Option 1(Previous):

| geneA function | geneB function | SL |
|---|---|---|
| 9510 | 9231 | 1 |
| 9510 | 9159 | 0 |
| 9475 | 9511 | 0 |

One row in the table indicates which function is associated with each function in a gene pair. A particular gene may be associated with multiple functions and therefore multiple rows may be required for one gene pair.

Problem: Not encoded and therefore model may incorrectly assume numerical meaning of the function values.

Major flaw: Using the function combinations as different entries in the table instead of putting all functions in each entry results in loss of/inaccurate data.
Example:
gene HK has the functions "metabolism" and "catabolism". Gene AB has the function "cell growth." It is known from biological experiments that HK and AB form a synthetic lethal pair. The program sets up the data in the following table:

| geneA function | geneB function | SL |
|---|---|---|
| Metabolism | Cell growth | 1 |
| Catabolism | Cell growth | 1 |

From this, the model makes the assumption that a gene associated with metabolism and a gene associated with cell growth are likely to form a synthetic lethal pair. However, let's assume that the only reason HK an AB are synthetic lethal is because HK is associated with catabolism. Then, the model may make a false prediction because it was told that a gene associated with metabolism and a gene associated with cell growth are likely to form a synthetic lethal pair.

How big of a problem is this? If along with HK and AB, there are many gene pairs with the combination of metabolism and cell growth that are not synthetic lethal, perhaps the earlier mistake will not be weighted as heavily. In this case there is less of an issue.

Option 2:

| FunctA | FunctB | FunctC | FunctD | SL |
|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 | 0 |
| 0 | 1 | 2 | 1 | 0 |
| 1 | 2 | 0 | 0 | 1 |

One row in the table corresponds to all of the functions associated with either gene in a gene pair.
In a single table entry:

- A 0 indicates that neither gene in the pair has that function
- A 1 indicates that one of the genes has that function
- A 2 indicates that both genes are associated with that function

Minor problem: Many columns

Major flaw: does not specify which gene in a pair has which function.

Option 3:

| FunctA | FunctB | FunctC | FunctA | FunctB | FunctC | SL |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 | 1 | 1 | 1 |

Each row in the table corresponds to all of the functions associated with each gene in a gene pair. As can be seen in the table, the column titles for functions are repeated because the information about the genes in the pair are being kept separate from each other.

Con: Very large dataframe