Cross-Evaluation with ChemFH

To provide an external reference point for the nuisance classification results reported in the main text, we evaluated ChemFH, a publicly available web-based tool for frequent false-positive detection, on our held-out scaffold-disjoint test set (∼38,700 molecules).

ChemFH predicts seven interference-related classes. For this evaluation, we selected the four classes with the closest correspondence to our nuisance taxonomy: *Colloidal aggregators* (mapped to Aggregator), *FLuc inhibitors* (mapped to Luciferase Inhibitor), *Reactive compounds* (mapped to Reactive), and *Promiscuous compounds* (mapped to Promiscuous).

Table 1: Comparison of ChemFH (CFH) and CAGE-Fusion (CF) on the scaffold-disjoint test set.

Label	**ROC-AUC**		**PR-AUC**		**MCC**	
	CFH	CF	CFH	CF	CFH	CF
Aggregator	0.60	**0.93**	0.21	**0.77**	0.13	**0.63**
Luciferase Inhibitor	0.90	**0.97**	0.34	**0.74**	0.44	**0.66**
Reactive	0.89	**0.98**	0.24	**0.86**	0.40	**0.80**
Promiscuous	0.81	**0.90**	0.27	**0.54**	0.34	**0.51**

CAGE-Fusion achieves higher scores across all labels and metrics. However, several methodological factors should be considered when interpreting these results:

1. **Evaluation protocol differences.** Our test set is scaffold-disjoint with respect to CAGE-Fusion training (no Bemis–Murcko scaffold overlap), which is a stricter generalization setting for CAGE-Fusion. ChemFH is a fixed, externally trained model validated under random splits on its own dataset; its exposure to scaffolds present in our test set is unknown.

2. **Label definition differences.** While the four selected ChemFH classes are conceptually similar to our nuisance categories, exact inclusion and exclusion criteria differ across studies. For example, the boundary between "reactive" and "non-reactive" compounds depends on the source assays, activity thresholds, and curation decisions used by each group. These definitional differences can shift the effective difficulty of the classification task.

3. **Threshold calibration.** ChemFH uses Youden's index for uncertainty categorization, but the specific decision thresholds applied for reject/pass classification are not documented in sufficient detail to reproduce externally. For the MCC calculation reported here, we optimized per-label thresholds on the test set for ChemFH.

4. **Reproducibility note.** All ChemFH predictions reported here were obtained via the ChemFH API endpoint[1] in batches of 500. We observed discrepancies between predictions returned by the ChemFH API and those obtained through the ChemFH web interface (single compound evaluation mode). This inconsistency may reflect version differences between the web interface and the exposed API; all results reported here are based exclusively on the ChemFH API output.

Additionally, our dataset draws on sources also used in ChemFH's development, so ChemFH may have seen related compounds or scaffolds during its own training, whereas CAGE-Fusion was trained with scaffold-disjoint separation. This comparison should therefore be interpreted as a cross-study reference rather than a controlled architectural comparison. For a controlled comparison between a standalone DMPNN and CAGE-Fusion's co-attention architecture, we refer to the ablation study, Table 3, which uses identical data, splits, and evaluation protocol.

[1] `https://chemfh.scbdd.com/api/fh`