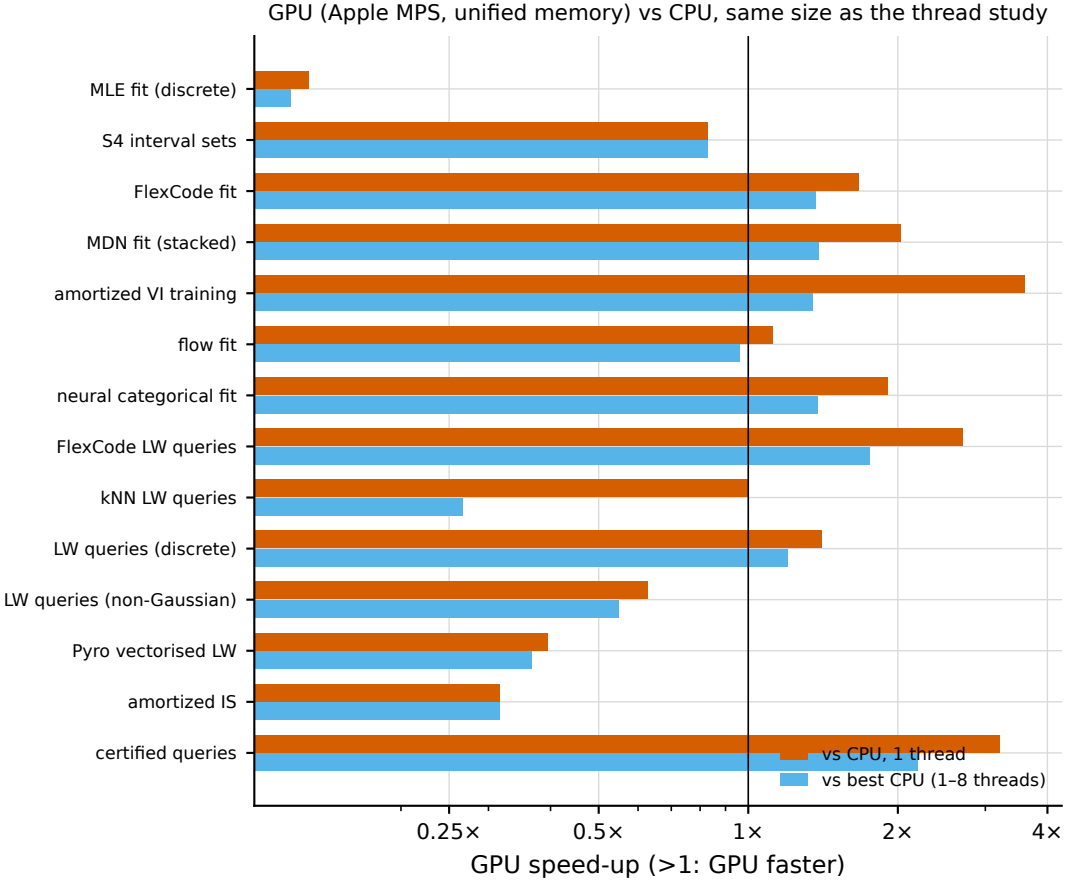

GPU (Apple MPS, unified memory) vs CPU, same size as the thread study
MLE fit (discrete)
S4 interval sets
FlexCode fit
MDN fit (stacked)
amortized VI training
flow fit
neural categorical fit
FlexCode LW queries
kNN LW queries
LW queries (discrete)
LW queries (non-Gaussian)
Pyro vectorised LW
amortized IS
certified queries
vs CPU, 1 thread
vs best CPU (1–8 threads)
0.25×
0.5×
1×
2×
4×
GPU speed-up (>1: GPU faster)