# Ligand hallucinations in cofolding methods

Dehaen, Wim1,2

1 Department of Organic Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 16 628 Prague 6, Czech Republic;

2 Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, 16 628 Prague 6, Czech Republic;

Cofolding methods such as the Nobel-prize winning AlphaFold3 and its closely related methods such as Boltz1, Chai and Protenix have revolutionized protein-ligand structure prediction. Unlike traditional pose prediction methods, these deep learning-based methods predict the structure of the protein and the ligand at the same time. On common benchmarks, these data-driven methods outperform or match molecular docking, the most commonly used traditional pose prediction method.

Nonetheless, some unusual issues plague these new methods. The authors of AlphaFold3 already point out that asymmetric carbons are frequently inverted in the output of their method, but otherwise they manage to pass standard validity checks. Surprisingly, we have found several examples where the input molecules (given as SMILES) did not match the output molecules (given as CIF). Some common problem motifs are: cyclohexanes (aromatized to benzenes), tetrahydrofurans (aromatized to furan), allenes, alkynes and nitriles (all 3 with non-linear structures). These can be considered an example of the “hallucination” phenomenon which is well known in other generative AI methods such as LLMs. These ligand hallucinations are also interesting because they subvert pose validation methods, being able to both pass PoseBusters as well as pass RMSD based checks compared to a reference.

We propose a method to systematically identify instances of cofolding ligand hallucinations by generating ligand structures, inferring their connectivity and comparing it to the input connectivity. We use this to compile a list of problematic functionalities. This will pave the way for correcting these issues (e.g. through data augmentation), which is of high priority to enhance the efficacy of the applications of cofolding methods on novel ligands.