## TVD Distributions on trainset Oracle 1 vs Oracle 2, $\mu$ =0.506 3.0 Human vs GPT2, $\mu$ =0.678 Human vs GPT2 FT, $\mu$ =0.628 Human vs GPT2 FTsub, $\mu$ =0.631 2.5 Density 1.5 1.0 0.5 0.0 0.2 0.4 0.6 8.0