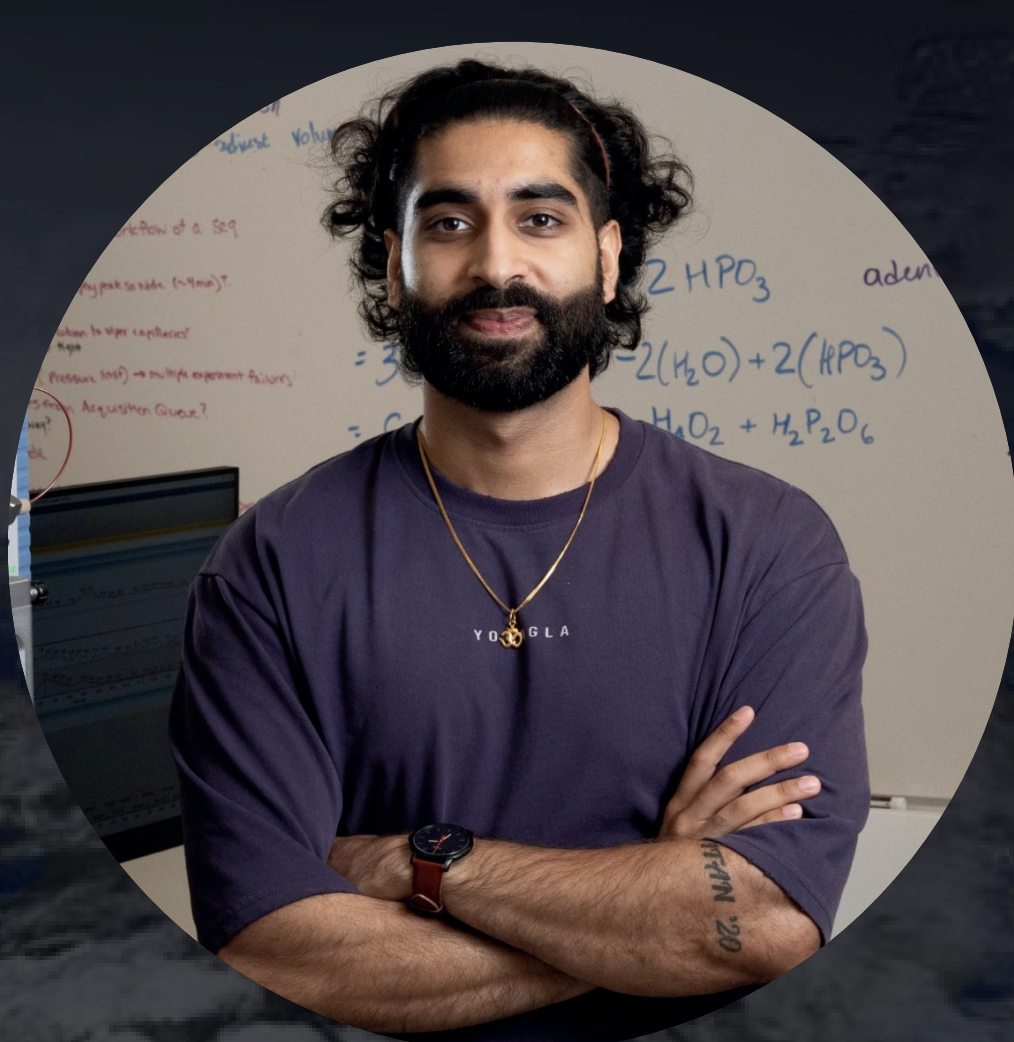




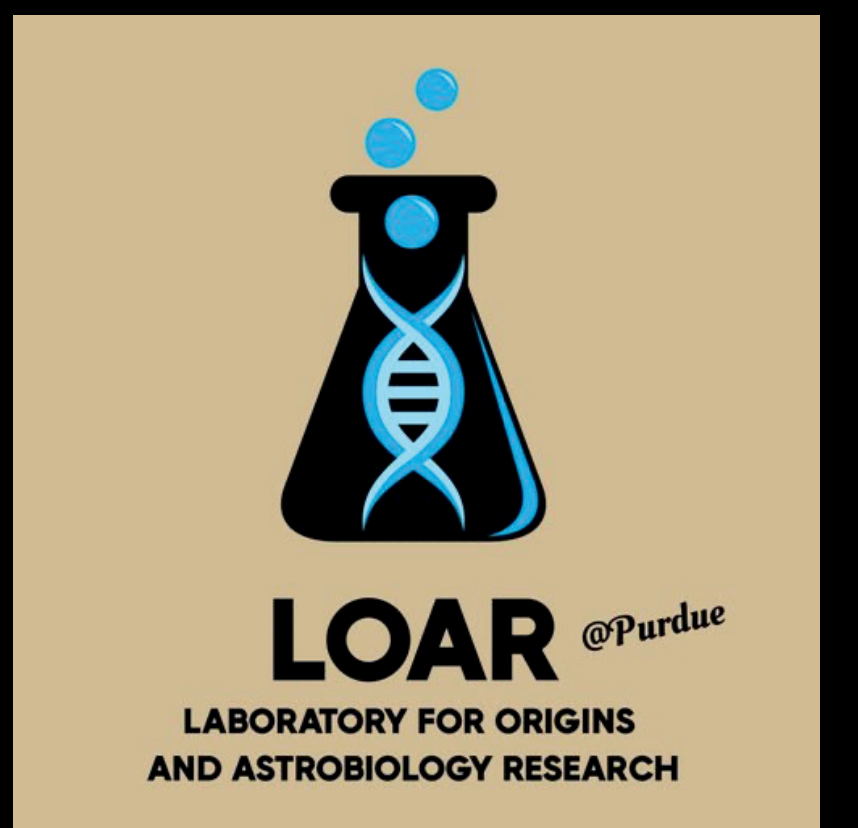
Manuscript in-prep
References can be
shared upon request

Why Did Life Choose A, C, G, U, and T in RNA and DNA?

Ishaan Madan and Ben K.D. Pearce



Department of Earth, Atmospheric,
and Planetary Sciences



BACKGROUND

- **Life uses a small genetic alphabet.** Five canonical nucleobases—A, T, C, G, and U—underpin information storage in DNA and RNA.
- **Prebiotic chemistry offered more possibilities.** Laboratory experiments and asteroid samples reveal diverse organic molecules, including alternatives to canonical bases.
- **Why these five?** What chemical factors could have favored their incorporation into life's genetic alphabet?
- **Could formation energetics help explain this selection?** We compare nucleotide formation free energies across five canonical and sixteen noncanonical base families.

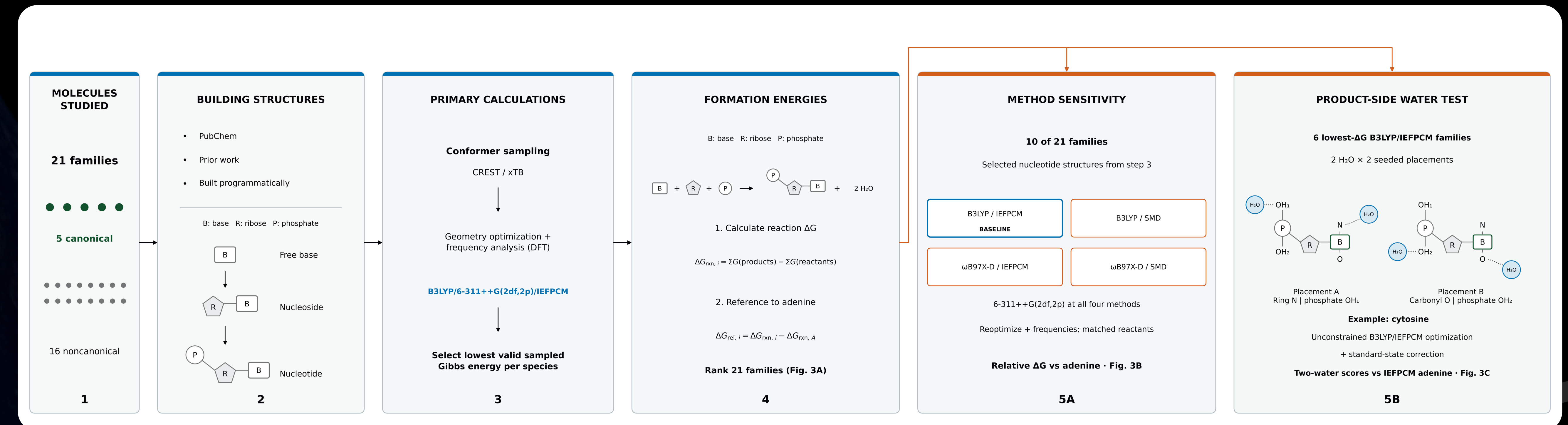


Figure 2. Computational workflow for comparing nucleotide formation energetics. Structures from 21 base families—five canonical and sixteen noncanonical—underwent conformer sampling, geometry optimization, and frequency analysis. The lowest valid sampled Gibbs energy for each species was used to calculate reaction free energies relative to adenine. Sensitivity tests examined four computational treatments for ten families and two explicit-water placements around nucleotide products for six families.

Figure 1. Potential stages of selection for life's genetic alphabet. Ten organizational levels span nucleobase formation, nucleotide and polymer assembly, replication, and biological function.

POTENTIAL LEVELS OF SELECTION	LITERATURE VERDICT
FUNCTION & LOCK-IN L8–L10	
Enzymatic Replication, Translation & LUCA	L10: Alphabet fixed; code optimization may reinforce lock-in, not select the original bases.
DNA Specialization & Repair Lock-in	L9: DNA uses T, so U from C deamination can be recognized as damage.
Folding, Catalysis, Function	L8: Noncanonical nucleosides expand function without replacing the core alphabet.
REPLICATION L7	
Nonenzymatic Copying & Extension	L7: Copying favors ribose and 3'-5' linkages; A/I/s ² U/s ² C copies more uniformly.
POLYMERS L5–L6	
Duplex Stability & Mismatch Discrimination	L6: Alternative alphabets can match or exceed canonical duplex stability and selectivity.
Oligomer / Polymer Assembly	L5: Polymer assembly creates backbone constraints; base identity remains largely unconstrained.
MONOMERS L1–L4	
Self-assembly / Base-pairing	L4: Canonical nucleobases do not pair in water; noncanonical systems enrich β-anomer.
Nucleotide Assembly / Persistence	L3: No privileged canonical formation-energy region; rankings overlap with noncanonical chemistry.
Nucleoside Assembly / Persistence	L2: Noncanonical bases form nucleosides readily; canonical formation is unresolved.
Nucleobase Assembly / Persistence	L1: Canonical bases recur, but show no consistent assembly or persistence advantage.

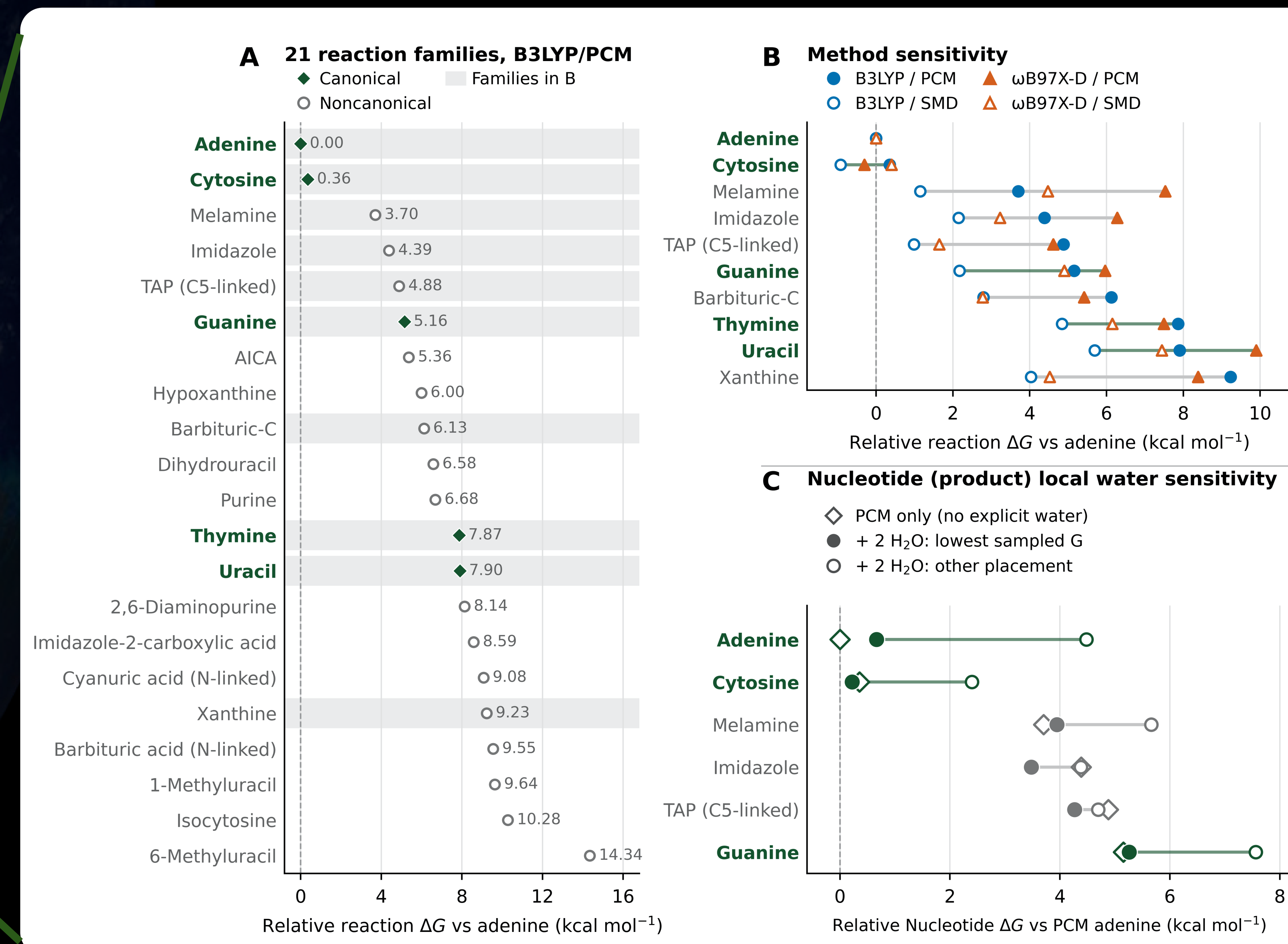


Figure 3. Nucleotide energetics and sensitivity. (A) Formation free energies for 21 neutral families at B3LYP/PCM, relative to adenine. Shading marks families tested in B. (B) Ten families compared across four computational treatments, each referenced to its corresponding adenine value. (C) Six nucleotide products compared without explicit water and with two placements of two water molecules, relative to PCM-only adenine. Lines connect treatments or placements, not uncertainty intervals. Canonical families appear in green; energies are in kcal mol⁻¹.

KEY TAKEAWAYS

Adenine and cytosine consistently led in the main comparisons.

Formation energetics *did not* uniquely favor the canonical alphabet.

Rankings and energy differences varied with computational treatment.

WHY THIS MATTERS FOR CONSCIOUSNESS

This work explores the chemical foundations of biological information processing; an early step in the evolutionary history that ultimately gave rise to cognition and consciousness.

Research & Contact

