**Methods**

1) **Logistic Regression**

Each atom is embedded using a hand-tuned naïve weighting algorithm that decays by 4x for every further distance from the atom. We then model the probability of a bond forming between a source atom $a_i$ and a sink atom $a_j$ under reagent $k$ as the following:

$$P((a_i, a_j), R_k) = g(x_i^T A_k x_j)$$

We want to learn $A_k$ for each reagent $k$ in the dataset. This model corresponds to a chemical intuition that each reagent corresponds to a linear transformation which maps source atom embeddings onto the direction of sink atom embeddings.

2) **Gaussian Discriminant Analysis**

Using the same embeddings as in logistic regression, we assume that bonding pairs $(x_i, x_j)$ are generated according a multivariate Gaussian with a Wishart prior. Since chemically, reactivity typically depends on the difference in electrical characteristics between two atoms, rather than their absolute values, we model the differences themselves:

$$x_i - x_j \sim \mathcal{N}(\mu_1, \Sigma_1)$$

3) **Graphical Convolutional Neural Network (GNN)**

To embed atoms, we implement a graphical convolutional neural network as described in [2]. In layer $t$, we update the embedding of a node $x_i$ as follows:

$$x_i^{[t]} := g \left( W_t^{[t]} x_i^{[t-1]} + \sum_{j \in \mathcal{N}(i)} \sum_{l=1}^{n} W_l^{[t]} x_j^{[t-1]} \right)$$

We then apply the same linear transformation as in logistic regression for prediction.

**Results**

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Log. Reg.</th>
<th>GDA</th>
<th>GNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>48.5%</td>
<td>16.4%</td>
<td>65.3%</td>
</tr>
<tr>
<td>01</td>
<td>42.7%</td>
<td>15.7%</td>
<td>62.9%</td>
</tr>
<tr>
<td>02</td>
<td>60.0%</td>
<td>22.1%</td>
<td>59.7%</td>
</tr>
<tr>
<td>03</td>
<td>59.7%</td>
<td>18.1%</td>
<td>69.4%</td>
</tr>
<tr>
<td>04</td>
<td>83.2%</td>
<td>30.4%</td>
<td>78.2%</td>
</tr>
<tr>
<td>Total</td>
<td>52.2%</td>
<td>17.9%</td>
<td>65.2%</td>
</tr>
</tbody>
</table>

**Discussion**

- Generally speaking, GNN performs better than log. reg., which performs better than the GDA.
- Log. reg. occasionally performs better for reactions for which the naïve embedding is well-suited (typically strongly acidic or basic reagents, e.g. HCl). However, the naïve manual embedding also makes performance heterogeneous.
- GDA uniformly performed poorly, possibly due to a Gaussian being a poor model.

**Future Work**

- Transfer learning: using already-learned embeddings to speed up learning on future reagents
- More sophisticated GNN with bonds as nodes; also, more complex spectral methods for faster learning