The node2vec Algorithm
- Represents network nodes as lower dimensional vectors
- Generalized from word2vec model for chain graphs
- Randomly walks graph to approximate similarity between nodes
- Encodes similarity between nodes as cosine between vectors
  - d - number of dimensions
  - p - return to previous node
  - q - Walk away from (DFS) or around (BFS) source
- 4 binary operators tested:
  - L1
  - L2
  - Hadamard
  - Average
- Best performance: L2 operator

Evaluating Performance
Using the best set of parameters above, I attained the following confusion matrix and predictions

<table>
<thead>
<tr>
<th>CHEMBL ID</th>
<th>Entrez ID</th>
<th>Drug Name</th>
<th>Gene name</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHEMBL120790</td>
<td>57624</td>
<td>METHYPRYLON</td>
<td>VAP2</td>
<td>0.9999998745</td>
</tr>
<tr>
<td>CHEMBL860</td>
<td>720985</td>
<td>MEPHENTHYON</td>
<td>PM1A8A</td>
<td>0.9999999476</td>
</tr>
<tr>
<td>CHEMBL452</td>
<td>6234</td>
<td>CLONAZEPAM</td>
<td>FAL</td>
<td>0.9999999441</td>
</tr>
<tr>
<td>CHEMBL1522</td>
<td>6860</td>
<td>ESZOPICLONE</td>
<td>TGA9</td>
<td>0.9999999168</td>
</tr>
<tr>
<td>CHEMBL591</td>
<td>63990</td>
<td>DAP000163</td>
<td>BRIP1</td>
<td>0.999999155</td>
</tr>
<tr>
<td>CHEMBL591</td>
<td>14873</td>
<td>DAP000163</td>
<td>Selo1</td>
<td>0.999999032</td>
</tr>
<tr>
<td>CHEMBL452</td>
<td>26352</td>
<td>CLONAZEPAM</td>
<td>OR10H3</td>
<td>0.9999998376</td>
</tr>
<tr>
<td>CHEMBL1213252</td>
<td>748</td>
<td>CLORAZEPATE</td>
<td>JXK</td>
<td>0.9999998061</td>
</tr>
<tr>
<td>CHEMBL1213252</td>
<td>96785</td>
<td>CLORAZEPATE</td>
<td>CYP2C1</td>
<td>0.9999997669</td>
</tr>
<tr>
<td>CHEMBL285674</td>
<td>59662</td>
<td>ESTAZOLAM</td>
<td>ECHDC1</td>
<td>0.9999997263</td>
</tr>
</tbody>
</table>

Possible Improvements
- Different performance metrics
- K-Fold cross-validation
- Concatenating Additional Features

Conclusion
While this data does not achieve the performance detailed in the original node2vec paper, it indicates performance better than chance. The results of this project corroborate the potential of link prediction in biological networks using node embeddings generated by node2vec.

References
- https://snap.stanford.edu
- Yan Zhang (United States Food and Drug Administration), Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (p. 855–864).

Acknowledgements
- NIH R21-M011372-01: Big Data for Indiana State University (Grant)
- Kevin Coombes (Ohio State University Department of Biomedical Informatics), Yan Zhang (United States Food and Drug Administration)