

Figure 1 is a line plot showing the performance of various methods on the Pain dataset. The x-axis is labeled 'pain' and ranges from 0.0 to 20.0. The y-axis represents performance, with a scale from 0.0 to 1.0. Three lines represent different values of  $p$ :  $p=0.1$  (orange),  $p=0.5$  (green), and  $p=0.9$  (black). The plot shows that performance generally increases with 'pain' and is higher for larger values of  $p$ . The Hamid-20 method is the top performer across all  $p$  values.

Method	$p=0.1$	$p=0.5$	$p=0.9$
Hamid-20	12.8	12.8	19.0
UDInfoDFP-47	9.5	9.5	9.5
huds sna-29	8.0	8.0	8.0
UDInfoSEP-46	7.5	7.5	7.5
run2-32	7.0	7.0	7.0
UDInfoDFP-47	6.5	6.5	6.5
run1-31	6.0	6.0	6.0
CriticRunBiAm-21	5.5	5.5	5.5
UDInfoSEP-46	5.0	5.0	5.0
WaterlooBaseline-50	4.5	4.5	4.5
run3-13	4.0	4.0	4.0
UDInfoDFP-47	3.5	3.5	3.5
huds sna-29	3.0	3.0	3.0
run2-32	2.5	2.5	2.5
QUExpT-39	2.0	2.0	2.0
QUExpT-39	1.5	1.5	1.5
run1-11	1.0	1.0	1.0
QUExpT-39	0.5	0.5	0.5
run2-12	0.2	0.2	0.2
UDInfoDFP-47	0.1	0.1	0.1
runBM25-43	0.0	0.0	0.0

