Supplementary File

A. Proteins and Peptides Complete Data Github URL

The field of therapeutics encompasses a wide range of molecules that are used for medicinal purposes. The U.S. Food and Drug Administration (FDA) plays a crucial role in ensuring the safety and efficacy of these therapeutic molecules. To facilitate their regulatory processes, the FDA classifies therapeutics into distinct categories. One such classification is based on the molecule's composition and structure, explicitly dividing them into two main classes: proteins and peptides. In this categorization, molecules with an amino acid sequence length below 40 are considered peptides rather than therapeutic proteins. This classification helps distinguish and analyze these molecules based on their unique characteristics and potential therapeutic applications. For a comprehensive analysis, data for proteins and peptides have been compiled into separate data sets, allowing for a detailed examination of their properties and attributes.

A.1 Proteins Data

Protein therapeutics represent a significant portion of the FDA-regulated molecules used in medical treatments. The data set dedicated to protein therapeutics consists of 188 unique molecules. Each entry in the data set provides valuable information, including the name, accession number, BLA/NDA status (Biologics License Application/New Drug Application), brand name, and the type of molecule.
Additionally, essential characteristics such as the amino acid count, molecular weight, and amino acid sequence are included. To further aid in understanding these proteins, scores from advanced structural prediction algorithms, AlphaFold2 and ESMFold are provided. These scores, such as AlphaFold2 pLDDT Score and AlphaFold pTM Score, assess the structural reliability and protein domain predictions, respectively. Other properties such as hydrophobicity (GRAVY), isoelectric point, extinction coefficients, and instability index contribute to a comprehensive analysis of protein therapeutics.

Link: Please refer to the proteins data for more information.

A.2 Peptides Data

Peptide therapeutics constitute a distinct subset within the realm of therapeutic molecules, characterized by their relatively shorter amino acid sequences. The data set dedicated to peptides comprises 16 molecules that fall under this category. These peptides, with their unique characteristics and potential therapeutic applications, are of great interest to researchers and clinicians. Similar to the protein data set, each entry provides relevant details, including the name, accession number, BLA/NDA status, brand name, and molecule type. Additionally,

important information such as amino acid count, molecular weight, and amino acid sequence is available for analysis. While peptides generally exhibit shorter sequences, they can still possess diverse properties crucial for their therapeutic potential. By examining properties such as AlphaFold2 and ESMFold scores, hydrophobicity (GRAVY), isoelectric point, extinction coefficients, and instability index, a comprehensive understanding of peptide therapeutics can be obtained.

Link: Please refer to the peptides data for more information.

B. Biosimilars - Rank order

The prediction scores indicate the reliability of predicted structures and take into account the potential variability in structural characteristics that can arise from pre-translation modifications during multiple batch productions in in-vivo systems. These scores can be utilized to rank and prioritize biosimilar candidates for reduced testing, with those having higher scores being given preference. The rankings of therapeutic proteins using the scores from AF2 algorithm have been documented. These rankings were designed to assess the risk associated with pre-translation modifications during multiple batch productions and provide a basis for minimizing testing requirements for biosimilar candidates with higher scores.

Link: Please refer to the pLDDT for rank order using AF2 pLDDT scores

Link: Please refer to the pTM for rank order using AF2 pTM scores

C. Physiochemical Attributes computation Source Code

A Python script has been developed to facilitate the retrieval of physicochemical attributes of proteins and peptides using the Protparam tool from Expasy. The Protparam tool is a valuable resource that provides essential information about various properties of therapeutics using their amino acid sequence only. This script automates the process of retrieving these attributes, simplifying the task for analyzing proteins and peptides. To access the script and take advantage of its capabilities, the link is ProtParam-Expasy.