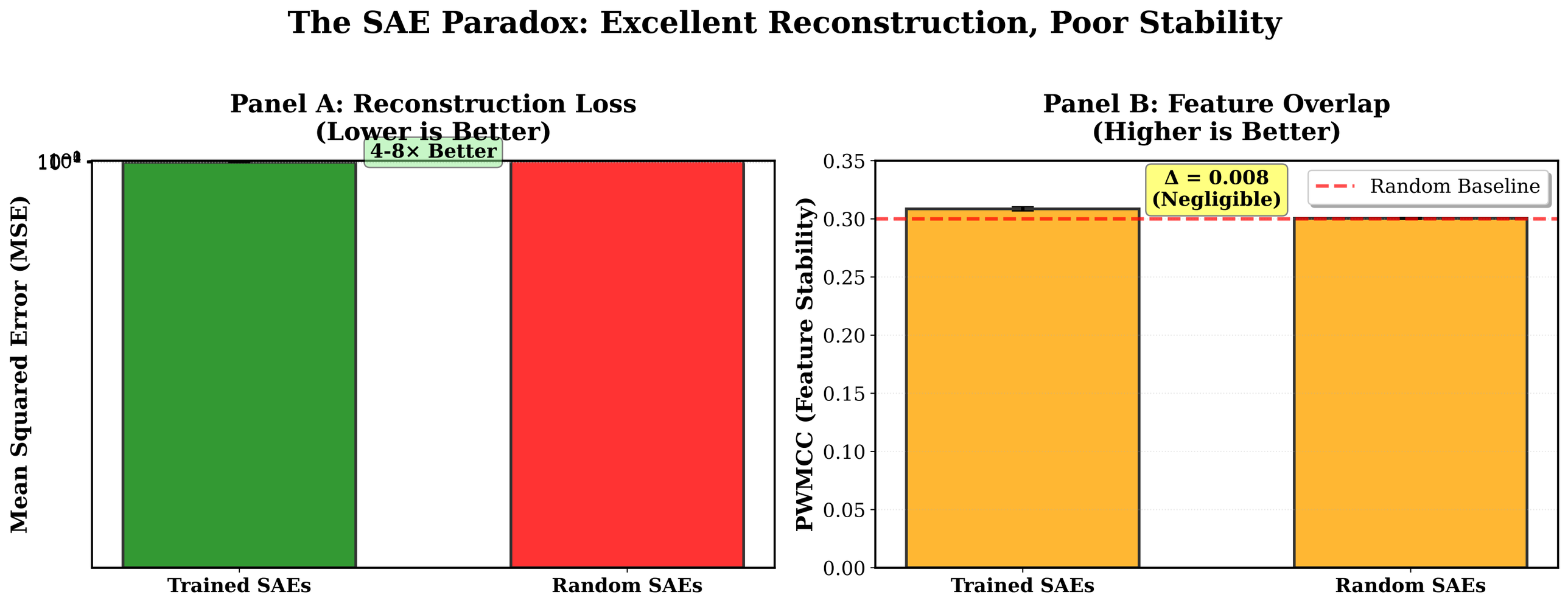
The SAE Paradox: Excellent Reconstruction, Poor Stability
Panel A: Reconstruction Loss
(Lower is Better)
4-8× Better
Mean Squared Error (MSE)
Trained SAEs
Random SAEs
Panel B: Feature Overlap
(Higher is Better)
Δ = 0.008
(Negligible)
Random Baseline
PWMCC (Feature Stability)
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Trained SAEs
Random SAEs