Supplementary Material

for

Integrating Protein Localization with Automated Signaling Pathway Reconstruction

Ibrahim Youssef[1,2], Jeffrey Law[3], and Anna Ritz[1]

[1]Biology Department, Reed College, Portland, OR 97202, USA, [2]Biomedical Engineering Department, Cairo University, Giza 12613, Egypt, and [3]Genetics, Bioinformatics and Computational Biology, Virginia Tech, Blacksburg, VA 24061, USA.

1 Illustrative Example of the Dynamic Program

In this section we give a simple, illustrative example of the dynamic program to break tied paths described in Section 2.4 in the main manuscript. We first summarize the dynamic program as described in the text, then modify it to work with the original localization scores for ease of exposition. We then walk through scoring a four-edge path with this approach.

Consider the set $V = \{v_1, v_2, \ldots\}$ of proteins that contain localization information (e.g. localization scores for at least one of *ExtMem, Cytosol, and Nucleus*). For each protein v, we use $\ell_v^{ext}, \ell_v^{cyt}$, and ℓ_v^{nuc} to denote these scores, where $0 \leq \ell \leq 1$ for all scores. We log-transform these scores, that is, $\mathcal{T}_v^c = -\log \ell_v^c$ for each protein v and each cellular compartment c. Our goal is to find a selection of compartments that maximize the path score by (by summing log-transformed scores) while respecting the signaling flow structure outlined in Section 2.3 in the main manuscript. Let $v_1, v_2, \ldots, v_m$ be the m proteins in path P_i.We aim to compute the optimal signaling score of the entire path ending in the nucleus, which we denote by $s(v_m, nuc)$. This score can be computed as:

$$s(v_m, nuc) = \min\left[s(v_{m-1}, cyt), s(v_{m-1}, nuc)\right] + \mathcal{T}_{v_m}^{nuc}. \tag{1}$$

The largest score of the entire path (to v_m) ending in the nucleus is the score of protein v_m in the nucleus plus the maximum of *either* (a) the largest score of the path up to v_{m-1} in the nucleus, or (b) the largest score of the path up to v_{m-1} in the cytosol. This is consistent with our assumptions that (1) the path must end in the nucleus, and (2) the last interaction must either be in the nucleus or must involve a protein in the cytosol and a protein in the nucleus. In general, at node v_j, $j = 2, 3, \ldots, (m-1)$, the series of equations for the scores are:

$$s(v_j, ext) = s(v_{j-1}, ext) + \mathcal{T}_{v_j}^{ext} \tag{2}$$

$$s(v_j, cyt) = \min\left[s(v_{j-1}, ext), s(v_{j-1}, cyt)\right] + \mathcal{T}_{v_j}^{cyt} \tag{3}$$

$$s(v_j, nuc) = \min\left[s(v_{j-1}, cyt), s(v_{j-1}, nuc)\right] + \mathcal{T}_{v_j}^{nuc}. \tag{4}$$

Note that we can only reach a protein in *ExtMem* from another protein in *ExtMem*, we can reach a protein in *cytosol* from another protein in either *ExtMem* or *cytosol*, and we can reach a protein in *nucleus* from another one in either *cytosol* or *nucleus*.

To ensure that the path starts with the cellular compartment *ExtMem*, the base case for these recurrence relations are:

$$s(v_1, ext) = \mathcal{T}_{v_1}^{ext} \tag{5}$$

$$s(v_1, cyt) = \infty \tag{6}$$

$$s(v_1, nuc) = \infty. \tag{7}$$

For ease of exposition, we will work with the localization scores in their original format without log-transforming them. In this case, addition in Equations (1–4) will be multiplication and *min* will be *max*. Moreover (∞) in Equations (6–7) will be $(-\infty)$. Here are the equations in their new structure.

$$s(v_m, nuc) = \max\left[s(v_{m-1}, cyt), s(v_{m-1}, nuc)\right] * \ell_{v_m}^{nuc}. \tag{8}$$

$$s(v_j, ext) = s(v_{j-1}, ext) * \ell_{v_j}^{ext} \tag{9}$$
$$s(v_j, cyt) = \max\left[s(v_{j-1}, ext), s(v_{j-1}, cyt)\right] * \ell_{v_j}^{cyt} \tag{10}$$
$$s(v_j, nuc) = \max\left[s(v_{j-1}, cyt), s(v_{j-1}, nuc)\right] * \ell_{v_j}^{nuc}. \tag{11}$$

$$s(v_1, ext) = \ell_{v_1}^{ext} \tag{12}$$
$$s(v_1, cyt) = -\infty \tag{13}$$
$$s(v_1, nuc) = -\infty. \tag{14}$$

These recurrence relations can be efficiently calculated using a dynamic program, filling an $m \times 3$ table denoting the number of nodes (m) by the three compartments. The final score taken will be $s(v_m, nuc)$, since we require that the path terminates in the nucleus.

The following is an example of a path of five nodes/proteins, $\langle v_1, v_2, \ldots, v_5 \rangle$, and four edges/interactions. The 5×3 table below represents the table used to iteratively compute the signaling score. Each column represents a protein and each row represents a cellular compartment. The localization scores (probabilities of a protein to be found in each of the three cellular compartments) are shown above the proteins. The red cells in the table represent the cells that do not affect computing the signaling score. For example, at the first protein in the path v_1, we ignore the *Cytosol* and the *Nucleus* compartments to force the path to start with a protein at either the extracellular domain or the cell membrane, and hence both cells take extreme values like $(-\infty)$.

At v_1:

To initialize the dynamic program, we apply Equations (12–14) as shown below in the first column of the table.

$Pr\{\text{Ext}\}$:	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{2}$
$Pr\{\text{Cyt}\}$:	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
$Pr\{\text{Nuc}\}$:	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$
	v_1	v_2	v_3	v_4	v_5
ExtMem	$\frac{1}{2}$	—	—	—	
Cytosol	$-\infty$	—	—	—	
Nucleus	$-\infty$	$-\infty$	—	—	—

At v_2:

From v_2 to the end of the path, we use Equations (9–11) to compute the signaling score at the intermediate proteins.

$$s(v_2, ext) = s(v_1, ext) * \ell_{v_2}^{ext} = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}.$$
$$s(v_2, cyt) = \max\left[s(v_1, ext), s(v_1, cyt)\right] * \ell_{v_2}^{cyt} = \max\left[\frac{1}{2}, -\infty\right] * \frac{3}{4} = \frac{3}{8}.$$
$$s(v_2, nuc) = \max\left[s(v_1, cyt), s(v_1, nuc)\right] * \ell_{v_2}^{nuc} = \max\left[-\infty, -\infty\right] * \frac{1}{4} = -\infty.$$

Tails of the blue arrows in the table indicate the compartment in the previous step that was used to calculate the current step and the current compartment.

$Pr\{\text{Ext}\}$:	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{2}$
$Pr\{\text{Cyt}\}$:	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
$Pr\{\text{Nuc}\}$:	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$

	v_1	v_2	v_3	v_4	v_5
ExtMem	$\frac{1}{2}$	$\frac{1}{4}$	—	—	
Cytosol	$-\infty$	$\frac{3}{8}$	—	—	
Nucleus	$-\infty$	$-\infty$	—	—	—

At v_3:

Following the same procedure used at v_2, we get the next equations.

$$s(v_3, ext) = s(v_2, ext) * \ell_{v_3}^{ext} = \frac{1}{4} * \frac{1}{4} = \frac{1}{16}.$$
$$s(v_3, cyt) = \max\left[s(v_2, ext), s(v_2, cyt)\right] * \ell_{v_3}^{cyt} = \max\left[\frac{1}{4}, \frac{3}{8}\right] * \frac{1}{2} = \frac{3}{16}.$$
$$s(v_3, nuc) = \max\left[s(v_2, cyt), s(v_2, nuc)\right] * \ell_{v_3}^{nuc} = \max\left[\frac{3}{8}, -\infty\right] * \frac{1}{4} = \frac{3}{32}.$$

$Pr\{\text{Ext}\}$:	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{2}$
$Pr\{\text{Cyt}\}$:	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
$Pr\{\text{Nuc}\}$:	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$

	v_1	v_2	v_3	v_4	v_5
ExtMem	$\frac{1}{2}$	$\frac{1}{4}$ →	$\frac{1}{16}$	—	
Cytosol	$-\infty$	$\frac{3}{8}$ →	$\frac{3}{16}$	—	
Nucleus	$-\infty$	$-\infty$	$\frac{3}{32}$	—	—

At v_4:

$$s(v_4, ext) = s(v_3, ext) * \ell_{v_4}^{ext} = \frac{1}{16} * \frac{1}{10} = \frac{1}{160}.$$

$$s(v_4, cyt) = \max\left[s(v_3, ext), s(v_3, cyt)\right] * \ell_{v_4}^{cyt} = \max\left[\frac{1}{16}, \frac{3}{16}\right] * \frac{1}{4} = \frac{3}{64}.$$

$$s(v_4, nuc) = \max\left[s(v_3, cyt), s(v_3, nuc)\right] * \ell_{v_4}^{nuc} = \max\left[\frac{3}{16}, \frac{3}{32}\right] * \frac{1}{8} = \frac{3}{128}.$$

$Pr\{\text{Ext}\}$:	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{2}$
$Pr\{\text{Cyt}\}$:	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
$Pr\{\text{Nuc}\}$:	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$

	v_1	v_2	v_3	v_4	v_5
ExtMem	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{16}$ →	$\frac{1}{160}$	
Cytosol	$-\infty$	$\frac{3}{8}$	$\frac{3}{16}$ →	$\frac{3}{64}$	
Nucleus	$-\infty$	$-\infty$	$\frac{3}{32}$ ↘	$\frac{3}{128}$	—

At v_5:

This is the last node, so we use Equatin (8).

$$s(v_5, nuc) = \max\left[s(v_4, cyt), s(v_4, nuc)\right] * \ell_{v_5}^{nuc}$$
$$= \max\left[\frac{3}{64}, \frac{3}{128}\right] * \frac{1}{4} = \frac{3}{256}.$$

$Pr\{\text{Ext}\}$:	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{2}$
$Pr\{\text{Cyt}\}$:	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
$Pr\{\text{Nuc}\}$:	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$

	v_1	v_2	v_3	v_4	v_5
ExtMem	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{160}$	$\frac{1}{320}$
Cytosol	$-\infty$	$\frac{3}{8}$	$\frac{3}{16}$	$\frac{3}{64}$	$\frac{9}{256}$
Nucleus	$-\infty$	$-\infty$	$\frac{3}{32}$	$\frac{3}{128}$	$\frac{3}{256}$

Recovering the Most Probable Compartments:

We can now trace back the most probable sequence of the cellular compartments for this path as shown with the sequence of the blue arrows below. This sequence is {*ExtMem*–*Cytosol*–*Cytosol*–*Cytosol*–*Nucleus*}. The signaling score for this path is $\frac{3}{256}$.

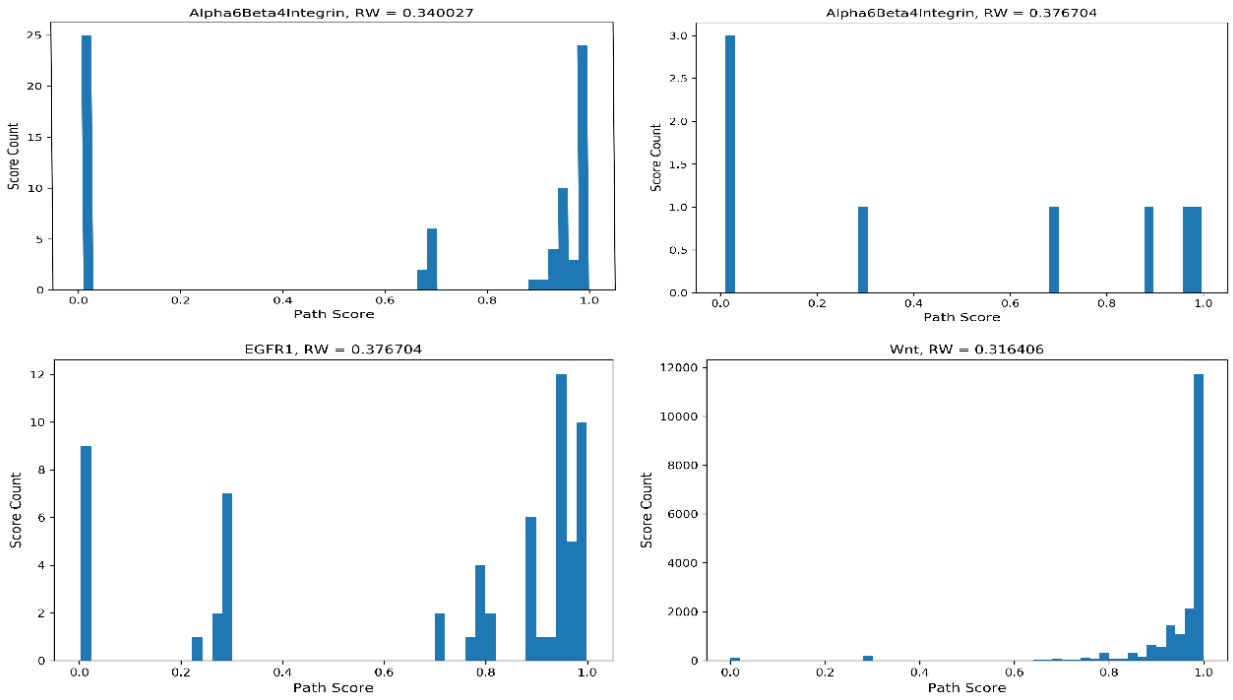
$Pr\{\text{Ext}\}$:	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{2}$
$Pr\{\text{Cyt}\}$:	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
$Pr\{\text{Nuc}\}$:	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$

	v_1	v_2	v_3	v_4	v_5
ExtMem	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{160}$	$\frac{1}{320}$
Cytosol	$-\infty$	$\frac{3}{8}$	$\frac{3}{16}$	$\frac{3}{64}$	$\frac{9}{256}$
Nucleus	$-\infty$	$-\infty$	$\frac{3}{32}$	$\frac{3}{128}$	$\mathbf{\frac{3}{256}}$

ExtMem → *Cytosol* → *Cytosol* → *Cytosol* → *Nucleus*

2 Signaling Scores Histogram

Tied paths share the same reconstruction cost, but they have different signaling score values. We use the signaling scores to re-prioritize these tied paths (Section 2.1). S.Figure 1 shows four histogram examples of the signaling scores for different NetPath pathways [Kandasamy et al., 2010]. Although all the paths in each example have the same reconstruction cost, they differ in their signaling scores, which span a large dynamic range from almost zero to almost unity. We rely on these distinct signaling scores of the tied paths to reorder the paths and re-prioritize them. A path with a higher signaling score will be ranked up early in the paths list, while another path with a lower signaling score will be pushed down the list.



S.Figure 1. Signaling scores histogram for paths with tied reconstruction score. The title of each sub-figure indicates the pathway name and the tied reconstruction score for a group of paths, while the sub-figure shows the histogram of the signaling scores for those paths that share that reconstruction cost.

3 Incorporating the Mitochondria Compartment in the Signaling Model

The mitochondria compartment, denoted here as *Mt*, was added to the signaling model as an intermediate compartment, like the *Cytosol*, between the two terminal compartments *ExtMem* and *Nucleus*. S.Figure 2 shows a simplified diagram for the interrelationships between the different signaling compartments. We can reach a protein in the *ExtMem* only from another protein in the *ExtMem*. We can reach a protein in the *Cytosol* from another protein in one of the three compartments: *ExtMem*, *Cytosol*, or *Mt*. We can reach a protein in the *Mt* from another protein in one of the three compartments: *ExtMem*, *Cytosol*, or *Mt*. Finally, we can end up with a protein in the *Nucleus* from a protein in one of the three compartments: *Cytosol*, *Mt*, or *Nucleus*. This signaling model allows for having cyclic paths because of having two intermediate compartments: *Cytosol* and *Mt*. To adhere to the signaling assumptions outlined in Section 2.3 in the main text, a path has to start with a protein in the *ExtMem*, end with a protein in the *Nucleus*, and have at least one protein in one of the intermediate compartments. So it is not necessary for a single path to have proteins in all the four cellular compartments, and consequently a path may have either three or four cellular compartments.

Equations (1-7) are re-written here to consider the mitochondria compartment. The path final signaling score is computed as:

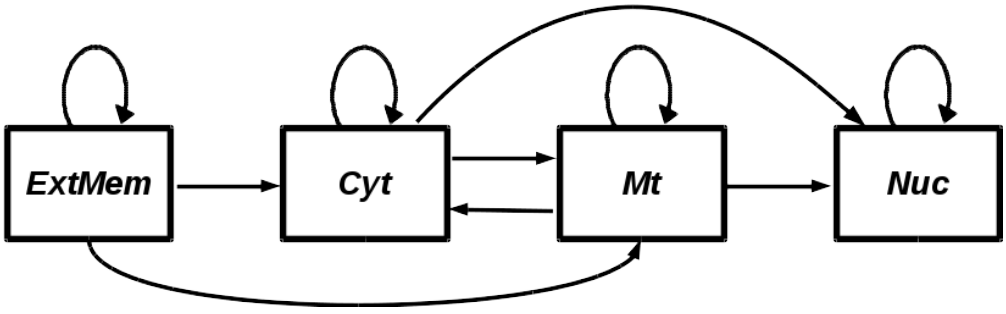
$$s(v_m, nuc) = \min\left[s(v_{m-1}, cyt), s(v_{m-1}, Mt), s(v_{m-1}, nuc)\right] + \mathcal{T}_{v_m}^{nuc}.$$

At node v_j, $j = 2, 3, \ldots, (m-1)$, the series of equations for the scores are:

$$s(v_j, ext) = s(v_{j-1}, ext) + \mathcal{T}_{v_j}^{ext} \tag{15}$$
$$s(v_j, cyt) = \min\left[s(v_{j-1}, ext), s(v_{j-1}, cyt), s(v_{j-1}, Mt)\right] + \mathcal{T}_{v_j}^{cyt} \tag{16}$$
$$s(v_j, Mt) = \min\left[s(v_{j-1}, ext), s(v_{j-1}, cyt), s(v_{j-1}, Mt)\right] + \mathcal{T}_{v_j}^{Mt} \tag{17}$$
$$s(v_j, nuc) = \min\left[s(v_{j-1}, cyt), s(v_{j-1}, Mt), s(v_{j-1}, nuc)\right] + \mathcal{T}_{v_j}^{nuc}. \tag{18}$$

The base case for these recurrence relations are:

$$s(v_1, ext) = \mathcal{T}_{v_1}^{ext} \tag{19}$$
$$s(v_1, cyt) = \infty \tag{20}$$
$$s(v_1, Mt) = \infty \tag{21}$$
$$s(v_1, nuc) = \infty. \tag{22}$$



S.Figure 2. Signaling model when incorporating the mitochondria compartment.

4 Evaluation of Multiple Pathways

We start by defining the precision and recall for individual pathways, and then extend this to the case of multiple pathways—the aggregate pathways. For each pathway, we compute its precision and recall (PR) values using its set of positives P, its set of negatives N, and its set of predicted interactions X. The interactions in X are ranked by the number/rank of the first path an interaction appears in (ascending order). Let X_i denote the set of unique predicted interactions up to the path i. The precision and recall for X_i are computed as:

$$Precision_i = \frac{|X_i \cap P|}{|X_i|} \quad and \quad Recall_i = \frac{|X_i \cap P|}{|P|}, \tag{23}$$

where $|S|$ means the number of elements in set S.

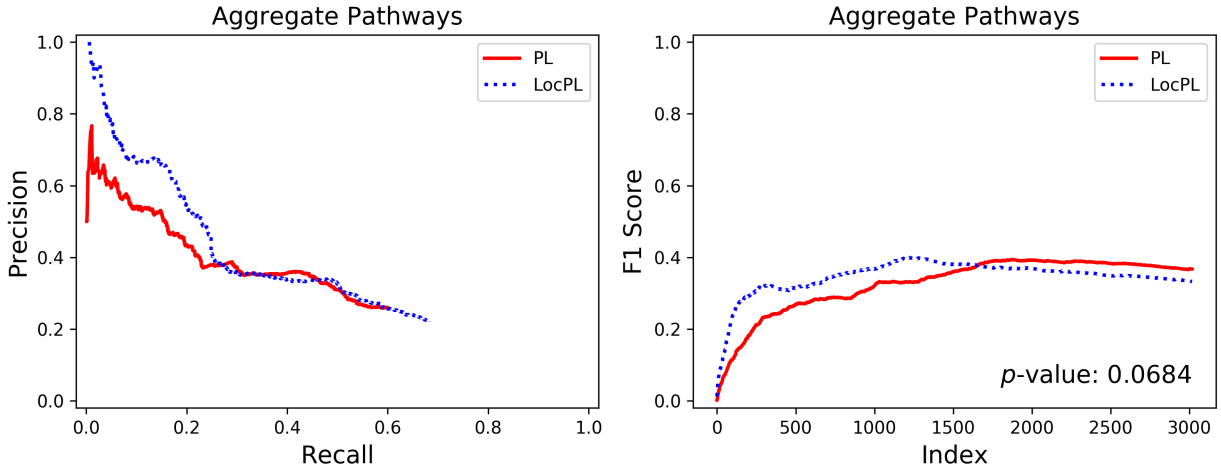
For the case of computing the PR values for m pathways $p_1, p_2, \ldots, p_m$, we have m distinct collections of positive interactions, negative interactions, and ranked predicted interactions, denoted as P^j, N^j, and X^j, respectively. We aggregate the sets of the ranked predictions as:

$$X = \bigcup_{j=1}^{m} \left[(e,k) \quad \text{for} \quad e,k \in X^j\right], \tag{24}$$

where e is a predicted interaction and k is its rank in X^j. We then rank the elements in X by the value k. Similarly, we aggregate the sets of positives and negatives as:

$$P = \bigcup_{j=1}^{m} \left[p \quad \text{for} \quad p \in P^j\right] \quad and \quad N = \bigcup_{j=1}^{m} \left[n \quad \text{for} \quad n \in N^j\right]. \tag{25}$$

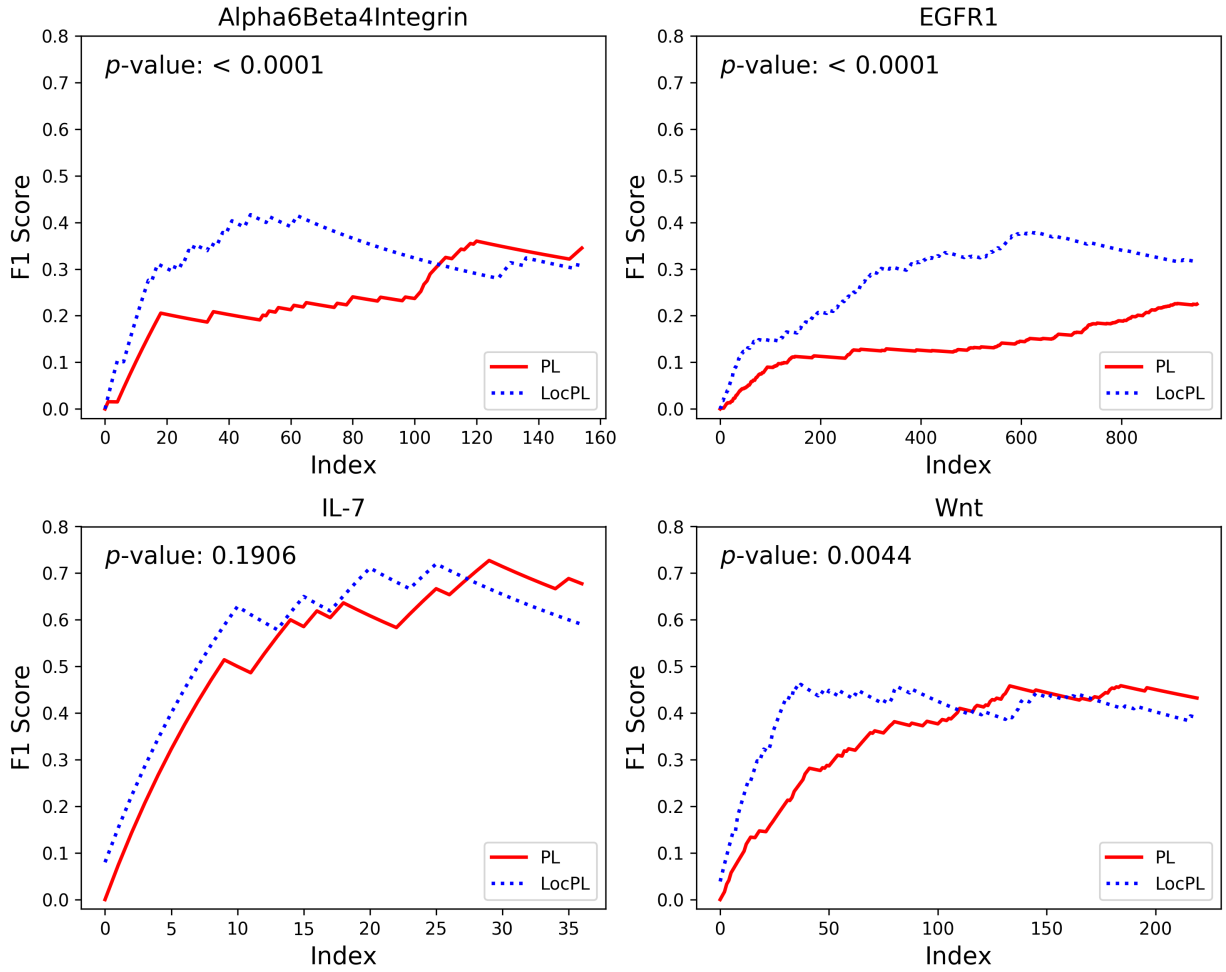
We used these three aggregate sets, X, P, and N to compute the precision and recall values using Equation (23). The size of the negatives set is 50 times the size of the positives set for the individual pathways as what was done in the original PathLinker study [Ritz et al., 2016]. But the negatives size drops to 25 times the positives size for the aggregate case due to insufficient number of interactions, after excluding the positives, to subsample negatives from. S. Figure 3 (Left) shows the aggregate PR curve for eight NetPath pathways for the updated PathLinker interactome $PLNet_2$.



S.Figure 3. $PLNet_2$: (**Left**) Aggregate PR curve, and (**Right**) F_1 score curve over eight signaling pathways from the NetPath database compared for PL and $LocPL$. The p-value is for the Mann-Whitney U test (alternative: $LocPL > PL$).

5 F_1 Score Plots for the $PLNet_2$ Interactome

S.Figure 4 plots the F_1 scores for the four NetPath pathways: $\alpha 6\beta 4$Integrin, EGFR1, IL-7, and Wnt. *LocPL* provides better performance than that of *PL* for all the four pathways, although performance of IL-7 is not statistically superior.



S.Figure 4. $PLNet_2$: F_1 scores for the individual NetPath pathways. These values are fed to the MWU test to check for difference significance. The p-value is for the MWU test (alternative: $LocPL > PL$).

6 Signaling Pathways

We used a set of eight NetPath pathways [Kandasamy et al., 2010] to evaluate the proposed method. S.Table 1 summaries the number of protein-protein interactions (PPIs), receptors, transcription regulators (TRs) for each pathway.

S.Table 1. Signaling pathways used in this study and numbers of their interactions, receptors, and transcription regulators (TRs) for both the $PLNet_2$ and HIPPIE [Alanis-Lobato et al., 2017] interactomes and for the condition of the complete interactome and the condition of the interactome intersected with the ComPPI database [Veres et al., 2015].

Pathways	Interactome Condition	$PLNet_2$			HIPPIE Interactome		
		PPIs	Receptors	TRs	PPIs	Receptors	TRs
$\alpha 6\beta 4$ Integrin	Complete	192	7	3	—	—	—
	∩ ComPPI	108	7	3	—	—	—
EGFR1	Complete	1308	6	33	36	2	14
	∩ ComPPI	633	6	33	35	2	14
IL3	Complete	145	2	9	21	1	5
	∩ ComPPI	135	2	9	21	1	5
IL6	Complete	124	4	14	18	1	6
	∩ ComPPI	107	4	14	16	1	6
IL7	Complete	42	2	3	7	2	2
	∩ ComPPI	38	2	3	6	2	2
RANKL	Complete	126	2	12	—	—	—
	∩ ComPPI	95	2	12	—	—	—
TGF_β_Receptor	Complete	782	5	77	—	—	—
	∩ ComPPI	640	5	77	—	—	—
Wnt	Complete	347	14	14	—	—	—
	∩ ComPPI	168	12	14	—	—	—

7 Protein Compartments

S.Table 2. Protein compartment information in $PLNet_2$.

Cellular Compartment	# Proteins with a single compartment	# Proteins with multiple compartments	Average localization score
ExtMem	5	12719	0.8253
Cytosol	654	10089	0.8664
Nucleus	1096	8376	0.8814
Mitochondria	66	2210	0.7681
Secretory	18	4414	0.7657

8 Number of Ties

S.Table 3 reports the number of path groups that share the same reconstruction score after applying *PL* on the original $PLNet_2$ interactome and the filtered interactome using the cellular localization information. Filtering the interactome by keeping only the spatially coherent interactions reduced the number of ties in most of the pathways (6 our of 8). However, ties still dominate the reconstructions, and this urges the need for a way for breaking these ties.

S.Table 3. The number of ties (path groups sharing the same reconstruction score) for *PL* applied on the original $PLNet_2$ interactome and the filtered interactome.

Pathway	Original Interactome	Filtered Interactome
$\alpha 6\beta 4$ Integrin	82	38
EGFR1	17	99
IL7	30	48
Wnt	43	10
IL3	25	11
IL6	102	53
RANKL	24	12
TGF_β_Receptor	153	76

9 The First 10 and 100 Paths

S.Table 4 reports the number of paths that have at least one positive interaction (partially positive) and the number of paths whose all interactions are positives (completely positive) in the first 10 and 100 paths that are reconstructed by *PL* and *LocPL* for PLNet_2. In general, *LocPL* pushes paths with higher percentages of positive interactions up in the k-shortest paths list, and pushes down paths with smaller percentages of positives. *LocPL* ties to or exceeds the number of paths with positive interactions of *PL* for all the pathways except for three instances out of 16 ones in total (bold numbers in S.Table 4).

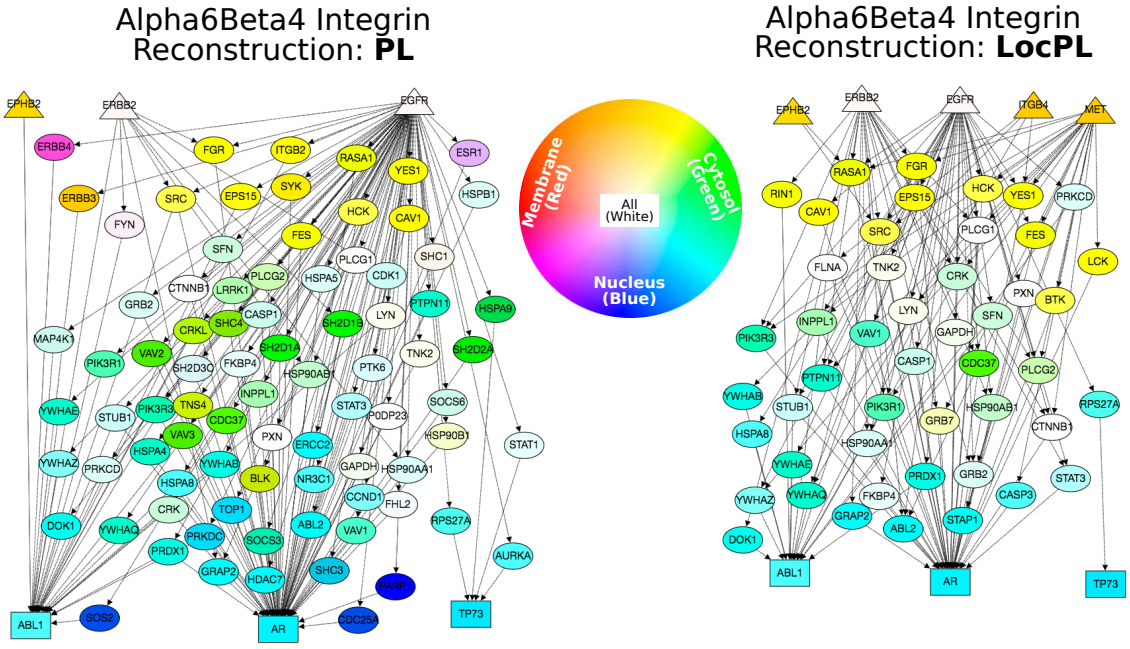
S.Table 4. PLNet_2: The number of partially positive and completely positive paths by *PL* and *LocPL* among the first 10 and 100 paths.

Pathway	Method	First 10 Paths		First 100 Paths	
		Partial	Complete	Partial	Complete
$\alpha 6\beta 4$ Integrin	*PL*	0	0	0	0
	LocPL	2	0	3	0
EGFR1	*PL*	0	0	11	3
	LocPL	3	1	31	3
IL7	*PL*	6	2	33	4
	LocPL	8	2	60	5
Wnt	*PL*	3	0	32	6
	LocPL	6	4	34	6
IL3	*PL*	6	0	**84**	**6**
	LocPL	9	1	**78**	**4**
IL6	*PL*	0	0	7	0
	LocPL	0	0	19	1
RANKL	*PL*	7	1	83	5
	LocPL	10	3	100	6
TGF_β_Receptor	*PL*	0	0	**36**	6
	LocPL	1	0	**25**	9

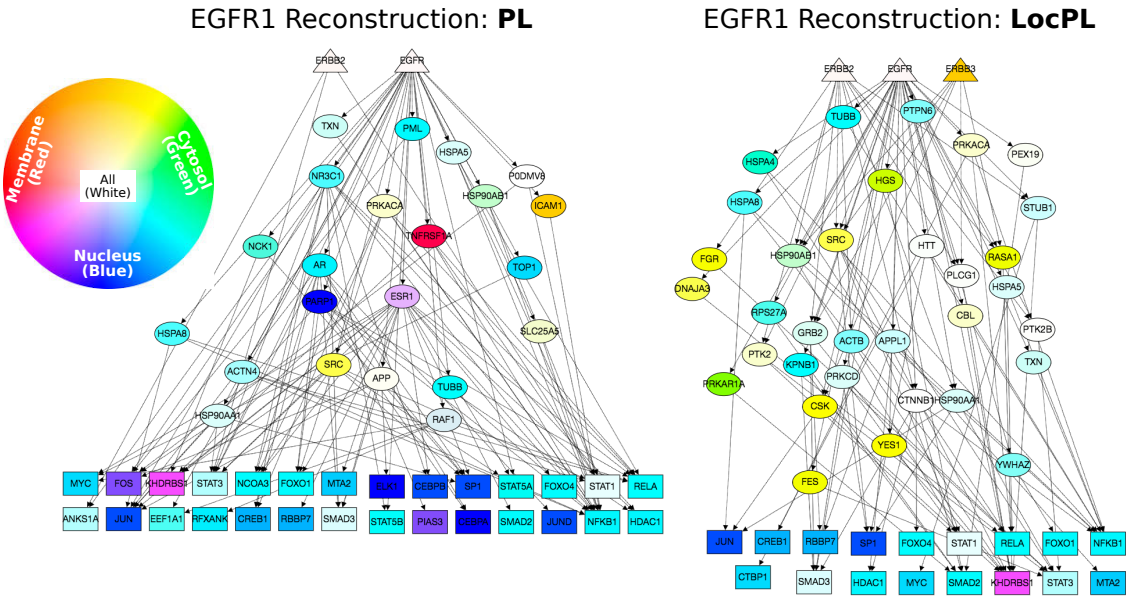
10 Pathway Reconstructions for $\alpha 6\beta 4$ Integrin, EGFR1, IL7, and Wnt

S.Table 5. Number of nodes and edges for the first 100 paths in each pathway reconstruction. Numbers in parentheses denote the percentage of nodes and edges that appear in only one method.

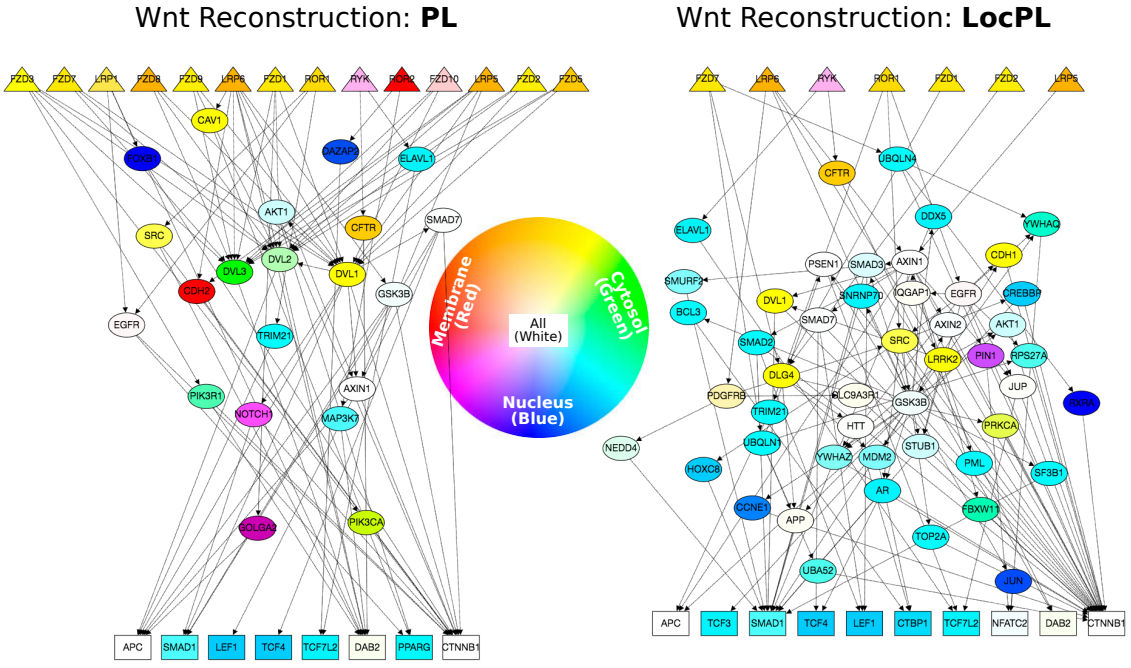
Pathway	Method	# Nodes (% Unique)	# Edges (% Unique)
$\alpha 6\beta 4$ Integrin	*PL*	89 (47%)	182 (52%)
	LocPL	56 (10%)	141 (38%)
EGFR1	*PL*	52 (50%)	122 (78%)
	LocPL	53 (52%)	112 (76%)
IL7	*PL*	40 (45%)	117 (65%)
	LocPL	41 (48%)	116 (65%)
Wnt	*PL*	43 (44%)	90 (72%)
	LocPL	65 (95%)	134 (81%)



S.Figure 5. Pathway reconstructions (first 100 paths) for $\alpha 6\beta 4$ Integrin using *PL* (left) compared to *LocPL* (right). Receptors are labeled as triangles; transcriptional regulators are rectangles, intermediary proteins are ellipses. Color denotes compartment localization; proteins may belong to multiple compartments (and will be lighter shades). Networks generated using GraphSpace [Bharadwaj et al., 2017].



S.Figure 6. Pathway reconstructions (first 100 paths) for EGFR1 using *PL* (left) compared to *LocPL* (right). Nodes are described as in S.Figure 5.



S.Figure 7. Pathway reconstructions (first 100 paths) for Wnt using *PL* (left) compared to *LocPL* (right). Nodes are described as in S.Figure 5.

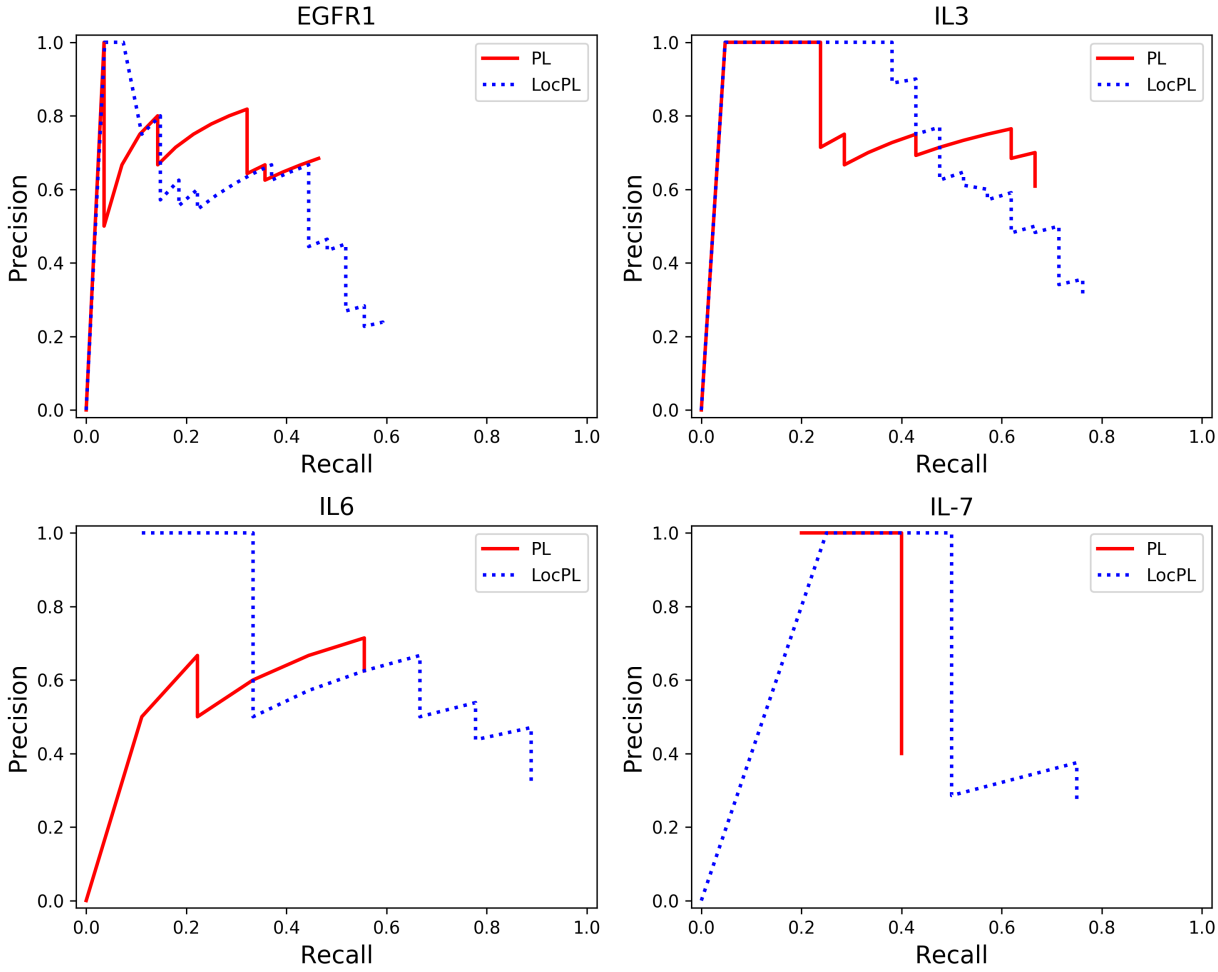
11 HIPPIE Interactome Analysis

11.1 HIPPIE Interactome

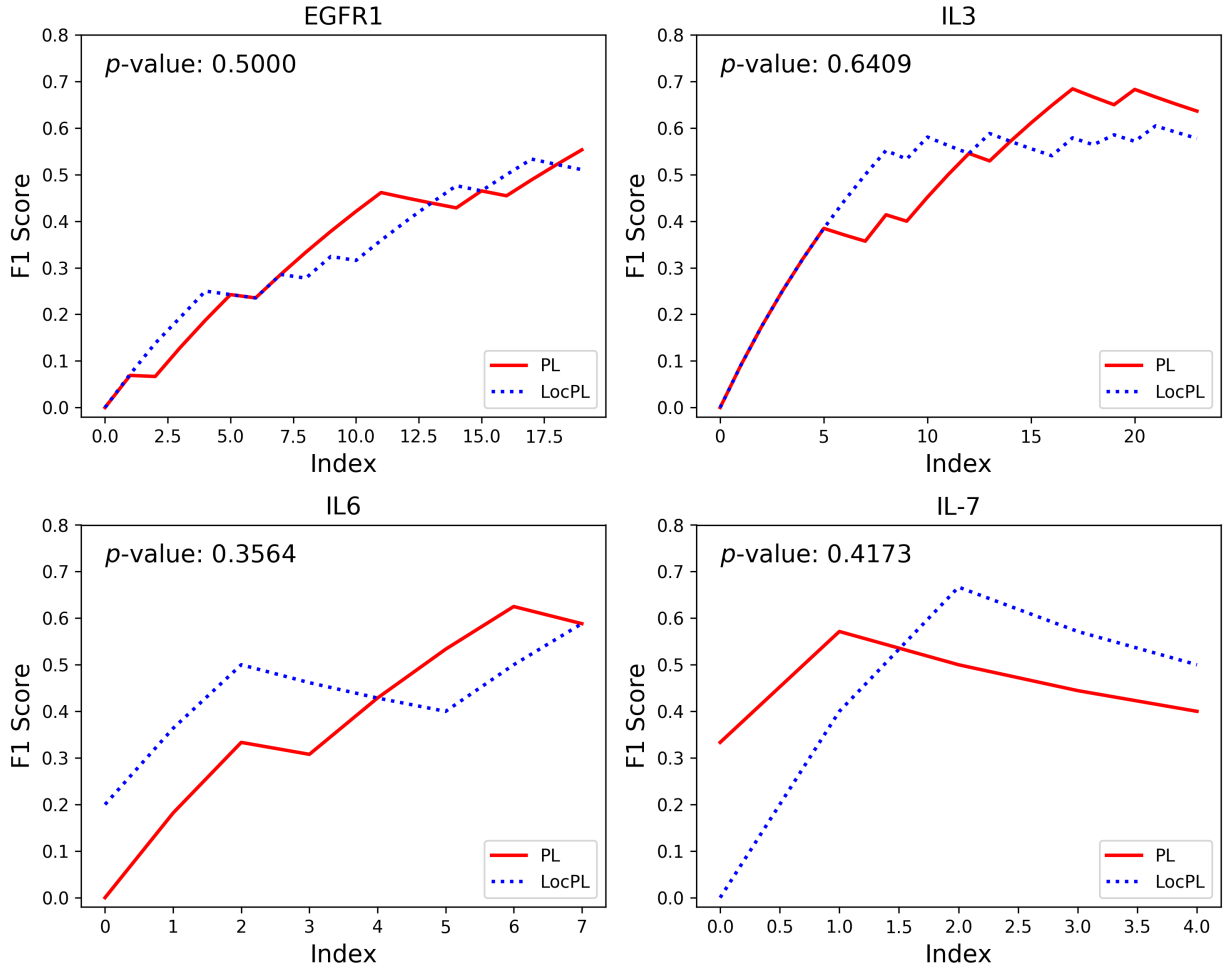
HIPPIE (Human Integrated Protein Protein Interaction rEference) is a repository of 16,707 proteins and 315,484 PPIs (version 2.1, July 18^{th}, 2017). Each interaction has a confidence score calculated as a weighted sum of the number of studies detecting the interaction, the number and quality of experimental techniques used in these studies to measure the interaction, and the number of non-human organisms in which the interaction was reproduced [Alanis-Lobato et al., 2017].

11.2 HIPPIE Results

We extended our experiments to four NetPath signaling pathways (EGFR1, IL-7, IL3, and IL6) [Kandasamy et al., 2010] that contain enough NetPath positive interactions within the HIPPIE interactome [Alanis-Lobato et al., 2017] to assess our method. S.Figure 8 shows the PR curves for these pathways and S.Figure 9 shows their F_1 score curves. Though the MWU test indicates that performance difference between the two methods is statistically insignificant, but the earlier paths of *LocPL* have more positive interactions than those of *PL*. Taking the IL6 pathway as an example, the first 500 paths by *PL* have only 4 paths with at least one positive interaction, leading to a recall of 0.22 representing only 22% of the positive edges that should be discovered. On the other hand, the first 500 paths for *LocPL* have 58 paths with at least one positive interaction, achieving a recall value of 0.67 (67% of all the positives). In general, final recall value is higher for *LocPL* meaning higher proportions of positives are discovered (S.Figure 8).

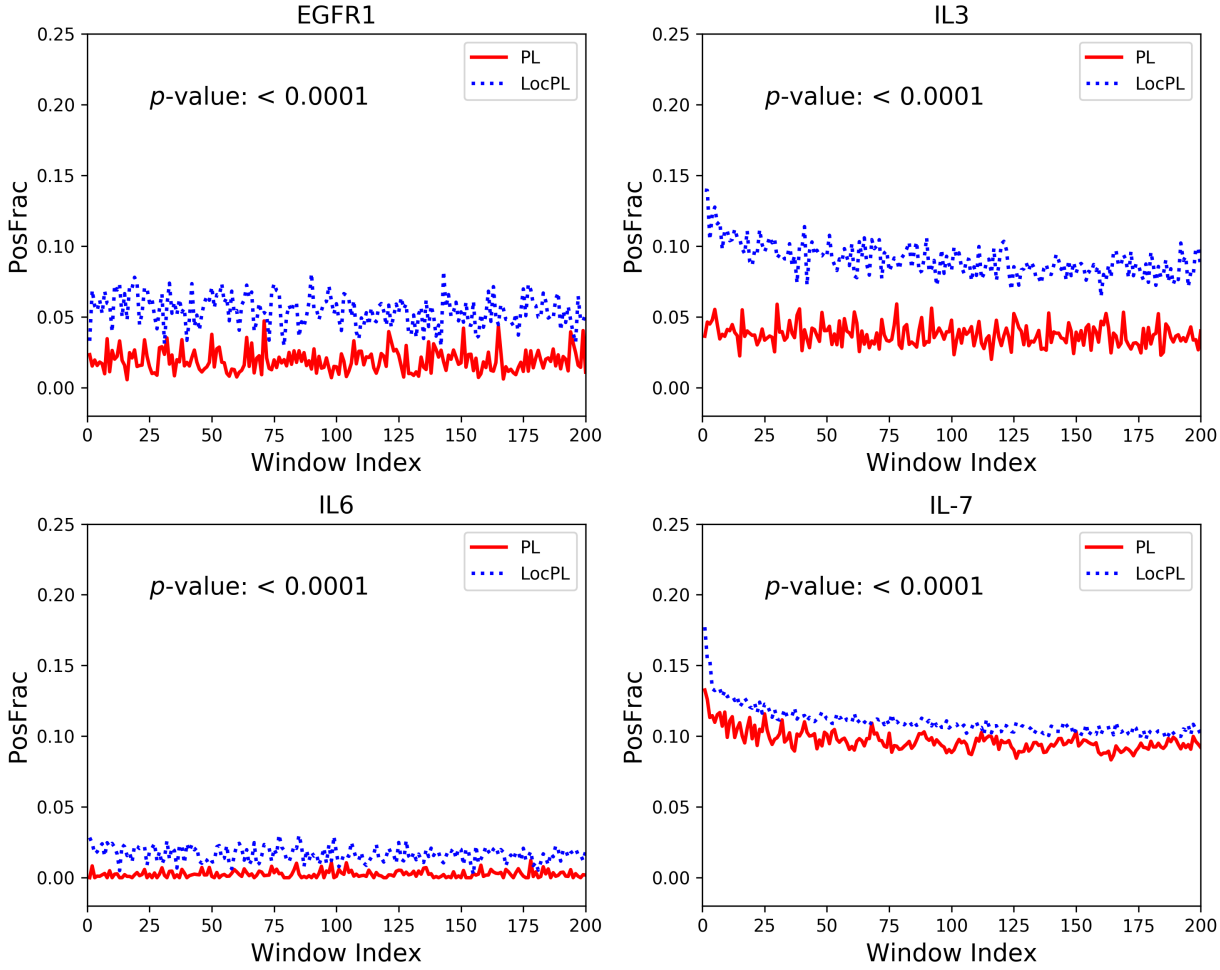


S.Figure 8. HIPPIE Interactome: Assessing the *global* performance through the PR curves on four signaling pathways from the NetPath database.



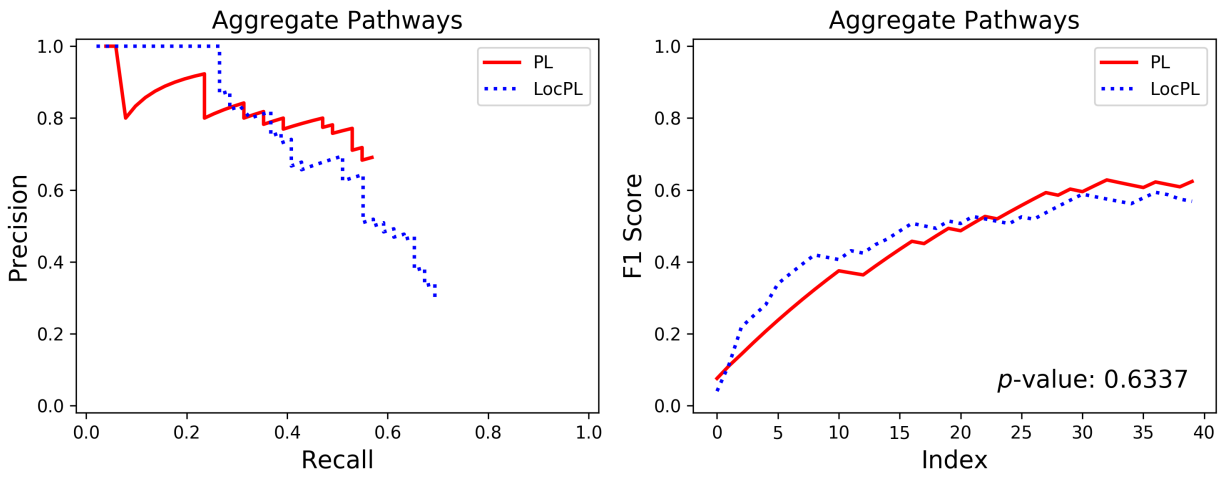
S.Figure 9. HIPPIE Interactome: The F_1 score curves for the individual NetPath pathways. Values of these curves are fed to the MWU test to check for difference significance. The p-value is for the MWU test (alternative: $LocPL > PL$).

Moreover, *PosFrac* is consistently higher for *LocPL* for all the paths as shown in S.Figure 10. It is worth noting that the PR and F_1 curves for HIPPIE are not as smooth and the *PosFrac* curves are not as high as those for PLNet_2 curves because some of the interactions of the NetPath dataset are not present in the HIPPIE interactome; these positives are ignored in our analysis.



S.Figure 10. HIPPIE Interactome: Assessing the *local* performance through the averaged positives percentage metric on two signaling pathways from the NetPath database. Each window contains 100 non-overlapping paths. The p-value is for the MWU test (alternative: $LocPL > PL$).

S.Figure 11 shows the aggregate PR curve for the HIPPIE interactome. The PR values were computed over four NetPath pathways: EGFR1, IL3, IL6, and IL7. The proposed method performs better than the original technique at the early recall values, but precision deteriorates with increasing recall. In addition, the proposed method collects higher percentage of the positives than the original method.



S.Figure 11. HIPPIE Interactome: Aggregate PR curve over four signaling pathways from the NetPaths database compared for the original PL technique and its enhanced version LocPL. The p-value is for the Mann-Whitney U test (alternative: LocPL > PL).

References

K. Kandasamy et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biology*, 11(1):R3, Jan 2010.

Anna Ritz et al. Pathways on demand: automated reconstruction of human signaling networks. *npj Systems Biology and Applications*, 2:16002, 2016.

G Alanis-Lobato et al. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Research*, 45:D408–D414, 2017.

D. Veres et al. ComPPI: a cellular compartment-specific database for protein–protein interaction network analysis. *Nucleic Acids Research*, 43(D1):D485–D493, 2015.

Aditya Bharadwaj et al. Graphspace: stimulating interdisciplinary collaborations in network biology. *Bioinformatics*, 33(19):3134–3136, 2017.