Supporting Information

Data Set Assembly: The starting data for this analysis were obtained from the tabulation of NIH NME approvals between 2010-2016 in the supplementary materials reported by Cleary (Galkina Cleary et al., 2018). This starting data set includes the common names for 210 NMEs and the search terms that were used in the Cleary study to find targets in the published literature.

Using the Cleary starting data we have identified the corresponding LMW NMEs in the PDB dictionary of chemical components (Westbrook et al., 2015) and protein NMEs either by comparison with a UniProt (The UniProt Consortium, 2017) reference protein sequence or by protein name. For the 151 targets in the Cleary data set we have identified accession codes for a reference protein sequence with human taxonomy corresponding to a target with structural coverage in the PDB, or lacking a reference sequence we have identified specific examples of the target in PDB using search services of the RCSB/PDB(Burley et al., 2018).

For the corresponding LMW NMEs we have tabulated either the PDB chemical component identifiers for each molecule. For corresponding protein NMEs, the PDB accession code containing the NME and the reference protein sequence, if available were tabulated. Similarly, for corresponding targets we have tabulated either the accession code for the relevant reference sequence or lacking a reference sequence the PDB accession code containing an instance of the target.

All of the correspondence information described here is included in the supplemental comma-separated values (CSV) format document (pdb_drug_approvals_si.csv). This document includes columns containing the NME name, target search terms, NME identifiers, and target identifiers. NME identifiers are either PDB accession codes (4-characters) or PDB chemical component identifiers (3-characters). Target identifiers are either UniProt accession codes or PDB accession codes. PDB 4-character accession codes may include additional information to specific polymer entity in the structure. This includes an underscore separated entity identifier and a colon separated sequence identity (e.g. 50). The latter designates the PDB sequence is beyond a 95% identity threshold from the target reference sequence.

Data Analysis: Using the correspondence information described in the previous section as seed data, we evaluated the coverage PDB structures within a sequence identity threshold of 95% (Steinegger and Soding, 2018). This broader coverage also includes all PDB structures containing an example of the NME matching a PDB chemical component definition. Mapping UniProt reference sequences to sequences in PDB structure data

has been performed using the mapping information produced during PDB Biocuration (Young et al., 2018) and updated weekly by the SIFTS service(Velankar et al., 2013).

For our analysis of the PDB primary citation impact analysis we have incorporated the 'times cited' information provided by the Thomson Reuters Links Article Match Retrieval Service (Reuters, 2018). The time interval between PDB structure deposition and FDA Approval incorporates approval dates for the NMEs obtained from the FDA(FDA, 2018).

References:

Burley, S.K., Berman, H.M., Christie, C., Duarte, J., Feng, Z., Westbrook, J., Young, J., and Zardecki, C. (2018). RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. Protein Sci *27*, 316–330.

FDA (2018). Drugs@FDA Data Files.

Galkina Cleary, E., Beierlein, J.M., Khanuja, N.S., McNamee, L.M., and Ledley, F.D. (2018). Contribution of NIH funding to new drug approvals 2010-2016. Proc Natl Acad Sci U S A *115*, 2329-2334.

Reuters, T. (2018). Links Article Match Retrieval Service (Links AMR).

Steinegger, M., and Soding, J. (2018). Clustering huge protein sequence sets in linear time. Nature communications *9*, 2542.

The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res *45*, D158-D169.

Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.J., and Kleywegt, G.J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res *41*, D483-489.

Westbrook, J.D., Shao, C., Feng, Z., Zhuravleva, M., Velankar, S., and Young, J. (2015). The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. Bioinformatics *31*, 1274-1278.

Young, J.Y., Westbrook, J.D., Feng, Z., Peisach, E., Persikova, I., Sala, R., Sen, S., Berrisford, J.M., Swaminathan, G.J., Oldfield, T.J.*, et al.* (2018). Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data. Database (Oxford) *2018*.