Statistical Orthogonality Between Thermodynamic Predictions and Experimental DNA Melt Profiles

Authors: Haoxiang Sun, Qinguo Liu

Affiliation: Westlake University, Hangzhou, Zhejiang, China

Abstract

Universal Digital High-Resolution Melt (U-dHRM) assays provide a rapid, non-sequencing-based platform for infectious disease screening and microbial identification [1]. However, clinical implementation currently relies on extensive supervised machine learning models trained on empirical species-specific data [1]. Bridging experimental melt curve space to canonical sequence-based phylogenetic trees remains a major computational challenge.

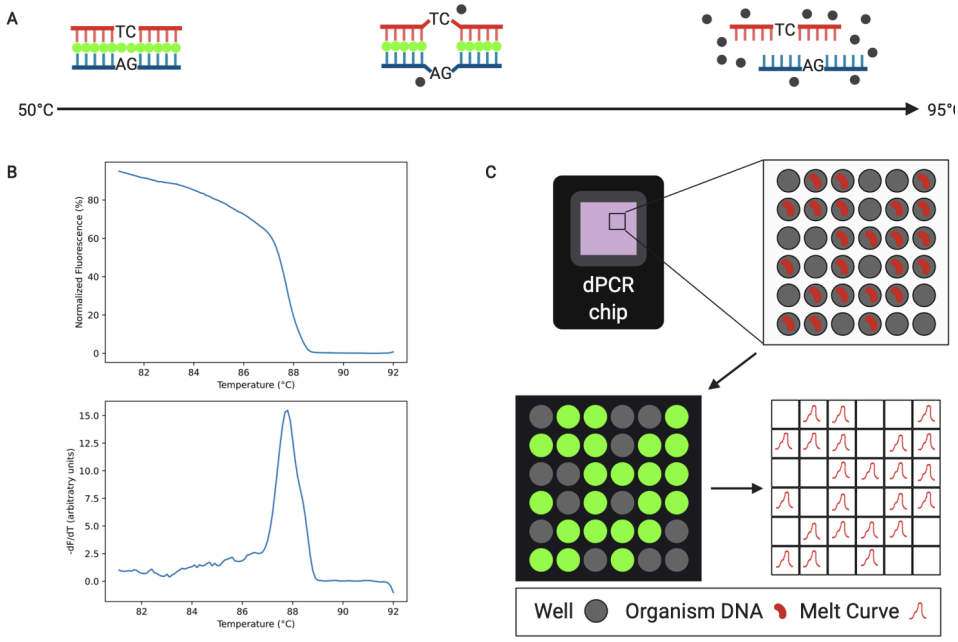


Figure 1. U-dHRM leverages melting dynamics of double-stranded DNA to detect differences in sequences [1].

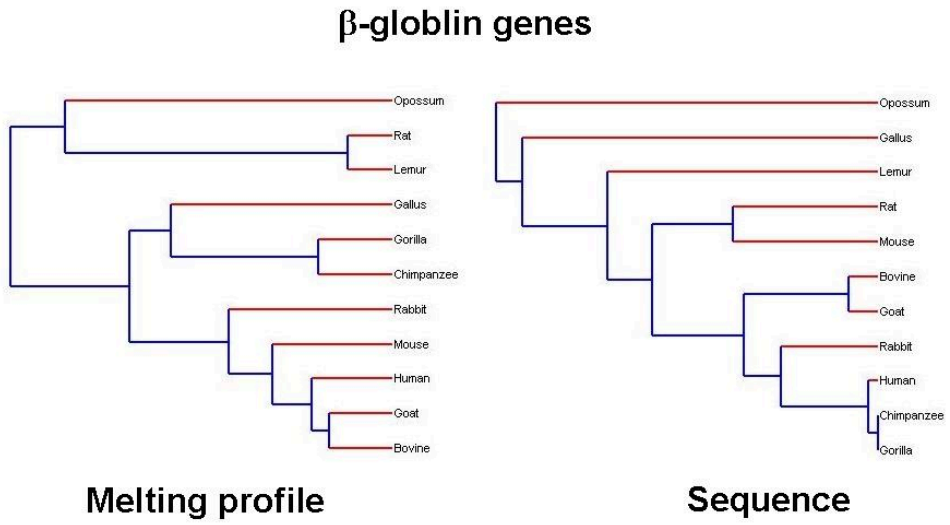
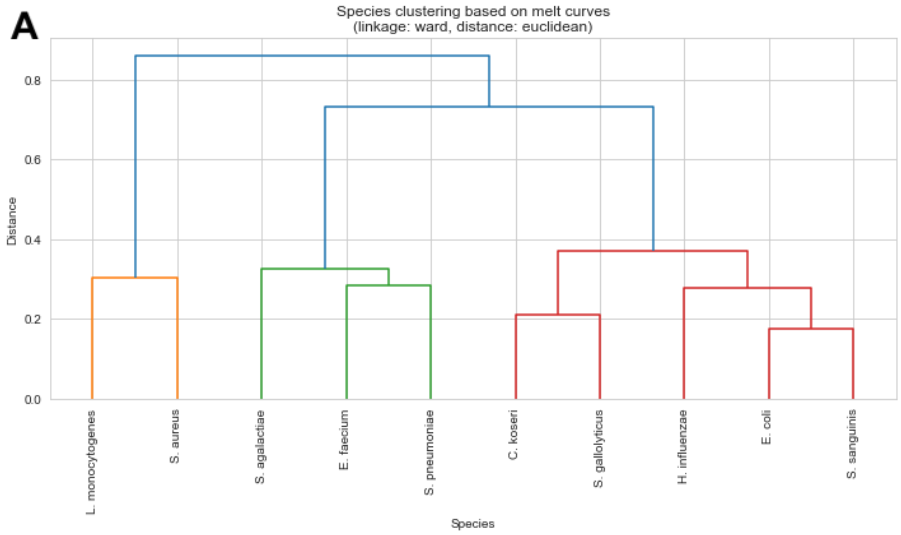
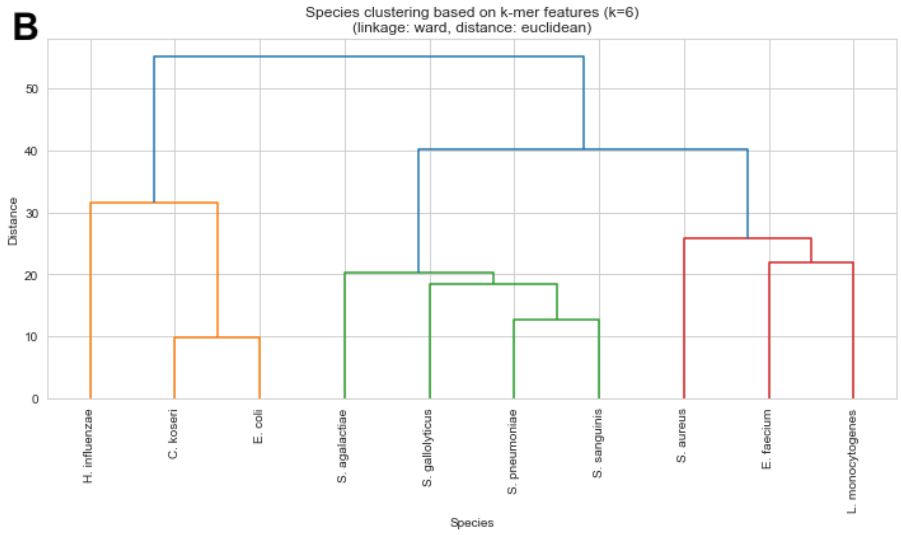


Figure 2. The melting profile was previously characterized to be capable of generating phylogenetic trees, albeit different from the ground truth generated by sequence [2].

In this study, we evaluate whether sequence-to-structure feature transformations—specifically Weak/Strong (W/S) k-mer representations and thermodynamic modeling—can bridge the gap between primary 16S rRNA gene sequence space (SILVA reference database), thermodynamic predictions (uMelt), and empirical U-dHRM experimental melt curves across 10 bacterial pathogen species. Using matrix-level Mantel test correlations, Fast Fourier Transforms (FFT), dimensional reduction (PCA and t-SNE), and dependent correlation significance testing via Steiger's Z-test, we quantify the structural relationship across these feature spaces.


A
Species clustering based on melt curves
(linkage: ward, distance: euclidean)
Distance
0.0
0.2
0.4
0.6
0.8
L. monocytogenes
S. aureus
S. agalactiae
E. faecium
S. pneumoniae
C. koseri
S. gallolyticus
H. influenzae
E. coli
S. sanguinis
Species


B
Species clustering based on k-mer features (k=6)
(linkage: ward, distance: euclidean)
Distance
0
10
20
30
40
50
H. influenzae
C. koseri
E. coli
S. agalactiae
S. gallolyticus
S. pneumoniae
S. sanguinis
S. aureus
E. faecium
L. monocytogenes
Species

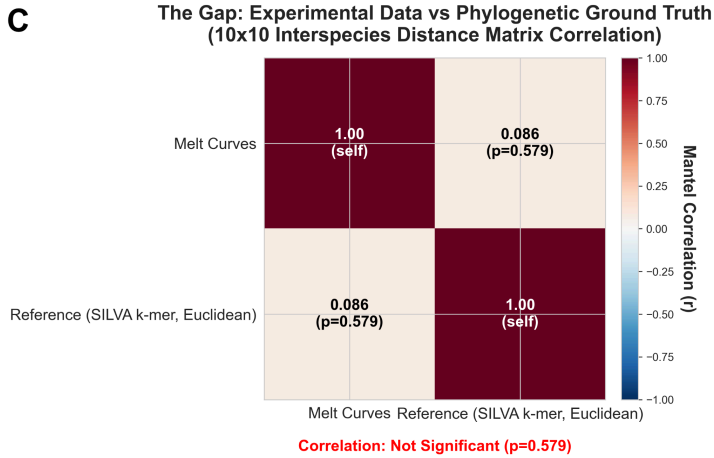

C
The Gap: Experimental Data vs Phylogenetic Ground Truth
(10x10 Interspecies Distance Matrix Correlation)
Melt Curves
Reference (SILVA k-mer, Euclidean)
1.00
(self)
0.086
(p=0.579)
0.086
(p=0.579)
1.00
(self)
Mantel Correlation (r)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Melt Curves
Reference (SILVA k-mer, Euclidean)
Correlation: Not Significant (p=0.579)

Figure 3. The evolutionary dendrograms created by melt curves (A) & corresponding reference sequences (B, ground truth) are remarkably different, with virtually no correlation between them.

We show that while W/S k-mer representations exhibit a minor descriptive increase in correlation with thermodynamic predictions ($\Delta r=+0.0237$) and synthetic melting curves ($\Delta r=+0.1102$) compared to raw SILVA sequences, this advantage is statistically non-significant ($p=0.4331$ and $p=0.2514$, respectively). Furthermore, experimental melt profiles remain strictly orthogonal to sequence-derived features. These findings characterize key limitations in current nearest-neighbor thermodynamic modeling and outline necessary constraints for future sequence-to-melt predictive frameworks.

Keywords: Universal Digital High-Resolution Melt (U-dHRM), 16S rRNA, DNA Thermodynamics, Weak/Strong Nucleotide Encoding, Mantel Test, Steiger's Z-Test.

Materials and Methods

2.1 Dataset and Species Selection

Experimental U-dHRM melting curves and corresponding 16S rRNA reference sequences were collected for 10 distinct bacterial species: Citrobacter koseri, Escherichia coli, Enterococcus faecalis, Haemophilus influenzae, Listeria monocytogenes, Klebsiella pneumoniae, Proteus mirabilis, Pseudomonas aeruginosa, Streptococcus pneumoniae, and Staphylococcus aureus.

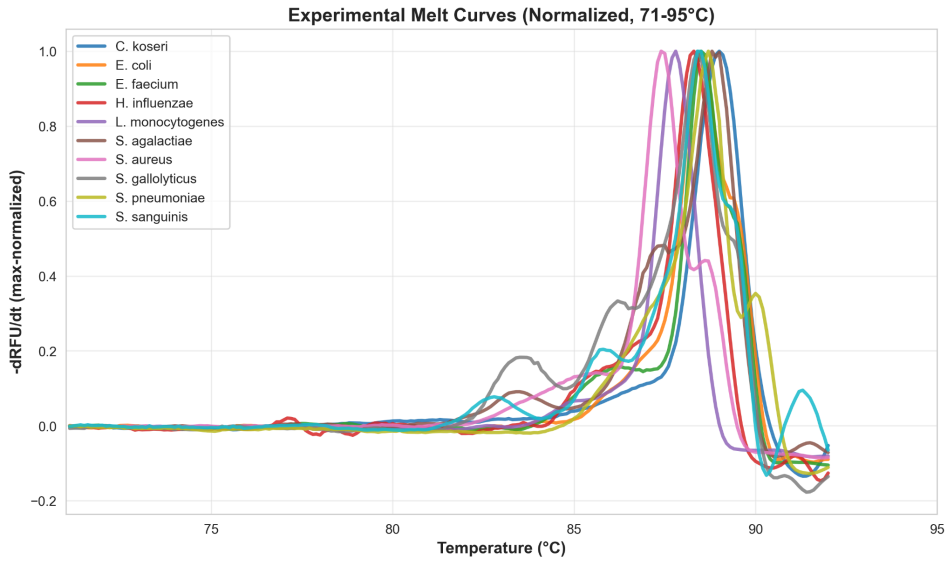


Figure 4. Average of experimental U-dHRM melt curves for each of the 10 reference species [1].

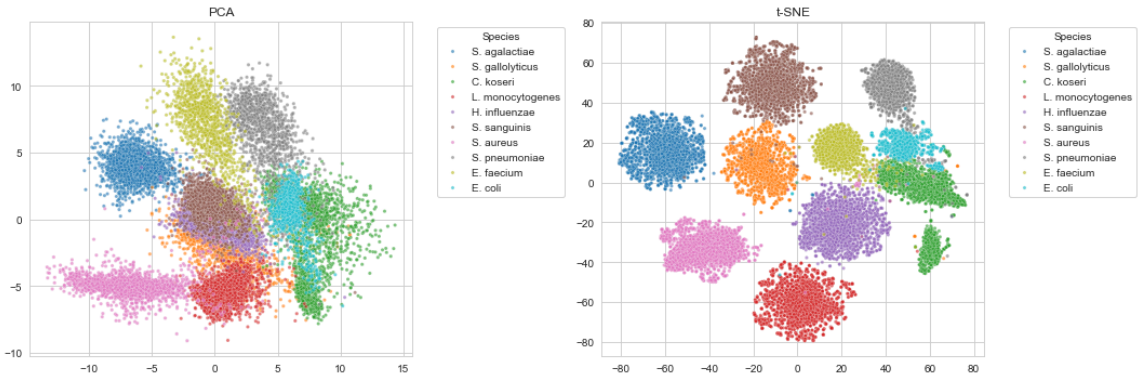


Figure 5. PCA & t-SNE clustering of melt curves [1].

2.2 Feature Space Transformations

To capture structural and functional properties beyond raw primary sequences, 5 distinct feature representations were generated:

Reference SILVA k-mers: Primary 16S sequence representations calculated via Euclidean distances of normalized k-mer frequency distributions (k=6).

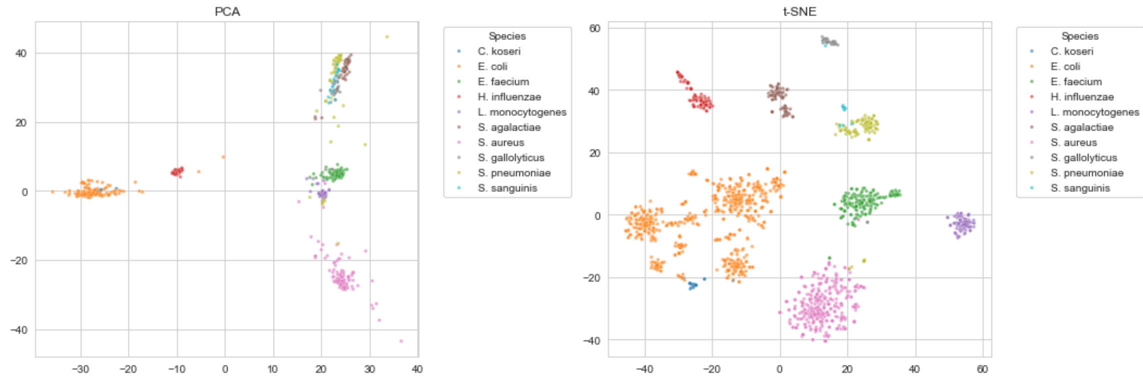


Figure 6. PCA & t-SNE clustering of k-mer (k=6) features of all discovered SILVA sequences [3].

Weak/Strong (W/S) k-mers: DNA sequences mapped to a binary alphabet where A, T→W and G, C→S, followed by k-mer frequency vector generation.

A

```
======================================================================
CONVERTING SEQUENCES TO W/S ALPHABET
======================================================================
Original sequences: 1670
Converted to W/S alphabet: 1670

--- Example Conversions ---

Species: C. koseri
Original (first 60 bp):  TGTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCAT
W/S format (first 60):   WSWSWSSSWWWSWSSSWSWWSSWSSSSSWWWWSWWSWSSWWWSSSWWSSWWWWSSSSWW

Species: C. koseri
Original (first 60 bp):  TGTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCAT
W/S format (first 60):   WSWSWSSSWWWSWSSSWSWWSSWSSSSSWWWWSWWSWSSWWWSSSWWSSWWWWSSSSWW

Species: C. koseri
Original (first 60 bp):  TGTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCAT
W/S format (first 60):   WSWSWSSSWWWSWSSSWSWWSSWSSSSSWWWWSWWSWSSWWWSSSWWSSWWWWSSSSWW

--- Overall W/S Composition ---
W (Weak - A/T): 744,125 (46.74%)
S (Strong - G/C): 847,926 (53.26%)
Total bases: 1,592,051
```

B

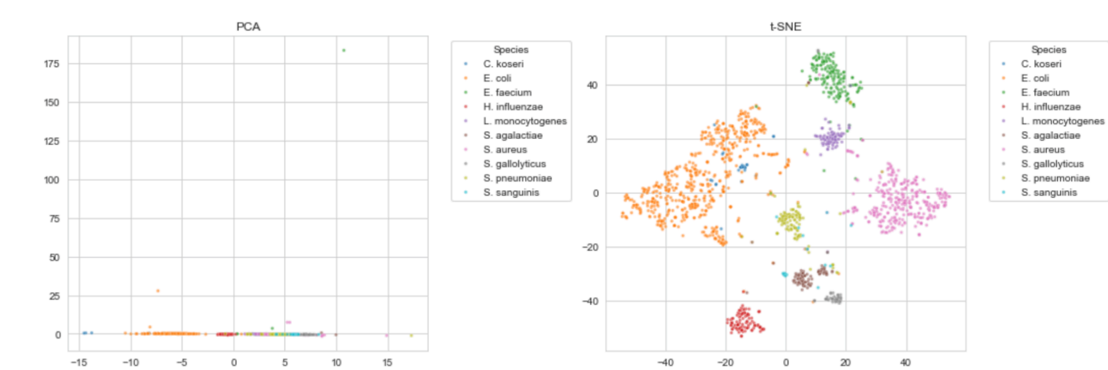


Figure 7. Logic of ATGC-to-WS conversion (A) & PCA & t-SNE clustering of k-mer (k=6) features of the ATGC-to-WS-conversion of all discovered SILVA sequences (B).

Fast Fourier Transform (FFT) Features: Experimental melt curves (dF/dT over 71°C to 95°C) converted to frequency domain spectra to isolate fundamental shape harmonics.

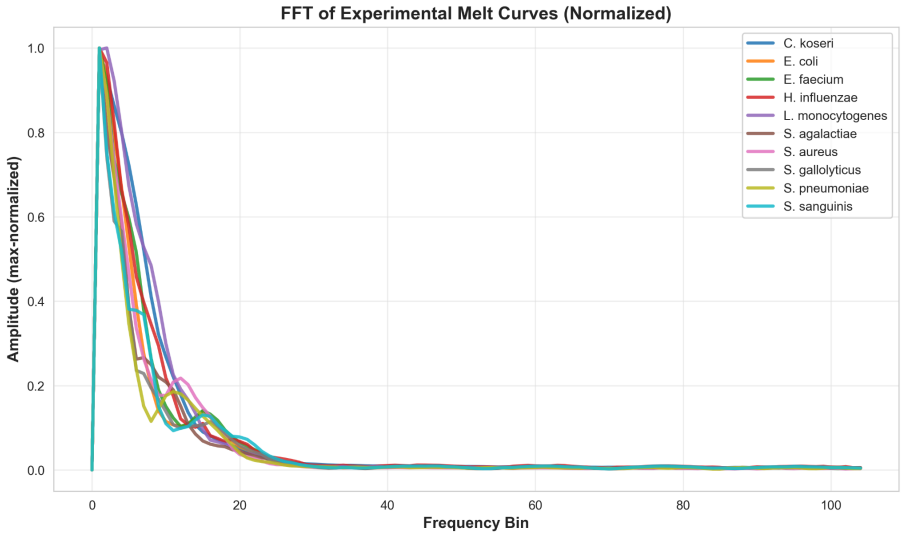


Figure 8. Average of the FFT of experimental U-dHRM melt curves for each of the 10 reference species.

uMelt Thermodynamic Parameters: Simulated denaturation enthalpy (ΔH), entropy (ΔS), free energy (ΔG), and melting transitions (Tm) computed via nearest-neighbor thermodynamic algorithms [4].

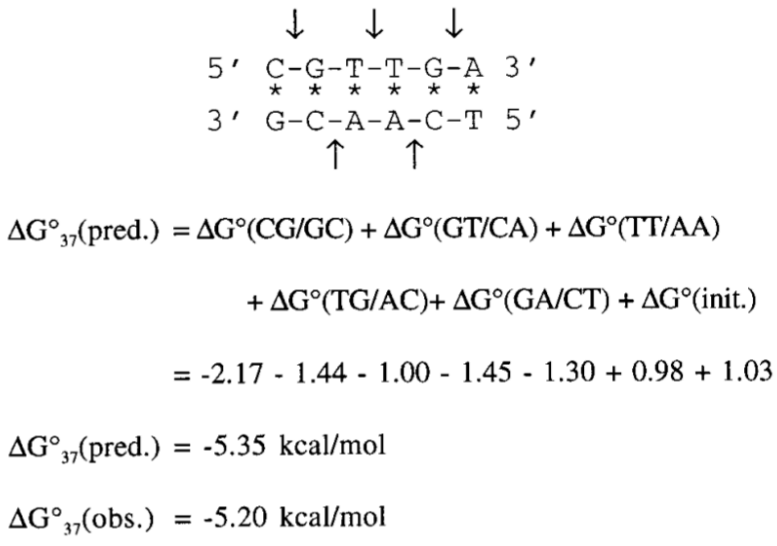


Figure 9. Diagram of nearest-neighbor thermodynamic model assumptions used in this simulation [4].

Synthetic Melt Curves: Simulated derivative melting curves generated synthetically using the uMelt v3 software [5].

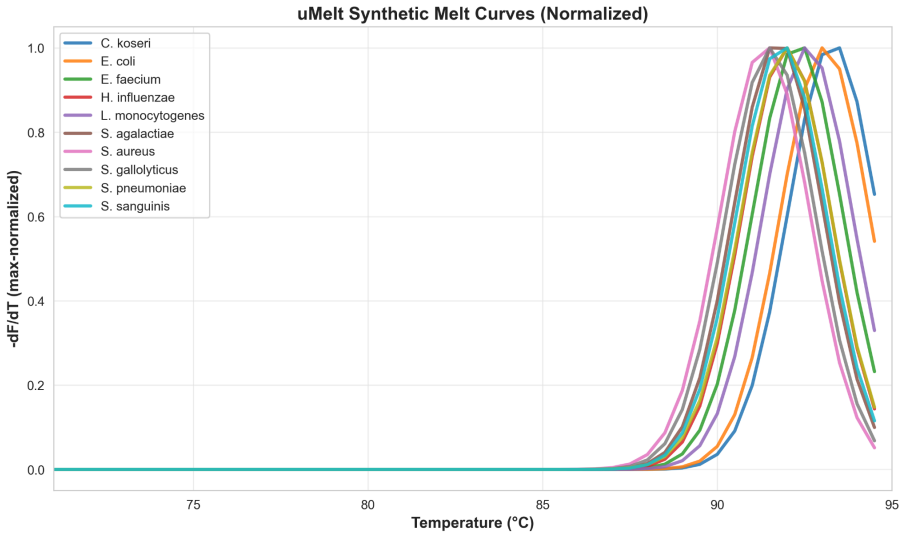


Figure 10. Average of the uMelt v3 simulated synthetic melt curves for each of the 10 reference species [5].

2.3 Matrix Similarity and Mantel Tests

Interspecies distance matrices (10×10) were constructed for each feature representation. Non-parametric Mantel permutation tests (999 permutations) were performed to calculate the matrix correlation coefficient (r) between pair-wise distance representations:

$$r = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where xi and yi represent upper-triangular elements of the distance matrices, and N=n*(n-1)/2 for n=10 species (N=45).

2.4 Steiger's Z-Test for Dependent Correlations

To evaluate whether W/S k-mer representations (X1) correlate significantly better with thermodynamic/synthetic features (Y) than primary SILVA sequences (X2), we performed Steiger's Z-test for overlapping dependent correlations measured on the same species sample set:

$$Z_1 = \frac{1}{2}\ln\left(\frac{1+r_{1y}}{1-r_{1y}}\right), \quad Z_2 = \frac{1}{2}\ln\left(\frac{1+r_{2y}}{1-r_{2y}}\right)$$

The standard error incorporating the inter-feature correlation r12=r(X1,X2) is defined by:

$$\mathrm{SE} = \sqrt{\frac{2(1-r_{12})}{n-3}\cdot h}$$

Where:

$$h = \frac{1-f\bar{r}^2}{1-\bar{r}^2}, \quad \bar{r}^2 = \frac{r_{1y}^2 + r_{2y}^2}{2}, \quad \text{and} \quad f = \frac{1-r_{12}}{2(1-\bar{r}^2)}$$

The test statistic Z=(Z1-Z2)/SE was evaluated against a standard normal distribution under the one-tailed hypothesis H1:r1y>r2y.

Results

3.1 Classification and Novelty Detection Benchmarks

Supervised machine learning classifiers evaluated on empirical U-dHRM data demonstrated high classification accuracy and novelty detection ROC AUC across species, replicating previously claimed results in the U-dHRM publication.

Melt Curve Analysis Summary: Classification & Novelty Detection Results

Species	Cls Train Size	Cls Test Size	Cls Fit Time (s)	Cls Predict Time (s)	Cls Accuracy	Nov Train Size	Nov Unknown Test Size	Nov Fit Time (s)	Nov Predict Time (s)	Nov ROC AUC
C. koseri	1828	16453	0.765	0.128	0.9913	1622	2052	0.672	0.112	0.9533
E. coli	1828	16453	0.765	0.128	0.9144	1743	843	0.890	0.202	0.9796
E. faecium	1828	16453	0.765	0.128	0.9717	1714	1137	0.843	0.207	0.9928
H. influenzae	1828	16453	0.765	0.128	0.9961	1601	2265	0.683	0.115	0.9990
L. monocytogenes	1828	16453	0.765	0.128	0.9970	1603	2244	0.699	0.102	0.9374
S. agalactiae	1828	16453	0.765	0.128	0.9995	1603	2249	0.719	0.113	0.9965
S. aureus	1828	16453	0.765	0.128	0.9852	1625	2028	0.727	0.100	0.9804
S. gallolyticus	1828	16453	0.765	0.128	0.9867	1627	2006	0.700	0.103	0.9947
S. pneumoniae	1828	16453	0.765	0.128	0.9956	1702	1255	0.718	0.121	0.9837
S. sanguinis	1828	16453	0.765	0.128	0.9945	1607	2202	0.704	0.112	0.9963
OVERALL AVG	**1828**	**16453**	**0.765**	**0.128**	**0.9832**	**1644**	**1828**	**0.735**	**0.129**	**0.9814**

Classification Metrics Novelty Detection Metrics

Figure 11. Classification & leave-one-out cross-validation-based novelty detection tests on the experimental U-dHRM melt curves show high accuracy & ROC AUC [1].

3.3 Mantel Matrix Correlations and Feature Space Alignment

Pairwise Mantel test comparisons across all feature types generated a comprehensive distance correlation matrix. Feature similarity relative to canonical 16S SILVA sequences revealed a clear gradient: W/S k-mers exhibited the highest affinity to SILVA ($r>0.95$), followed by uMelt thermodynamic parameters and synthetic melt curves. Experimental melt curves and FFT-transformed features demonstrated virtually zero correlation ($r\approx0.05$) with sequence-based representations.

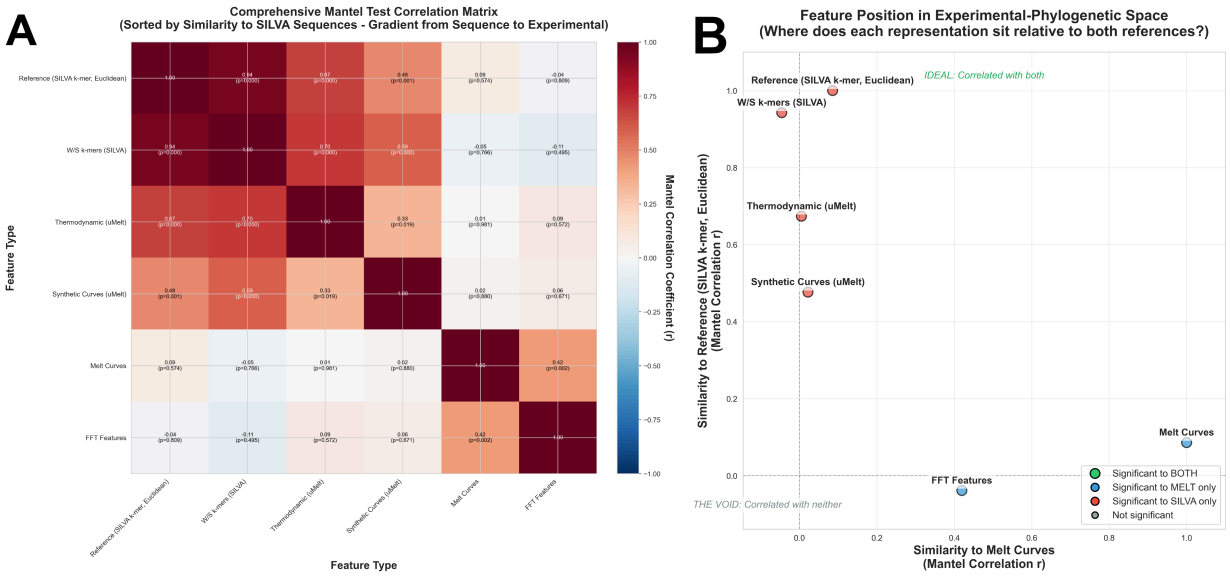


Figure 12. Mantel test comparing between the dendrograms/similarity matrices generated using each type of transform, shown in heatmap (A) & scatterplot (B) formats.

3.4 Statistical Evaluation of Weak/Strong (W/S) Transformation Advantage

To test whether converting primary sequences into hydrogen-bonding W/S encodings provides a statistically significant advantage in capturing thermodynamic parameters, Steiger's Z-tests were executed.

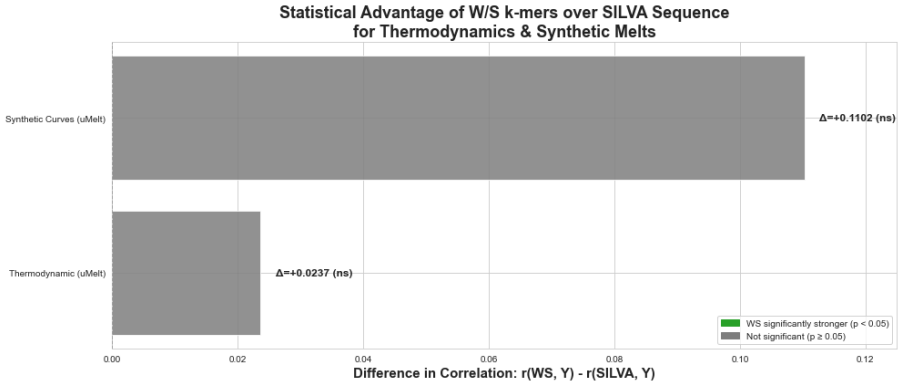


Figure 13. Steiger's Z-test results of the correlation between dendrograms/similarity matrices generated using ATGC or WS references with the dendrograms/similarity matrices generated using thermodynamic & uMelt simulations.

For uMelt thermodynamic predictions (ΔH,ΔS,ΔG,Tm), W/S k-mers achieved a correlation of r=0.6973 compared to r=0.6737 for raw SILVA sequences (Δr=+0.0237, Z=0.168, p=0.4331, non-significant). For synthetic melt curves, W/S k-mers achieved r=0.5866 versus r=0.4764 for SILVA sequences (Δr=+0.1102,Z=0.670,p=0.2514, non-significant).

Discussion

The primary objective of this study was to determine whether simple chemical transformations (W/S encodings) could bridge the gap between sequence-based phylogenetic references and physical DNA melting curves.

Our statistical evaluation reveals two major insights:

1. **Experimental-Sequence Orthogonality:** Empirical U-dHRM melt curves remain strictly orthogonal to sequence-derived thermodynamic predictions. This discrepancy stems from factors unmodeled in standard nearest-neighbor algorithms, including ionic strength fluctuations, dye-binding kinetics, tertiary DNA conformations, and microfluidic thermal dissipation.
2. **Non-Significance of W/S Advantage**: Although W/S k-mer encodings slightly increase linear correlation with nearest-neighbor thermodynamic simulations, Steiger's Z-test confirms that this increase is statistically non-significant ($p>0.05$). This indicates that binary hydrogen-bonding models alone do not capture the higher-order interactions present in thermodynamic algorithms.

Discussion

While W/S sequence transformations fail to statistically bridge the gap between primary 16S sequences and physical melt curves, this work provides a rigorous baseline characterizing the limitations of current thermodynamic models. Future efforts aiming to enable sequence-free U-dHRM diagnostic predictions must incorporate deep geometric learning architectures or explicit physical microfluidic parameters rather than relying on linear sequence-to-thermodynamic mappings.

Furthermore, the power of the negative results shown in this study are severely limited by the availability of 16S rRNA reference sequences. Because the reference sequences used to generate the U-dHRM curves are never given in the original publication [1], I used the collection of 16S rRNA sequences of the 10 species of interest inside the SILVA database [3], which has significant data heterogeneity that likely blurred the melt curve-reference sequence correlation (**Figure 6**).

Nr.	Species	Amplicon Length	Melt Curves (Nr.)	Mean Peak (°C)	Std. Peak (°C)	Median DTW
1	*Citrobacter koseri*	969	2052	89.06	0.42	2.31
2	*Enterococcus faecium*	978	1137	88.53	0.10	1.68
3	*Escherichia coli*	981	843	88.49	0.13	2.60
4	*Haemophilus influenzae*	967	2265	88.29	0.10	2.14
5	*Listeria monocytogenes*	994	2244	87.79	0.08	2.23
6	*Staphylococcus aureus (MSSA)*	981	2028	87.43	0.19	1.51
7	*Streptococcus gallolyticus*	978	2006	88.43	0.78	3.83
8	*Streptococcus 'group B' (GBS)*	978	2249	88.90	0.10	2.37
9	*Streptococcus pneumoniae*	978	1255	88.71	0.15	1.62
10	*Streptococcus sanguinis*	972	2202	88.45	0.09	3.00
	Average	977	1828	88.41	0.21	2.33

Figure 14. Expected amplicon size of the U-dHRM targets, reference sequences or source databases are not given in the original publication [1].

References

1. Langouche, L. (2021). Advancing Rapid Infectious Disease Screening Using a Combined Experimental/Computational Approach. Ph.D. dissertation, University of California San Diego.

2. Reese, E., & Krishnan, V. V. (2010). Classification of DNA sequences based on thermal melting profiles. Bioinformation, 4(10), 463–470.
3. Chuvochina, M., et al. (2026). "SILVA in 2026: a global core biodata resource for rRNA within the DSMZ digital diversity." Nucleic Acids Research.
4. SantaLucia J. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proceedings of the National Academy of Sciences. 95(4):1460–1465. doi:10.1073/pnas.95.4.1460.
5. Dwight Z. uMelt – Melting curve prediction software for single, batch, and multiplex PCR products [Online]. DNA-UTAH.ORG. Available from: https://dna-utah.org/umelt/