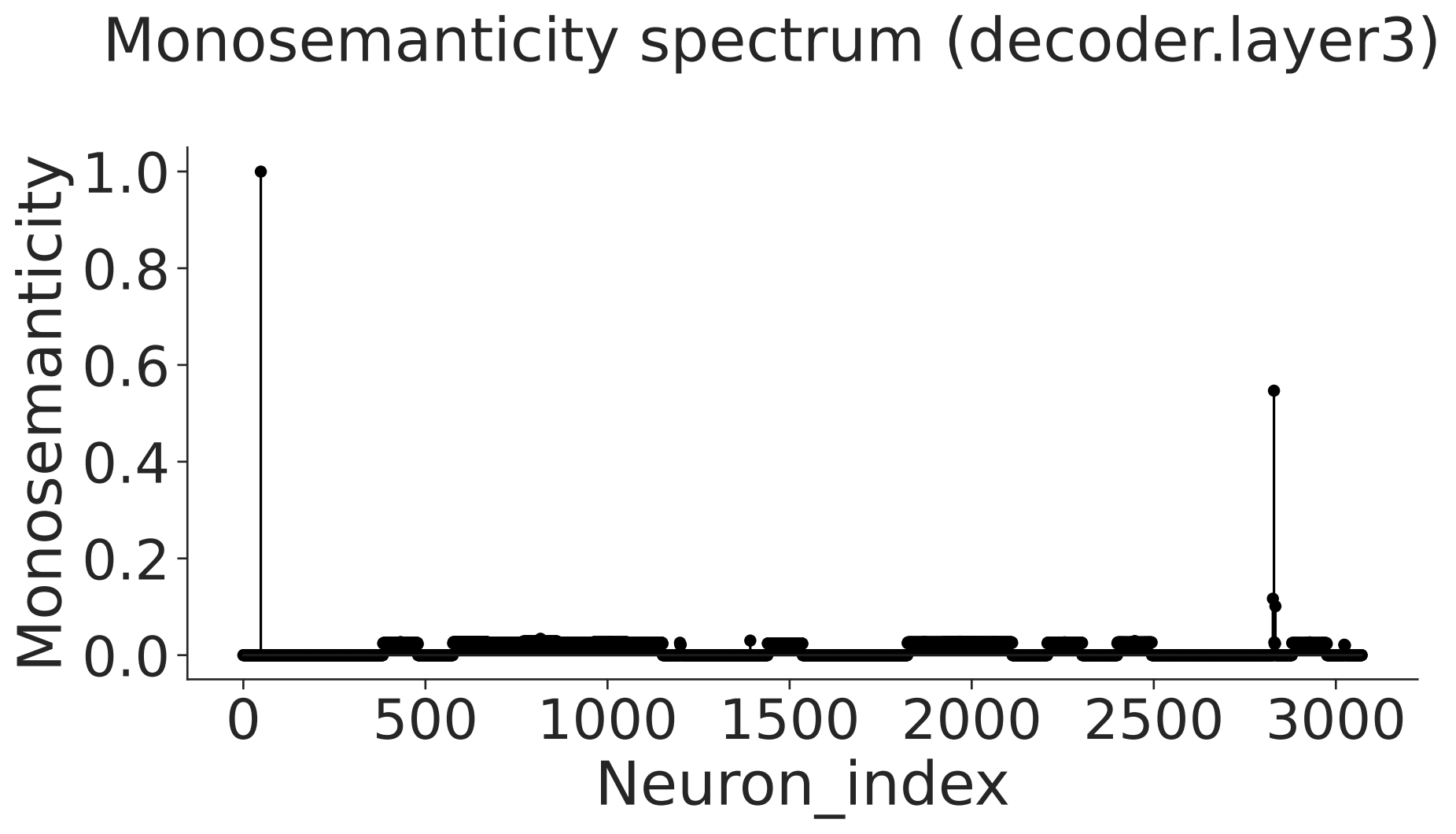
Monosemanticity spectrum (decoder.layer3)
Monosemanticity
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
3000
Neuron_index