Workshop data problem

Wednesday, May 8, 2024 3:29 PM

Design of Biopharmaceutical Formulations Accelerated by Machine Learning | Molecular Pharmaceutics (acs.org)
Formulation dataset, may not have enough data samples

We applied our approach to three different variants of a tandem single-chain variable fragment (scFv) derived from the antibody Humira

Eight factors (pH, sodium chloride, l-arginine, l-lysine, l-proline, trehalose, mannitol, and Tween 20) were considered as independent variables to maximize T_m.

We noted that we trained a different surrogate model for each of the variants since we performed the optimization of formulation conditions independently for the different variants. However, since the initial design is independent of the response and was designed with the purpose of filling the design space, a common design was used for all of the variants.

Question on the dataset of paper "Design of Biopharmaceutical Formulations Accelerated by Machine Learning"

Hi Harini, my name is Yan Chen, a Post-Doc fellow at Merck. I came across your paper "Design of Biopharmaceutical Formulations Accelerated by Machine Learning". I noticed that it mentioned in the paper that eight factors (pH, sodium chloride, l-arginine, l-lysine, l-proline, trehalose, mannitol, and Tween 20) were considered to maximize T_m. However, in the supplemental material, only five factors were provided. I am wondering where to find the data for the remaining three factors (l-arginine, l-lysine, l-proline)? Thanks a lot!

https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-023-06416-w#data-availability
Dataset development of pre-formulation tests on fast disintegrating tablets (FDT): data aggregation

Optimization of controlled release nanoparticle formulation of verapamil hydrochloride using artificial neural networks with genetic algorithm and response surface methodology - ScienceDirect
A different area of formulation, but has explicit y = f(x) functions

https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.0c00629
Machine Learning Models of Antibody–Excipient Preferential Interactions for Use in Computational Formulation Design
Contain categorical? Yes, aHelix, Bsheet, loop, number of levels, number of categorical variables

Prediction Machines: Applied Machine Learning for Therapeutic Protein Design and Development - ScienceDirect
Pharmaceutical drug development requires comprehensive biophysical characterization for physical stability and comparability assessment. Several groups have used machine learning models to predict drug stability and biosimilarity during stressed conditions including extremes of temperature and pH. Perhaps the most comprehensive application of statistical learning to protein biophysical stability and comparability over the past decade has been the application of

statistical learning to protein biophysical stability and comparability over the past decade has been the application of singular value decomposition and novel two dimensional visualization by the Middaugh lab.[33,69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80]

A dataset of formulation compositions for self-emulsifying drug delivery systems | Scientific Data (nature.com)
OSF | A dataset of formulation compositions for self-emulsifying drug delivery systems
OSF | A dataset of formulation compositions for self-emulsifying drug delivery systems Wiki

Table 2 List of features in the SEDDS dataset and their related formulation component and description. (nature.com)

Dataset consists of 20 drugs, each with a unique API_mol_wt

Other datasets that may not be immediately useable

Predicting Antibody Developability Profiles Through Early Stage Discovery Screening - PMC (nih.gov)

DOTAD: A Database of Therapeutic Antibody Developability | Interdisciplinary Sciences: Computational Life Sciences (springer.com)

Machine learning prediction of antibody aggregation and viscosity for high concentration formulation (tandfonline.com)

Developability assessment at early-stage discovery to enable development of antibody-derived therapeutics | Antibody Therapeutics | Oxford Academic (oup.com)

Machine Learning for Biologics: Opportunities for Protein Engineering, Developability, and Formulation: Trends in Pharmacological Sciences (cell.com)

Automated optimisation of solubility and conformational stability of antibodies and proteins | Nature Communications

FireProtDB: database of manually curated protein stability data | Nucleic Acids Research | Oxford Academic (oup.com)

https://pymoo.org/problems/many/dtlz.html

https://pymoo.org/problems/multi/sym_part.html
The SYM-PART [51] problem suite is a multi-modal multi-objective optimization problem (MMOP). **In MMOPs, a solution :math:`y` in the objective space may have several inverse images in the decision space**. For this reason, an MMOP could have more than one Pareto subsets.