Why full-space KL can be large while held-out prediction is stable





