Detecting irregularities in scientific datasets with a domain-agnostic Bayesian framework of numerical distributions

Uday Abu-Shehab[1,2★], Matthias Welsch[1,2,3★], Johannes Kirchmair[1,3✤]

[1]*Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, Josef-Holaubek-Platz 2, 1090, Vienna, Austria.*

[2]*Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, Josef-Holaubek-Platz 2, 1090, Vienna, Austria.*

[3]*Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, Josef-Holaubek-Platz 2, 1090, Vienna, Austria.*

Data is the cornerstone of modern research, but its true value depends on its integrity. Benford's law is a promising yet underexplored diagnostic for flagging irregularities across large collections of datasets.[1] Benford's law posits that smaller digits occur more frequently as leading digits in naturally occurring datasets.[1] Therefore, deviations from Benford's law signal potential data irregularities. However, quantifying adherence to Benford's law and estimating the associated uncertainty remain challenging.[2] We introduce Simulation-Based Benfordness Estimation (SBBE), a Bayesian framework that estimates the proportion of a dataset that conforms to Benford's law, along with the associated uncertainty, yielding interpretable, comparable "Benfordness" estimates across datasets of varying sizes.[2]

When applied to three major bioactivity databases (Figure 1), SBBE Benfordness scores align with known curation standards: PDBbind (96%), ChEMBL (76%), and PubChem (42%).[2] Moreover, SBBE assay-level analysis of ChEMBL enables tracing six concrete mechanisms underlying data irregularities, including quantization artifacts and database import errors.[2] This work demonstrates that SBBE can serve as a domain-agnostic proxy for data quality and a practical tool for prioritizing expert review of data.[2]

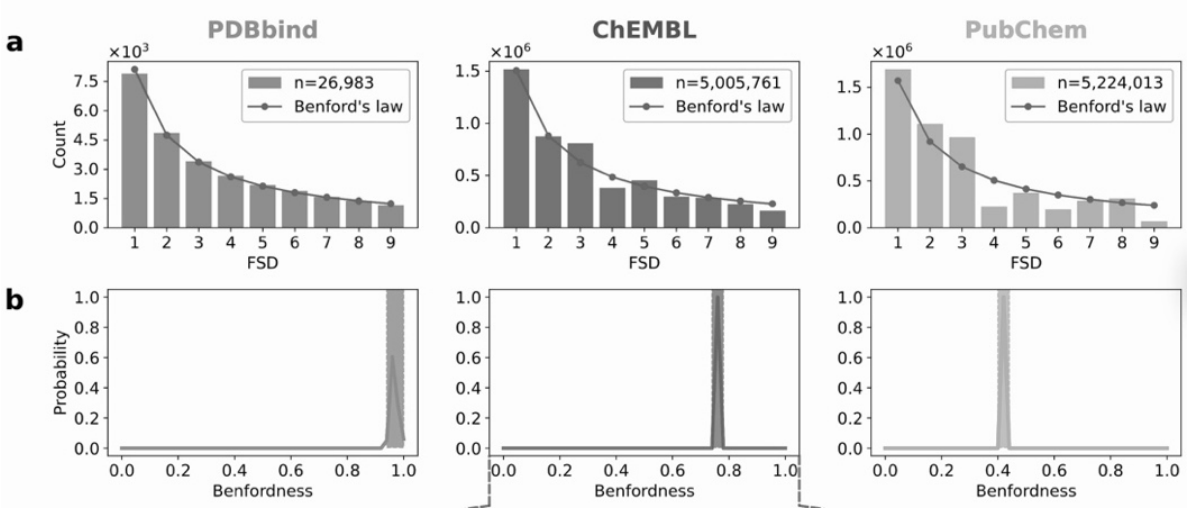

Figure 1. Analysis of the Benfordness of data from PDBbind, ChEMBL, and PubChem. (a) Distributions of the first significant digits across the three databases, demonstrating different alignments with Benford's law. (b) Estimated Benfordness by SBBE for the individual databases. Datasets compiled with higher curation standards align more closely with Benford's law.

Bibliography:

[1] F. Benford. (1938) Proc. Am. Phil. Soc. 78, 551–572
[2] U. Abu-Shehab, M Welsch, J Kirchmair. (2026) ChemRxiv DOI: 10.26434/chemrxiv.15000459/v2