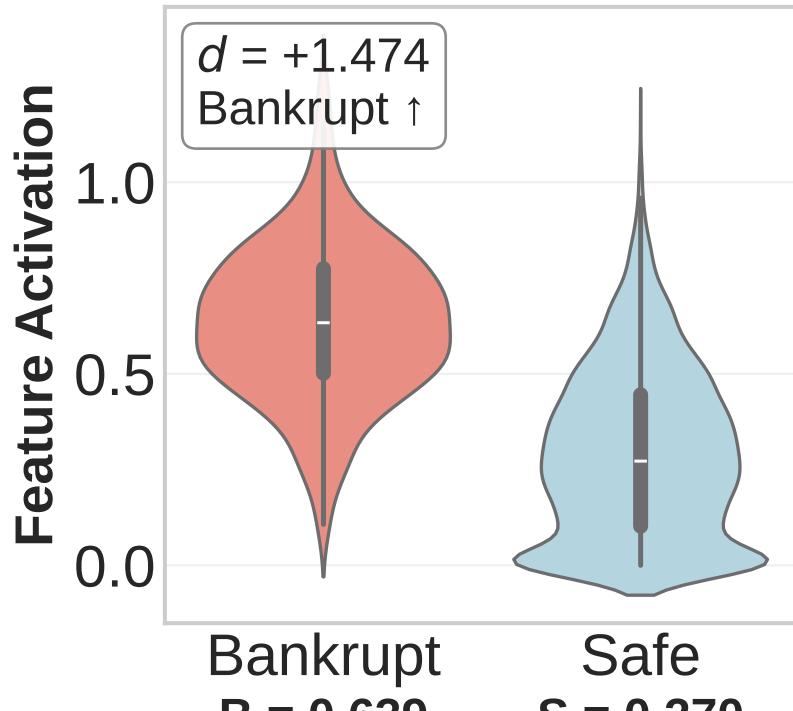
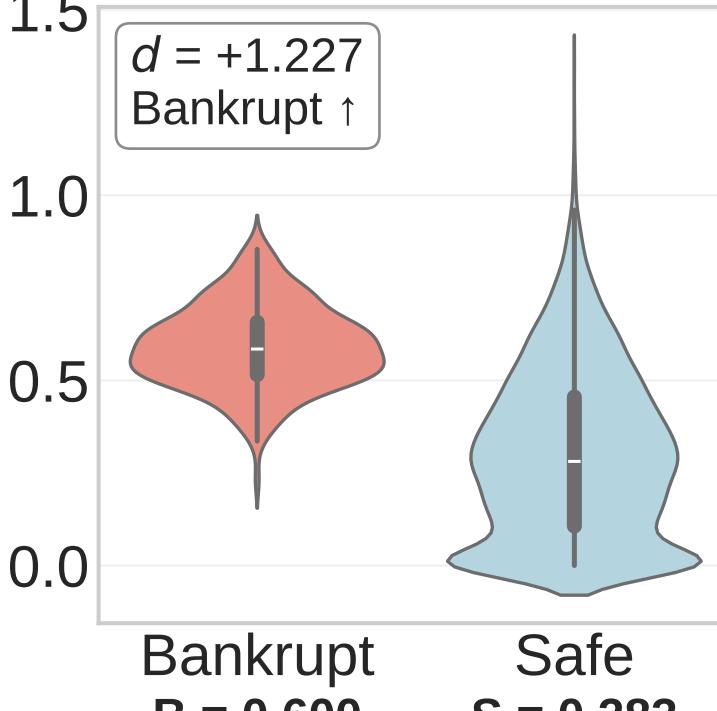


SAE Feature Separation Analysis

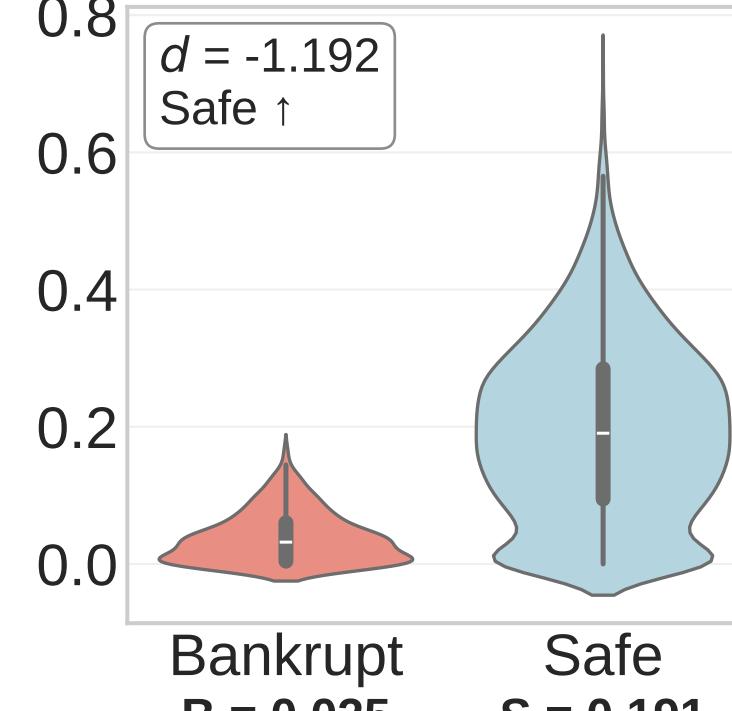
Layer 25 Feature 13464



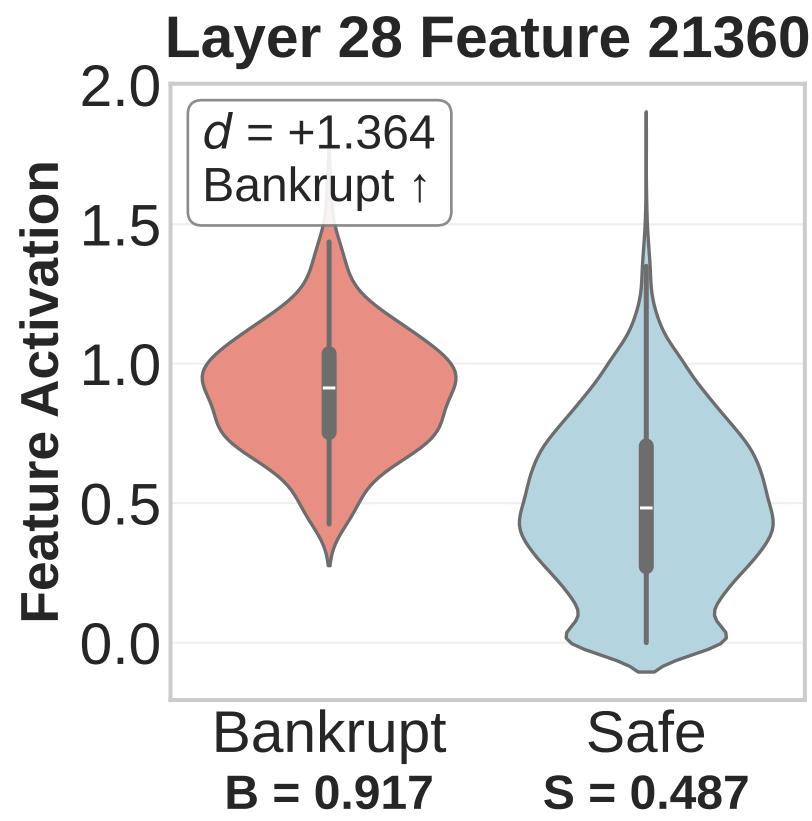
Layer 26 Feature 6259



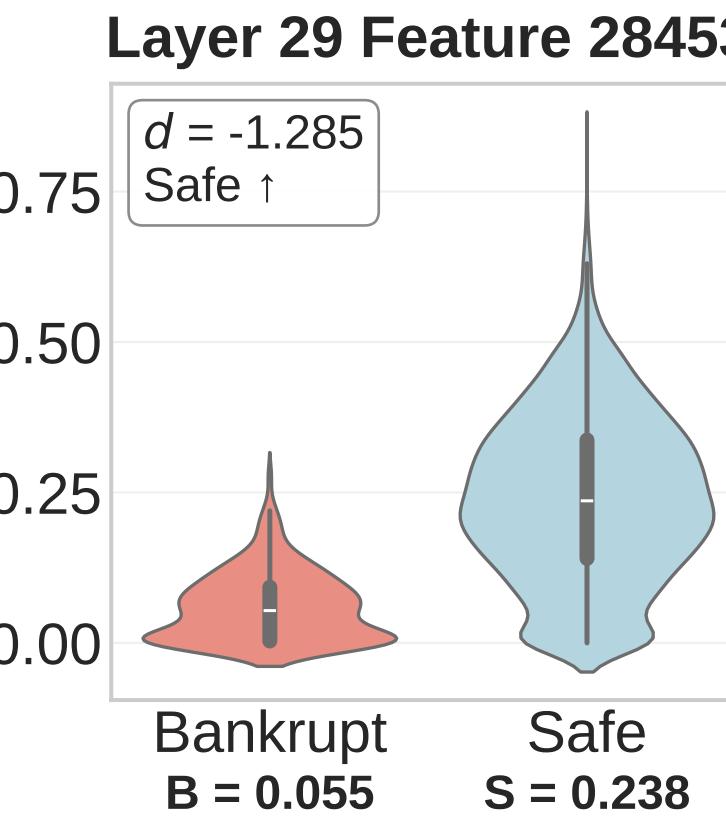
Layer 27 Feature 9866



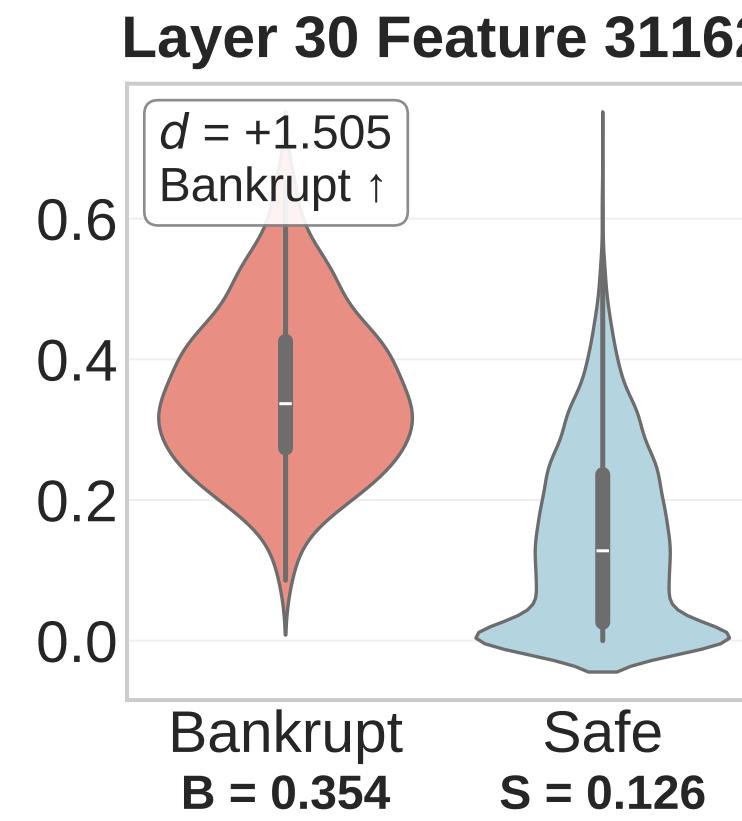
Layer 28 Feature 21360



Layer 29 Feature 28453



Layer 30 Feature 31162



Layer 31 Feature 22798

