Exploring Chemical Biosynthetic Design Space with Transform-MinER

Jonathan D. Tyzack[a*], Antonio J. M. Ribeiro[a], Neera Borkakoti[a], Janet M. Thornton[a]

[a] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI),

Wellcome Genome Campus, CB10 1SD, United Kingdom

[*] Correspondence: tyzack@ebi.ac.uk

Supplementary Information - KEGG Pathway analysis

The following analysis provides more background information around the KEGG pathway fragment dataset. Firstly, the similarity of intermediate pathway nodes to the pathway fragment target is analyzed to investigate how the KEGG pathway fragments progress towards their target product. Secondly, the change in similarity to the target product across steps in the pathway fragments is analyzed to further investigate how the KEGG pathway fragments progress towards their target product. Thirdly, an analysis of the exploration of nodes using different time cut-offs and minimum similarity thresholds is explored to investigate the effect of these parameters on the search.

Analysis of Node Similarity to Pathway Fragment Target

Supplementary Table 1 shows that similarity to target increases for 73% of intermediate steps in the KEGG pathway fragments, where an intermediate step is defined as one that doesn't terminate at the target molecule. Terminal steps are excluded from the analysis as by definition they all move closer to the target. When combining the intermediate steps into the full pathway fragment the similarity to target increases across all intermediate steps for 29%, showing that most contain a step which moves further from the target.

Path fragments		(a) Pathway fragments where similarity to target increases across all intermediate steps		Intermediate steps	(b) Intermediate steps where similarity to target increases	
Length	Count	Count	%	Count	Count	%
5	167	32	19	668	473	71
4	57	24	42	171	135	79
3	48	23	48	96	70	73
Total	272	79	29	935	678	73

Supplementary Table 1: Counts and percentages of: (a) pathway fragments where similarity to target increases across all intermediate steps; and (b) intermediate steps where similarity to target increases. An intermediate step is defined as one that doesn't terminate at the target molecule.

Figures S1-S3 show an analysis of the similarity of each node in a pathway fragment to the pathway fragment target using a Morgan fingerprint method implemented in RDKit for the pathway fragments of length 5, 4 and 3 respectively. The graphs were produced using Microsoft Excel Box and Whisker plots.

The graphs show that on average, as a pathway fragment progresses, the node molecule becomes more similar to the ultimate pathway fragment target. When exploring a pathway, Transform-MinER prioritizes nodes that move towards the target as quickly as possible using steps that are most similar to native, so this result broadly supports this strategy. However, it should be noted that there is significant variation across the pathway fragments in the dataset, with those paths that vary significantly providing the greatest challenge for this prioritization method. Inevitably, there must be some prioritization method to select the next node to explore, and the logical approach in Transform-MinER (to prioritize paths that move closest to the target using steps most similar to native) makes the best use of the data available on similarity to target and similarity to native pathways to make an informed choice.

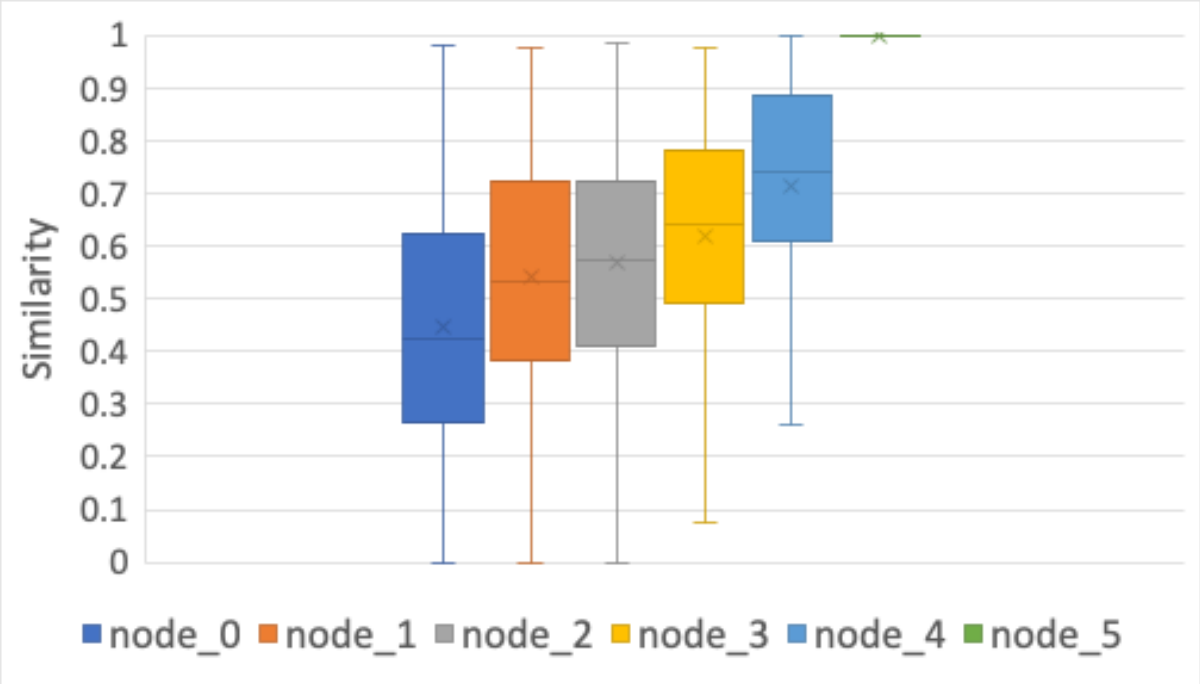

Figure S1: Similarity of nodes to target molecules in 5-step KEGG pathway fragments

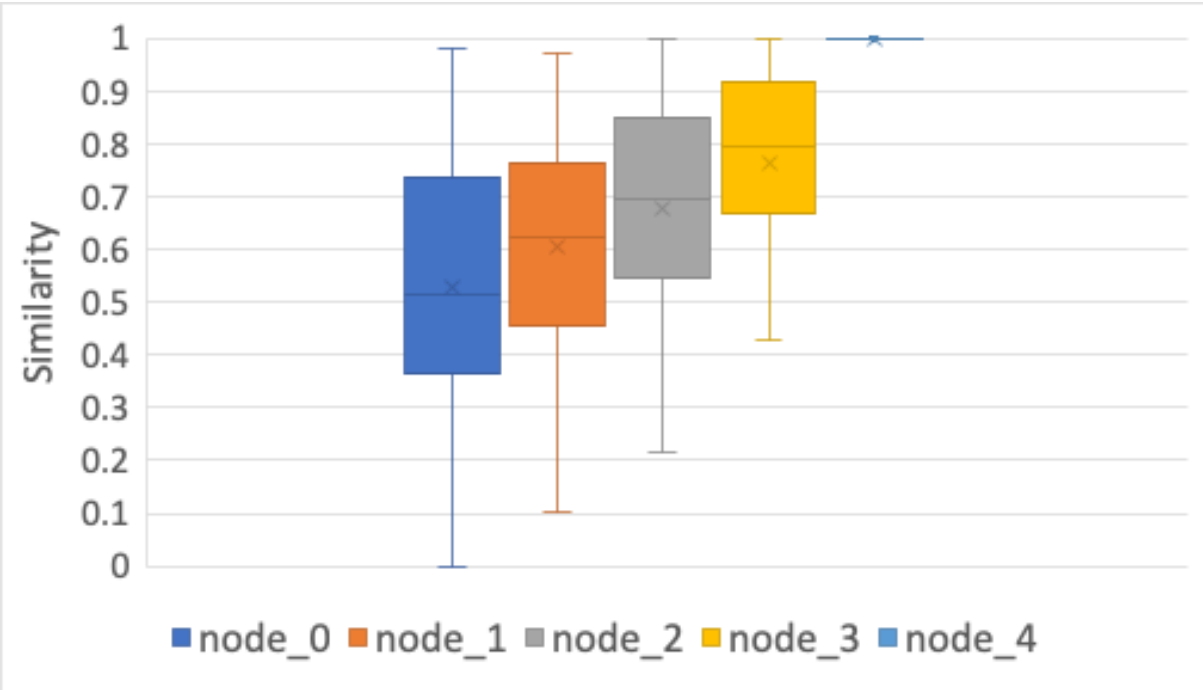

Figure S2: Similarity of nodes to target molecules in 4-Step KEGG pathway fragments

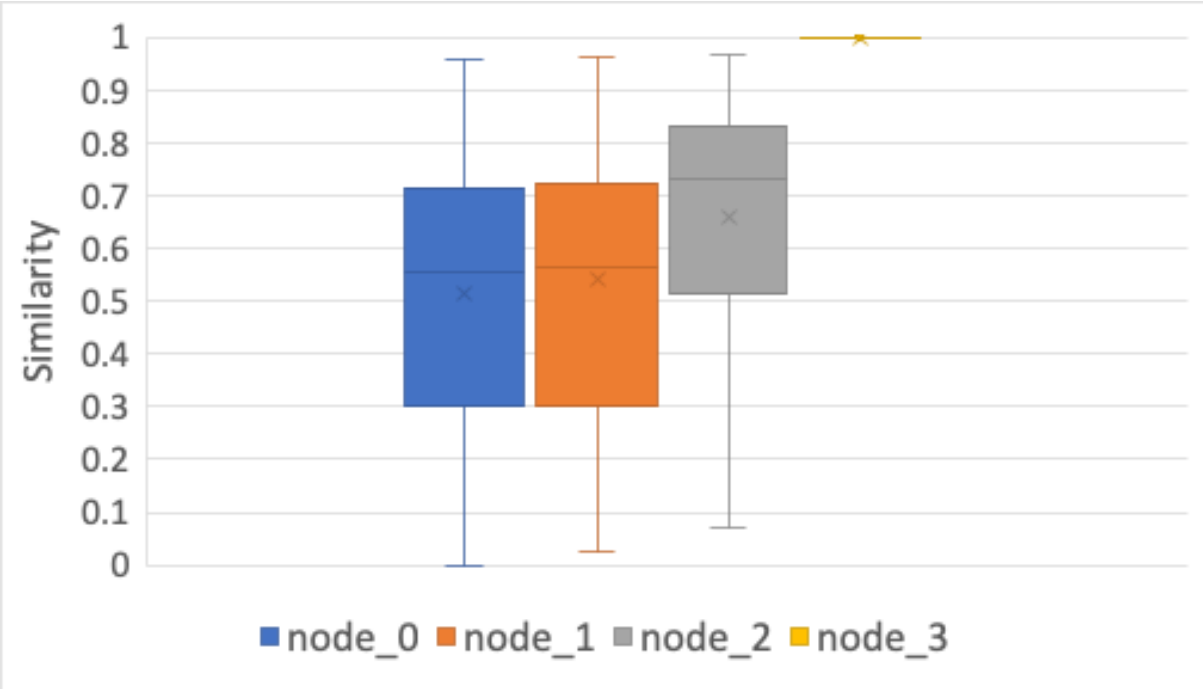

Figure S3: Similarity of nodes to target molecules in 3-step KEGG pathway fragments

Analysis of Change in Similarity towards the Target Molecule across Steps in the Pathway Fragments

Figures S4-S6 show an analysis of the distribution of change in similarity towards the target molecule across steps in the pathway fragments of length 5, 4 and 3 respectively. These figures reinforce the points made previously that a significant number of steps move further away from the target product in terms of chemical similarity, highlighting that the algorithm employed to search chemical space, prioritizing paths that appear to move most directly towards the target molecule, does not model the complexities of natural evolution.

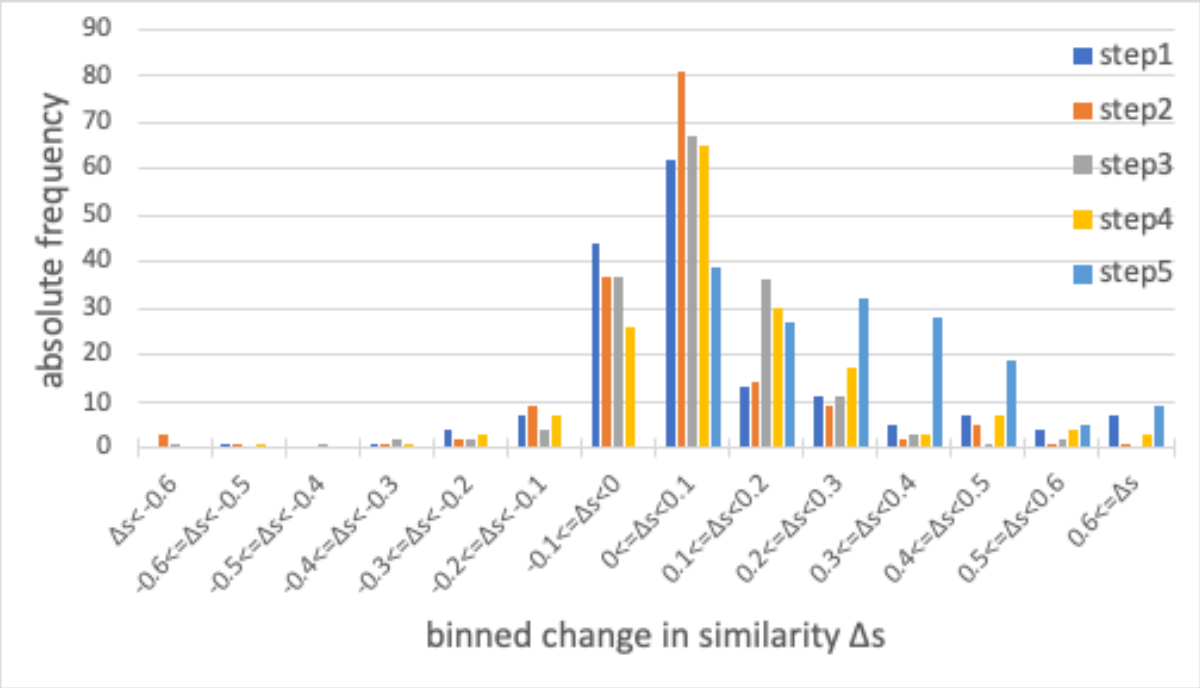

Figure S4: Distribution of change in similarity towards the target molecule across steps in 5-step KEGG pathway fragments

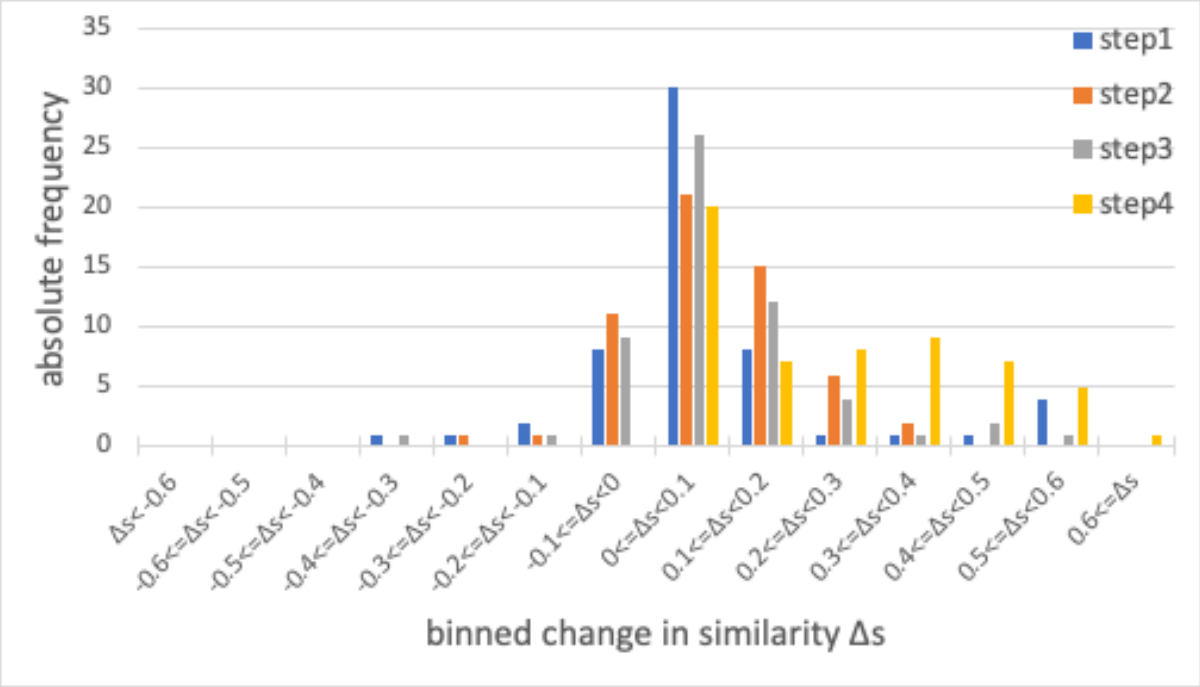

Figure S5: Distribution of change in similarity towards the target molecule across steps in 4-step KEGG pathway fragments

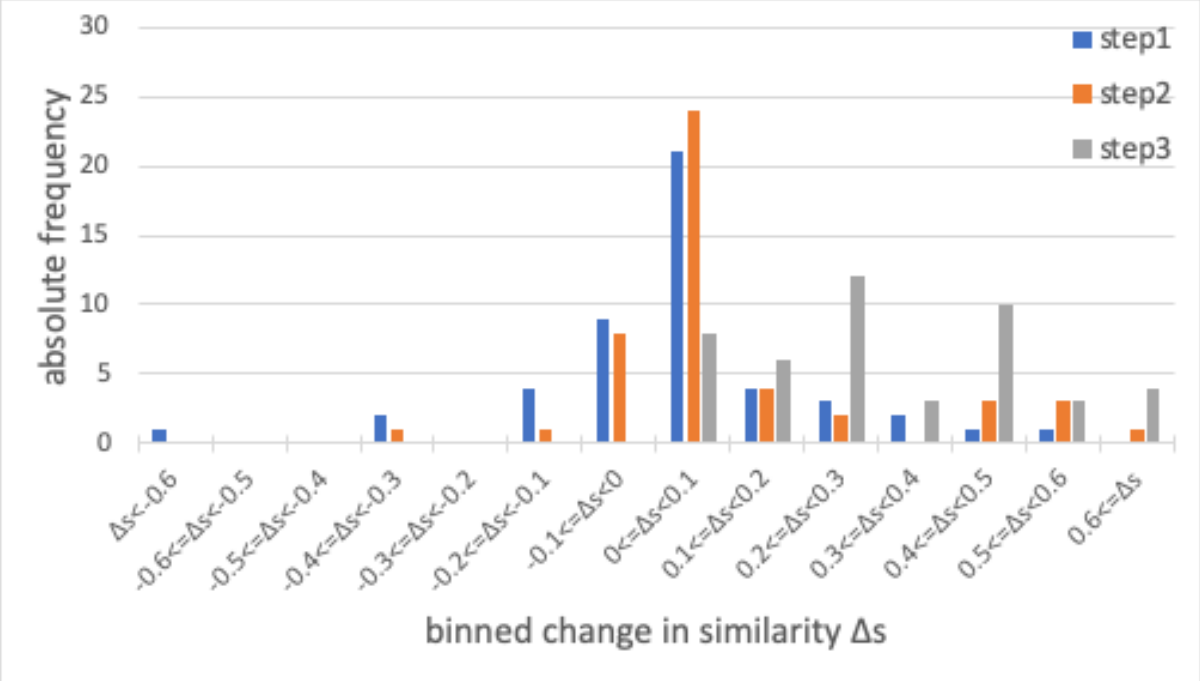

Figure S6: Distribution of change in similarity towards the target molecule across steps in 3-step KEGG pathway fragments

Scenario Analysis of Node Exploration and Success Rates

Figures S7-S9 show an analysis of the absolute and relative number of successful and unsuccessful nodes explored for different scenarios for the pathway fragments of length 5, 4 and 3 respectively. The different scenarios involve setting the time cut-off at 60, 90 and 120 seconds and varying the minimum similarity threshold (minSimThresh) between 1.0 and 0.5 at steps of 0.1. A node is classified as successful if at the end of the search it is part of a pathway to the target whereas it is classified as 'failing' if not.

A number of trends become apparent from the graphs in S7-S9. From parts (A) of S7-S9 it is apparent that as the time cut-off is increased more nodes can be explored, although a slight deterioration in the relative proportion of successful nodes (comparing data points with the same minSimThresh) is observable in part (B) supporting the effectiveness of the prioritization algorithm. From part (C) of S4-S6 it is also apparent that increasing the time cut-off increases the proportion of the data set where at least one successful path is found, since the algorithm has more time to perform the search.

As the minSimThresh is decreased from 1.0 to 0.5, it can be seen from parts (B) of S7-S9 that the relative proportion of successful nodes increases, which is to be expected since looser matches are being allowed. However, from part (A) a side-effect of reducing the minSimThresh becomes apparent since fewer nodes can be explored in a given time period. This is due to the extra computational burden at each node since more matches between query and native reaction center molecular environments will be above threshold, meaning that more in silico transformations and similarity calculations need to be performed at each step.

Transform-MinER has functionality to allow search parameters to be varied (within pre-defined limits such as a maximum run time of 120 seconds and maximum number of steps of 5) to give users flexibility. It is anticipated that users will be able to tune the parameters for individual searches if too few or too many results are being returned, depending on the characteristics of the molecules being input.

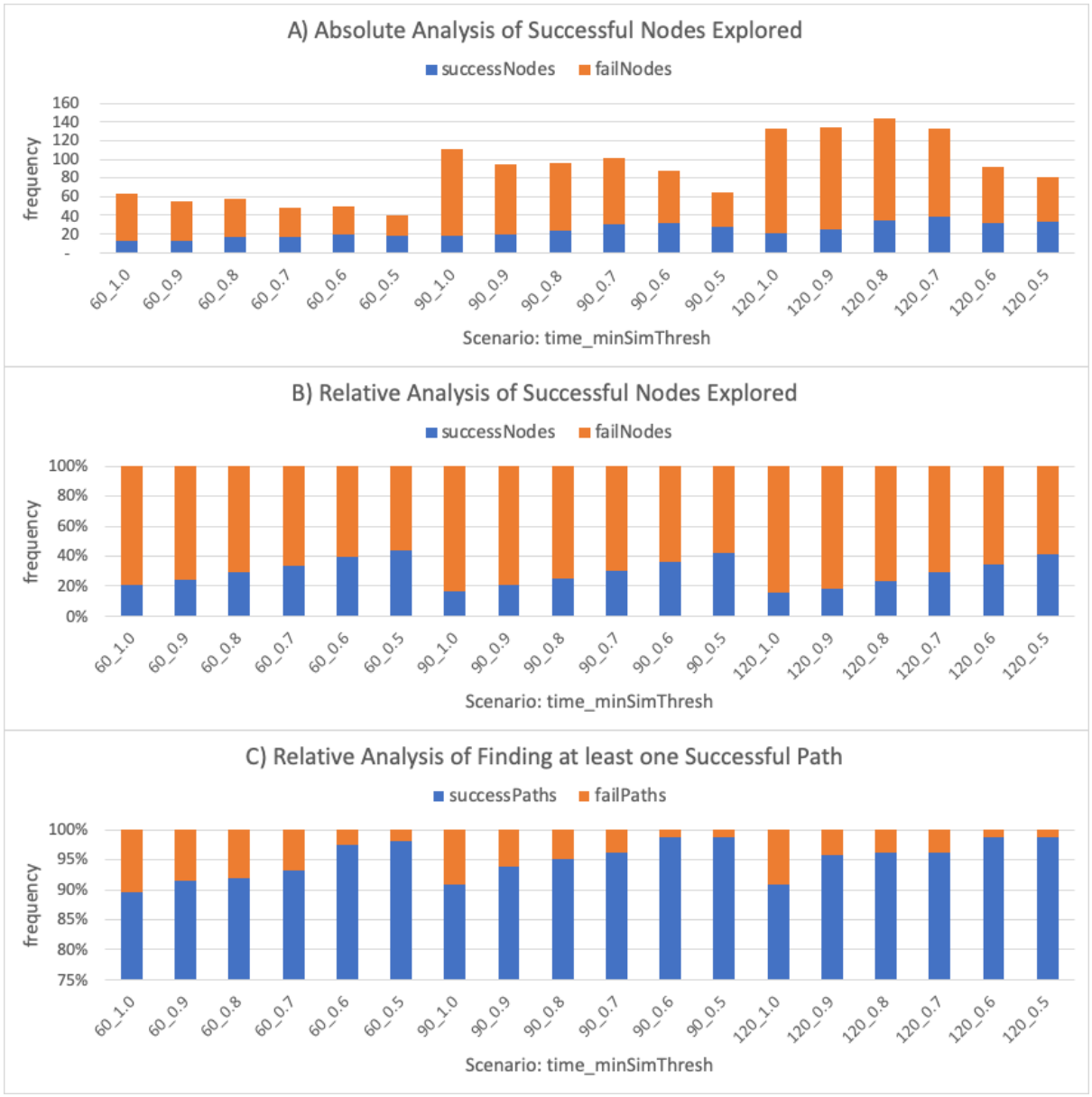

Figure S7: Successful nodes exploration analysis for pathway fragments of length 5. Across different scenarios of runtime cut-off and minimum similarity threshold: Part (A) shows an absolute analysis of successful and failing nodes explored; Part (B) shows a relative analysis of successful and failing nodes explored; and part (C) shows a relative analysis of the proportion of the data set where at least one successful path is found. A successful node is defined as one involved in a path to the target at the end of the search, otherwise it is classified as failing.

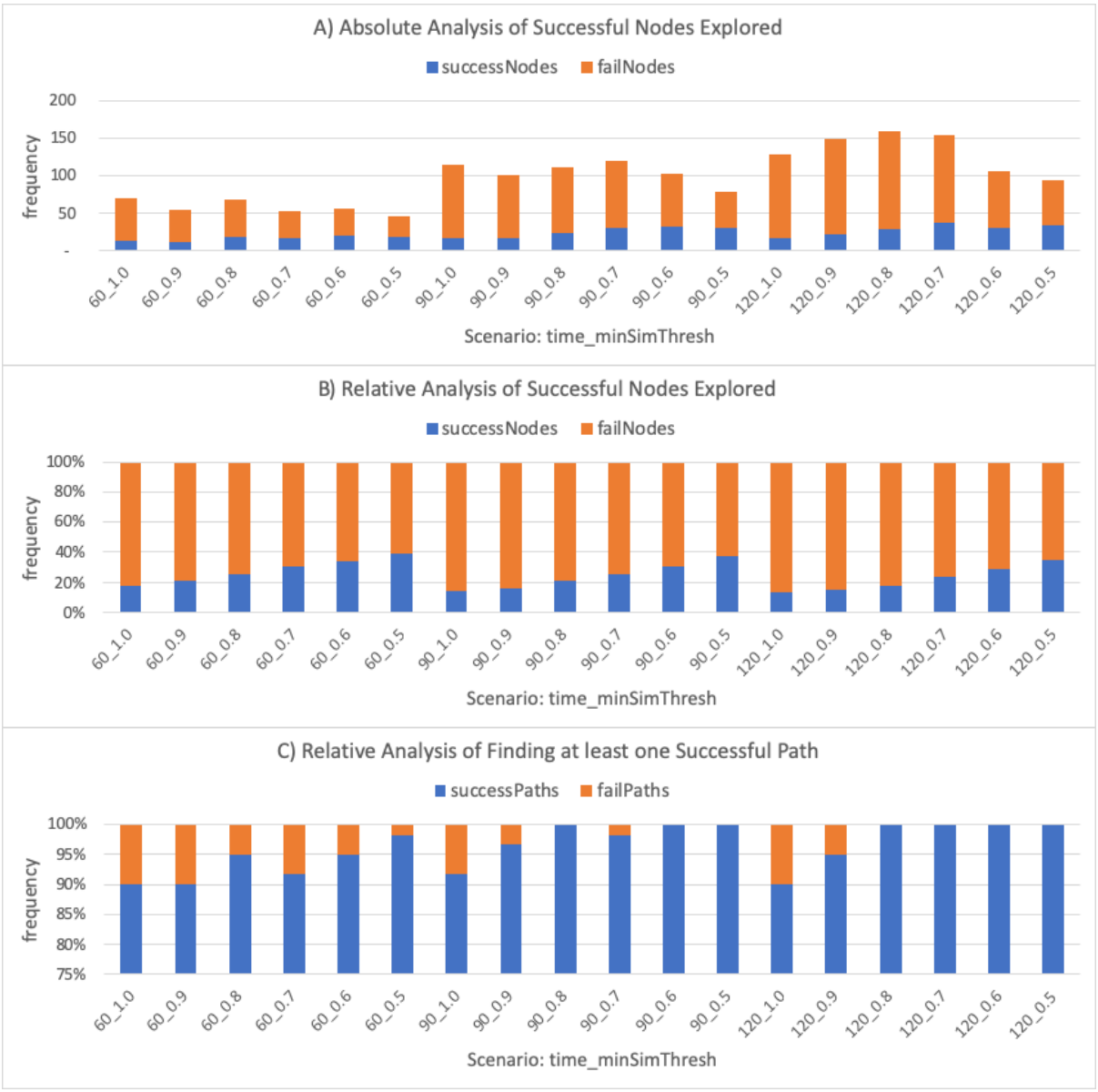

Figure S8: Successful nodes exploration analysis for pathway fragments of length 4. Across different scenarios of runtime cut-off and minimum similarity threshold: Part (A) shows an absolute analysis of successful and failing nodes explored; Part (B) shows a relative analysis of successful and failing nodes explored; and part (C) shows a relative analysis of the proportion of the data set where at least one successful path is found. A successful node is defined as one involved in a path to the target at the end of the search, otherwise it is classified as failing.

Figure S9: Successful nodes exploration analysis for pathway fragments of length 3. Across different scenarios of runtime cut-off and minimum similarity threshold: Part (A) shows an absolute analysis of successful and failing nodes explored; Part (B) shows a relative analysis of successful and failing nodes explored; and part (C) shows a relative analysis of the proportion of the data set where at least one successful path is found. A successful node is defined as one involved in a path to the target at the end of the search, otherwise it is classified as failing.