% gaps, 62.31% invariant columns, \$104\$ unique patterns), whereas MAFFT v7.526 L-INS-i compresses those loop indels into denser polymorphic columns (1,803 bp, 14.34% gaps, 49.20% invariant columns, \$100\$ unique patterns). Because Felsenstein pruning treats gaps as missing data ($L_i(x) = [1, 1, 1, 1]$) while penalizing mismatched nucleotides across compressed columns, MAFFT's tighter column packing shifts the HKY log-likelihood downward by $\Delta \ln L = -330.50$ log-units in Native ModelTest (-5073.85 $\rightarrow$ -5404.35) and by an almost identical $\Delta \ln L = -330.81$ log-units in jModelTest v2.1.9 (-5068.65 $\rightarrow$ -5399.45).

* AIC Sensitivity to Loop Compression: Under the lighter AIC penalty (2k), both Native ModelTest (Cell 1) and jModelTest v2.1.9 (Cell 3 and Cell 4) favor the 6-rate GTR matrix to accommodate cross-phylum transversion asymmetry in *B. subtilis**, whereas under BIC ($k \ln n$), all four cells converge on HKY.

BIC Rank	Cell 1 (Native+Native) Model	Cell 1 -lnL	Cell 1 BIC	Cell 2 (MAFFT+Native) Model	Cell 2 -lnL	Cell 2 BIC	Cell 3 (Native+jMT) Model	Cell 3 -lnL	Cell 3 BIC	Cell 4 (MAFFT+jMT) Model	Cell 4 -lnL	Cell 4 BIC
1	HKY	5073.8473	10230.5535	HKY	5404.3508	10891.1709	HKY	5068.6458	10227.6831	HKY	5399.4535	10888.8735
2	HKY+G	5074.8495	10240.0905	HKY+G	5405.9733	10901.9131	HKY+G	5068.2160	10234.3561	HKY+G	5399.3735	10896.2107
3	K80	5091.7990	10243.8590	GTR	5406.8260	10926.1101	HKY+I	5068.2377	10234.3995	HKY+I	5399.4601	10896.3839
4	GTR	5068.1834	10249.3562	K80	5434.0160	10928.0097	GTR	5058.2347	10236.9914	HKY+I+G	5399.4139	10903.7887
5	K80+G	5092.9514	10253.6964	GTR+G	5409.1927	10938.3407	HKY+I+G	5068.2025	10241.8617	GTR	5393.8932	10907.7417
6	GTR+G	5070.3144	10261.1508	K80+G	5435.4790	10938.4329	GTR+I	5058.0108	10244.0762	GTR+G	5393.8096	10915.0717

Generated by: Native Rust Column-Compressed Felsenstein ModelTest Engine (modeltest.rs) * Native Rust Gotoh 3-State Affine-Gap & 9-mer Anchor Multiple Sequence Aligner (phylo.rs) * MAFFT v7.526 (L-INS-i)

5. Cohort III: Fungal ITS1-5.8S-ITS2 -- Anchor Locking & Rate-Heterogeneity (+G vs. +I) Partitioning

Cohort III evaluates the nuclear ribosomal ITS1-5.8S-ITS2 cistron across five fungal reference species (*Aspergillus fumigatus* , *A. flavus* , *A. niger* , *Candida albicans* , and *Saccharomyces cerevisiae* ; 524-596 bp). This tripartite locus pairs two hypervariable non-coding spacers (ITS1 and ITS2 , gap fraction 22.20%-23.21%) with the conserved 158 bp 5.8S ribosomal subunit ($H = 0.34$ bits).

Across all four cells of the 2 x 2 factorial matrix (Table 4), both aligners successfully lock the conserved 5.8S subunit in phase, and both model-selection engines converge on the 2-parameter Kimura K80 base matrix coupled with strong among-site rate heterogeneity (+G or +I):

* Cell 1 (Native MSA + Native ModelTest): Selects K80+G ($\alpha = 0.95$, $-\ln L = 2294.4525$, BIC = 4648.6756, w BIC = 0.7513), with K80+I ($p_{inv} = 0.4374$, $\Delta BIC = 2.2238$, w BIC = 0.2471) in second place -- together capturing 99.84% of total BIC posterior weight.

* Cell 2 (MAFFT L-INS-i + Native ModelTest): Because MAFFT compresses the flanking ITS1 / ITS2 indels into 756 bp (\$187\$ unique patterns vs. \$174\$ in Native Gotoh), the relative BIC ranking between the top two models swaps by $\Delta BIC = 3.02$ units: K80+I ($p_{inv} = 0.4149$, $-\ln L = 2440.4441$, w BIC = 0.8145) edges out K80+G ($\alpha = 0.55$, $-\ln L = 2441.9542$, w BIC = 0.1799), with the pair together capturing 99.44% of total BIC weight.

* Cell 3 (Native MSA + jModelTest v2.1.9) & Cell 4 (MAFFT L-INS-i + jModelTest v2.1.9): Both select K80+G (#1 BIC, w BIC = 0.7997 and \$0.9173\$) followed immediately by K80+I (#2 BIC, w BIC = 0.1611 and \$0.0440\$). On the MAFFT alignment (Cell 2 vs. Cell 4), Native ModelTest and jModelTest v2.1.9 achieve a 24-model log-likelihood correlation of $r = 0.9833$, while the alignment-induced log-likelihood shift between Native Gotoh (766 bp) and MAFFT (756 bp) is $\Delta \ln L = 145.99$ units in Native ModelTest and $\Delta \ln L = 146.12$ units in jModelTest v2.1.9 .

BIC Rank	Cell 1 (Native+Native) Model	Cell 1 -lnL	Cell 1 BIC	Cell 2 (MAFFT+Native) Model	Cell 2 -lnL	Cell 2 BIC	Cell 3 (Native+jMT) Model	Cell 3 -lnL	Cell 3 BIC	Cell 4 (MAFFT+jMT) Model	Cell 4 -lnL	Cell 4 BIC
1	K80+G	2294.4525	4648.6756	K80+I	2440.4441	4940.5406	K80+G	2284.2192	4634.8502	K80+G	2430.3382	4926.9568
2	K80+I	2295.5644	4650.8994	K80+G	2441.9542	4943.5608	K80+I	2285.8212	4638.0542	K80+I	2433.3763	4933.0330
3	K80+I+G	2297.3378	4661.0874	K80+I+G	2442.1335	4950.5474	K80+I+G	2284.2054	4641.4638	K80+I+G	2430.2328	4933.3741
4	SYM+G	2290.6516	4667.6386	HKY+I	2439.5690	4958.6745	SYM+G	2275.5325	4644.0416	SYM+G	2423.7095	4940.2116
5	HKY+G	2294.5321	4668.7584	SYM+I	2437.8540	4961.8725	SYM+I	2277.3986	4647.7738	HKY+G	2428.4743	4943.1131

| 6 || HKY+I || 2294.9425 || 4669.5792 || HKY+G || 2441.3669 || 4962.2703 ||
 SYM+I+G || 2275.5124 || 4650.6425 || SYM+I+G || 2423.6311 || 4946.6828 |

Generated by: Native Rust Column-Compressed Felsenstein ModelTest Engine (modeltest.rs) * Native Rust Gotoh 3-State Affine-Gap & 9-mer Anchor Multiple Sequence Aligner (phylo.rs) * MAFFT v7.526 (L-INS-i)

6. Focal Alignment Architecture & Phylogenetic Tree Reconstruction

Below is the focal alignment comparison across the 5.8S rRNA anchor boundary and adjacent ITS2 Helix III transition, illustrating why locking the invariant 158 bp 5.8S core preserves bimodal rate-heterogeneity estimation across all four factorial combinations, whereas excising the 5.8S anchor causes both aligners to exceed 46.6% gap saturation (312 bp, 47.12% gaps in Native Gotoh vs. 309 bp, 46.60% gaps in MAFFT L-INS-i) and collapses downstream model selection onto uninformative uniform/invariant noise models.

Generated by: Native Rust Column-Compressed Felsenstein ModelTest Engine (modeltest.rs) * Native Rust Gotoh 3-State Affine-Gap & 9-mer Anchor Multiple Sequence Aligner (phylo.rs) * MAFFT v7.526 (L-INS-i)

Generated by: Native Rust Column-Compressed Felsenstein ModelTest Engine (modeltest.rs) * Native Rust Gotoh 3-State Affine-Gap & 9-mer Anchor Multiple Sequence Aligner (phylo.rs) * MAFFT v7.526 (L-INS-i)

Generated by: Native Rust Column-Compressed Felsenstein ModelTest Engine (modeltest.rs) * Native Rust Gotoh 3-State Affine-Gap & 9-mer Anchor Multiple Sequence Aligner (phylo.rs) * MAFFT v7.526 (L-INS-i)

Discussion & Key Takeaways

This 2 x 2 factorial synthesis provides a direct empirical answer to whether substituting native Rust alignment and model-selection algorithms for external legacy binaries (MAFFT v7.526 and jModelTest v2.1.9) alters downstream phylogenomic conclusions:

* Complete Independence on Protein-Coding Loci: On ungapped protein-coding barcodes such as vertebrate mitochondrial COX1 (1,545 bp), Cell 1 (Native MSA + Native ModelTest) and Cell 2 (MAFFT + Native ModelTest) yield identical log-likelihoods to four decimal places ($-\ln L = 4780.6745$, $r = 1.0000$), and all four cells select HKY+I with 267.74x faster execution in the all-native stack.

* Predictable Likelihood Translation Under Ribosomal Loop Compression: On multi-operon bacterial 16S rRNA and fungal ITS1-5.8S-ITS2 cistrons, MAFFT L-INS-i packs hypervariable expansion loops into slightly fewer columns (1,803 vs. 1,868 bp and \$756\$ vs. 766 bp). While this shifts the absolute log-likelihood baseline by $\Delta \ln L = 330.8$ and 146.0 log-units respectively, both Native ModelTest and jModelTest v2.1.9 respond identically to the shift, maintaining 100% concordance on the HKY base model for 16S rRNA and the bimodal K80+G / K80+I rate-heterogeneity family for ITS1-5.8S-ITS2.

* Scope & Operational Guidance: Automated bioinformatic calibration surfaces statistical correlations and algorithmic performance boundaries across curated

reference cohorts. Defaulting to the all-native phylogenomic workflow eliminates 99.2% to 99.6% of wall-clock compute time across systematic sweeps while retaining full interoperability with external iterative aligners and likelihood optimizers for targeted cross-engine audits.

7. Provenance Ledger & Verified NCBI RefSeq Accessions

All 20 biological sequences analyzed across the 2 x 2 factorial calibration were retrieved from NCBI RefSeq and cached in content-addressable local storage (Database Verification: NCBI RefSeq Verified).

| Cohort | | Accession ID | | Organism | | Genomic Locus | | Length (bp) | | Database Verification |

| Cohort I | | NC_028302.1 | | Panthera leo | | Mitochondrial COX1 cds | | 1545 | | NCBI RefSeq Verified |

| Cohort I | | NC_010642.1 | | Panthera tigris | | Mitochondrial COX1 cds | | 1545 | | NCBI RefSeq Verified |

| Cohort I | | NC_010641.1 | | Panthera pardus | | Mitochondrial COX1 cds | | 1545 | | NCBI RefSeq Verified |

| Cohort I | | NC_022842.1 | | Panthera onca | | Mitochondrial COX1 cds | | 1545 | | NCBI RefSeq Verified |

| Cohort I | | NC_028313.1 | | Lynx canadensis | | Mitochondrial COX1 cds | | 1545 | | NCBI RefSeq Verified |

| Cohort I | | NC_001700.1 | | Felis catus | | Mitochondrial COX1 cds | | 1545 | | NCBI RefSeq Verified |

| Cohort I | | NC_002008.4 | | Canis lupus | | Mitochondrial COX1 cds | | 1545 | | NCBI RefSeq Verified |

| Cohort I | | NC_003427.1 | | Ursus arctos | | Mitochondrial COX1 cds | | 1545 | | NCBI RefSeq Verified |

| Cohort I | | NC_003428.1 | | Ursus maritimus | | Mitochondrial COX1 cds | | 1545 | | NCBI RefSeq Verified |

| Cohort I | | NC_003426.1 | | Ursus americanus | | Mitochondrial COX1 cds | | 1545 | | NCBI RefSeq Verified |

| Cohort II | | NC_000913.3 | | Escherichia coli K-12 MG1655 | | 16S rRNA rrsA , rrsB , rrsH operons | | 1542 | | NCBI RefSeq Verified |

| Cohort II | | NC_003197.2 | | Salmonella enterica LT2 | | 16S rRNA rrsA operon | | 1544 | | NCBI RefSeq Verified |

| Cohort II | | NC_000964.3 | | Bacillus subtilis subsp. subtilis 168 | | 16S rRNA rrnO operon | | 1554 | | NCBI RefSeq Verified |

| Cohort III | | NR_121481.1 | | Aspergillus fumigatus | | ITS1-5.8S-ITS2 rRNA cistron | | 596 | | NCBI RefSeq Verified |

| Cohort III | | NR_111250.1 | | Aspergillus flavus | | ITS1-5.8S-ITS2 rRNA cistron | | 595 | | NCBI RefSeq Verified |

| Cohort III | | NR_131276.1 | | *Aspergillus niger* | | ITS1-5.8S-ITS2 rRNA cistron |
| 596 | | NCBI RefSeq Verified |

| Cohort III | | NR_149334.1 | | *Candida albicans* | | ITS1-5.8S-ITS2 rRNA cistron |
| 536 | | NCBI RefSeq Verified |

| Cohort III | | NR_111007.1 | | *Saccharomyces cerevisiae* | | ITS1-5.8S-ITS2 rRNA
cistron | | 524 | | NCBI RefSeq Verified |

Generated by: NCBI Entrez E-Utilities (`efetch.fcgi`) * CRV Science Labs
PoliteCacheManager (`fetch.rs`) * Native Rust Column-Compressed Felsenstein
ModelTest Engine (`modeltest.rs`)

8. Primary Literature & Methodological References

* Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* , 30(4), 772-780. <https://doi.org/10.1093/molbev/mst010>

* Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* , 9(8), 772. <https://doi.org/10.1038/nmeth.2109>

* Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* , 59(3), 307-321. <https://doi.org/10.1093/sysbio/syq010>

* Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of Molecular Biology* , 162(3), 705-708. [https://doi.org/10.1016/0022-2836\(82\)90398-9](https://doi.org/10.1016/0022-2836(82)90398-9)

* Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* , 17(6), 368-376. <https://doi.org/10.1007/BF01734359>

* Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* , 39(3), 306-314. <https://doi.org/10.1007/BF00160154>