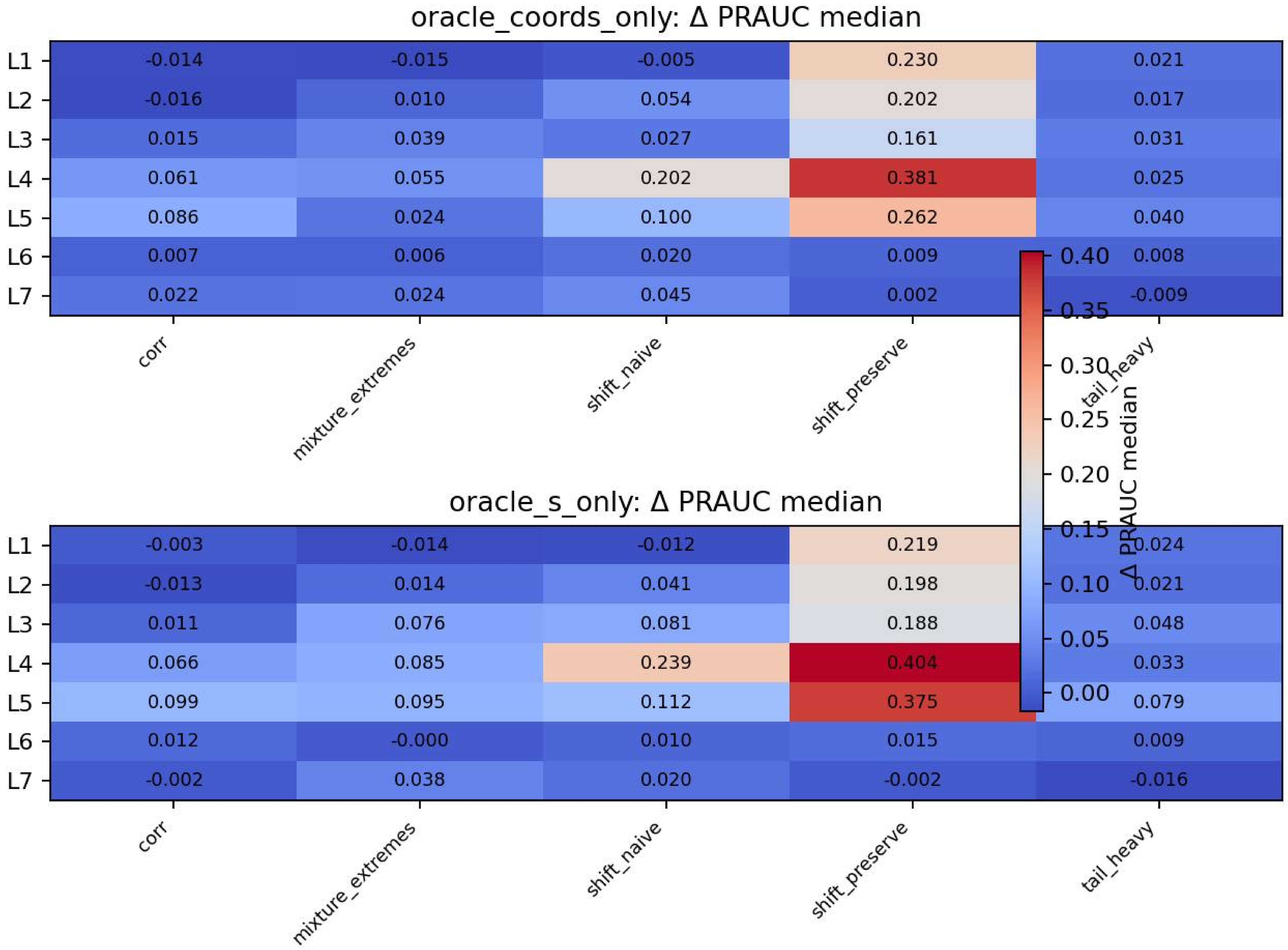
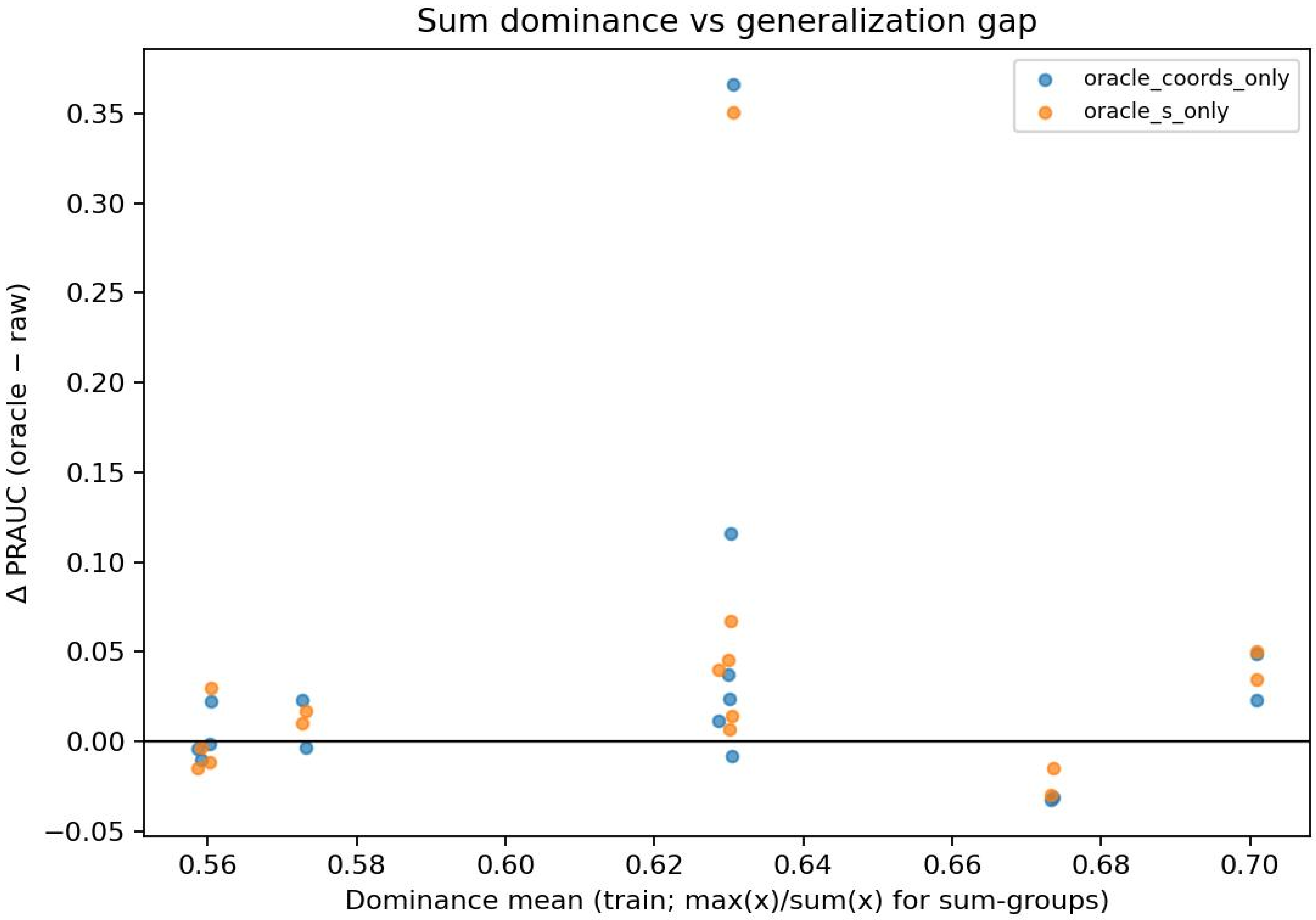
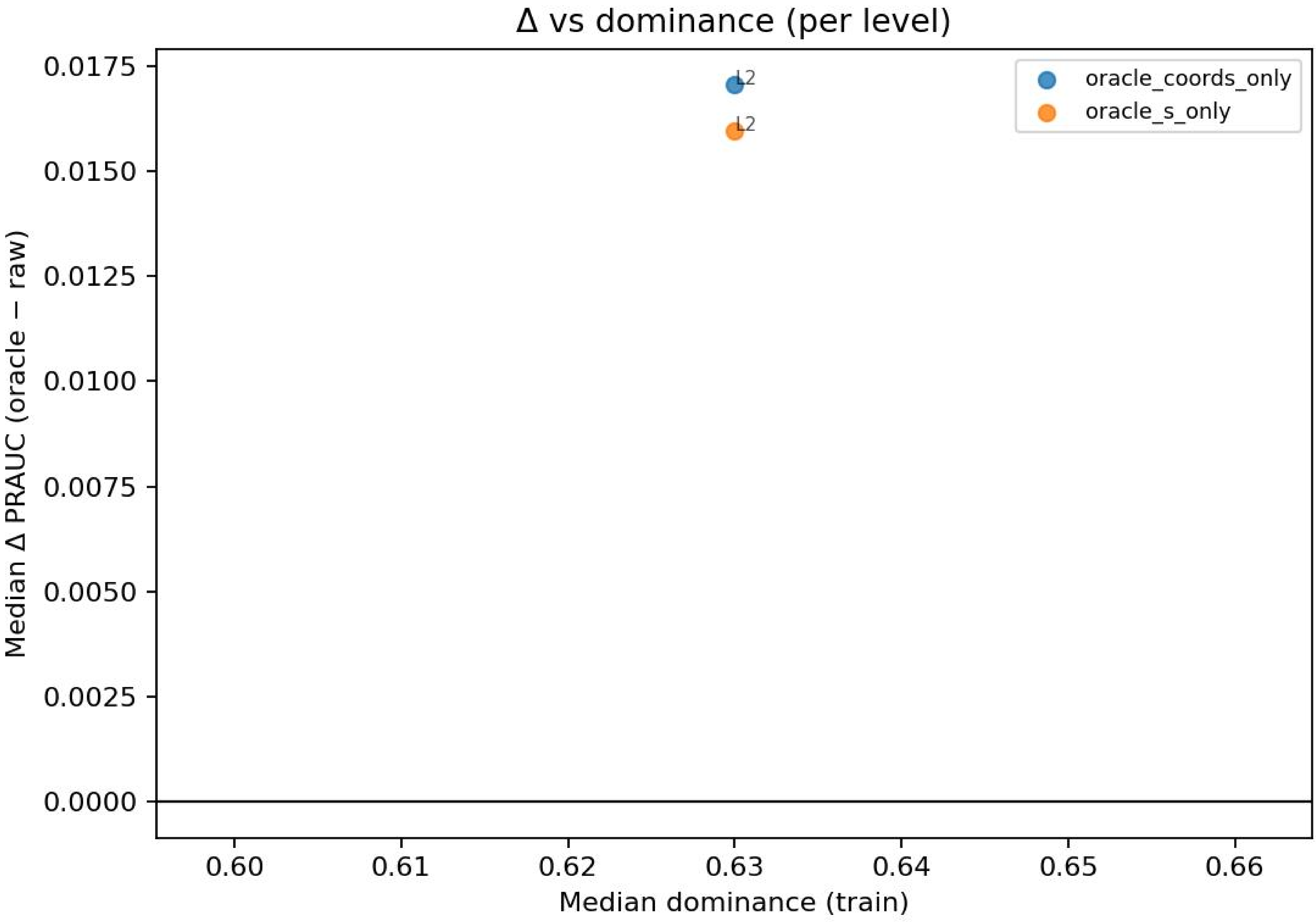


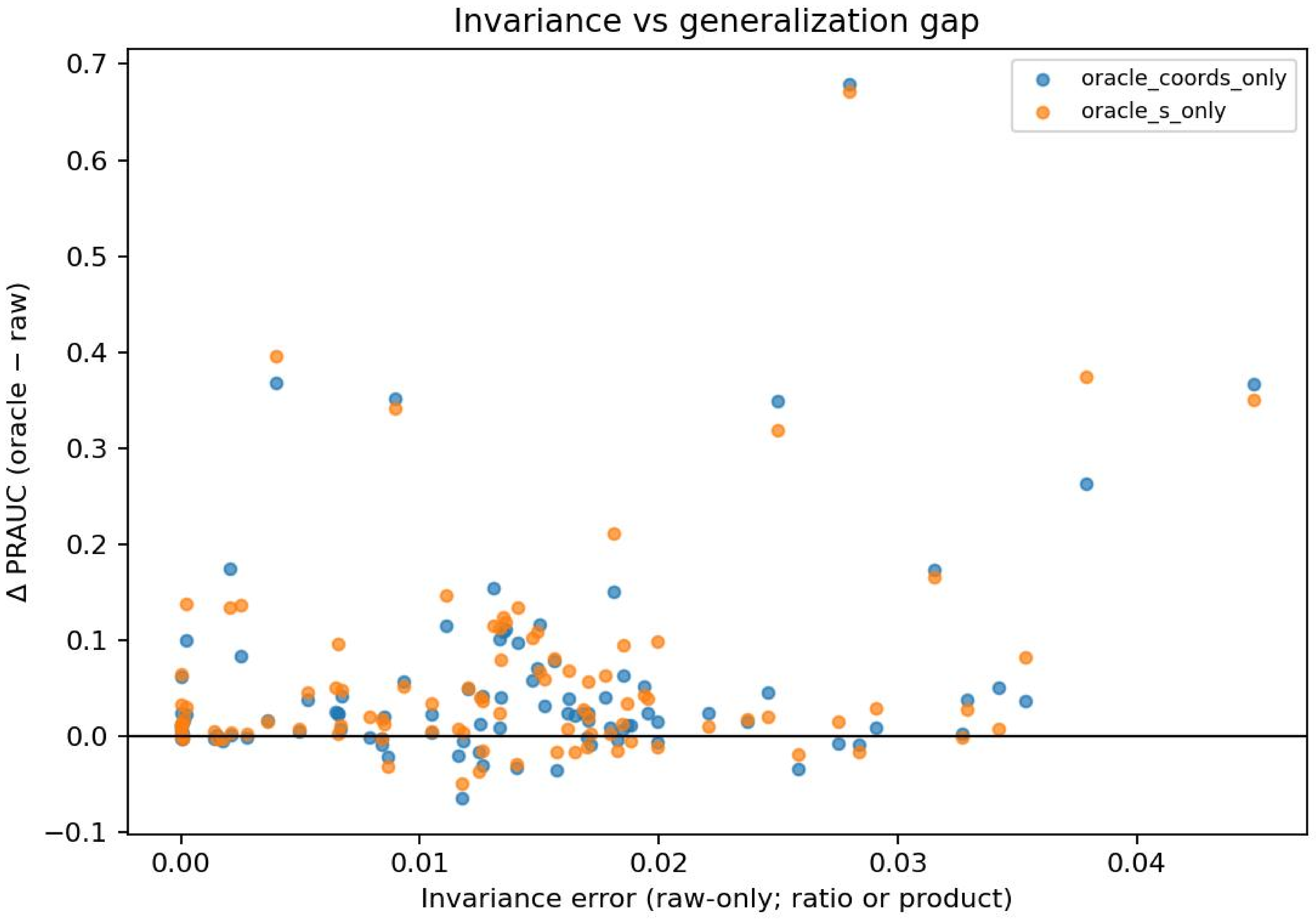
oracle_coords_only: Δ PRAUC median

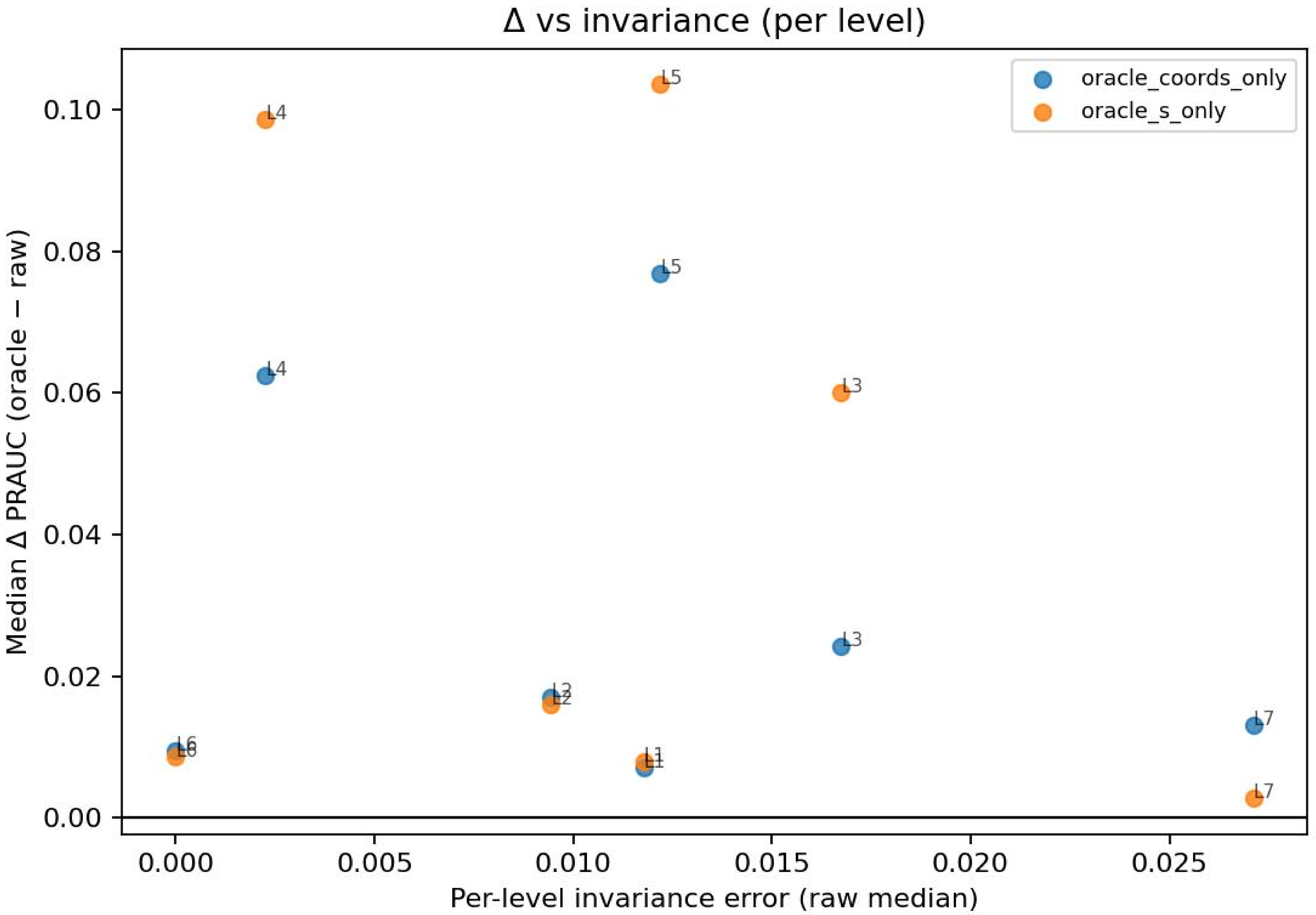
oracle_s_only: Δ PRAUC median

Δ PRAUC (oracle − raw) by task and oracle mode
Δ PRAUC
0.3
0.2
0.1
0.0
l1_product | oracle_coords_only
l1_product | oracle_s_only
l1_ratio | oracle_coords_only
l1_ratio | oracle_s_only
l2_product_of_sums | oracle_coords_only
l2_product_of_sums | oracle_s_only
l2_ratio_of_sums | oracle_coords_only
l2_ratio_of_sums | oracle_s_only
l3_product_diff | oracle_coords_only
l3_product_diff | oracle_s_only
l3_ratio_diff | oracle_coords_only
l3_ratio_diff | oracle_s_only
l4_product_x_product | oracle_coords_only
l4_product_x_product | oracle_s_only
l4_ratio_x_ratio | oracle_coords_only
l4_ratio_x_ratio | oracle_s_only
l5_ratio_x_product | oracle_coords_only
l5_ratio_x_product | oracle_s_only
l6_nonmonotone_product | oracle_coords_only
l6_nonmonotone_product | oracle_s_only
l6_nonmonotone_ratio | oracle_coords_only
l6_nonmonotone_ratio | oracle_s_only
l7_gated_ratio_vs_product | oracle_coords_only
l7_gated_ratio_vs_product | oracle_s_only

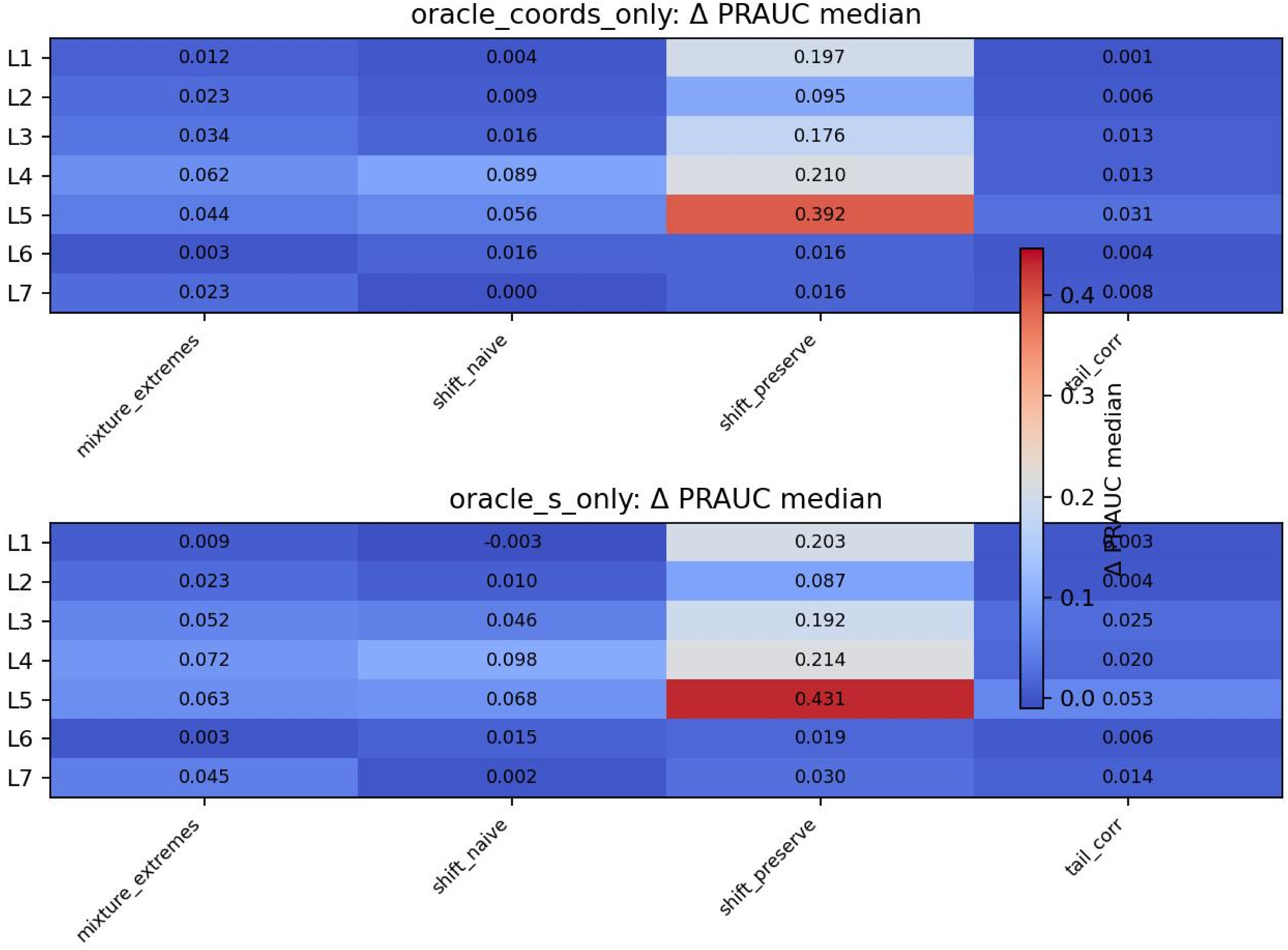

Sum dominance vs generalization gap
oracle_coords_only
oracle_s_only
Δ PRAUC (oracle − raw)
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
0.56
0.58
0.60
0.62
0.64
0.66
0.68
0.70
Dominance mean (train; max(x)/sum(x) for sum-groups)


Δ vs dominance (per level)
Median Δ PRAUC (oracle − raw)
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Median dominance (train)
0.60
0.61
0.62
0.63
0.64
0.65
0.66
L2
L2
oracle_coords_only
oracle_s_only

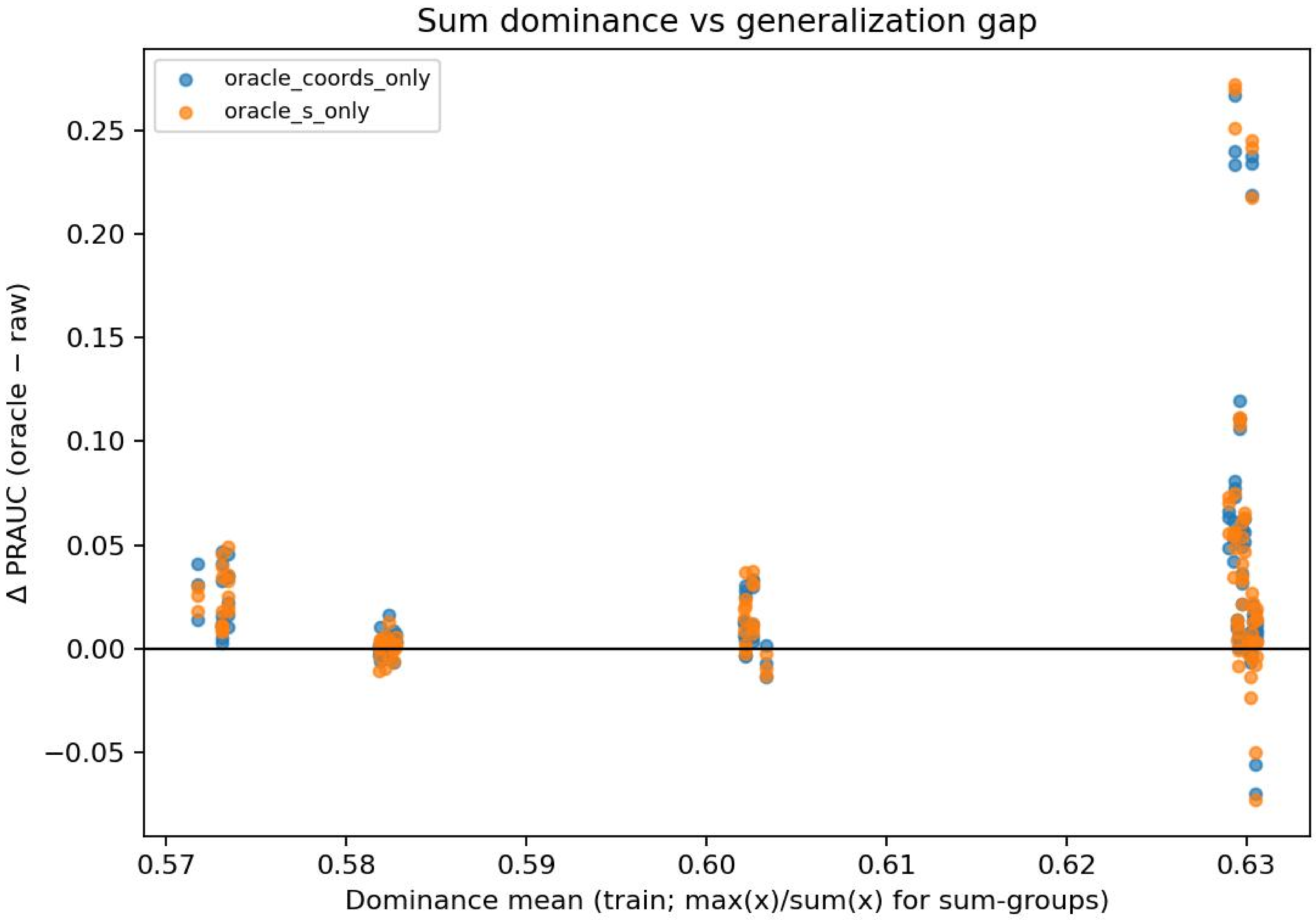

Invariance vs generalization gap
Δ PRAUC (oracle − raw)
Invariance error (raw-only; ratio or product)
oracle_coords_only
oracle_s_only
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0.00
0.01
0.02
0.03
0.04

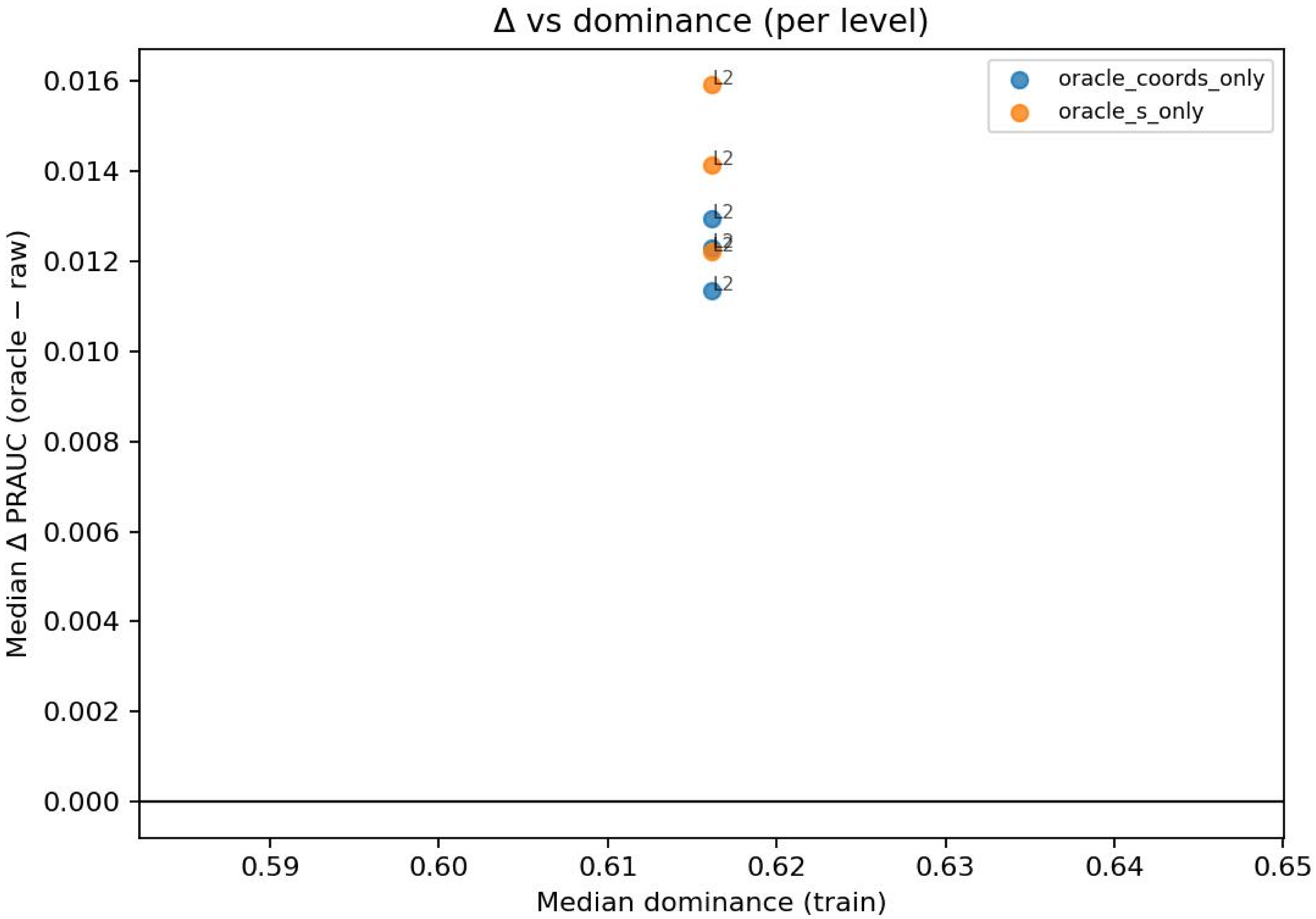

Δ vs invariance (per level)
Median Δ PRAUC (oracle − raw)
Per-level invariance error (raw median)
0.00
0.02
0.04
0.06
0.08
0.10
0.000
0.005
0.010
0.015
0.020
0.025
oracle_coords_only
oracle_s_only
L5
L4
L5
L4
L3
L3
L2
L2
L7
L6
L0
L1
L1
L7

No n-sweep data

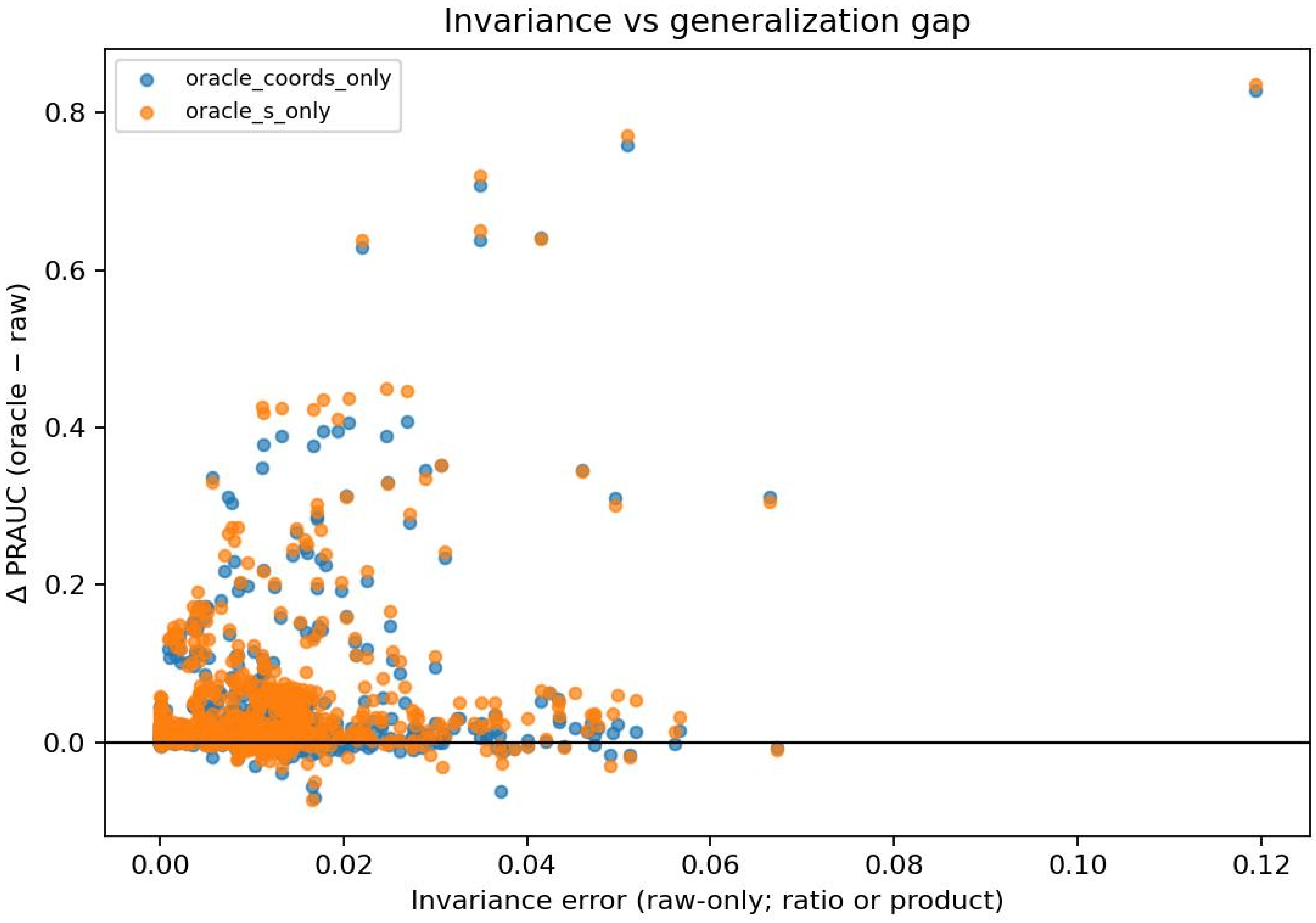

oracle_coords_only: Δ PRAUC median
L1 0.012 0.004 0.197 0.001
L2 0.023 0.009 0.095 0.006
L3 0.034 0.016 0.176 0.013
L4 0.062 0.089 0.210 0.013
L5 0.044 0.056 0.392 0.031
L6 0.003 0.016 0.016 0.004
L7 0.023 0.000 0.016 0.008
mixture_extremes
shift_naive
shift_preserve
tail_corr
oracle_s_only: Δ PRAUC median
L1 0.009 -0.003 0.203 0.003
L2 0.023 0.010 0.087 0.004
L3 0.052 0.046 0.192 0.025
L4 0.072 0.098 0.214 0.020
L5 0.063 0.068 0.431 0.053
L6 0.003 0.015 0.019 0.006
L7 0.045 0.002 0.030 0.014
mixture_extremes
shift_naive
shift_preserve
tail_corr
Δ PRAUC median
0.0
0.1
0.2
0.3
0.4

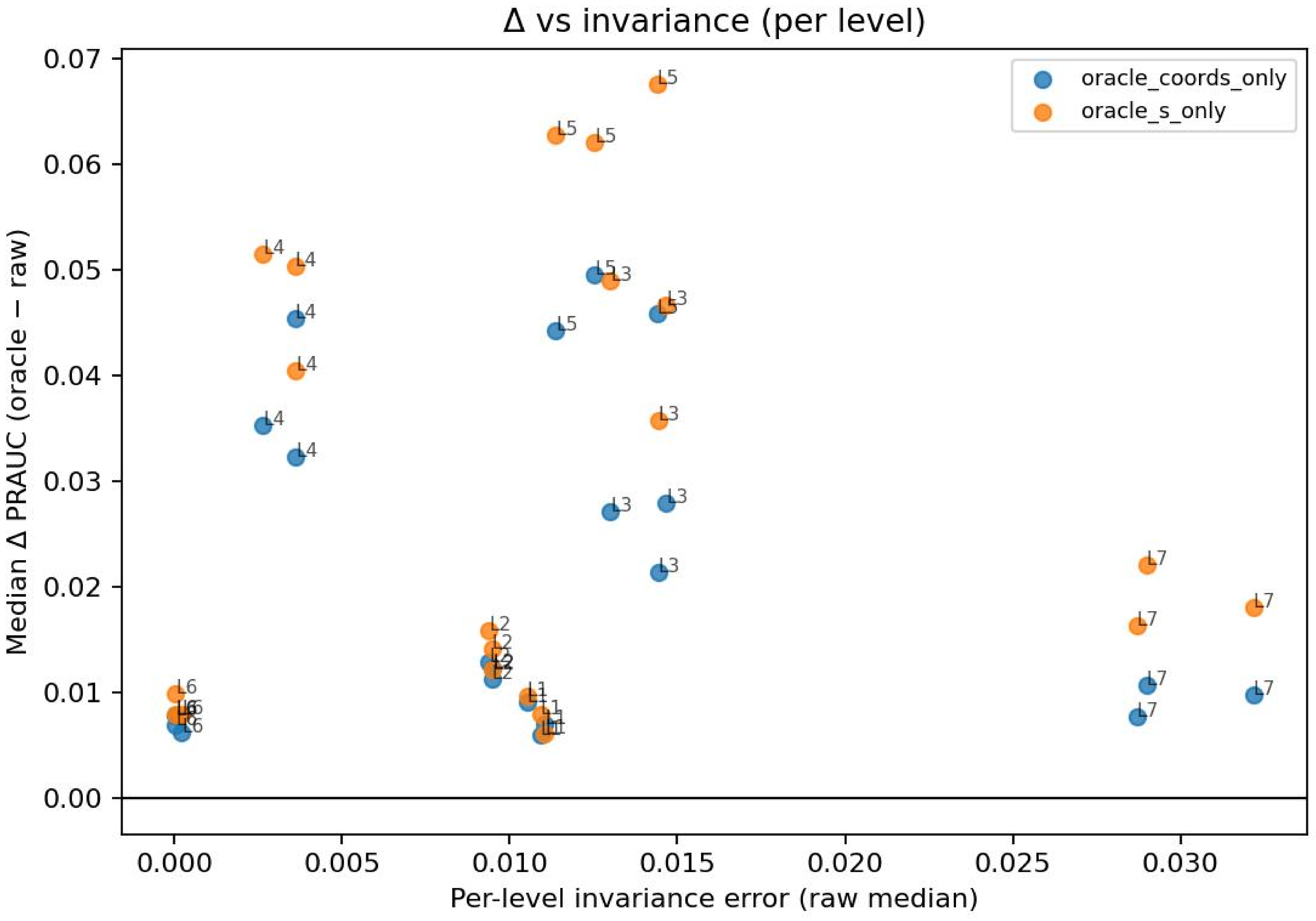
Δ PRAUC (oracle − raw) by task and oracle mode
Δ PRAUC
0.2
0.1
0.0
l1_product | oracle_coords_only
l1_product | oracle_s_only
l1_ratio | oracle_coords_only
l1_ratio | oracle_s_only
l2_product_of_sums | oracle_coords_only
l2_product_of_sums | oracle_s_only
l2_ratio_of_sums | oracle_coords_only
l2_ratio_of_sums | oracle_s_only
l3_product_diff | oracle_coords_only
l3_product_diff | oracle_s_only
l3_ratio_diff | oracle_coords_only
l3_ratio_diff | oracle_s_only
l4_product_x_product | oracle_coords_only
l4_product_x_product | oracle_s_only
l4_ratio_x_ratio | oracle_coords_only
l4_ratio_x_ratio | oracle_s_only
l5_ratio_x_product | oracle_coords_only
l5_ratio_x_product | oracle_s_only
l6_nonmonotone_product | oracle_coords_only
l6_nonmonotone_product | oracle_s_only
l6_nonmonotone_ratio | oracle_coords_only
l6_nonmonotone_ratio | oracle_s_only
l7_gated_ratio_vs_product | oracle_coords_only
l7_gated_ratio_vs_product | oracle_s_only


Sum dominance vs generalization gap
oracle_coords_only
oracle_s_only
Δ PRAUC (oracle − raw)
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
0.57
0.58
0.59
0.60
0.61
0.62
0.63
Dominance mean (train; max(x)/sum(x) for sum-groups)

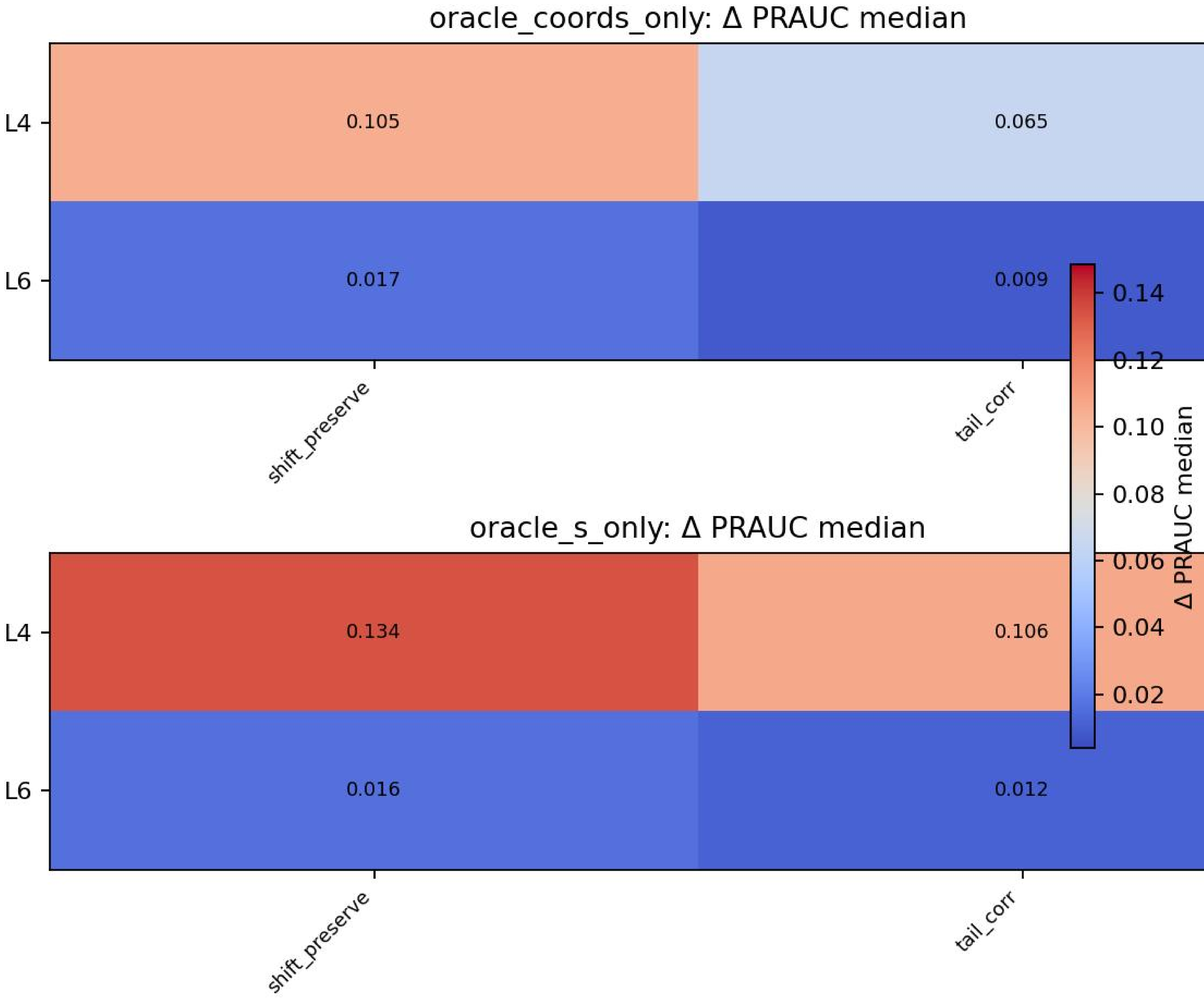


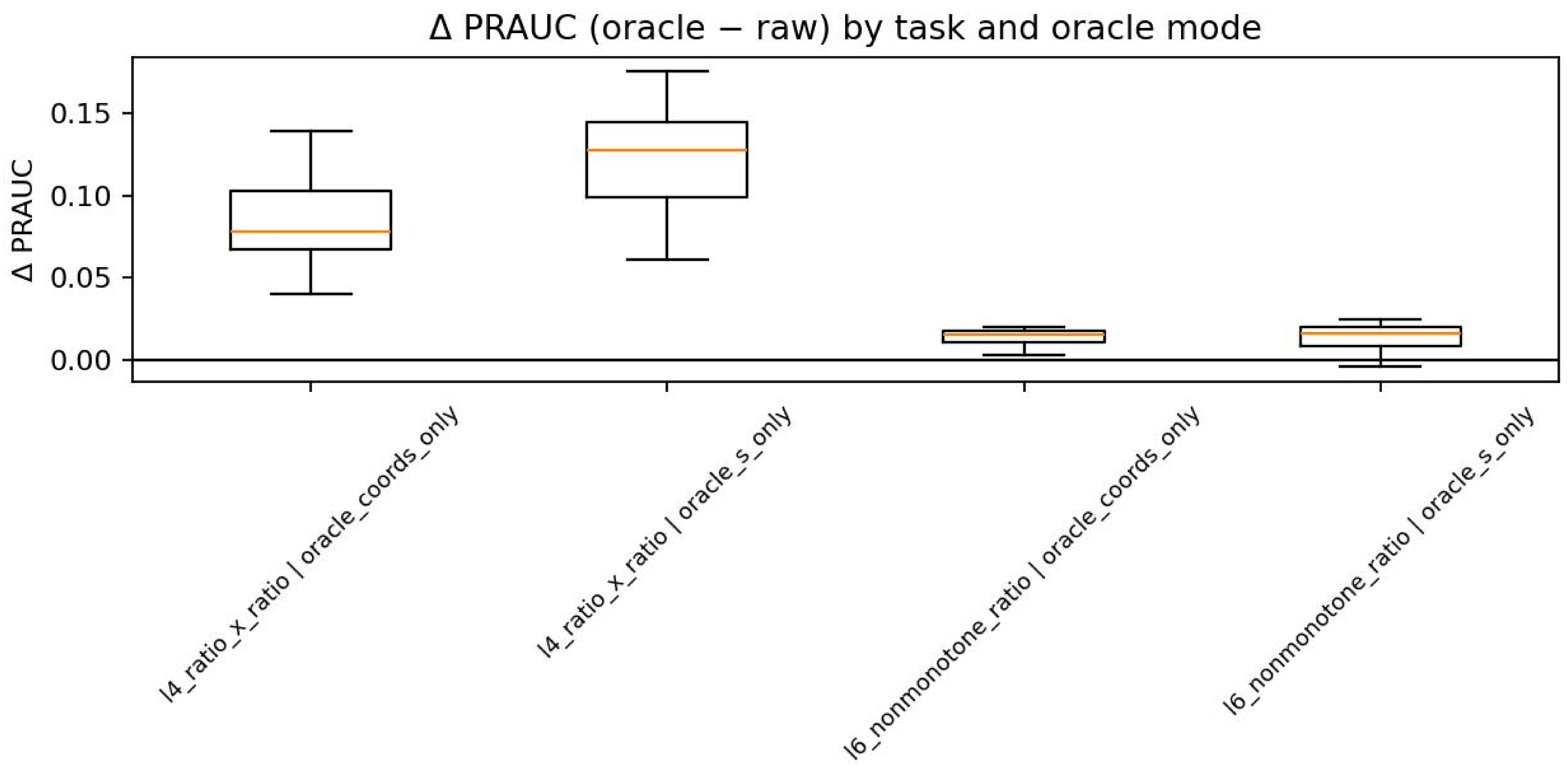
Δ vs dominance (per level)
Median Δ PRAUC (oracle − raw)
Median dominance (train)
oracle_coords_only
oracle_s_only
L2
L2
L2
L2
L2
L2
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0.59
0.60
0.61
0.62
0.63
0.64
0.65

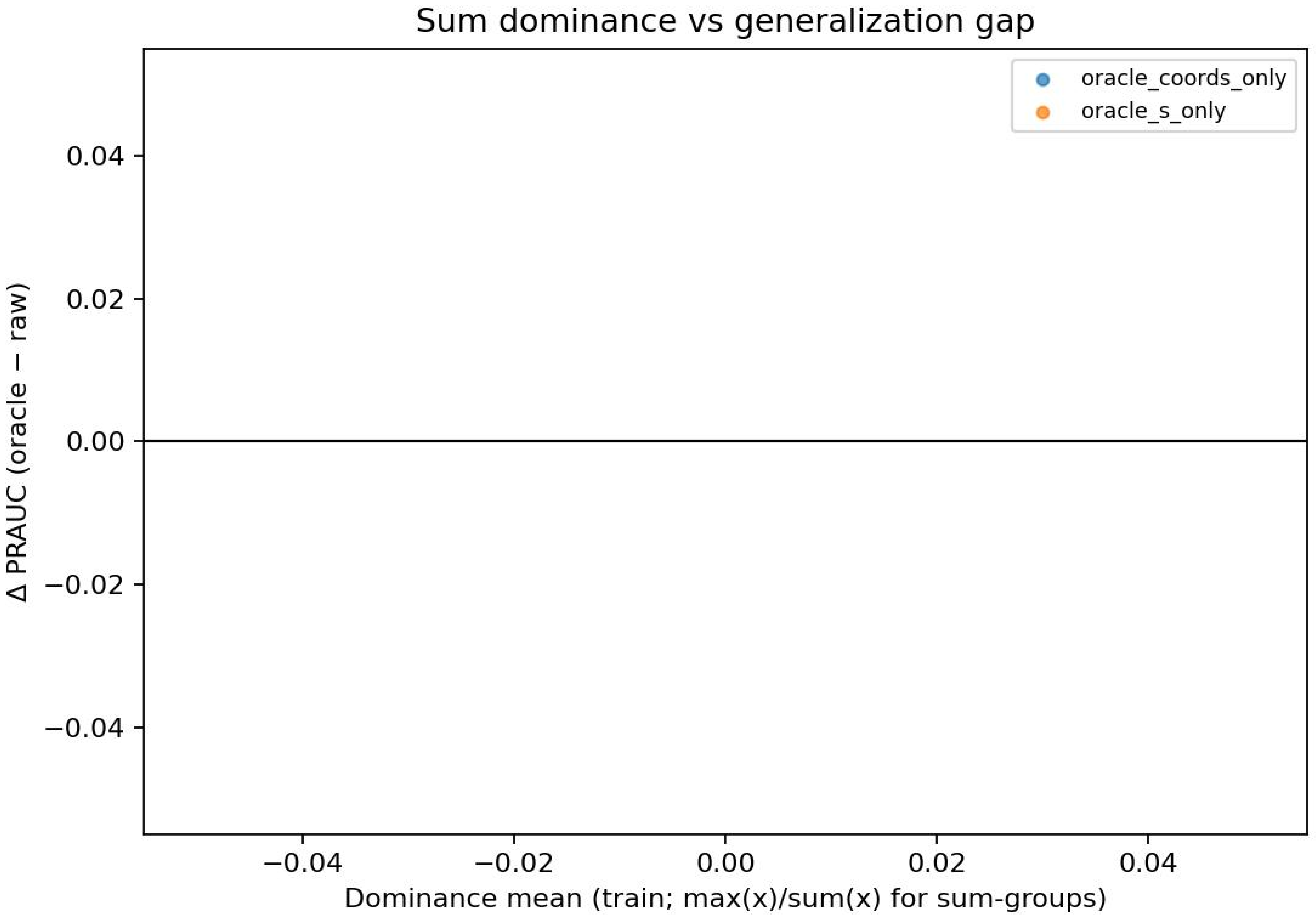

Invariance vs generalization gap
oracle_coords_only
oracle_s_only
Δ PRAUC (oracle − raw)
0.8
0.6
0.4
0.2
0.0
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Invariance error (raw-only; ratio or product)


Δ vs invariance (per level)
oracle_coords_only
oracle_s_only
Median Δ PRAUC (oracle − raw)
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Per-level invariance error (raw median)
L1
L2
L3
L4
L5
L6
L7

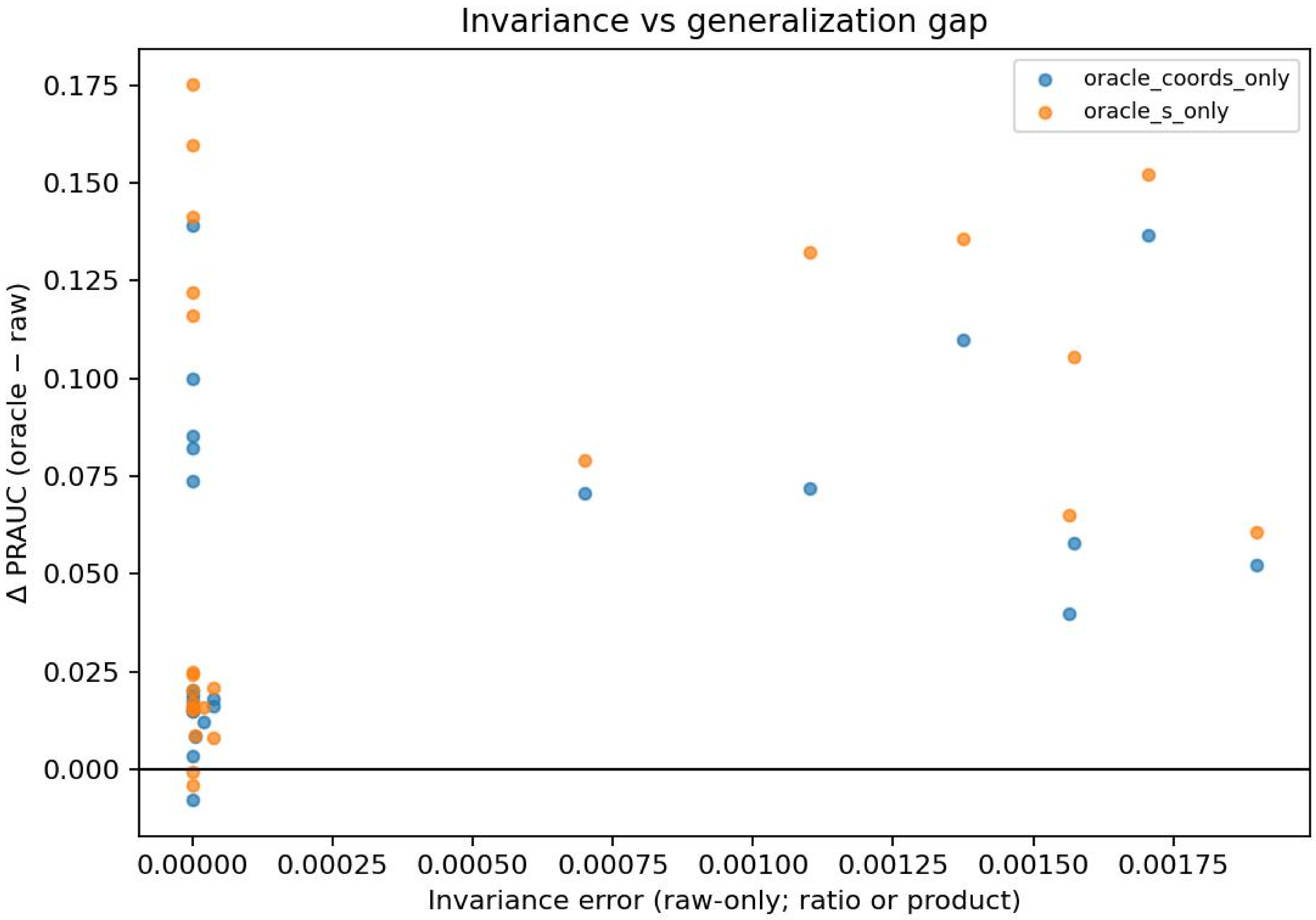
No n-sweep data

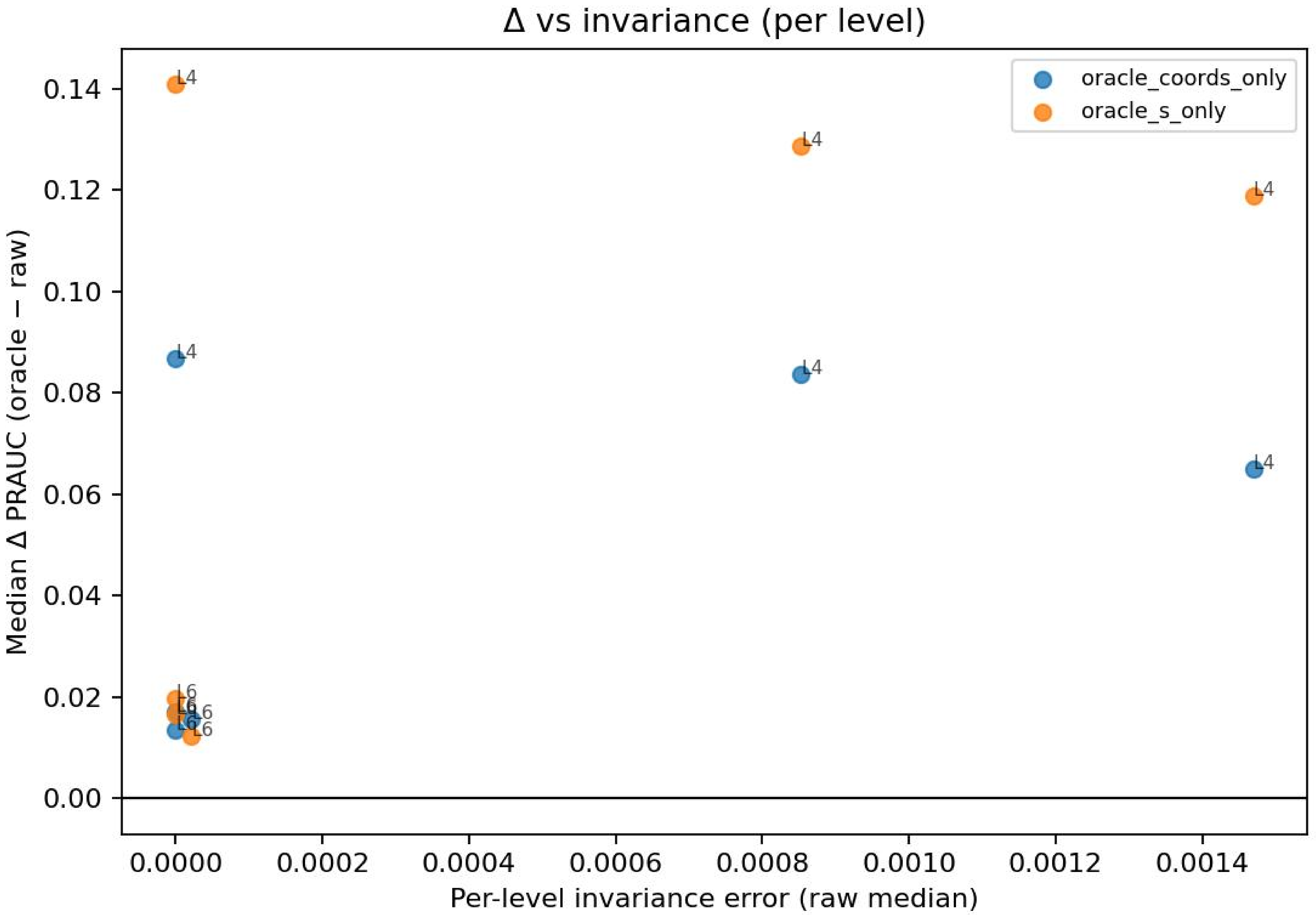

oracle_coords_only: Δ PRAUC median
L4
L6
0.105
0.065
0.017
0.009
shift_preserve
tail_corr
oracle_s_only: Δ PRAUC median
L4
L6
0.134
0.106
0.016
0.012
shift_preserve
tail_corr
Δ PRAUC median
0.14
0.12
0.10
0.08
0.06
0.04
0.02


Δ PRAUC (oracle − raw) by task and oracle mode
Δ PRAUC
0.15
0.10
0.05
0.00
I4_ratio_x_ratio | oracle_coords_only
I4_ratio_x_ratio | oracle_s_only
I6_nonmonotone_ratio | oracle_coords_only
I6_nonmonotone_ratio | oracle_s_only


Sum dominance vs generalization gap
oracle_coords_only
oracle_s_only
0.04
0.02
0.00
−0.02
−0.04
Δ PRAUC (oracle − raw)
−0.04
−0.02
0.00
0.02
0.04
Dominance mean (train; max(x)/sum(x) for sum-groups)

No dominance metrics available


Invariance vs generalization gap
oracle_coords_only
oracle_s_only
Δ PRAUC (oracle − raw)
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.00000
0.00025
0.00050
0.00075
0.00100
0.00125
0.00150
0.00175
Invariance error (raw-only; ratio or product)


Δ vs invariance (per level)
Median Δ PRAUC (oracle − raw)
Per-level invariance error (raw median)
oracle_coords_only
oracle_s_only
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
0.0014
L4
L4
L4
L4
L4
L4
L6
L6
L6
L6
L6

No n-sweep data