Accepted Manuscript

Title: Redundancy in two major compound databases

Authors: Dimitar Yonchev, Dilyana Dimova, Dagmar Stumpfe, Martin Vogt, J*ü*rgen Bajorath

PII: S1359-6446(18)30027-8
DOI: https://doi.org/10.1016/j.drudis.2018.03.005
Reference: DRUDIS 2204

To appear in:

Please cite this article as: Yonchev, Dimitar, Dimova, Dilyana, Stumpfe, Dagmar, Vogt, Martin, Bajorath, J*ü*rgen, Redundancy in two major compound databases.Drug Discovery Today https://doi.org/10.1016/j.drudis.2018.03.005

Redundancy in two major compound databases

Dimitar Yonchev, Dilyana Dimova, Dagmar Stumpfe, Martin Vogt and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53113 Bonn, Germany

*_Corresponding author_: Bajorath, J. (bajorath@bit.uni-bonn.de).

Highlights:

- Public compound databases are of particular relevance for drug discovery research
- Crosstalk between databases is desirable but also leads to redundancy
- ChEMBL and PubChem contain large numbers of shared compounds and activity records
- More than half of the 14.7 million activities in ChEMBL originate from PubChem
- More than 99% of the 1.25 million assays in PubChem are imported from ChEMBL
- Data exchange requires a high level of transparency