

Distribution of Toxicity Scores in Simulated Data

Toxicity Class anti-toxic neutral toxic

