We define an interaction between feature j and feature k to mean that changes in $g(\cdot)$ when we perturb both features are non-additive (for some instances) (equation (2) in the paper):

$$\left[g\left(\Delta\mathbf{x}_{j\wedge k}^{(i)}\right)-g\left(\mathbf{x}^{(i)}\right)\right] \not\approx \left[g\left(\Delta\mathbf{x}_{j}^{(i)}\right)-g\left(\mathbf{x}^{(i)}\right)\right]+\left[g\left(\Delta\mathbf{x}_{k}^{(i)}\right)-g\left(\mathbf{x}^{(i)}\right)\right]$$

For our simulated model under no noise, for non-interacting features j and k, $\forall i$ we have:

$$\left[g\left(\Delta\mathbf{x}_{j\wedge k}^{(i)}\right)-g\left(\mathbf{x}^{(i)}\right)\right] = \left[g\left(\Delta\mathbf{x}_{j}^{(i)}\right)-g\left(\mathbf{x}^{(i)}\right)\right]+\left[g\left(\Delta\mathbf{x}_{k}^{(i)}\right)-g\left(\mathbf{x}^{(i)}\right)\right] \tag{1}$$

$$\implies g\left(\Delta\mathbf{x}_{j\wedge k}^{(i)}\right) = g\left(\Delta\mathbf{x}_{j}^{(i)}\right)+g\left(\Delta\mathbf{x}_{k}^{(i)}\right)-g\left(\mathbf{x}^{(i)}\right)=c^{(i)}(\text{say}) \tag{2}$$

$$\implies L\left[y^{(i)},g\left(\Delta\mathbf{x}_{j\wedge k}^{(i)}\right)\right] = L\left[y^{(i)},g\left(\Delta\mathbf{x}_{j}^{(i)}\right)+g\left(\Delta\mathbf{x}_{k}^{(i)}\right)-g\left(\mathbf{x}^{(i)}\right)\right] \tag{3}$$

In the implementation, when testing outputs, we compare:

$$g\left(\Delta\mathbf{x}_{j\wedge k}^{(i)}\right)$$

and

$$g\left(\Delta\mathbf{x}_{j}^{(i)}\right)+g\left(\Delta\mathbf{x}_{k}^{(i)}\right)-g\left(\mathbf{x}^{(i)}\right)$$

Without noise, these are identical and equal to $c^{(i)}$. When noise is added, these become:

$$c^{(i)}+\gamma_1^{(i)}$$

and

$$c^{(i)}+\gamma_2^{(i)}+\gamma_3^{(i)}-\gamma_4^{(i)}$$

where $\gamma_j^{(i)} \sim N(0,\sigma^2)$. Thus these terms are distributed according to $N(c^{(i)},\sigma^2)$ and $N(c^{(i)},3\sigma^2)$ respectively. When we compare the two quantities in a paired statistical test, we take their difference, i.e.:

$$\gamma_1^{(i)}-\left[\gamma_2^{(i)}+\gamma_3^{(i)}-\gamma_4^{(i)}\right]$$

Since they both have equal mean, the statistical test returns an insignificant p-value.

When comparing losses, we're using RMSE, so for the i^{th}instance the loss is simply the absolute value of difference between the target $y^{(i)}$ and the model output. Hence we compare (without noise added):

$$\left|y^{(i)}-g\left(\Delta\mathbf{x}_{j\wedge k}^{(i)}\right)\right|$$

and

$$\left| y^{(i)} - \left[g\left(\Delta \mathbf{x}_j^{(i)}\right) + g\left(\Delta \mathbf{x}_k^{(i)}\right) - g\left(\mathbf{x}^{(i)}\right) \right] \right|$$

Let $y^{(i)} - c^{(i)} = d^{(i)} \quad \forall i$. Adding noise, we compare:

$$|d^{(i)} - \gamma_1^{(i)}|$$

and

$$|d^{(i)} - \gamma_2^{(i)} - \gamma_3^{(i)} + \gamma_4^{(i)}|$$

The terms inside the absolute value are distributed according to $N(d^{(i)}, \sigma^2)$ and $N(d^{(i)}, 3\sigma^2)$ respectively. But the absolute value results in a folded normal distribution, which results in a different mean for each term (since the mean of a folded normal incorporates the variance of its corresponding normal). Hence the statistical test always results in a significant p-value, even for non-interacting features.

One way around this to rearrange the terms in equation (2) before computing losses, so that each term being compared has equal variance. Namely:

$$\begin{aligned} g\left(\Delta \mathbf{x}_{j\wedge k}^{(i)}\right) &= g\left(\Delta \mathbf{x}_j^{(i)}\right) + g\left(\Delta \mathbf{x}_k^{(i)}\right) - g\left(\mathbf{x}^{(i)}\right) \\ \implies g\left(\Delta \mathbf{x}_{j\wedge k}^{(i)}\right) + g\left(\mathbf{x}^{(i)}\right) &= g\left(\Delta \mathbf{x}_j^{(i)}\right) + g\left(\Delta \mathbf{x}_k^{(i)}\right) \\ \implies L\left[y^{(i)}, g\left(\Delta \mathbf{x}_{j\wedge k}^{(i)}\right) + g\left(\mathbf{x}^{(i)}\right)\right] &= L\left[y^{(i)}, g\left(\Delta \mathbf{x}_j^{(i)}\right) + g\left(\Delta \mathbf{x}_k^{(i)}\right)\right] \end{aligned}$$