

Supplemental Figure 1: Process-specific performance of data integration. Each column in the heatmap represents one of four data integration schemes (classifiers trained and evaluated for individual processes, process-specific classifiers trained for specific processes and reintegrated, a single process-independent classifier trained globally and evaluated for individual processes, or an unregularized global classifier), and each scheme is evaluated using the entire genome or a set of ~6,000 holdout genes. Each cell in the heatmap represents the AUPRC over the precision (functional relationships) and recall (genes) in each of the 229 analyzed processes, sorted in order of decreasing number of genes. This performance evaluates predicted functional relationships using a gold standard derived from multiple curated functional catalogs (including the Gene Ontology, KEGG, and others); each process's performance is evaluated using the subset of this gold standard associated with the process (see Methods). While there is a clear correlation between process size and performance (due to the fact that more training data makes the machine learning task easier), certain processes are also predicted more accurately than others independent of size, such as *defense response to virus* or *protein polymerization*. The most specific classifiers overfit somewhat to the smallest terms, but this effect is mitigated in larger terms or by reintegration into the process-specific classifier. For medium and large processes, increased specificity provides increased performance, with individual process classifiers outperforming the process-specific classifier, which in turn outperforms the process-independent classifier. Finally, the unregularized classifier is outperformed in essentially every process.

