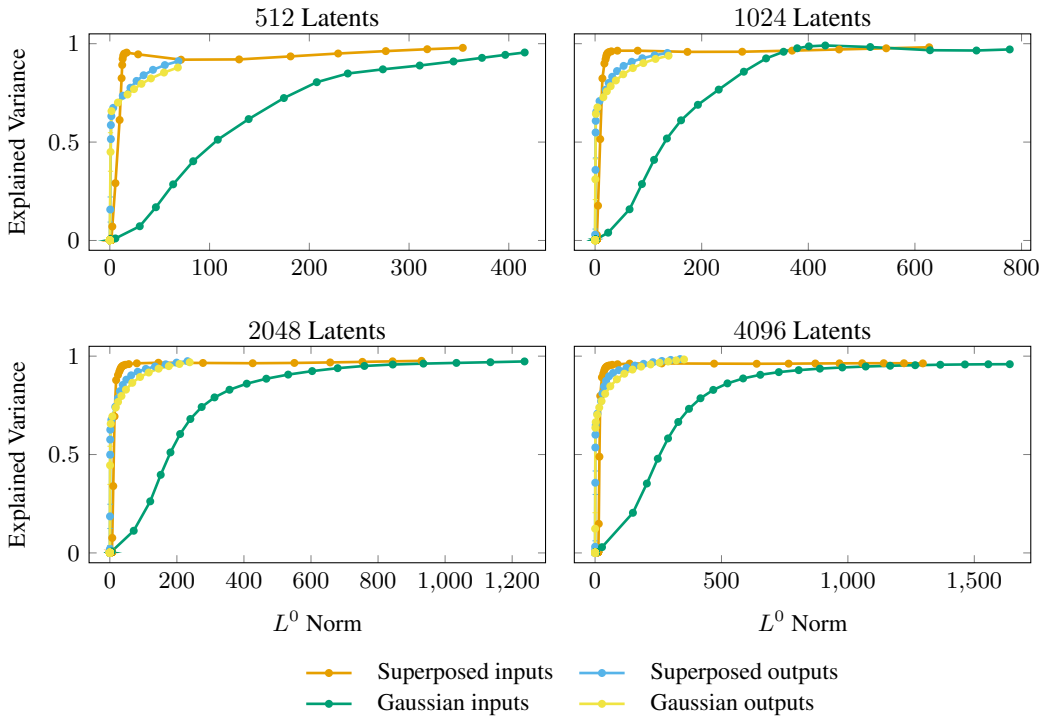

Figure 22: Pareto frontiers of the explained variance against the L^0 norm (sparsity) for toy datasets generated to exhibit superposition, Gaussian controls with the same mean and variance, and the corresponding outputs when these are passed to a randomly initialized two-layer MLP.

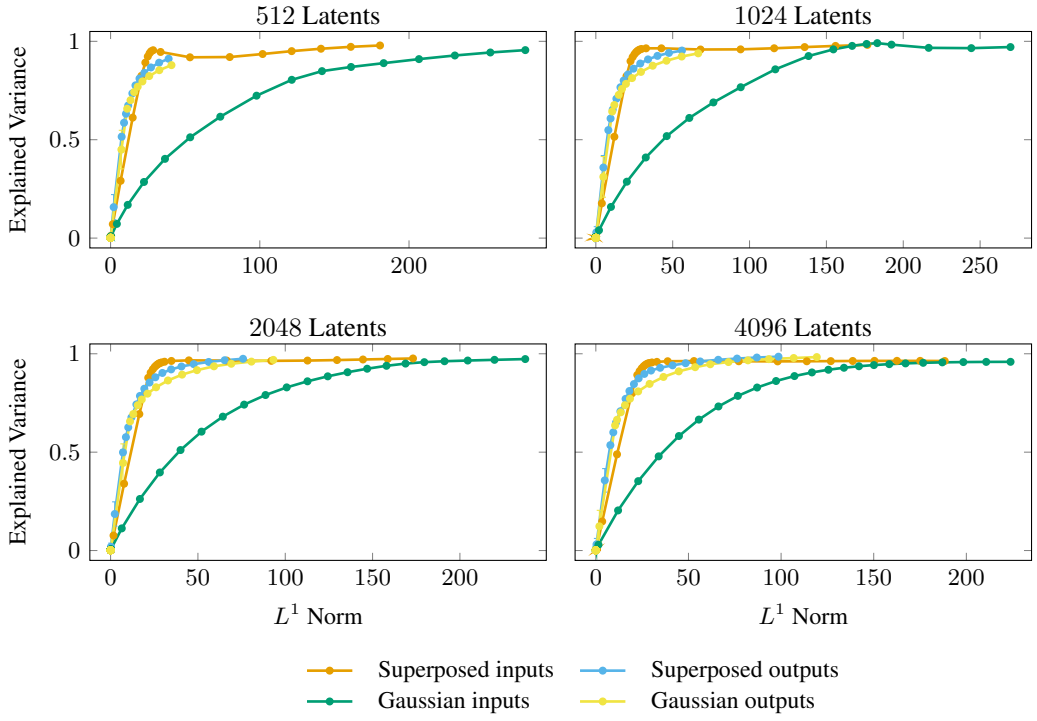

Figure 23: Pareto frontiers of the explained variance against the L^1 norm (sparsity) for toy datasets generated to exhibit superposition, Gaussian controls with the same mean and variance, and the corresponding outputs when these are passed to a randomly initialized two-layer MLP.

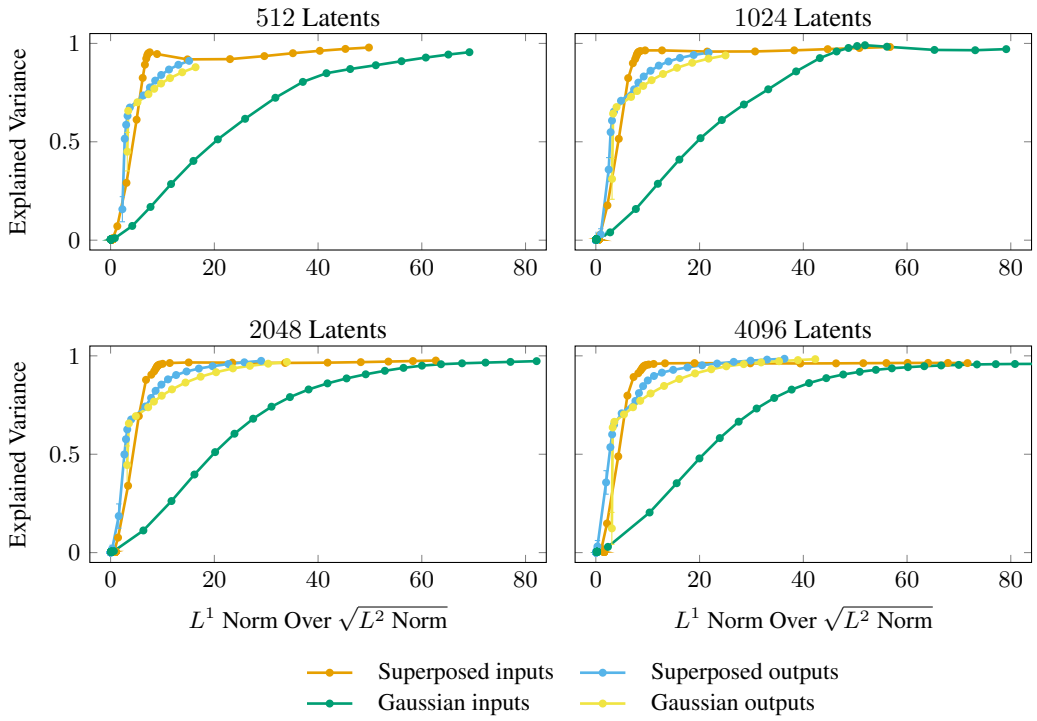

Figure 24: Pareto frontiers of explained variance against the L^1 norm over the square root of the L^2 norm (sparsity) for toy datasets generated to exhibit superposition, Gaussian controls with the same mean and variance, and the corresponding outputs when these are passed to a randomly initialized two-layer MLP.

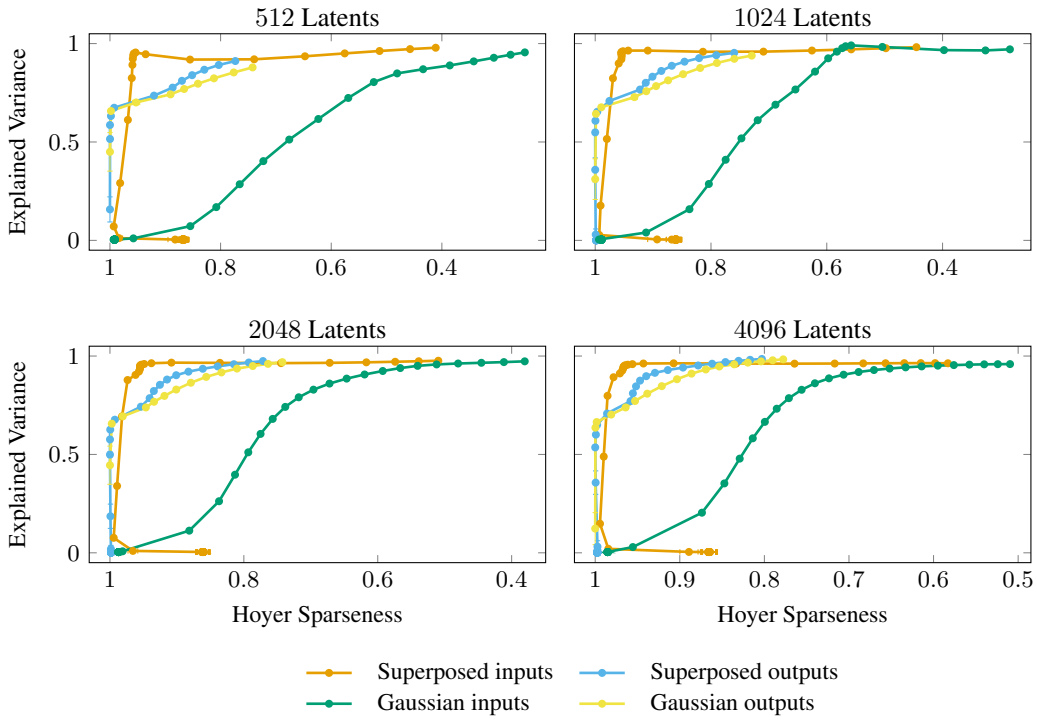

Figure 25: Pareto frontiers of explained variance against the Hoyer sparseness (sparsity) for toy datasets generated to exhibit superposition, Gaussian controls with the same mean and variance, and the corresponding outputs when these are passed to a randomly initialized two-layer MLP.