

# Supporting Information for GESim: Ultrafast Graph-Based Molecular Similarity Calculation via von Neumann Graph Entropy

Hiroaki Shiokawa,<sup>\*,†,‡,§</sup> Shoichi Ishida,<sup>\*,¶,‡,§</sup> and Kei Terayama<sup>\*,¶,‡</sup>

<sup>†</sup>*Center for Computational Sciences, University of Tsukuba, Tennodai 1-1-1, Tsukuba,  
Ibaraki, 305-8577, Japan*

<sup>‡</sup>*MolNavi LLC, #402 Wizard building 1-4-3 Sengen-cho Nishi-ku, Yokohama 220-0072  
Kanagawa, Japan*

<sup>¶</sup>*Graduate School of Medical Life Science, Yokohama City University, 1-7-29, Suehiro-cho,  
Tsurumi-ku, Yokohama 230-0045 Kanagawa, Japan*

<sup>§</sup>*These authors contributed equally to this work*

E-mail: shiokawa@cs.tsukuba.ac.jp; ishida.sho.nm@yokohama-cu.ac.jp;  
terayama@yokohama-cu.ac.jp

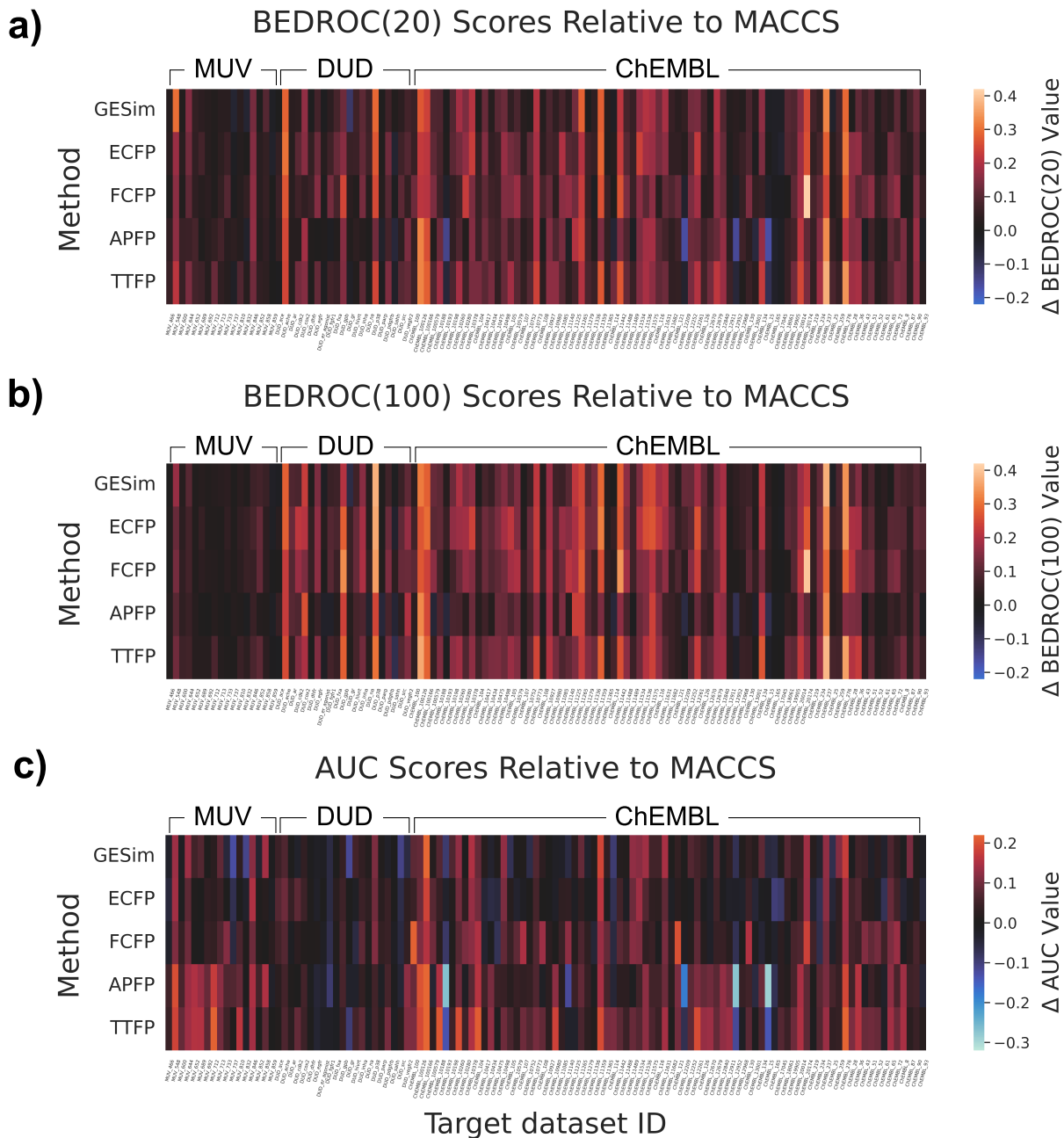


Figure S1: Average performance of six molecular similarity measures with (a) BEDROC( $\alpha = 20$ ), (b) BEDROC( $\alpha = 100$ ), and (c) AUC on the ligand-based virtual screening benchmark. The performance differences between each of the five similarity measures (GESim, ECFP, FCFP, APFP, and TTFP) and MACCS were visualized using a heatmap. Red indicates better performance compared with the MACCS, while blue indicates worse performance. The raw values used in the plots are available as CSV files in the Supporting Information.

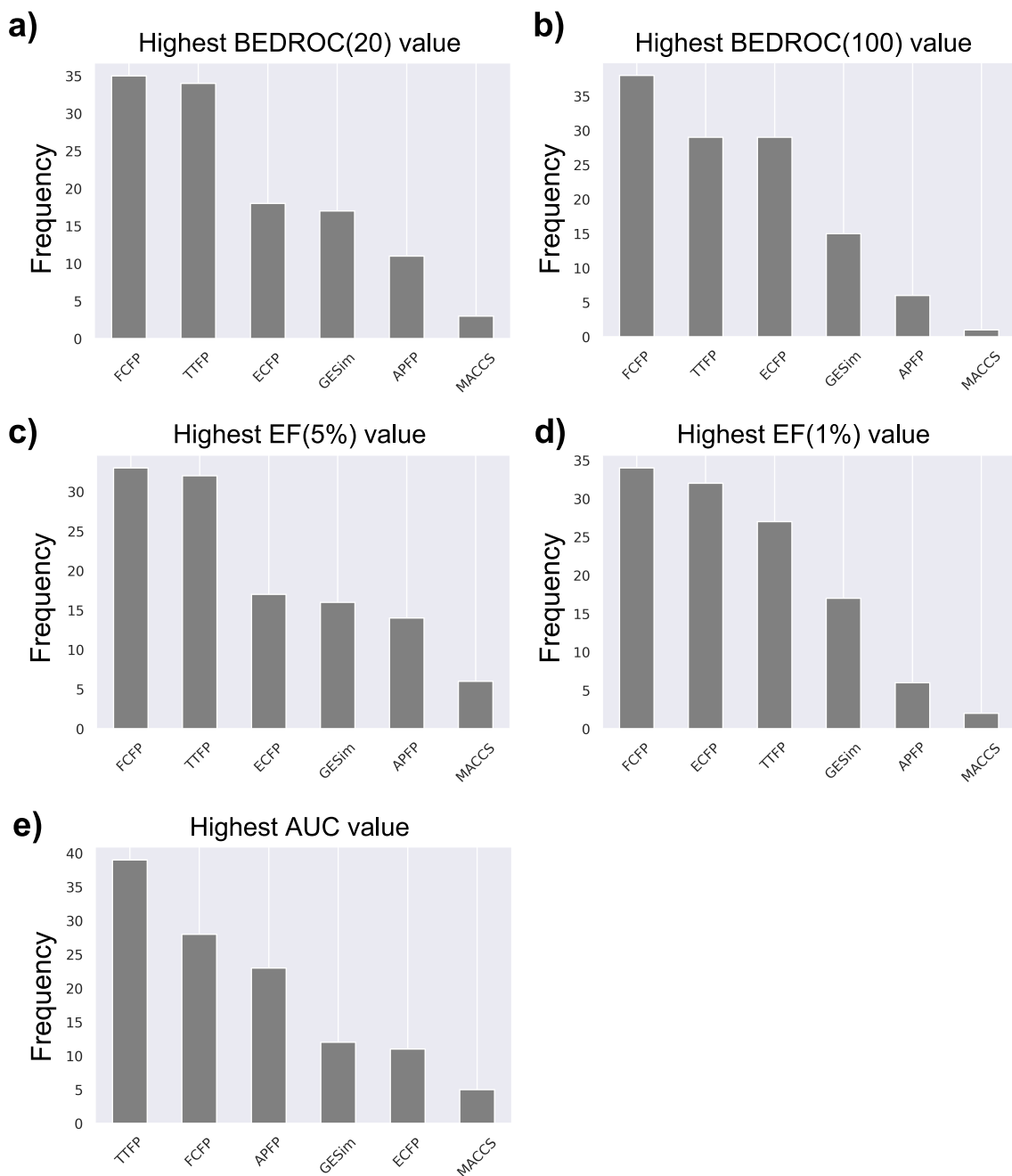
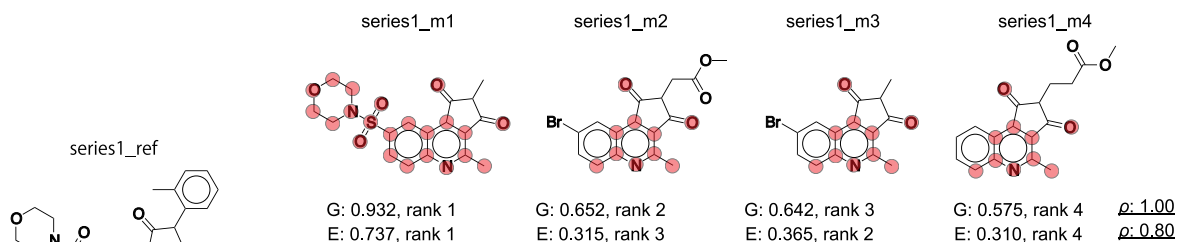


Figure S2: Statistical analysis of the ligand-based virtual screening benchmark. (a) The highest BEDROC( $\alpha = 20$ ) count, (b) highest BEDROC( $\alpha = 100$ ) count, (c) highest EF(5%) count, (d) highest EF(1%) count, and (e) highest AUC count across 118 targets are shown as bar plots.

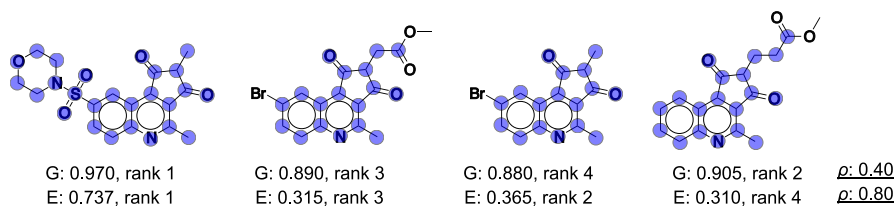
## Reference

## Four members of a series

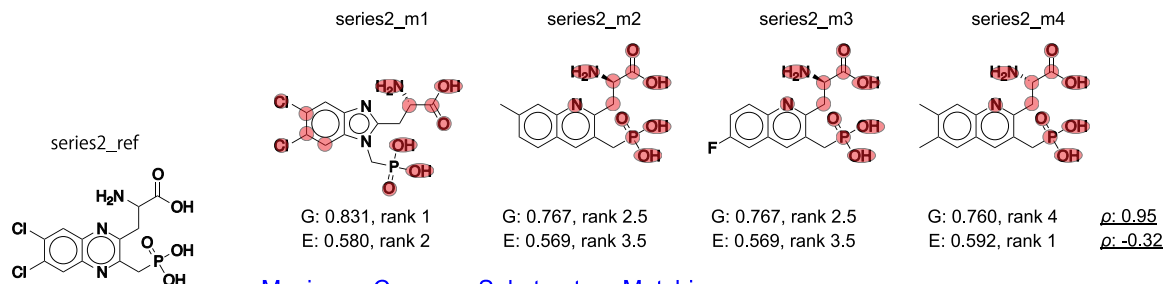
### Approximate Subgraph Matching



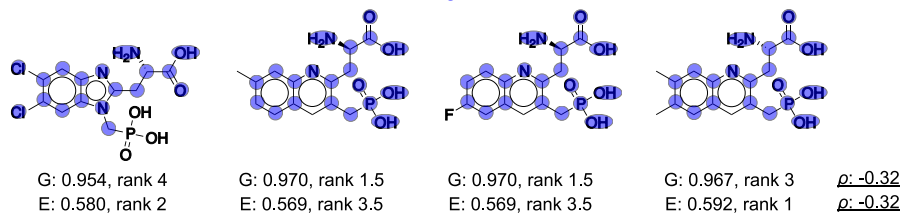
### Maximum Common Substructure Matching



### Approximate Subgraph Matching



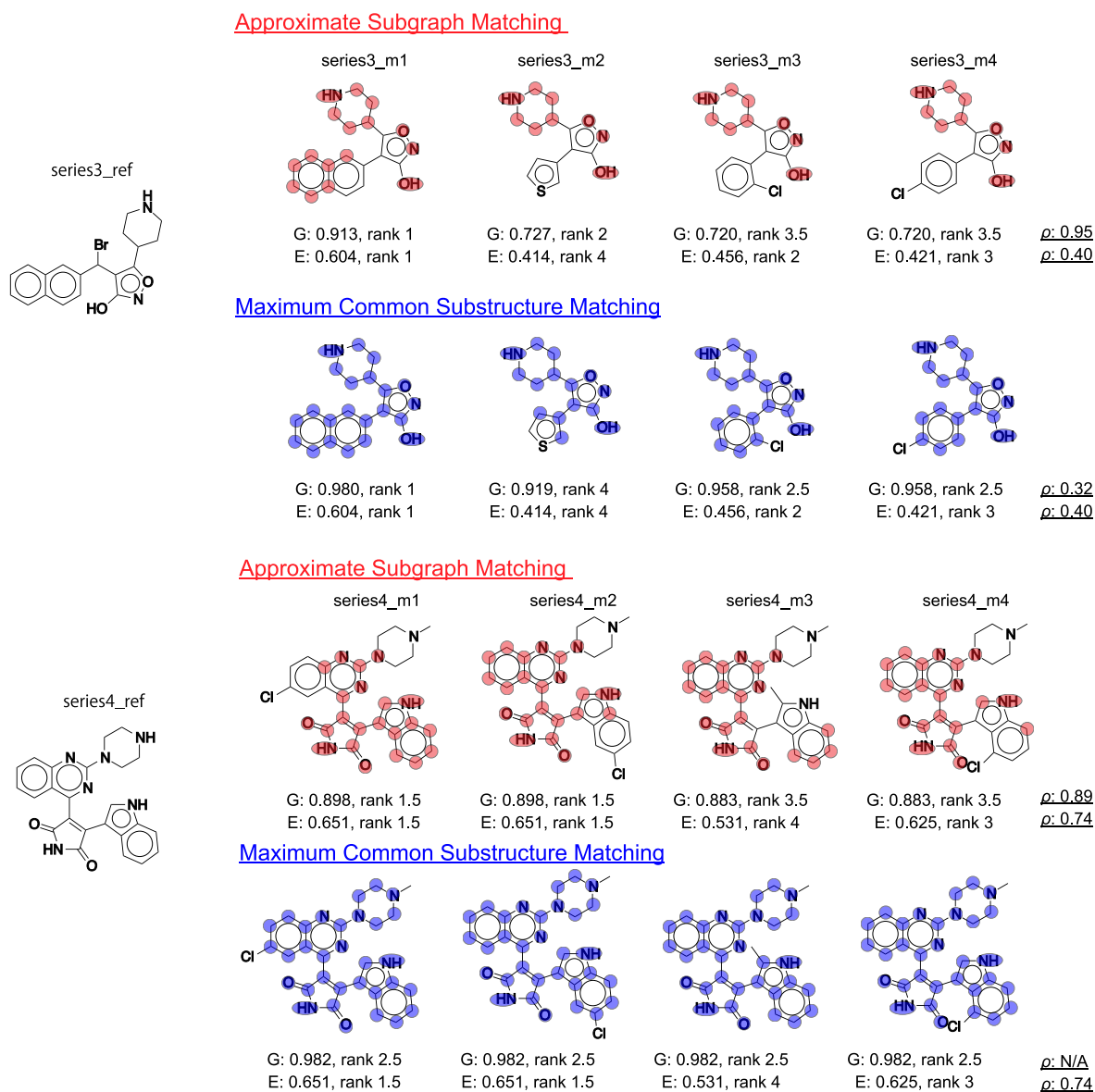
### Maximum Common Substructure Matching



Structurally similar to the reference molecule

## Reference

## Four members of a series



Structurally similar to the reference molecule

## Reference

## Four members of a series

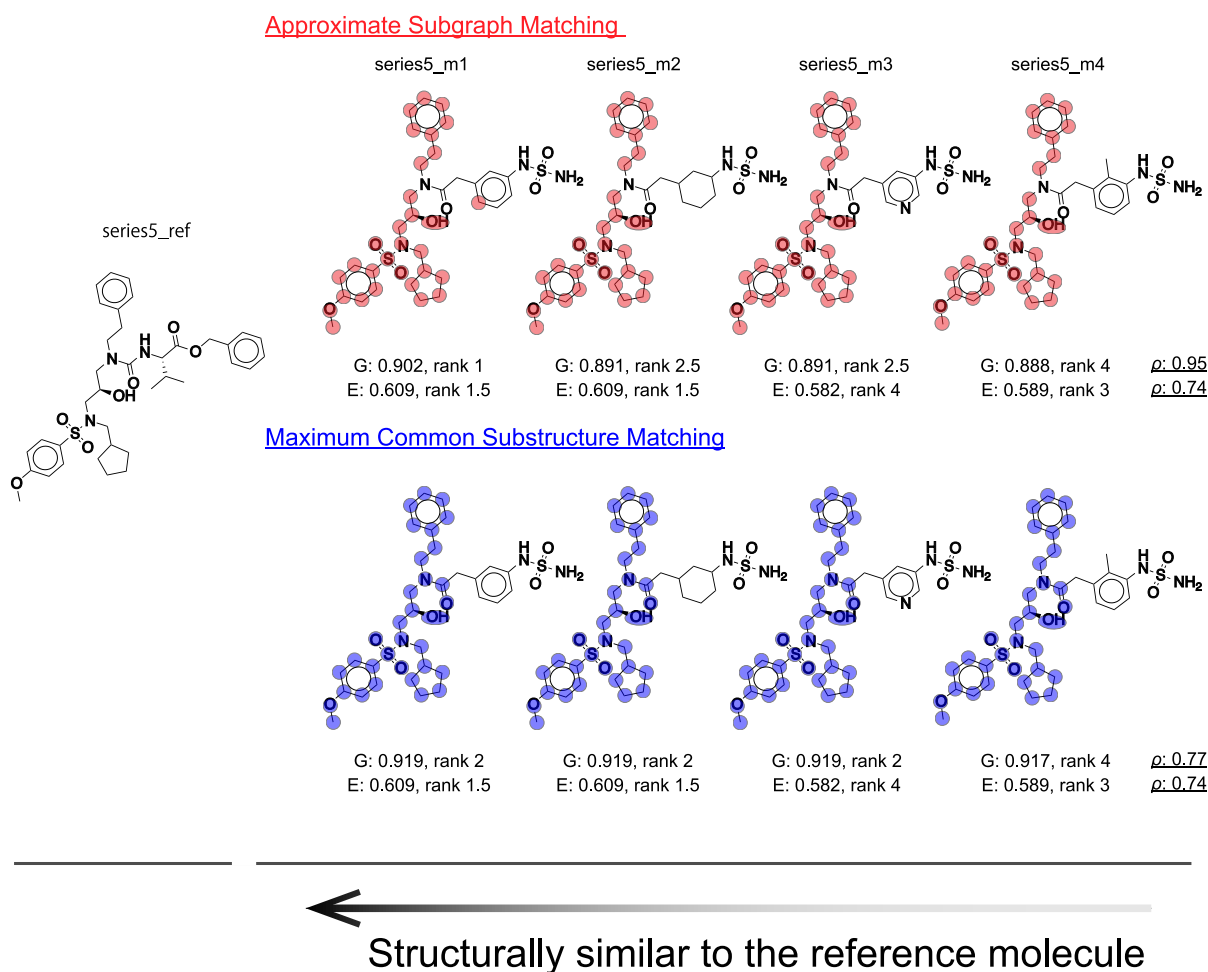


Figure S3: (pages ranging from 4 to 6) Comparison of atom-pair matching between GESim and its variant with MCS instead of GraphAligner (hereafter called MCS). The red- and blue-highlighted atoms within each molecule represent those that match atoms in the reference molecule using GraphAligner and MCS, respectively. Five molecules in a series from the structural similarity benchmark are positioned horizontally, where the first molecule serves as the reference and the next four are ordered on the basis of their similarity to this reference. The values labeled “G” and “E” below each molecule denote the similarity scores calculated by GESim and the Tanimoto similarity using ECFP, respectively, in relation to the reference molecule. The rankings in descending order based on these scores are also indicated, along with Spearman’s rank correlation coefficient ( $\rho$ ) computed from these rankings.

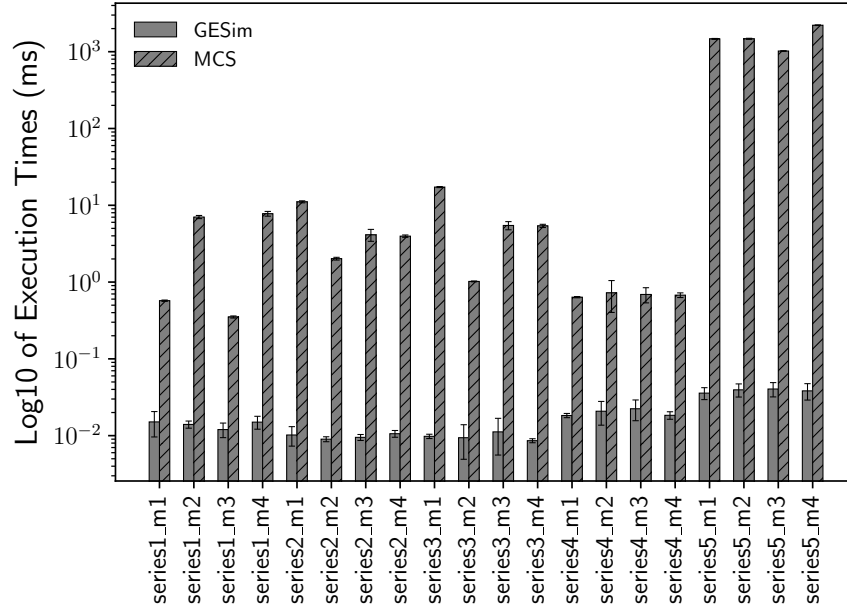


Figure S4: Calculation time comparison between GESim and its variant with MCS instead of GraphAligner (hereafter called MCS). The bar plots show the mean computation time with the standard deviation for each method across 10 trials. The vertical axis represents the logarithm of the total computation time in milliseconds.