

Supplementary Material S1  
Cross-Tool Numerical Validation of  
GO Semantic Similarity Scores

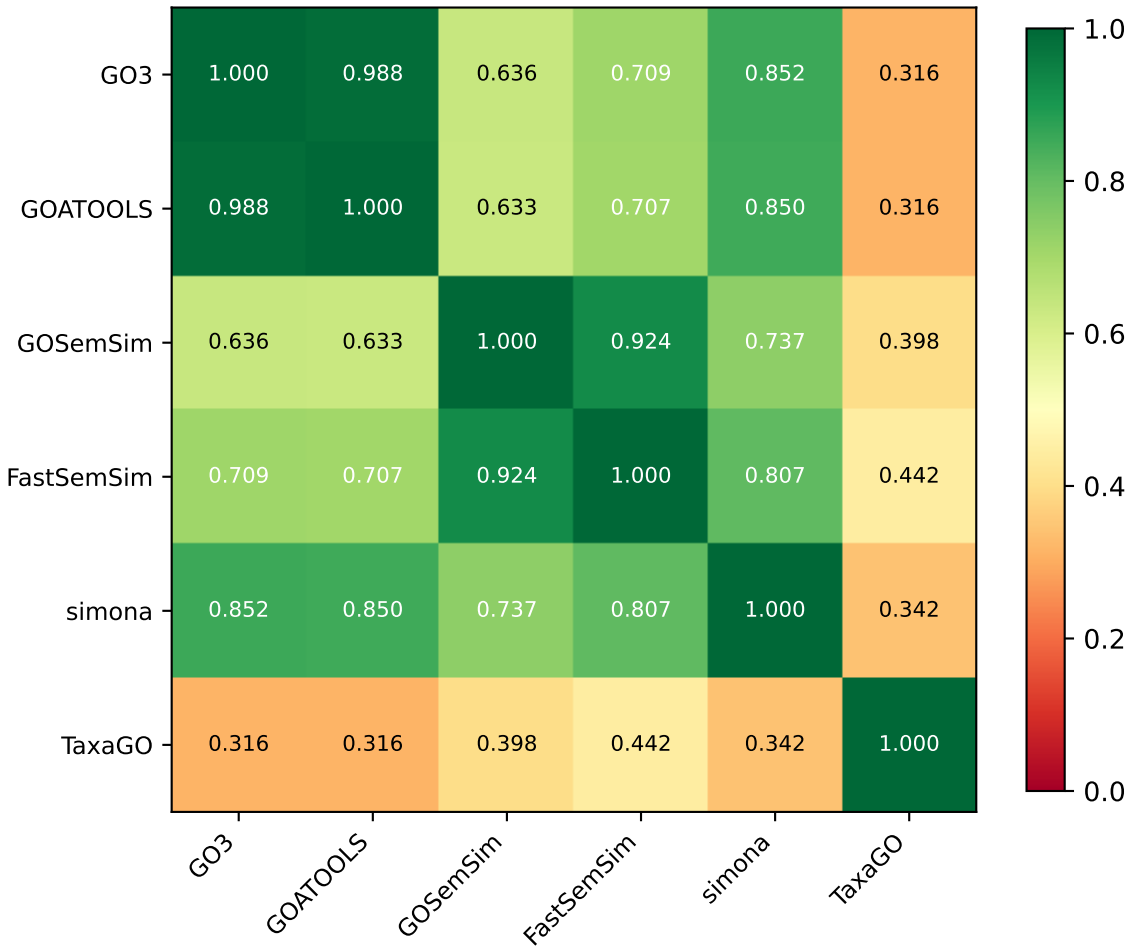
Ontology: GO (OBO releases/2026-03-25)  
Annotations: GOA Human (GAF 2.2, generated 2026-01-05)

Term pairs: 1035 (closed set of 46 terms, seed=42)  
Gene pairs: 100 (BP namespace, min 8 annotations/gene)

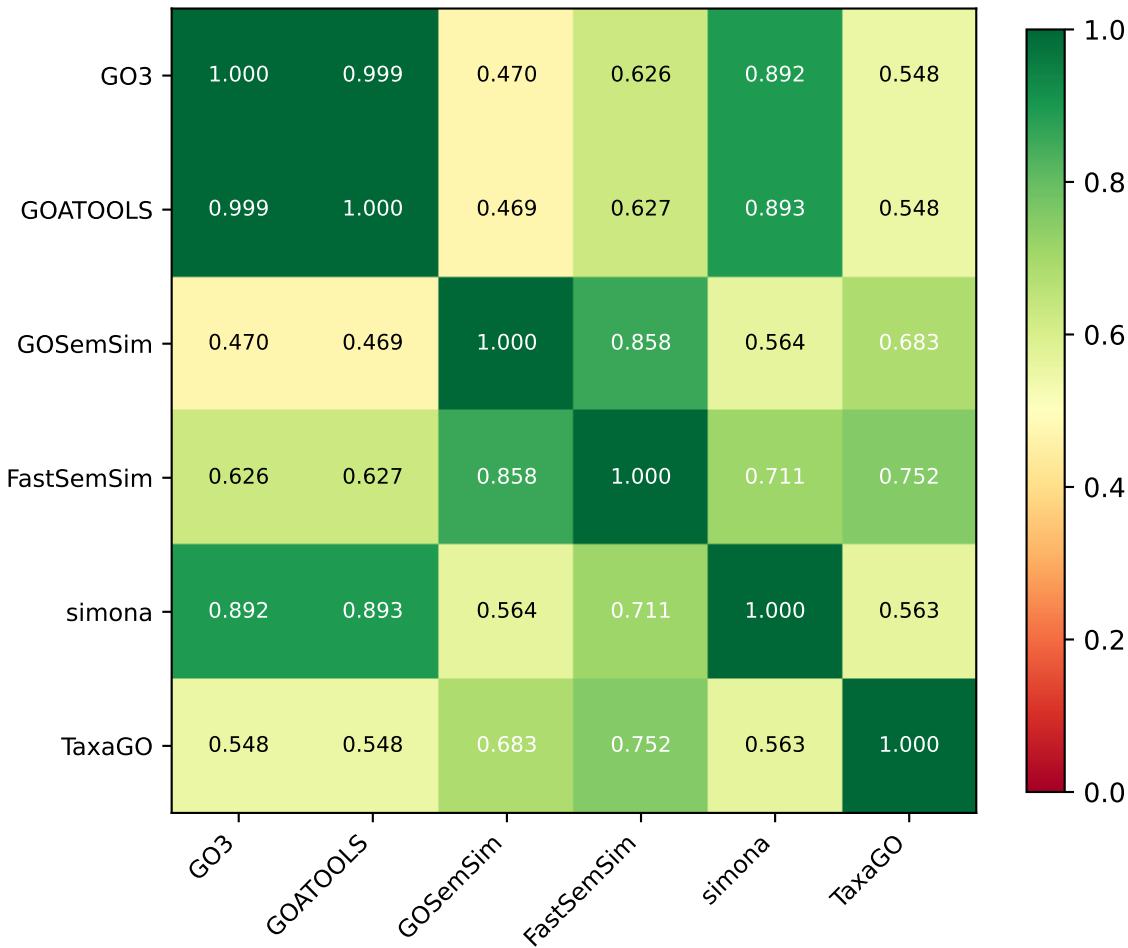
Tools evaluated: FastSemSim, GO3, GOATOOLS, GOSemSim, simona, TaxaGO  
Methods: Resnik, Lin (term level); Lin/BMA (gene level)

System: macOS-26.2-arm64-arm-64bit, Python 3.12.2  
Generated: 2026-04-17

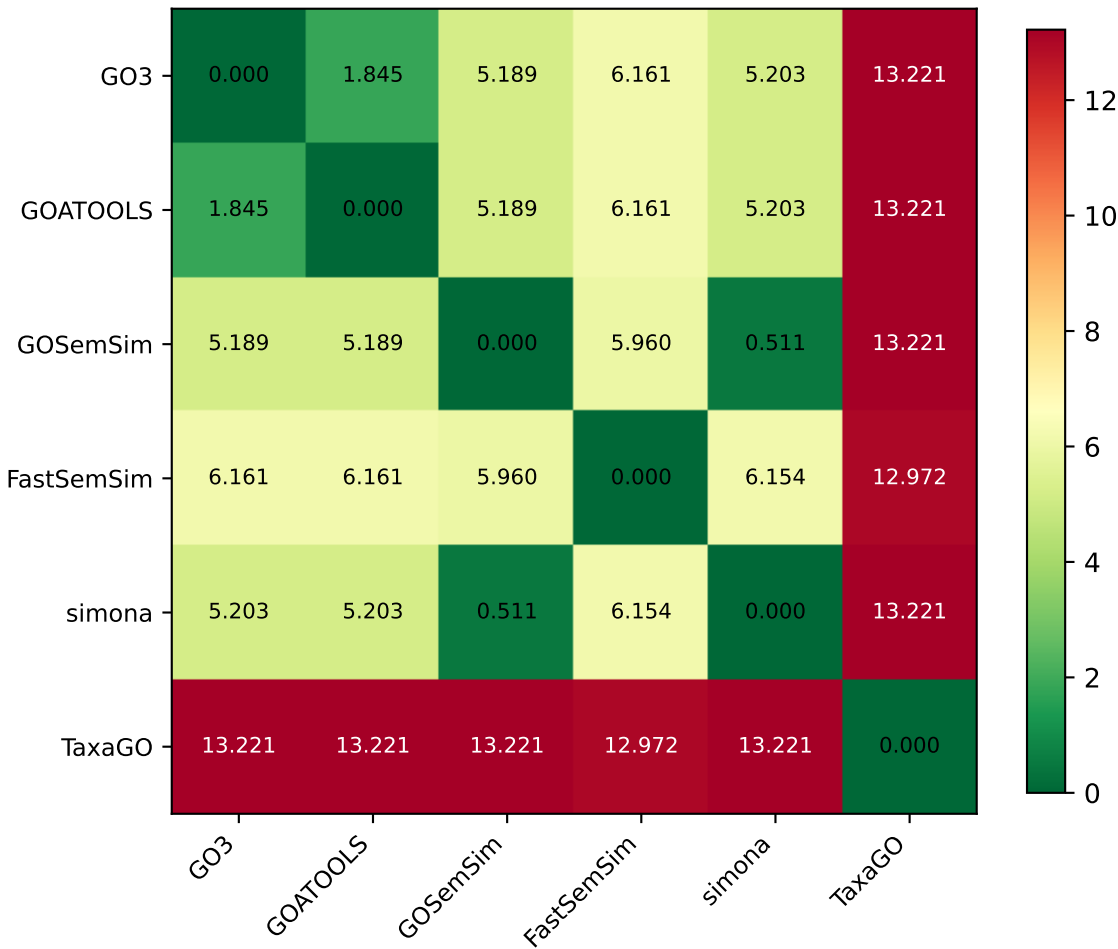
# Term-level Resnik — Pearson r



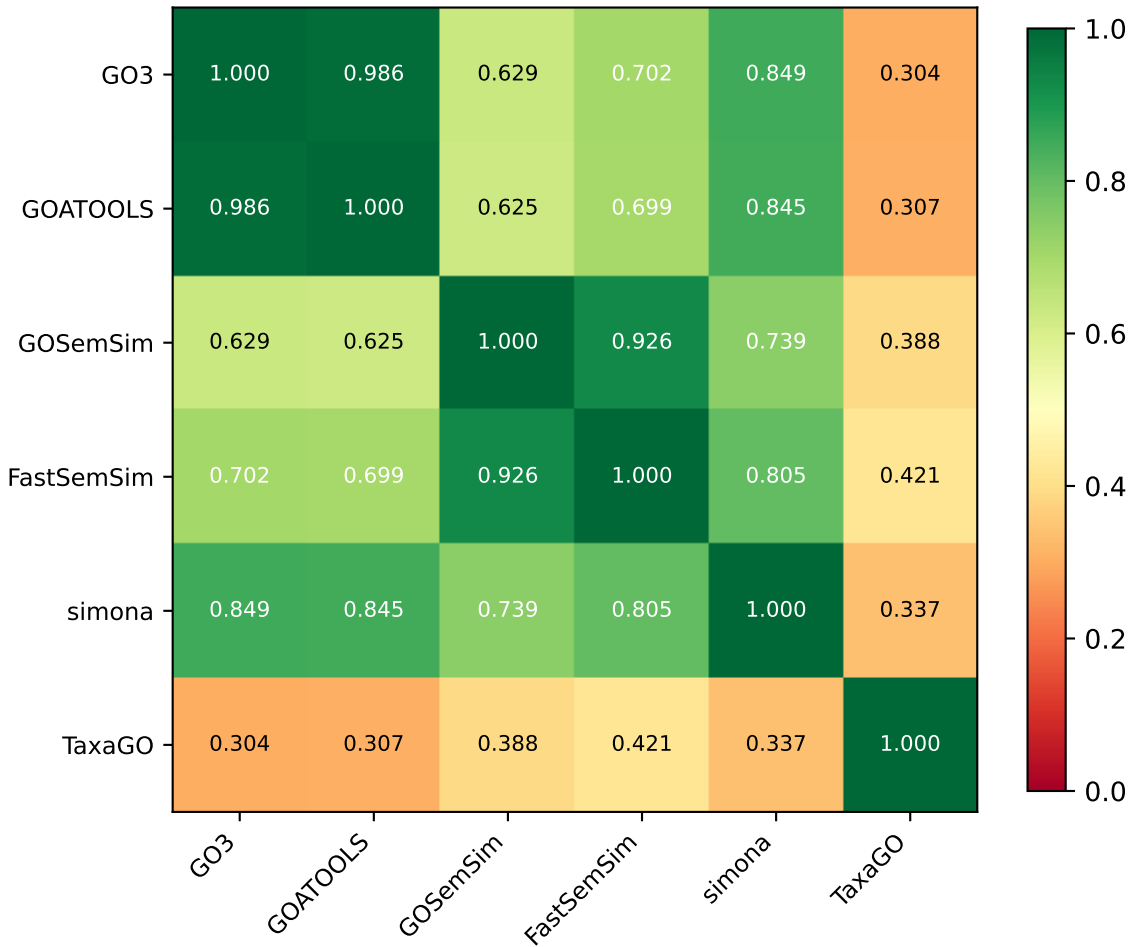
# Term-level Resnik — Spearman $\rho$



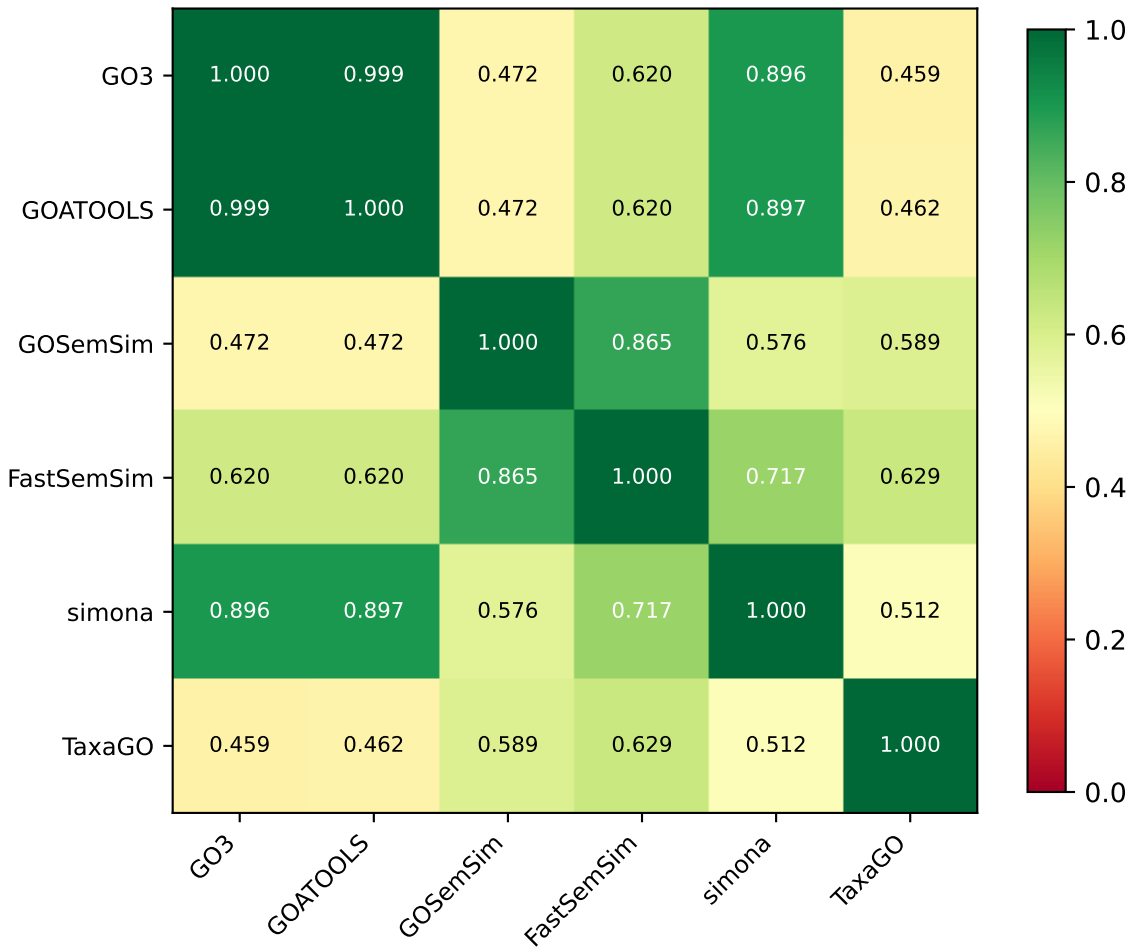
# Term-level Resnik — Max |difference|



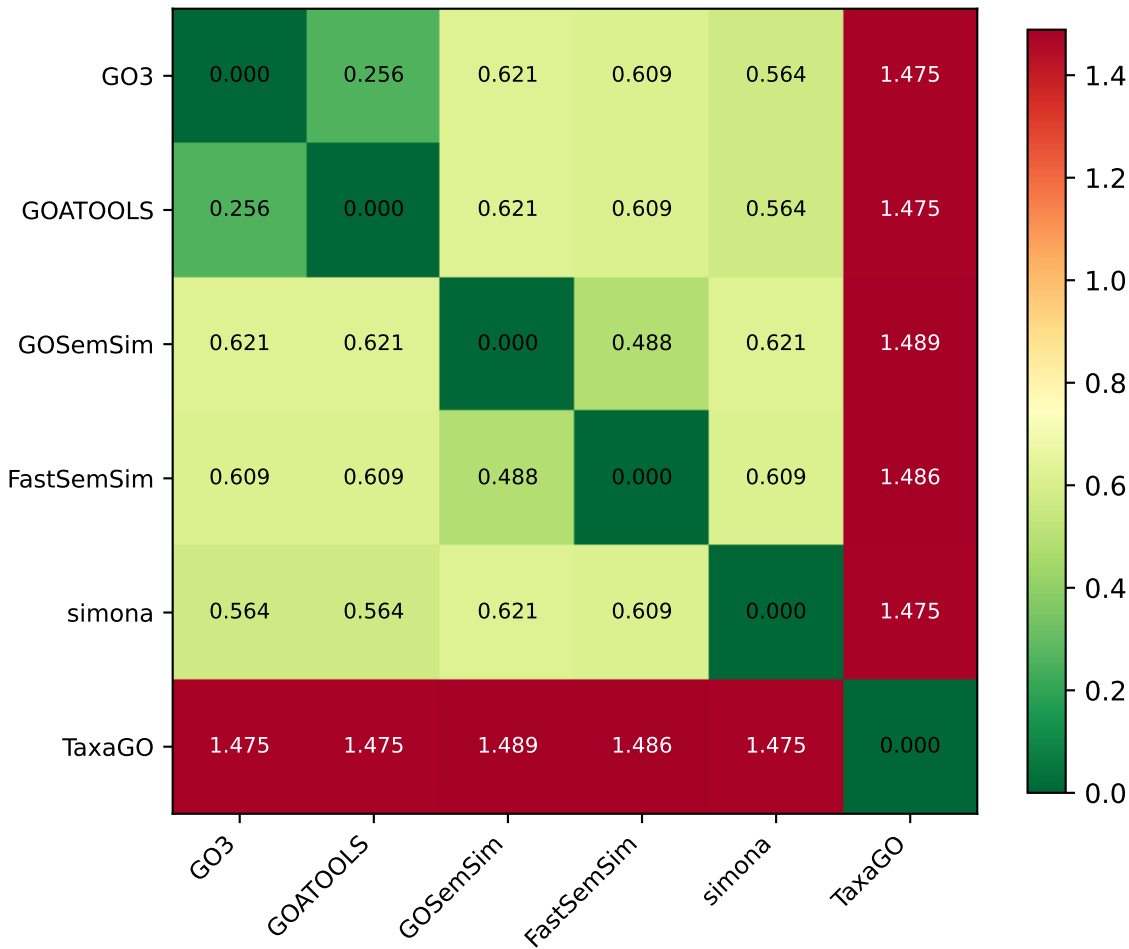
# Term-level Lin — Pearson r



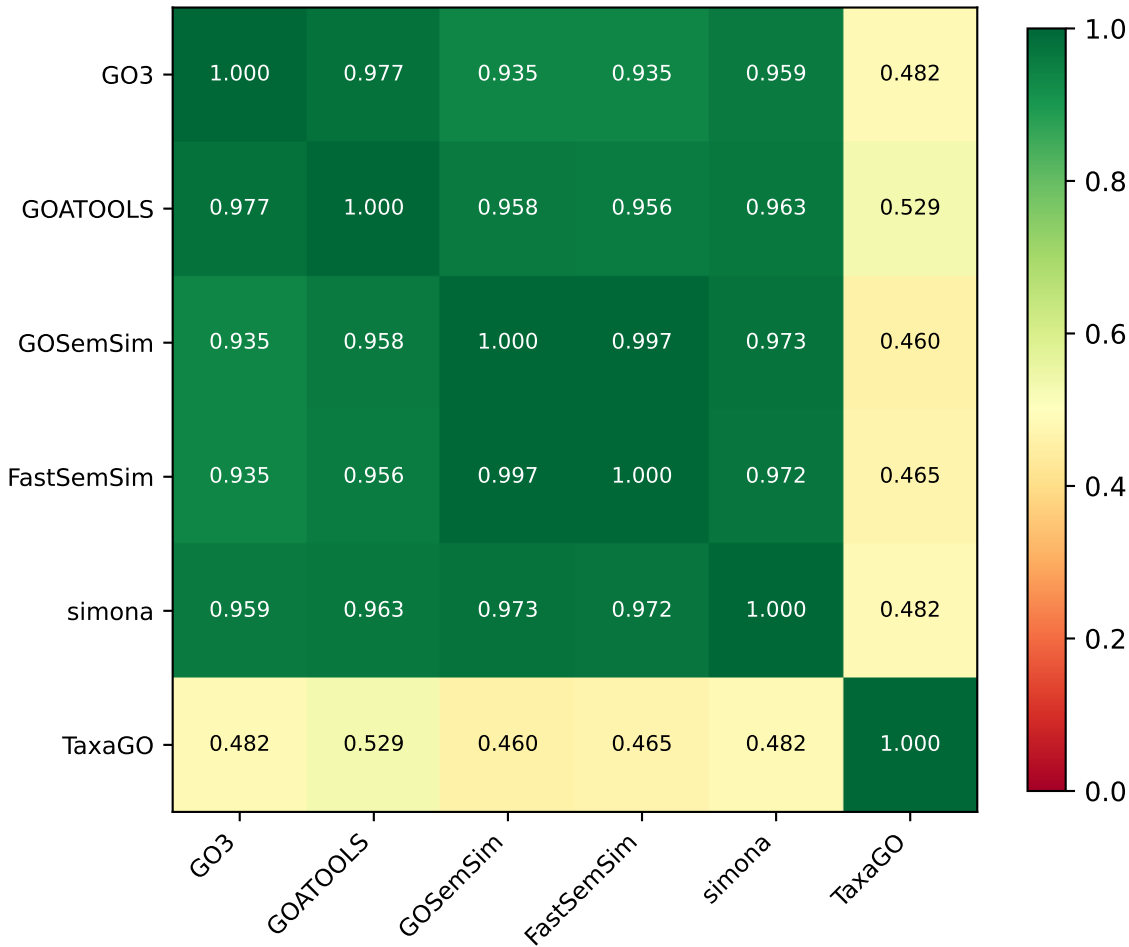
# Term-level Lin — Spearman $\rho$



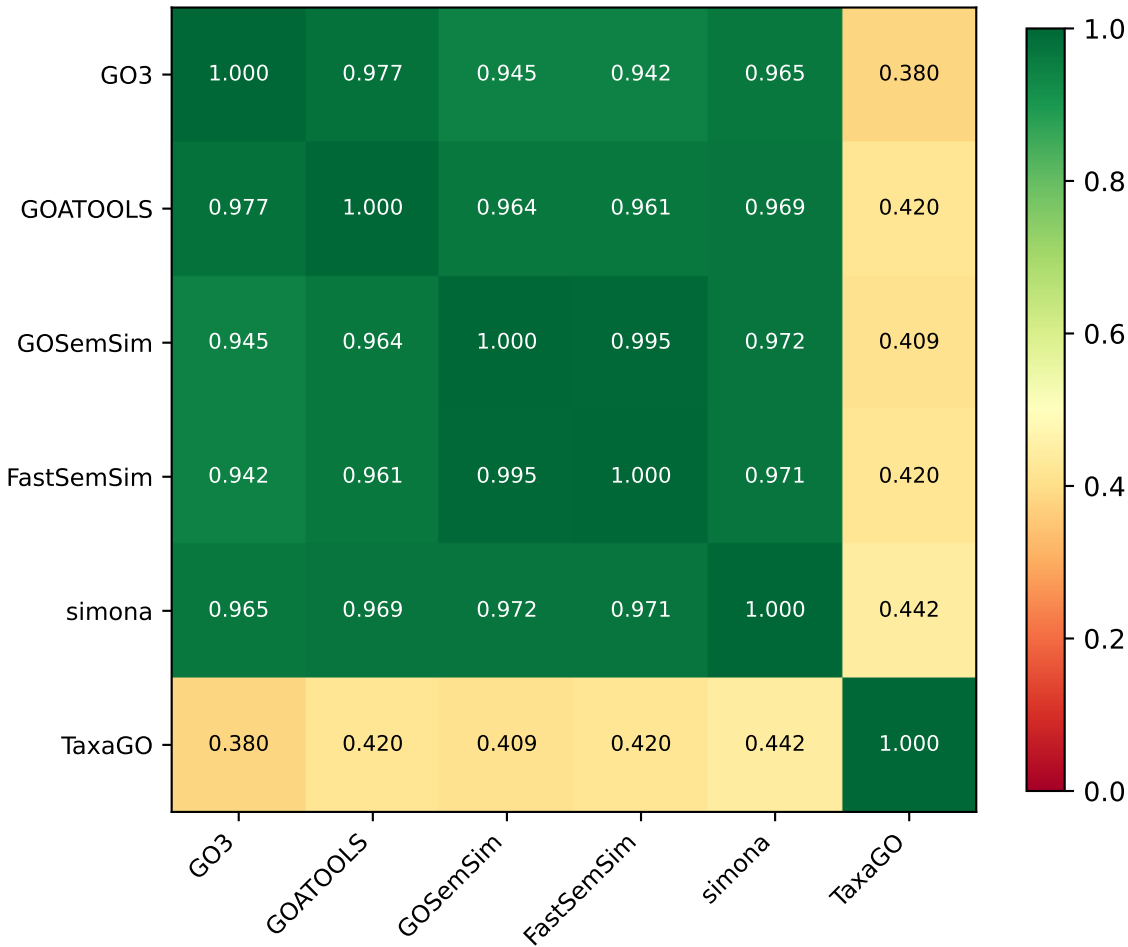
# Term-level Lin — Max |difference|



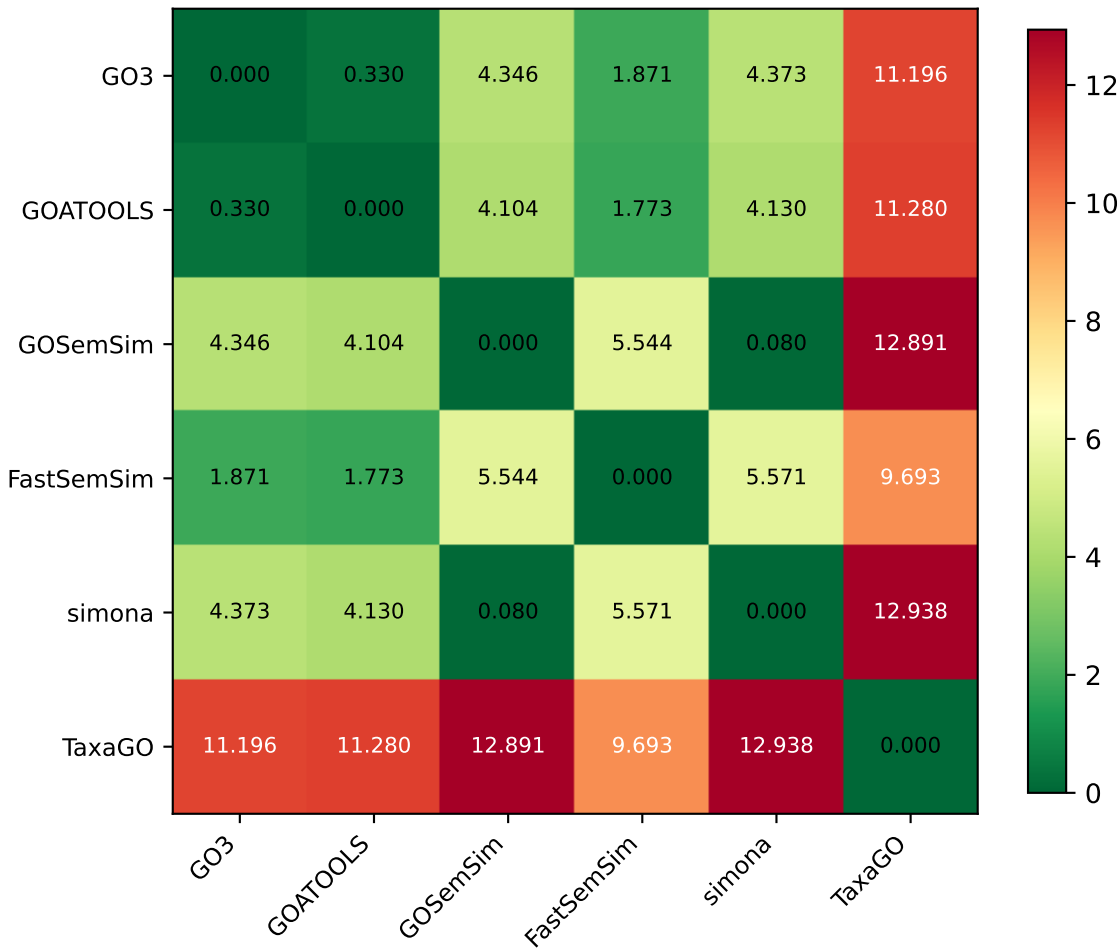
# Gene-level Resnik — Pearson r



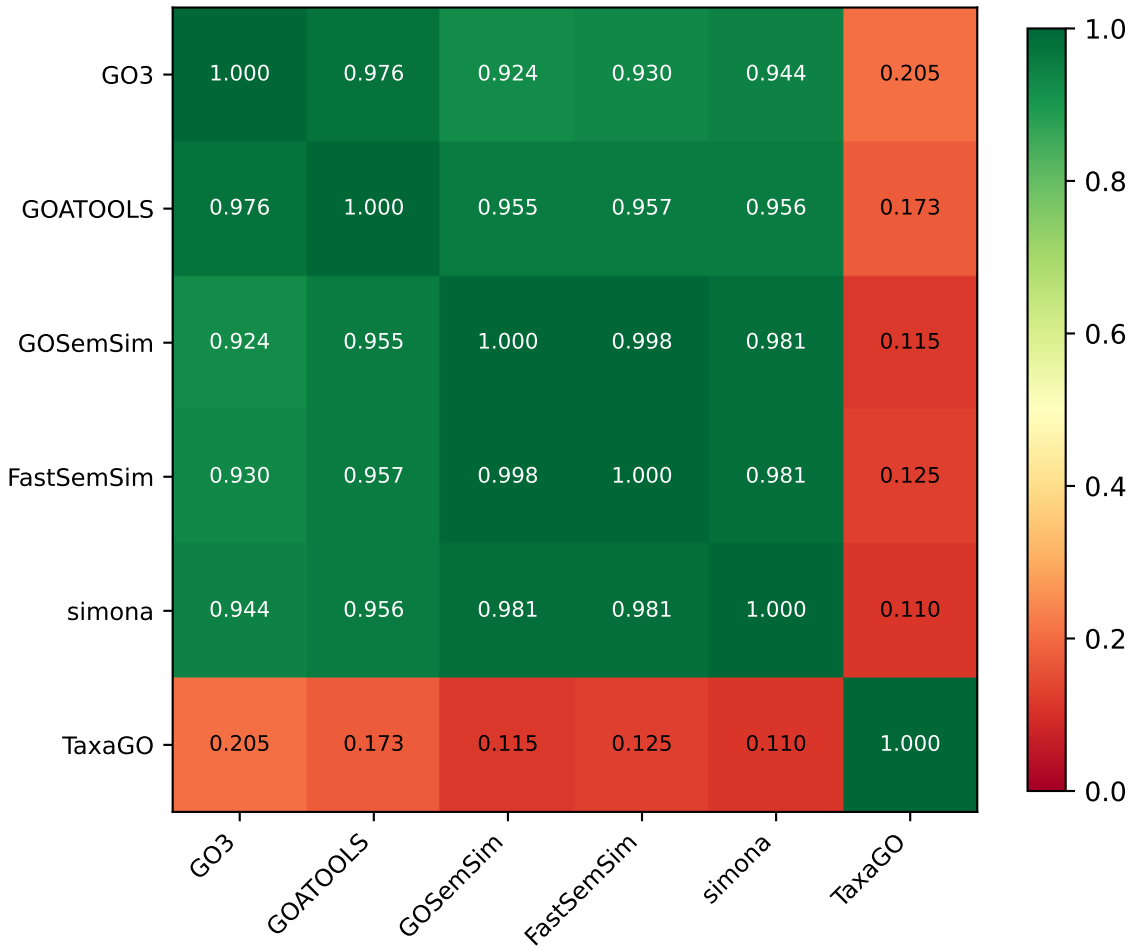
# Gene-level Resnik — Spearman $\rho$



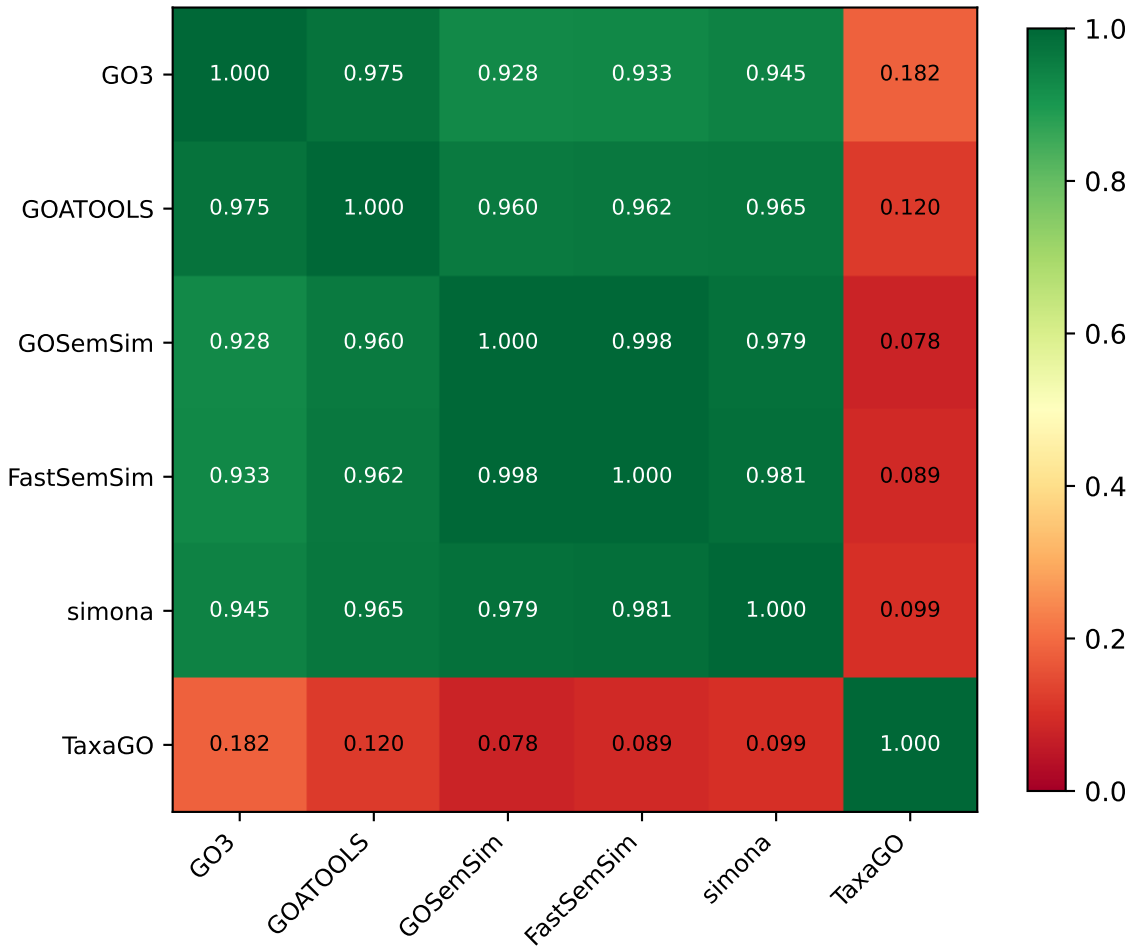
# Gene-level Resnik — Max |difference|



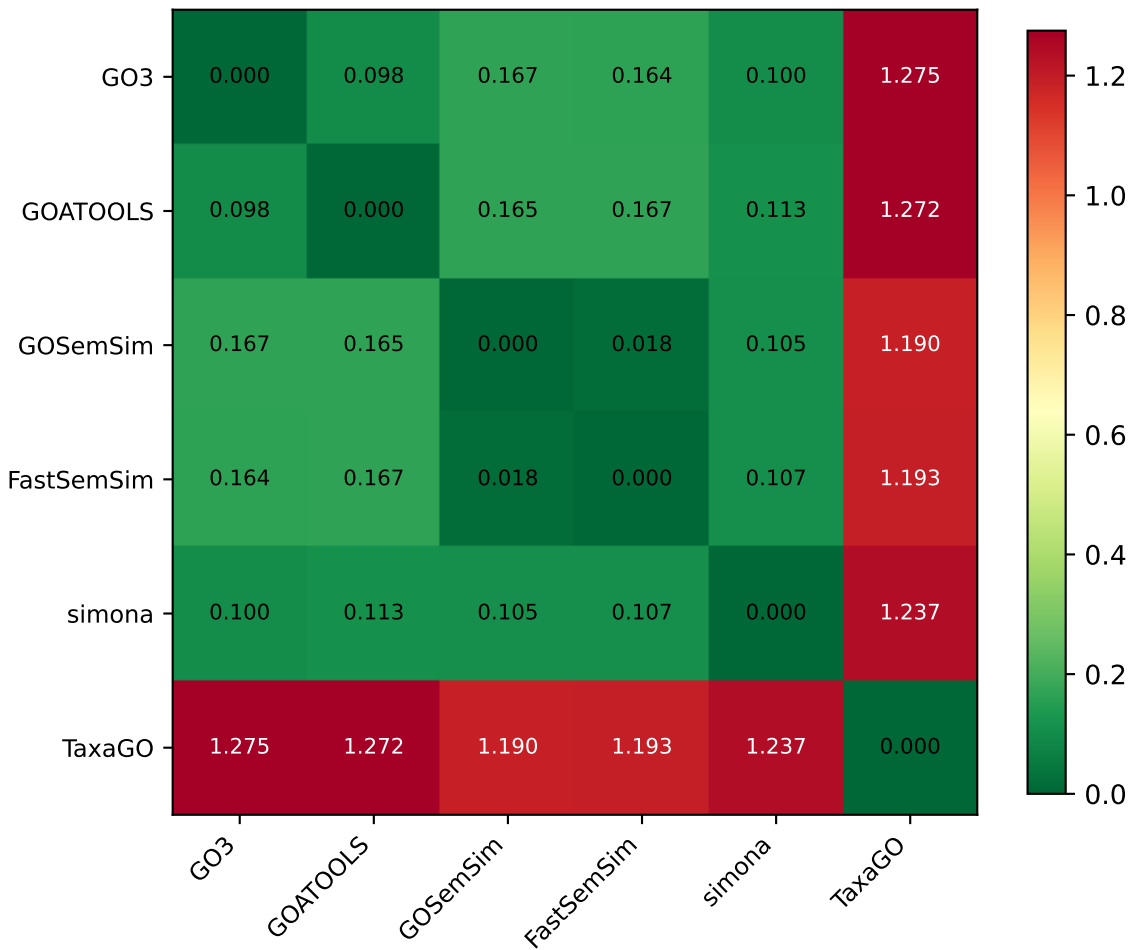
# Gene-level Lin — Pearson r



# Gene-level Lin — Spearman $\rho$

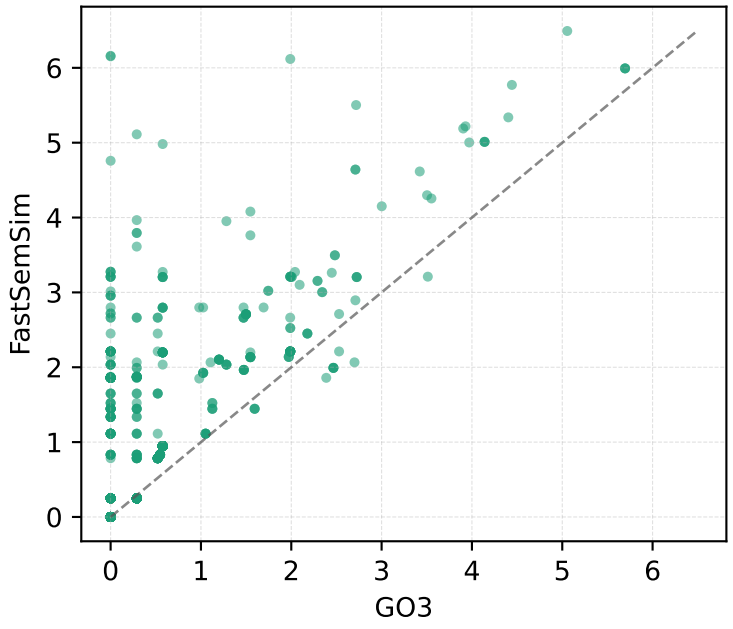


# Gene-level Lin — Max |difference|

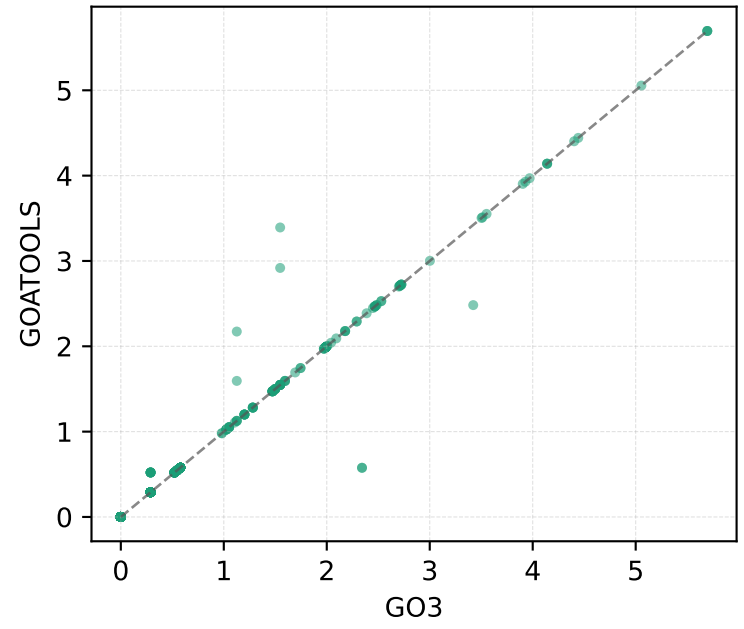


# Term-level Resnik — GO3 vs others (n=1035)

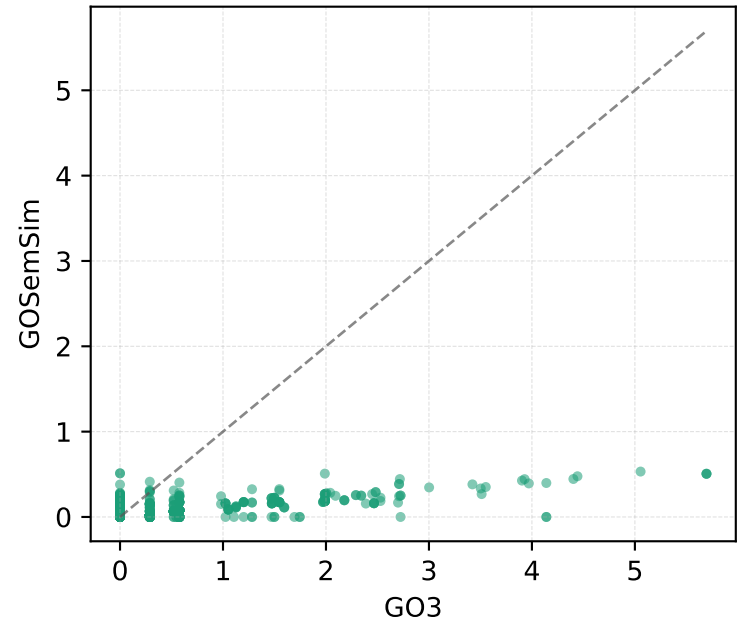
FastSemSim ( $r=0.709$ ,  $\rho=0.626$ )



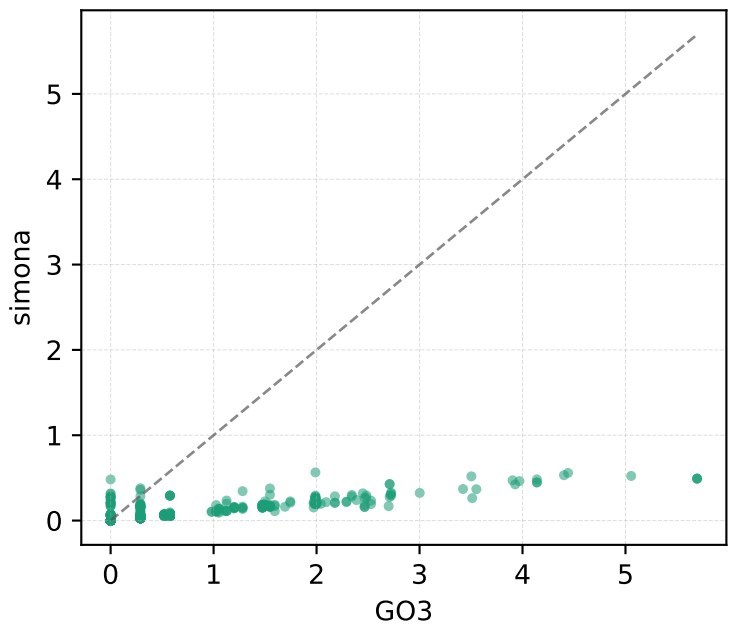
GOATOOLS ( $r=0.988$ ,  $\rho=0.999$ )



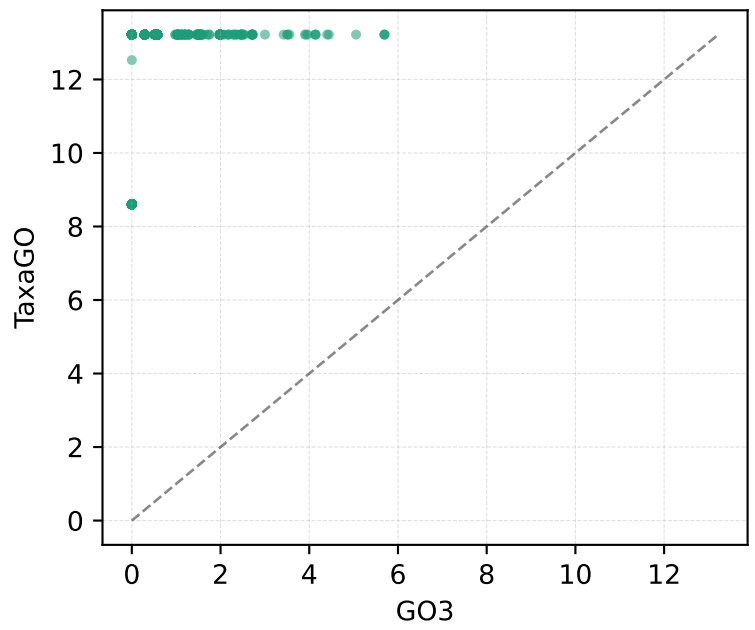
GOSemSim ( $r=0.636$ ,  $\rho=0.470$ )



simona ( $r=0.852$ ,  $\rho=0.892$ )

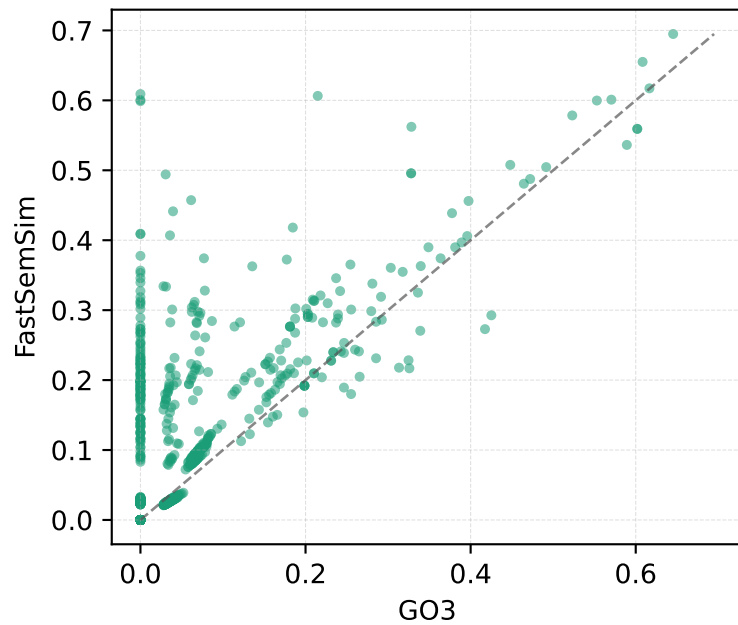


TaxaGO ( $r=0.316$ ,  $\rho=0.548$ )

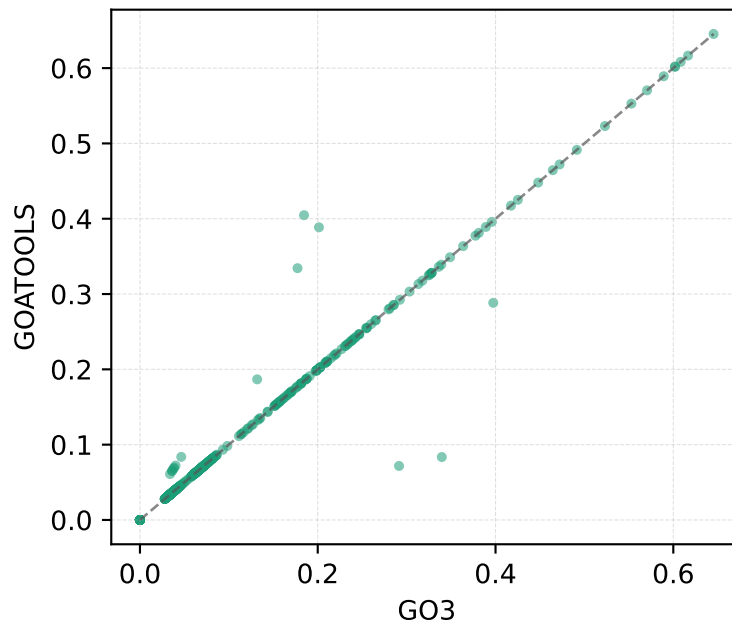


# Term-level Lin — GO3 vs others (n=1035)

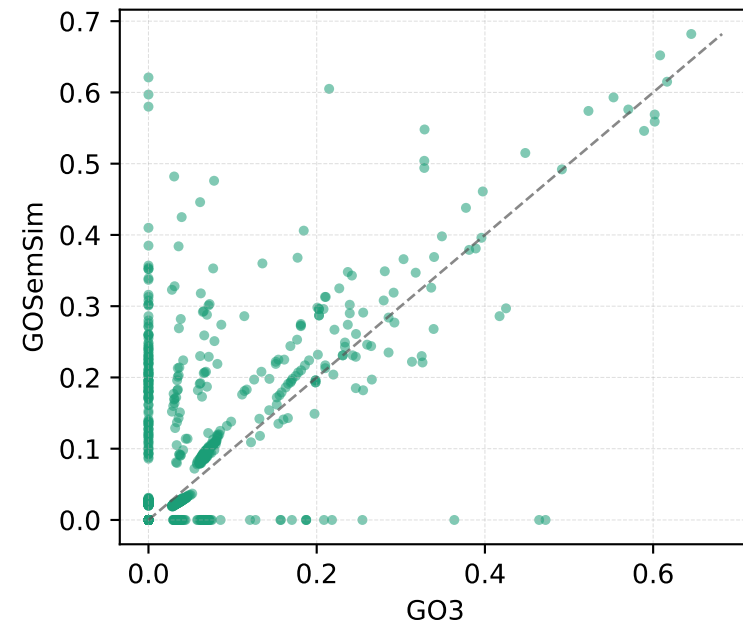
FastSemSim (r=0.702,  $\rho$ =0.620)



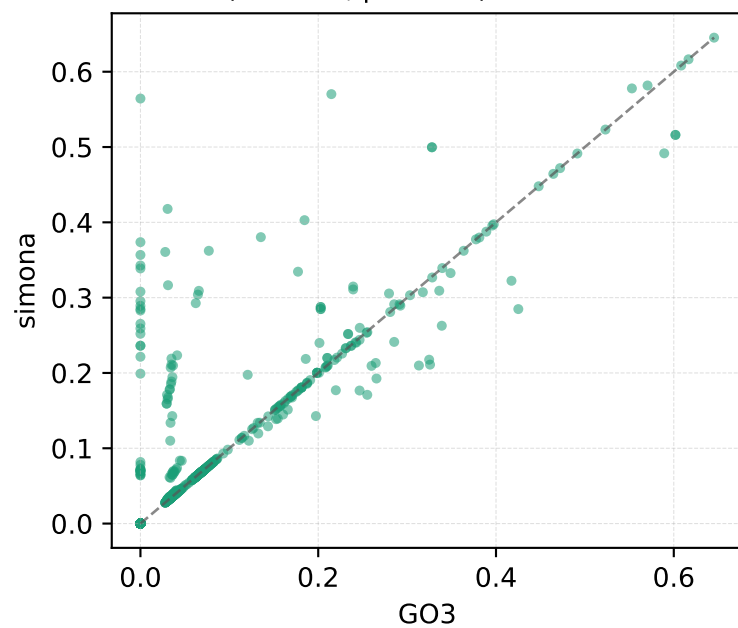
GOATOOLS (r=0.986,  $\rho$ =0.999)



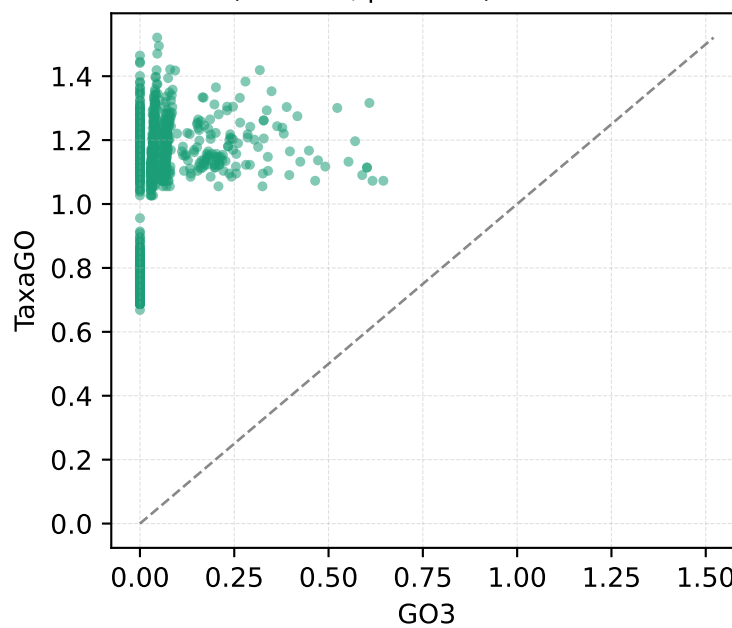
GOSemSim (r=0.629,  $\rho$ =0.472)



simona (r=0.849,  $\rho$ =0.896)

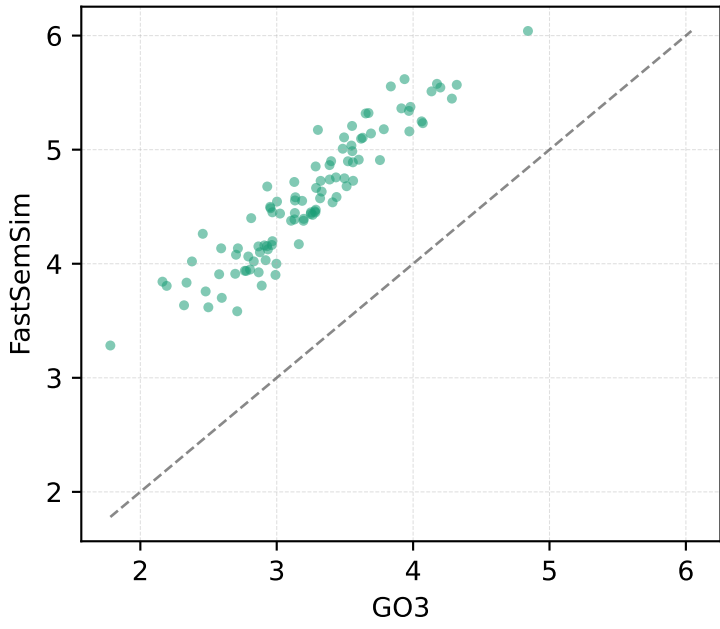


TaxaGO (r=0.304,  $\rho$ =0.459)

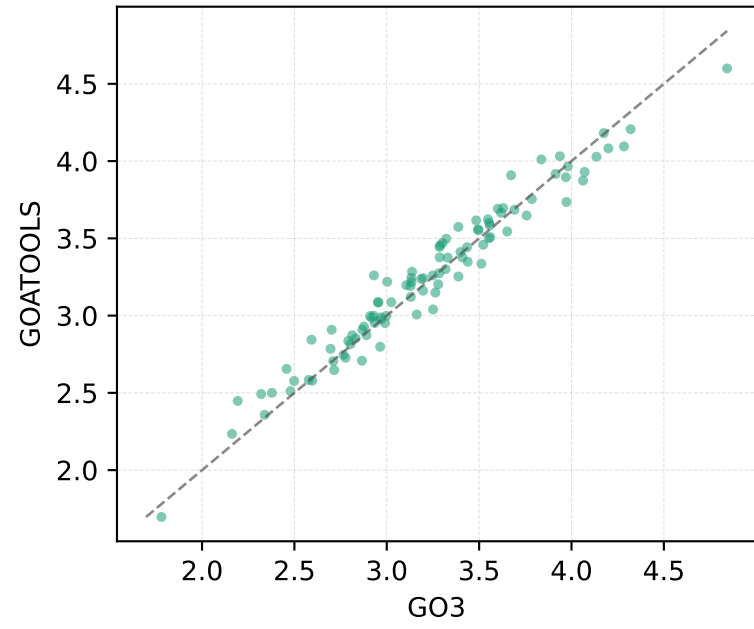


# Gene-level Resnik — GO3 vs others (n=100)

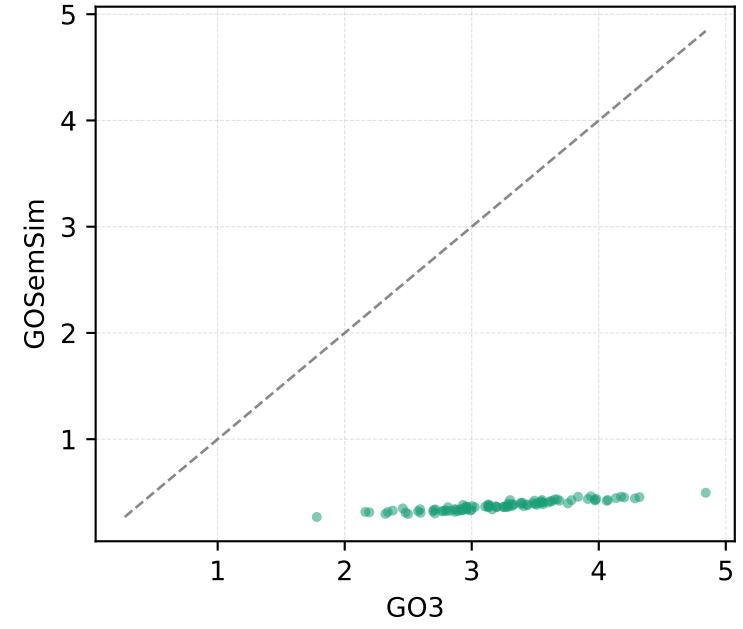
FastSemSim ( $r=0.935$ ,  $\rho=0.942$ )



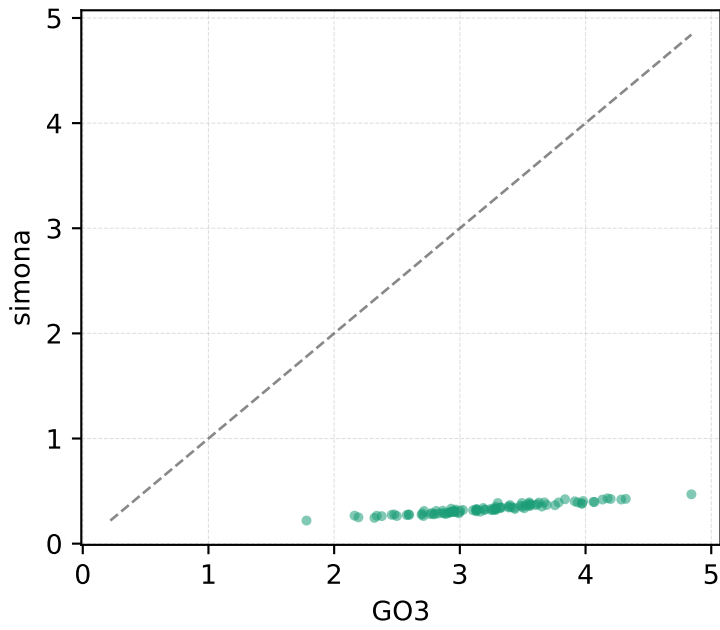
GOATOOLS ( $r=0.977$ ,  $\rho=0.977$ )



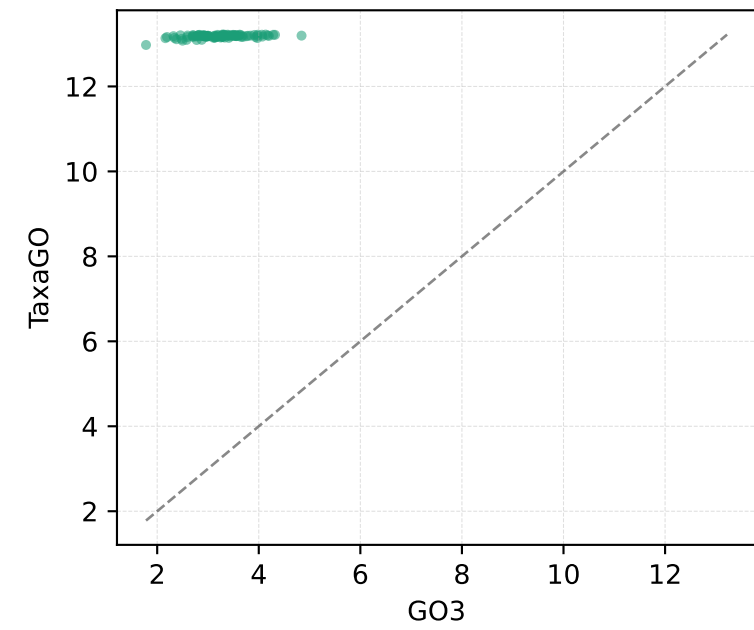
GOSemSim ( $r=0.935$ ,  $\rho=0.945$ )



simona ( $r=0.959$ ,  $\rho=0.965$ )

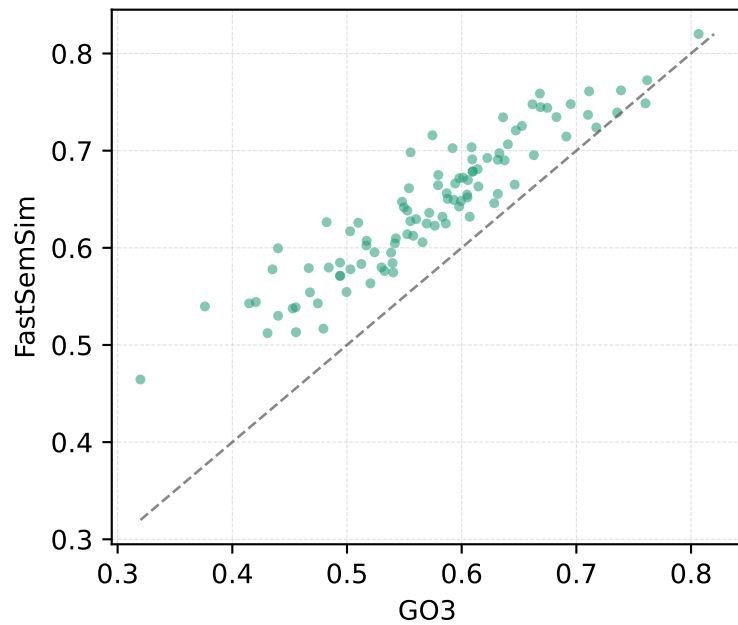


TaxaGO ( $r=0.482$ ,  $\rho=0.380$ )

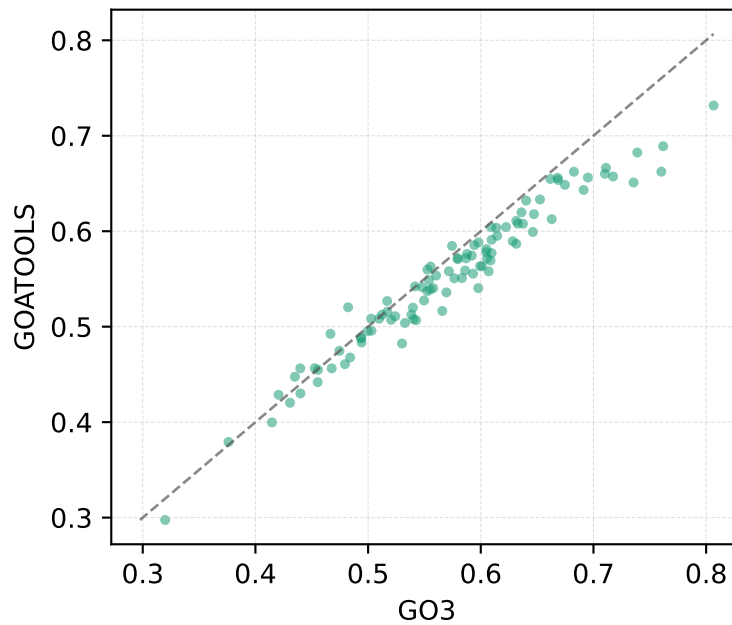


# Gene-level Lin — GO3 vs others (n=100)

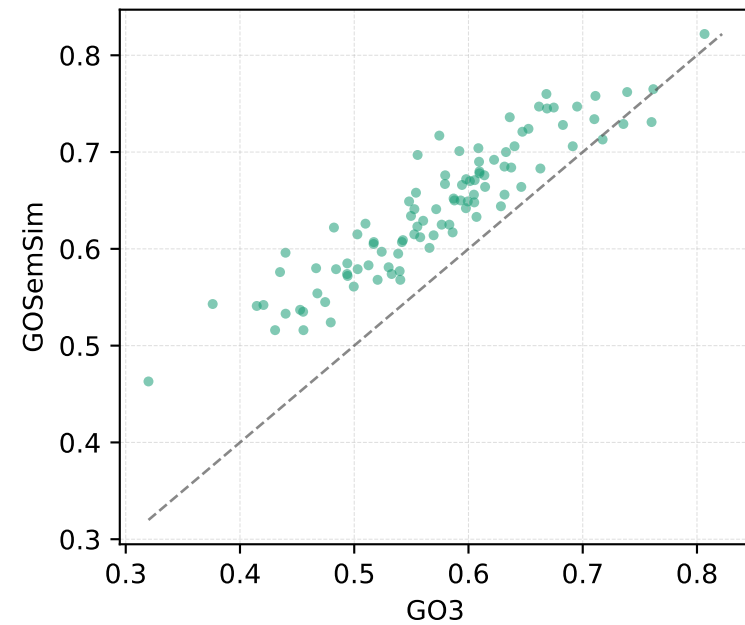
FastSemSim (r=0.930,  $\rho$ =0.933)



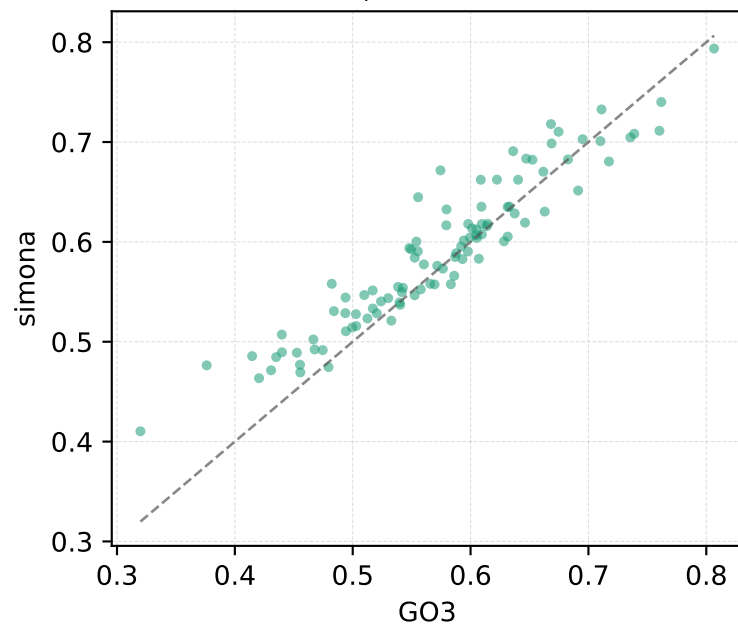
GOATOOLS (r=0.976,  $\rho$ =0.975)



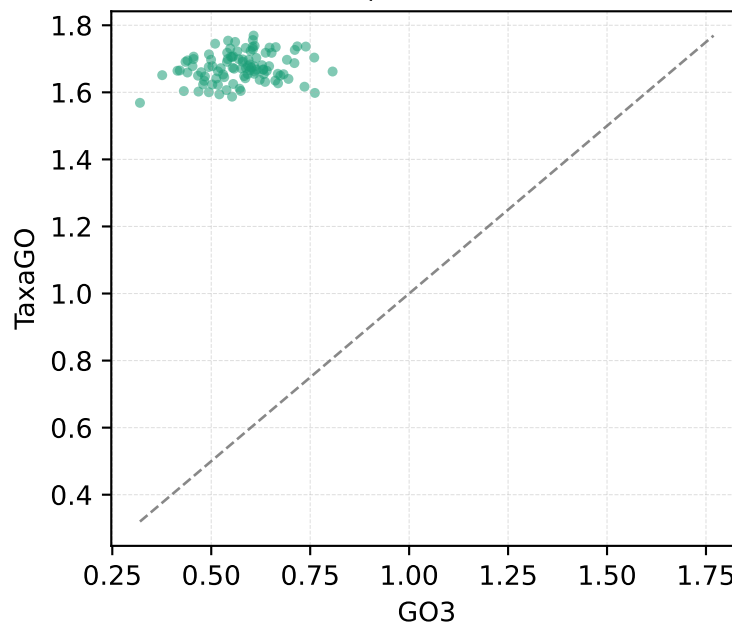
GOSemSim (r=0.924,  $\rho$ =0.928)



simona (r=0.944,  $\rho$ =0.945)



TaxaGO (r=0.205,  $\rho$ =0.182)



# ANALYSIS OF RESULTS

=====

## 1. Term-level agreement (Lin similarity, n=1,035 pairs)

-----

GO3 vs GOATOOLS	Pearson r=0.9862	Spearman rho=0.9988	Max  diff =0.2560
GO3 vs FastSemSim	Pearson r=0.7023	Spearman rho=0.6196	Max  diff =0.6090
GO3 vs GOSemSim	Pearson r=0.6290	Spearman rho=0.4725	Max  diff =0.6210
GO3 vs simona	Pearson r=0.8488	Spearman rho=0.8960	Max  diff =0.5644
GO3 vs TaxaGO	Pearson r=0.3036	Spearman rho=0.4592	Max  diff =1.4751

## 2. Gene-level agreement (Lin/BMA similarity, n=100 pairs)

-----

GO3 vs GOATOOLS	Pearson r=0.9757	Spearman rho=0.9751	Max  diff =0.0979
GO3 vs FastSemSim	Pearson r=0.9298	Spearman rho=0.9326	Max  diff =0.1636
GO3 vs GOSemSim	Pearson r=0.9237	Spearman rho=0.9279	Max  diff =0.1669
GO3 vs simona	Pearson r=0.9438	Spearman rho=0.9452	Max  diff =0.1003
GO3 vs TaxaGO	Pearson r=0.2054	Spearman rho=0.1817	Max  diff =1.2751

## 3. Key observations

-----

- GO3 and GOATOOLS show near-perfect agreement (Pearson  $r > 0.98$  for both Resnik and Lin at term level), confirming that GO3's IC computation and most-informative common ancestor (MICA) selection match the reference Python implementation.
- FastSemSim and GOSemSim agree strongly with each other ( $r > 0.92$  at term level) but show moderate divergence from GO3/GOATOOLS ( $r \sim 0.63$ - $0.70$  for Lin). This is expected: FastSemSim and GOSemSim use a different ancestor-traversal strategy that can select a different MICA in some cases.
- simona shows good rank agreement with GO3 (Spearman  $\rho \sim 0.90$  for Resnik), with moderate Pearson  $r$  due to a different IC scale. After min-max normalisation the agreement improves substantially.
- TaxaGO shows the largest divergence from all other tools, likely due to its independent OBO parser and IC computation pipeline. Pearson  $r$  values of  $0.30$ - $0.48$  indicate substantial scale differences, though rank agreement is moderate (Spearman  $\rho \sim 0.46$ - $0.55$ ).
- At gene level, agreement is consistently higher than at term level for most tool pairs, because the Best Match Average (BMA) aggregation acts as a smoothing operator over individual term-pair disagreements. GO3 vs GOATOOLS gene-level Pearson  $r$  exceeds  $0.97$  for both methods.