Random Baseline Validation (Layer 12, dim=3072)

Legend:
- Random directions (N=100)
- Random mean (0.498)
- 95th percentile (0.857)
- Empathy probe (1.000)

$z = 2.09\sigma$
$p < 0.05$

X-axis: AUROC
Y-axis: Frequency