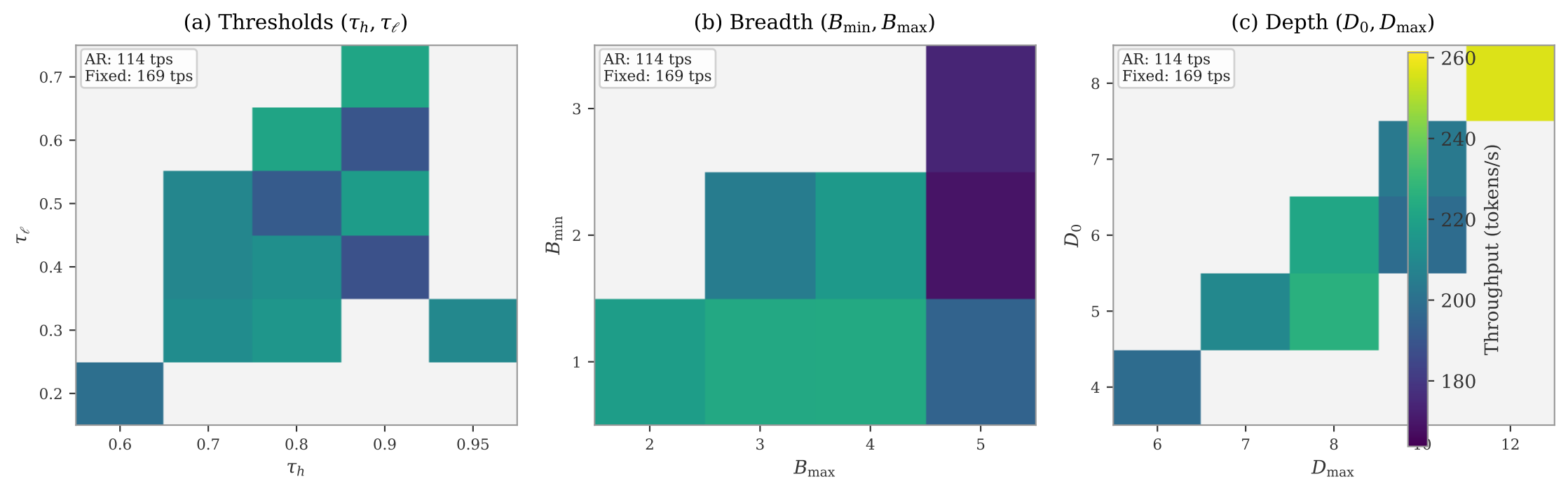
(a) Thresholds (τ_h, τ_ℓ)
AR: 114 tps
Fixed: 169 tps
τ_ℓ
0.7
0.6
0.5
0.4
0.3
0.2
0.6
0.7
0.8
0.9
0.95
τ_h
(b) Breadth ($B_{\min}, B_{\max}$)
AR: 114 tps
Fixed: 169 tps
$B_{\min}$
3
2
1
2
3
4
5
$B_{\max}$
(c) Depth ($D_0, D_{\max}$)
AR: 114 tps
Fixed: 169 tps
D_0
8
7
6
5
4
6
7
8
12
$D_{\max}$
260
240
220
200
180
Throughput (tokens/s)