**How Sarah Nexus Predicts**

1. **Query compound**
   - Check if this is an exact match
     - Yes: **Display prediction**
     - No: **Fragment structure**
2. **Fragment structure**
   - **Find hypotheses**
     - Yes: For each hypothesis find kNN over example set k=10
       - Calculate confidence level for each individual hypothesis based on the kNN confidence $c_h = |S_{h,x}|$
     - No: Out of domain
   - No: Out of domain
3. **Are all fragments adequately represented in the training set?**
   - Yes: Apply reasoning
     - Calculate overall confidence level $c_x = |S_x|$ and **Display prediction**
     - No: **Find kNN over whole training set k=8**
   - No: **Calculate confidence level $c_x = |S_x|$**

**Symbols:**
- $S_x$ overall signal for the query compound $x$
- $S_{h,x}$ signal strength of each individual hypothesis given the query $x$

*Process correct for Sarah Nexus v2.0.1*
1. Calculate the weighted signal for each hypothesis based on the kNN
$$W_{i,x} = \sqrt{\text{similarity}(x, e_i)}$$
$$WS_x = \frac{\sum_{i=1}^{k} W_{i,x} \times S_i}{\sum_{i=1}^{k} W_{i,x}}$$
$$S_i = -1$$ if $$e_i$$ is a negative example
$$S_i = +1$$ if $$e_i$$ is a positive example
2. Further moderate the signal to account for the average distance of the examples from the query structure
$$S_x = WS_x \times \frac{\sum_{i=1}^{k} W_{i,x}}{k}$$
3. Calculate confidence level for each individual hypothesis
$$\text{confidence}_{h,x} = |S_{h,x}|$$

$$k$$ number of nearest neighbours
$$e_i$$ examples in the training set
$$WS_{h,x}$$ weighted signal of each individual hypothesis given the query $$x$$
$$S_{h,x}$$ signal strength of each individual hypothesis given the query $$x$$
$$W_{i,x}$$ weighting factor
$$S_x$$ overall signal for the query compound $$x$$
$$S_i$$ signal of the nearest $$i^{th}$$ neighbour
$$m$$ number of relevant hypotheses
$$WS_x$$ weighted signal from nearest neighbours given the query $$x$$

* Process correct for Sarah Nexus v2.0.1