

How can pharmacologists know which drug structures are correct ?

C. Southan, E. Faccenda, S. D. Harding, J. L. Sharman, A. J. Pawson, J. A. Davies. IUPHAR/BPS Guide to PHARMACOLOGY, Edinburgh, UNITED KINGDOM.

Introduction: Human medicines represent the crown jewels of pharmacology. Paradoxically however; there is neither any “Gold Standard” set of approved chemical structures, nor agreement on totals. A 2009 comparison of three sets of approved drugs recorded only 807 exact structures-in-common from the expected ~1200 [1]. The IUPHAR/BPS Guide to Pharmacology (GtoPdb) team have grappled with this discordance issue for curating approved drugs and all ~ 6000 small-molecule ligands we deposit into PubChem [2]. Users have the same challenge of deciding correct structures when procuring compounds for experiments or navigating links between journals and databases. This work examines the problems and partial solutions.

Methods: We used PubChem to explore relationships for selected drugs already curated into GtoPdb. Tools included the “same connectivity” operator that records distinct compound record (CID) representations of the same carbon backbone. We divided structural multiplexing causes between stereo differences, mixtures and isotopic derivatives. We then performed Venn-type comparisons between DrugBank, ChEMBL, and the Therapeutic Target Database. Additional metrics were generated to dissect contributing factors to discordance between these three and other sources.

Results: Atorvastatin has 51 different single representations in PubChem and 248 mixtures with paclitaxel (taxol) having 142 and 330, respectively. Comparing three manually curated drug sets mentioned above inside PubChem showed the consensus was only 25% of the sum. Results comparing other drug sources also showed discordance. Causes for CID multiplexing discordance will be presented. Using PubChem tools we assessed a curation strategy of selecting CIDs with structures supported by the majority of submitting sources. While not infallible, comparison with INN documentation indicated its effectiveness. We will also show how tagging our own approved drug records facilitates easy retrieval of just these entries from PubChem but that vendor drug names sometimes mapped to different structures.

Conclusion: As PubChem pushes towards 100 million, we have examined problems of choosing correct structures of pharmacologically active compounds. The constitutive challenges of chemical representation and high levels of discordances we recorded indicate that definitive drug lists (even our own) will remain elusive until pharmaceutical companies submit their own records directly to open databases. In the meantime, we have optimised our GtoPdb curation for the submission of our own 1088 approved CID entries as both a partial solution and trusted reference set for the pharmacology community.

References: [1] Southan *et al.* (2009) *J Cheminform.* 1:1-10. [2] Southan *et al.* (2016). *Nucl. Acids Res.* 44 (Database Issue): D1054-68.