1 Histogram of Relative Solvent Accessibility of Putative SDPs

As part of the characterization of putative SDPs (residues in the $SDP_O$ set), we describe the relative solvent accessibility (RSA) distribution of putative SDPs and all alignment positions. Figure 1 shows a histogram of the distributions. The RSA values for each position were assigned to one of twenty equally sized bins.

Figure 1: Comparison of the relative solvent accessibility (RSA) of putative SDPs ($SDP_O$) to all positions. The RSA range is divided into 20 equal size bins, e.g., the first bin corresponds to RSA between 0 and 5%. Each point represents the fraction of the columns with RSA falling in the bin.

See Section 3.1.4 of the main text for a full discussion of these observations.
2 Results on $SDP_L$ and $SDP_A$ are similar to those on $SDP_O$.

Section 3.1.1 of the main text provides significant evidence that the definition of putative SDPs as the set $SDP_O$ is reasonable. This set includes all columns within 5Å of ligand relevant to the enzyme’s catalytic reaction that contain at least one specificity group with a conserved amino acid distribution that is significantly different from that of the other specificity groups in the column. In the main text, all methods are evaluated on this set of positives.

However, it is possible to perform evaluations of the SDP prediction methods using different sets of columns as the positives. Our other column filters (see Table 1 in the main text) provide several alternatives. The $SDP_L$ set provides a less strict definition of positive that includes any columns near ligands in which the specificity group amino acid distributions do not have significant overlap. Figures 2 and 3 show the box plots and PR-curves for this positive set.

Similarly, we can evaluate on the stricter set of positions provided by the $SDP_A$ set. All these columns are within 5Å of a relevant ligand and each specificity group’s amino acid distribution is conserved and different from the others. Figures 4 and 5 show the box plots and PR-curves for this positive set.

Our general conclusions from the main text—that $GroupSim$ is competitive with current methods and that $GroupSim+ConsWin$ is the best performing method—hold on both these alternative positive sets. Comparing the results across sets demonstrates that in general method performance
Figure 3: Precision-Recall curves for representative SDP prediction methods and two versions of GroupSim on SDP_L. The simple GroupSim is competitive with the other methods; SDPpred is the only method that significantly outperforms it. GroupSim+ConsWin outperforms all methods.

improves as the positive set becomes stricter; they do best on SDP_A and worst on SDP_L. One exception to this is the Sequence Harmony (SH) method. Its performance relative to other methods on SDP_A and itself on other positive sets is far worse on this strict set. This is in part because SH does not explicitly reward conservation within specificity groups and all columns in this positive set are required to have this property. However, SH’s handling of gaps, overlap between columns, and tie scores also harm its performance (data not shown).
Figure 4: Box plots for the SDP prediction methods on SDP_A. Each box shows the average over all alignments of the five-number summary (the minimum, lower quartile, median, upper quartile, and maximum) for a method. Lower averages indicate better performance. The simple GroupSim performs similarly to previous methods, and GroupSim+ConsWin provides the best results.

Figure 5: Precision-Recall curves for representative SDP prediction methods and two versions of GroupSim on SDP_A. The simple GroupSim is competitive with the other methods; SDPpred is the only method that significantly outperforms it. GroupSim+ConsWin significantly outperforms all methods.
3 Results on experimentally-derived data set are similar to those on EC-Pfam data set.

An experimentally-derived data set of 13 alignments was recently described in [1]. While we believe that the size of their data set limits its utility in method evaluation, we find that our overall conclusions are similar when this data set is used. Figures 6 and 7 provide box plots and PR curves for the representative SDP prediction methods on the experimentally-derived SDP data set. Table 3 provides more detail on the composition of this data set and the performance of the ConsWin heuristic.

Figure 6: Box plots for representative SDP prediction methods and two versions of GroupSim on the experimentally-derived data set of [1]. Each box shows the average over all alignments of the five-number summary (the minimum, lower quartile, median, upper quartile, and maximum) for a method. Lower averages indicate better performance. The simple GroupSim outperforms the previous methods, and GroupSim+ConsWin slightly improves on it.

In general, the difference between methods is less pronounced than in our results, but the general trends are similar. GroupSim provides competitive performance with all methods tested, and GroupSim+ConsWin provides improvement over GroupSim in 10 of the 13 alignments (Table 1). However, it is difficult to say that one method is superior to the others on this data set.

As noted in the main text, our analysis of SDP amino acid column patterns is in agreement with the column pattern distribution of the experimentally-determined SDPs.

The composition of the experimental data set (Table 1) is somewhat different from our large data set of enzymes. The experimental data set contains only five enzymes, and four of the alignments contain more than two specificity groups. These differences in composition explain some of the
Figure 7: Precision-Recall curves for representative SDP prediction methods and two versions of GroupSim on the experimentally-derived data set of [1]. The simple GroupSim is competitive with the existing method. GroupSim+ConsWin provides improvement for 10 of the 13 alignments.

variance in method performance relative to the EC-Pfam data set.

For example, the most notable contrast is the poor performance of MI relative to RE and SDPpred. MI’s change in performance is entirely the result of very poor performance on alignments with more than two specificity groups. However, we find that MI plus the shuffling normalization is competitive with SDPpred, further suggesting the importance of shuffling to the SDPpred method.

We were unable to evaluate Sequence Harmony on this data set because it cannot handle alignments with more than two groups.
Table 1: Summary of the experimentally-derived data set of [1]. The last three columns compare the PR-AUC performance of GroupSim and GroupSim+ConsWin. ConsWin provides improvement for 10 out of 13 families overall and 5 out of 8 non-enzyme families. See [1] for more information about the families and alignments.

<table>
<thead>
<tr>
<th>Family</th>
<th>Enzyme</th>
<th># Specificity Groups</th>
<th>GroupSim+ConsWin</th>
<th>GroupSim</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gprotein</td>
<td>0.737</td>
<td>0.698</td>
<td>0.038</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cd00333</td>
<td>0.533</td>
<td>0.517</td>
<td>0.016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cbm9</td>
<td>0.175</td>
<td>0.163</td>
<td>0.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cd00365</td>
<td>0.143</td>
<td>0.132</td>
<td>0.011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GST</td>
<td>0.737</td>
<td>0.727</td>
<td>0.010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ricin</td>
<td>0.219</td>
<td>0.211</td>
<td>0.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cd00423</td>
<td>0.475</td>
<td>0.470</td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cd00985</td>
<td>0.724</td>
<td>0.720</td>
<td>0.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cd00363</td>
<td>0.020</td>
<td>0.019</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cd00264</td>
<td>0.107</td>
<td>0.106</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cd00120</td>
<td>0.192</td>
<td>0.197</td>
<td>-0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNmyc</td>
<td>0.060</td>
<td>0.066</td>
<td>-0.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LacI</td>
<td>0.625</td>
<td>0.654</td>
<td>-0.029</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4 Conservation window heuristic improves all methods.

ConsWin, the conservation window heuristic we introduce, can be applied to any SDP scoring scheme that produces scores for each column of an alignment. ConsWin provides improvement over the raw version of all methods evaluated. Figures 8 and 9 illustrate the boost provided by ConsWin to RE, Xdet, and SH. Figure 1 in the main text shows the improvement ConsWin provides to GroupSim. Results are similar for all methods except SDPpred. The range of SDPpred scores for an alignment is often quite large and variable as a result of a few outlier column scores. This requires setting the window $\lambda$ parameter specially for each alignment.

Figure 8: Box plots comparing Xdet, SH, and RE with and without the ConsWin heuristic on the set of putative SDPs, $SDP_O$. Each box shows the average over all alignments of the five-number summary (the minimum, lower quartile, median, upper quartile, and maximum) for a method. Lower averages indicate better performance. The ConsWin heuristic improves all methods.

GroupSim+ConsWin was shown to outperform all current methods tested on the EC-Pfam data set in the main text. When the ConsWin is added to current methods, SDPpred+ConsWin, MI+ConsWin and RE+ConsWin become competitive with, but not better than, GroupSim+ConsWin.
Figure 9: Precision-Recall curves comparing Xdet, SH, and RE with and without the ConsWin heuristic on the set of putative SDPs, SDP₀. The ConsWin heuristic improves the performance of all methods tested.
5  *MI* and *RE* are sensitive to pseudocount used.

*Mutual information* (*MI*) and *relative entropy* (*RE*) both require the estimation of an amino acid probability distribution from a multiple sequence alignment column. A small, uniform pseudocount is commonly added to the observed counts in this distribution estimation step. We have found that the choice of pseudocount is important to the performance of *MI* and *RE*. Figures 10 and 11 show box plots for *MI* and *RE* across a range of pseudocount values.

![Box Plots - EC-Pfam Data Set Putative SDPs](image)

**Figure 10:** *Mutual information*’s performance on the set of putative SDPs (*SDP₀*) is sensitive to the pseudocount used. The number following “MI” gives the magnitude of the uniform pseudocount. A pseudocount of one gives the best performance.

There is a clear trend that favors larger pseudocounts all the way up to a pseudocount of one for this application. We also implemented more complex schemes for computing non-uniform pseudocounts [2], but found that the magnitude dominated the effect on performance. We use a pseudocount of one for both methods in all other analysis presented here and in the main text. This dependence on pseudocount magnitude is likely due to the sparsity of the distributions being compared in this context.
Figure 11: Relative entropy’s performance on the set of putative SDPs ($SDP_O$) is sensitive to the pseudocount used. The number following “RE” gives the magnitude of the uniform pseudocount. A pseudocount of one gives the best performance.
6 Use of Similarity Matrices with Xdet and GroupSim hurt performance.

Several methods for predicting SDPs from multiple sequence alignments (MSAs), allow the incorporation of the relationships between amino acids through the use of a similarity matrix. In our evaluation, we considered two such methods, Xdet [3] and our own GroupSim. Figures 12, 13, 14, and 15 provide box plots and PR-curves for Xdet and GroupSim with a range of similarity matrices.

Figure 12: Box plots for Xdet using the default McLachlan matrix [4] and using the identity matrix on the putative SDPs (SDP_O). Each box shows the average over all alignments of the five-number summary (the minimum, lower quartile, median, upper quartile, and maximum) for a method. Lower averages indicate better performance. Surprisingly Xdet with the identity matrix outperforms the use of a similarity matrix.

Both methods obtain the best performance when the identity matrix is used. The identity matrix is used with both methods in all other results presented here and in the main text. This surprising result is similar to an observation we made previously when evaluating methods for estimating amino acid conservation [6].

It seems that the matrix weights dominate the more relevant conservation and difference signals. We are currently investigating techniques for directly incorporating notions of amino acid similarity into SDP prediction methods.
Figure 13: Precision-Recall curves for $Xdet$ [3] using the default matrix, McLachlan, and using the identity matrix on the putative SDPs ($SDP_0$). $Xdet$ with the identity matrix is superior in PR analysis as well.

Figure 14: Box plots for $GroupSim$ using a range of BLOSUM [5] matrices and the identity matrix on the putative SDPs ($SDP_0$). Each box shows the average over all alignments of the five-number summary (the minimum, lower quartile, median, upper quartile, and maximum) for a method. Lower averages indicate better performance. Surprisingly $GroupSim$ with the identity matrix outperforms the direct use of a similarity matrices.
Figure 15: Precision-Recall curves for GroupSim using a range of BLOSUM [5] matrices and the identity matrix on the putative SDPs (SDP₀). GroupSim with the identity matrix is superior in this PR analysis as well.
Column Shuffling provides some improvement to methods.

SDPpred [7] incorporates a shuffling procedure into its mutual information (MI) based score. As we noted in the main text, this shuffling appears to provide a performance boost over MI in the Precision-Recall evaluation, but not in the box plot evaluation. Figures 16 and 17 give box plots and PR-curves for several methods with and without a column shuffling normalization. The results are similar for RE and GroupSim. The PR curves show consistent improvement, but the box plots show that the improvement is not consistent across the range of rank statistics. This provides evidence that shuffling may improve some predictions, but overall all predictions do not become more accurate. These results are based on 1000 iterations of the shuffling procedure of [7] without their linear transformation.

Figure 16: Box plots comparing GroupSim, RE, and MI with and without a column shuffling normalization on the set of putative SDPs, SDPΩ. Each box shows the average over all alignments of the five-number summary (the minimum, lower quartile, median, upper quartile, and maximum) for a method. Lower averages indicate better performance. Shuffling provides some improvement, but it is not consistent across statistics.
Figure 17: Precision-Recall curves comparing GroupSim, RE, and MI with and without a column shuffling normalization on the set of putative SDPs, SDP₀. Shuffling provides improvement for the methods in this evaluation.
References


