Machine learning driven acceleration of biopharmaceutical formulation development using Excipient Prediction Software (ExPreSo)

Estefania Vidal-Henriquez, Thomas Holder, Nicholas Franciss Lee, Cornelius Pompe, Mark George Teese*

Leukocare AG, Klopferspitz 19a, 82152 Martinsried/Munich, Germany

*corresponding author: mark.teese@leukocare.com

Declarations of interest: none

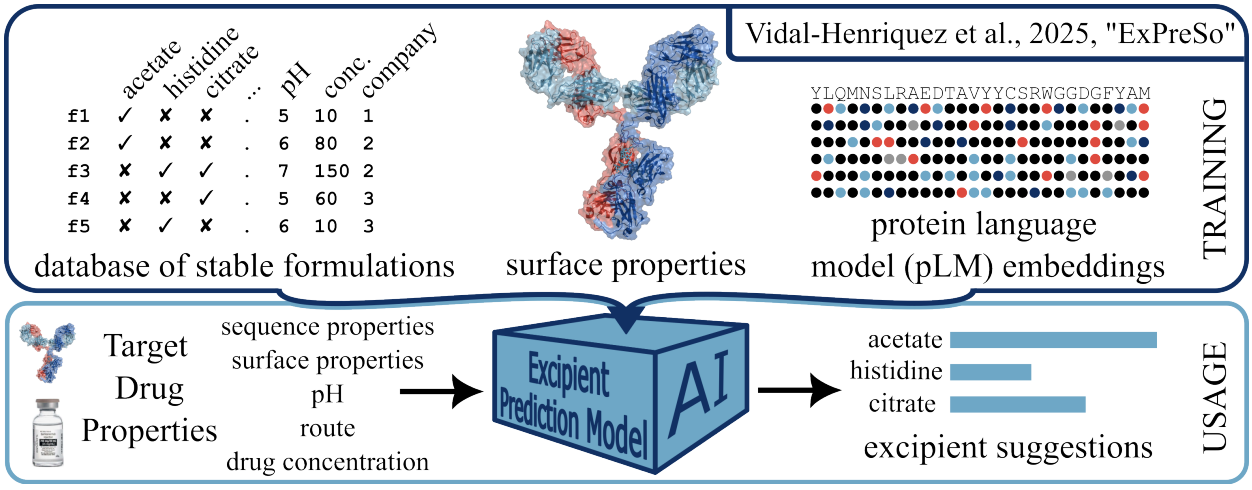
Abstract

Formulation development of protein biopharmaceuticals has become increasingly challenging due to new modalities and higher desired drug substance concentrations. The constraint in drug substance supply and the need for many analytical methods means that only a small selection of excipients can be thoroughly tested in the lab. There are few in-silico tools developed to refine the candidate excipients selected for wet lab testing. To fill this gap we developed the Excipient Prediction Software (ExPreSo), a machine learning algorithm that suggests inactive ingredients based on the properties of the protein drug substance and target product profile. A dataset of over 300 peptide/protein drug formulations with proven long-term stability was created. The dataset was augmented with predictive features including protein structural properties, protein language model embeddings, and drug product characteristics. Supervised machine learning was conducted to create a model that suggests excipients for each drug substance in the dataset. ExPreSo could successfully predict the presence of the nine most prevalent excipients, with validation scores well above a random prediction, and minimal overfitting. A fast variant of ExPreSo using only sequence-based input features showed similar prediction power to slower models that relied on molecular modeling. Interestingly, an ExPreSo variant with only protein-based input features also showed good performance, proving that the algorithm was resilient to the influence of platform formulations. To our knowledge, this is the first machine learning algorithm to suggest biopharmaceutical excipients. Overall, ExPreSo shows great potential to reduce the time, costs, and risks associated with excipient screening during formulation development.

Keywords: Formulation development, machine learning, excipients, inactive ingredients, biopharmaceuticals, monoclonal antibodies

Abbreviations used: ExPreSo, Excipient Prediction Software; ROC, Receiver Operating Characteristic; AUC, Area Under the Curve; TPP, Target Product Profile; FDA, Food and Drug Administration; PCA, Principal Component Analysis; LOGO-CV, Leave-One-Group-Out Cross Validation; mAb, monoclonal antibody; CDR, Complementarity-Determining Region; SHAP, SHapley Additive exPlanations; MOE, Molecular Operating Environment

Graphical Abstract



1. Introduction

Formulation development is the process in which inactive ingredients are chosen to be added to a drug substance in order to stabilize it for manufacturing, distribution, storage, and patient usage. For protein and peptide biopharmaceuticals, these inactive ingredients, known as excipients, must prevent protein degradation and configuration changes which might affect efficacy and safety[1]. From a regulatory and operational point of view, it is desirable that formulations have an assigned shelf-life at targeted storage temperature (e.g. 5°C ± 3°C or 25°C ± 2°C at 60% ± 5% relative humidity) of at least 12 months. This has to be enabled by the careful selection of the formulation pH and a handful of excipients (usually no more than 5). The large number of possible excipients and of effects that need to be accounted for make formulation development a highly complex field. Biologics license applications also need to outline qualitative and quantitative aspects regarding the use of each excipient contained in the drug product (EMEA/CHMP/QWP/396951/2006).

In order to reduce the above mentioned complexity, some companies have relied on *platform formulations* for developing their biopharmaceutical products.[2,3] Typically, this comprises a standard buffer and a list of excipients that are tested against individual drug substances. This process is combined with pre-screening to select molecules with a low amount of chemical liabilities and aggregation prone regions, and presumably, compatibility with the preferred base formulation. Further incremental improvements are usually made by exchanging individual excipients with alternatives that have the same mechanism of action, while keeping the remaining excipients constant. This approach can save time and effort[2] at the cost of not testing a wider variety of excipients that might be better suited for the protein of interest.

There is currently a strong interest in the development of biologics drugs with high drug concentrations.[4–7] This is driven by the increase in biologics with subcutaneous application,[8] which offers many advantages to patients but usually requires smaller injection volumes. Formulation development of high concentration protein drugs is extremely challenging due to problems with aggregation and viscosity/syringeability.[5] Furthermore, there are other market trends that offer challenges in formulation development, such as the increasing development of bispecific antibodies[9,10] and antibody drug conjugates (ADCs).[11] More than ever, a good formulation design must take into account the specific characteristics and liabilities of the drug substance under study. This applies to

new drug substances under investigation and also for reformulation, for example to change from lyophilized to liquid formulations and to move from intravenous to subcutaneous applications.

In the development of biopharmaceuticals, there are in-silico prediction tools supporting many processes before and after formulation development, but few for the stage of formulation development itself. In early drug development, there are a growing number of prediction tools supporting lead development, particularly for antibodies[12], and also for developability assessments aiming to select stable candidates.[13–15] In later stages of biopharmaceutical development, physics-based methods such as digital twins are well established for real-time process optimization.[16] For biopharmaceutical formulation development itself, there exist only a few studies of in-silico excipient screening. Molecular docking can been used to identify excipients binding a protein,[17] however most docking algorithms are designed to identify strong binders to act as inhibitors, rather than transient interactions.

The long-term stabilization of a drug substance involves complex stochastic interactions of many elements, such as protein-protein interactions, excipient-protein interactions, and excipient-excipient interactions among others. In order to produce an in-silico prediction of successful long-term stabilization, all these elements need to be modeled simultaneously, for extremely long time scales. Complexity is increased even further by the fact that some detergents are used in concentrations above their critical micelle content, i.e., they exist as micelles in solution, and need to be modeled accordingly, thus increasing the minimum size of the simulation.[18] This size should also be large enough to account for other indirect excipient effects such as preferential exclusion (for example by sucrose).[19,20] Taking into account all these interactions is a computationally intensive task. With current technology, such simulations take a large amount of time, and are too expensive to be commercially viable. So far, this type of calculations have only been published in an academic context for a handful of excipients.[21–25] Approaches to improve performance range from coarse-grained molecular dynamics (MD) simulations[26] to AI powered MDs, where AI-derived properties are used to refine the force-fields used in molecular dynamics,[27] or machine learning models are trained to predict the outcome of MD simulations.[28]

A different approach to predict excipient binding in-silico is to conduct short all-atom MD simulations with the protein surrounded by a high concentration of the desired excipient. This has been referred to as fragment mapping, and can be used to rank excipients according to their ligand affinity.[29–32] All methods that look at overall protein-excipient interactions rely on the assumption that general excipient binding increases protein stability. However, it has been shown that in many cases excipient binding does not improve stability.[33] Also, the exact sites of protein-protein interaction or transient unfolding are usually unknown, making it impossible to analyze the excipient binding in only the most relevant regions.

An alternative solution is the development of machine learning algorithms to predict stable formulations. A big challenge in creating such an algorithm is the availability of data, since it would need to be trained with a large number of excipients and drug substances. Since this data is kept confidential, this would only be feasible for companies with a large and diverse drug substance portfolio. However, what is publicly available is the final formulation for drugs approved by regulatory authorities. This database of formulations is growing rapidly[7,34] and it has already enabled the first quantitative analyses of stabilizing excipients,[7,34–36] and trends over time.[3] To our knowledge, this data has not yet been utilized for the development of machine learning algorithms to predict stabilizing excipients in formulations.

This machine learning approach is also supported by recent advances in computational tools such as AlphaFold2 and protein language models (pLMs). AlphaFold2 is a machine learning algorithm that provided a breakthrough in the de-novo prediction of protein structures.[37] Whereas, pLMs are large language models that have been trained explicitly to predict protein sequences. A byproduct of pLMs are protein embeddings, which are the vectorial representation of each amino acid in the language model.[38] Once a pLM is trained, the protein embeddings can be rapidly generated from any input protein sequence. These embeddings encode information about the amino acid and their surroundings in the sequence. Their use has provided a leap forward in predictive power for different machine learning tasks such as the prediction of structure, function, and epitopes.[39–42]

In this study, we created the Excipient Prediction Software (ExPreSo) using the database of formulations of approved biopharmaceuticals. ExPreSo is a group of machine learning algorithms, each trained on predicting the presence or absence of a particular excipient. It receives as input a drug substance sequence and information about the target product profile (TPP), such as pH, stock keeping unit (liquid or lyophilized), and desired drug substance concentration. As output, ExPreSo produces a series of percentages, each representing the likelihood of a specific excipient to be present in a stable formulation of the drug substance. ExPreSo has predictive power for nine excipients, and it is to our knowledge the first machine learning algorithm of its type.

2. Materials and Methods

A database was made of all formulations of drug products approved by the FDA at the time of 27 September 2024. This database was filtered to retain only formulations that contained one active ingredient which had an available amino acid sequence. The formulations were then converted to a table where each row represented a formulation, and each column represented an excipient. For compounds that are identical in solution, such as mono- or di-basic versions of the same chemical, all the formulations containing at least one of the related compounds were classified as containing the excipient. For example, formulations were deemed to contain the excipient buffer 'sodium phosphate' whether the listed ingredient was the mono- or dibasic form, or whether the powdered form of the chemical compound was hydrated or anhydrous. Each cell in the table contained the value True or False, representing the presence or absence of a particular excipient in a formulation. Duplicate formulations were removed if they had the same International Nonproprietary Name (core name, excluding biosimilar suffixes) and the same excipients. Excipients found in less than 10% of the formulations were then removed, resulting in nine remaining excipients. The smaller number of excipients created new duplicate formulations, which were again dropped.

To this dataset of excipients, we added information on the drug product. Several categorical variables were one-hot encoded, including the route of administration and stock keeping unit (liquid/lyophilized). The most recurring companies (marketing authorization holders) in our dataset were kept, while the rest of the companies were grouped under “other company”. The number of kept companies was selected such that “other company” amounted to approximately 50% of the total formulations. The date of marketing start of the product, the pH, and the drug concentration were kept as numerical variables. Missing numerical data was imputed with the mean value of the remaining observations.

We then added a number of protein based features (descriptors), based on sequence and generated structures. For each protein, we calculated the frequency of each individual amino acid and each

possible dipeptide pair in the sequence. We also added the sequence-based isoelectric point and fraction of helix, sheet, and turn residues using the BioPython ProtParam module.[43] Protein language model embeddings were calculated using the ProtTrans model.[41] Specifically, we calculated the embeddings using full-length sequences and the prott5_embedder.py script supplied by the ProtTrans developers, using the default prot_t5_xl_half_uniref50 model. When a drug substance had more than one protein chain, the chains were concatenated for this analysis. Since ProtTrans produces a vector of length 1024 for each amino-acid in the protein sequence, the embeddings were aggregated for each protein by taking the mean along the entire sequence.

Full-length 3D protein structures were then generated for each active substance. Monoclonal antibody (mAb) structures were generated using the antibody modeler algorithm of Chemical Computing Group Inc Molecular Operating Environment (MOE) version 2022.02, with the default Fc glycosylation. Non-mAb drug substances were modeled using AlphaFold2[37] and then protonated using the quick-prep protocol of MOE. Surface properties were calculated at pH 7.0 and 0.1 mM NaCl using MOE software and added to the dataset. The dataset was further supplemented with drug substance type. The type was defined as a mAb if MOE detected a complementarity-determining region (CDR). The IgG type (IgG1, IgG2 or IgG4) was one-hot encoded. The full list of input features is given in Supplementary Table S1. Although the code and full dataset is proprietary, a reduced dataset containing formulations with a marketing start before 2020 is included in Supplementary Data.

The initial number of input features was many times the number of observations (formulations), and initially yielded models with high overfitting (data not shown). We first removed highly correlated features (R>0.8). We then applied principal component analysis (PCA) to reduce the number of features, while retaining much of the information, until the final number of features used in the model was approximately half the number of observations (formulations). In order to maintain interpretability, PCA was applied separately to groups of features. We retained ten dimensions (features) for amino acid frequency, 32 for dipeptide frequency, and 32 for embeddings. For the features derived from the protein structures, we first grouped the features by similarity, for example we made groups of general hydrophobicity, hydrophobicity near CDR, general negative patches, negative patches near CDR, and so on. We then applied PCA, and reduced each group to two dimensions. This reduced the initial 89 protein properties calculated with MOE to 39, reducing the noise generated by the large number of features while retaining model interpretability.

The nine different machine learning algorithms were trained independently of each other using the ensemble Extra-Trees algorithm[44] as part of the python Scikit-learn package version 1.6.0 using the default hyperparameters.[45] The Extra-Trees algorithm is very similar to Random Forest, but includes additional randomness in the trees by selecting the split threshold for each feature completely at random, rather than by optimized selection.[44] The Extra-Trees algorithm is therefore faster to train than random forest but still allows the calculation of feature importances, as applied recently to understand the properties underlying the developability of therapeutic proteins.[46] To improve accuracy we performed SMOTE oversampling[47] to balance the dataset. To reduce overfitting, the final model used only the top 20 features, as determined by feature importance in an initial model, measured using SHAP.[48] Each of the nine machine learning models used a different set of 20 final features.

For validation purposes we divided the formulations into groups (clusters), taking care that similar drugs by name or by sequence similarity would be included in the same cluster. Clustering by sequence similarity was performed using CD-HIT[49,50] on joined sequences containing all domains of the full

protein, with cutoff identity threshold of 95% for mAbs, and 80% for non-mAbs. This ensured, for example, that biosimilars were clustered with their original innovator drugs, antibody-drug-conjugates were clustered with their unconjugated forms, and that all peptide variants (e.g. insulin human, insulin aspart, insulin glargine) belonged to the same cluster. During machine learning validation, all members within a cluster were either part of the train set or the test set, but were never split between both.

We created a blind test set comprising approximately 20% of the observations/formulations, with the remaining formulations forming the train set. We used the same blind test set for all excipient models, ensuring that it contained at least two formulations containing each excipient, and at least two formulations lacking each excipient. To prevent the validation results of the blind test from being dominated by drugs that have been extensively reformulated, only two representatives from each cluster were retained in the blind test set, and the remainder were dropped. Cross-validation within the train set was performed using leave one group out validation, where the algorithm was trained iteratively on all clusters except one, and then tested on the remaining cluster. To counteract the inherent high variability in validation metrics due to the small dataset we ran the ExPreSo algorithm ten times using different random seeds. The random seeds primarily affected the formulations selected for the blind test set, but also were used in SMOTE oversampling and the Extra-Trees algorithm itself. The results presented here correspond to the average of all ten runs.

3. Results and Discussion

We created a dataset of FDA-approved formulations of protein or peptide biopharmaceuticals. The dataset contained 335 formulations, comprising 241 different drug substances. The dataset contained 150 formulations from monoclonal antibodies (mAbs, 45%). Of the mAb formulations in the dataset, the majority had heavy chains of type IgG1 (30% of the dataset, 68% of the mAbs, Figure 1A). An analysis of the FDA approval dates shows that the number of protein/peptide drug formulations is increasing rapidly (Figure 1D), and indeed a large proportion of the dataset corresponded to drug formulations approved in the last 5 years. The dataset included information about the drug product, such as whether it was liquid or lyophilized, the year of approval, and the company (marketing authorization holder). This dataset was supplemented with features derived from the in-silico modeled 3D structures of the drug substances, such as size of surface patches (positive, negative, ionic, hydrophobic), charge dipole, and isoelectric point (Table S1). The dataset was also supplemented with sequence-based features, which included amino acid frequency, dipeptide frequency, and protein embeddings extracted from the ProtT5 model.[41]

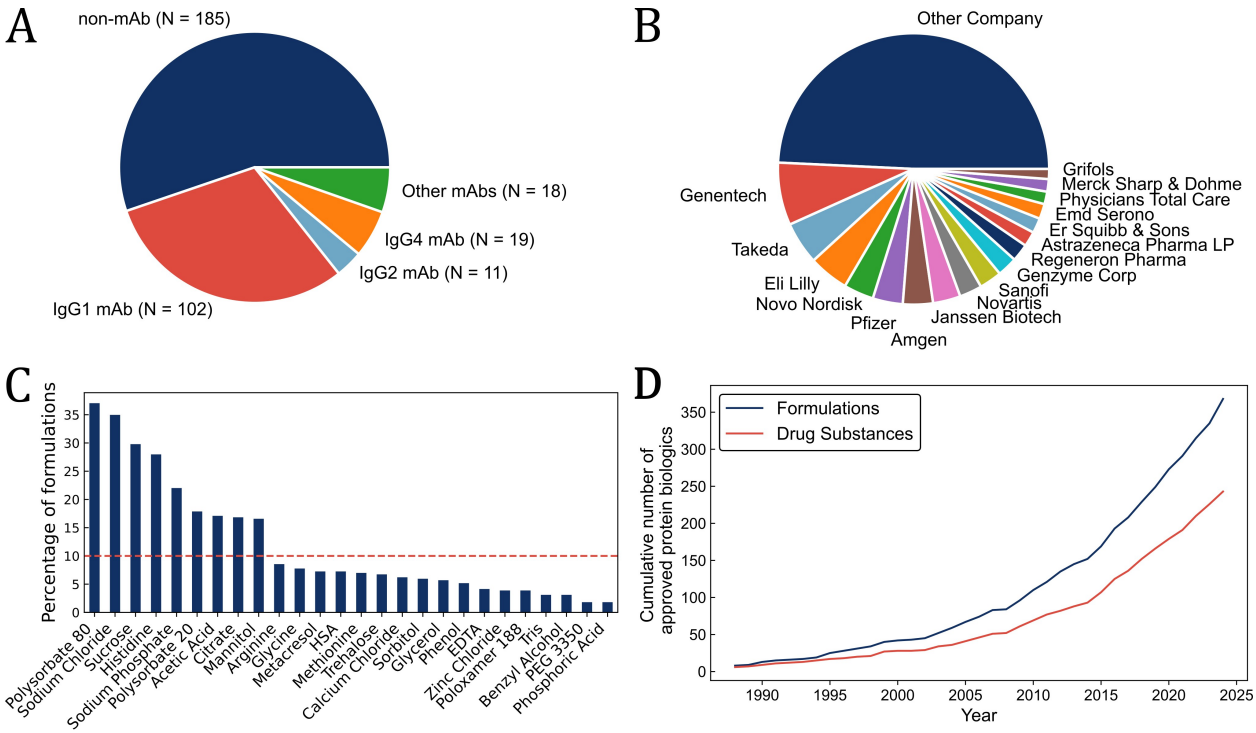


Figure 1. **Main characteristics of the dataset used for machine learning.** A) Composition of antibody type. B) Companies retained in the input features. The large "Other Company" category enables the use of this feature in ExPreSo as a bioinformatic tool (see Methods). C) Abundance of excipients in the dataset before selecting the top excipients used in ExPreSo and dropping duplicate formulations (see Methods). D) Cumulative number of approved formulations and drug substances in the ExPreSo dataset, before dropping duplicates (same data as in Figure 1C).

To predict excipients in formulations, we created a collection of independent machine learning algorithms. Each algorithm is an ExtraTrees classifier[44] (an algorithm similar to Random Forest[51]) that predicts the presence or absence of one excipient in the stable final drug product. The table containing the predictive features was therefore identical for all the different excipient targets, and comprised the protein properties and target product profile. To allow de-novo formulation suggestions, the predictive features did not contain any information regarding the presence of other excipients. Preliminary experiments revealed that prediction power was only reliable for more common excipients, therefore ExPreSo is limited to excipients present in more than 10% of formulations. This yielded nine excipient targets: acetic acid, citrate, histidine, mannitol, polysorbate 20, polysorbate 80, sodium chloride, sodium phosphate, and sucrose.

ExPreSo prediction power was best for acetic acid, histidine, and polysorbate 80, however prediction for all excipients was well above that of a random predictor (Figure 2). With the exception of acetic acid, the prediction power showed correlation to the abundance of the excipient in the dataset (R^2 = 0.54 without acetic acid, 0.13 with acetic acid, see Figure 2C). The more common excipients had a more balanced dataset, required less oversampling, and thus gave models with better prediction power. This also explains the lower prediction metrics seen for the less common excipients: citrate, mannitol, and polysorbate 20.

We validated the predictive power of ExPreSo using a Leave-One-Group-Out cross validation (LOGO-CV) methodology and a blind test validation (see Materials and Methods). In both cases we measured

the area under the curve (AUC) for the Receiver Operating Characteristic (ROC) curve and for the Precision-Recall (PR) curve. The dataset is small and unbalanced, in that there were many more formulations without each target excipient than with it. Because of this, there was high run-to-run variation in prediction performance, particularly for the blind test dataset. To counteract this we always show the average performance metrics for 10 experimental runs with different random seed values. ExPreSo showed high predicted power (ROC-AUC > 0.7) for 5/9 excipients in the LOGO cross-validation.

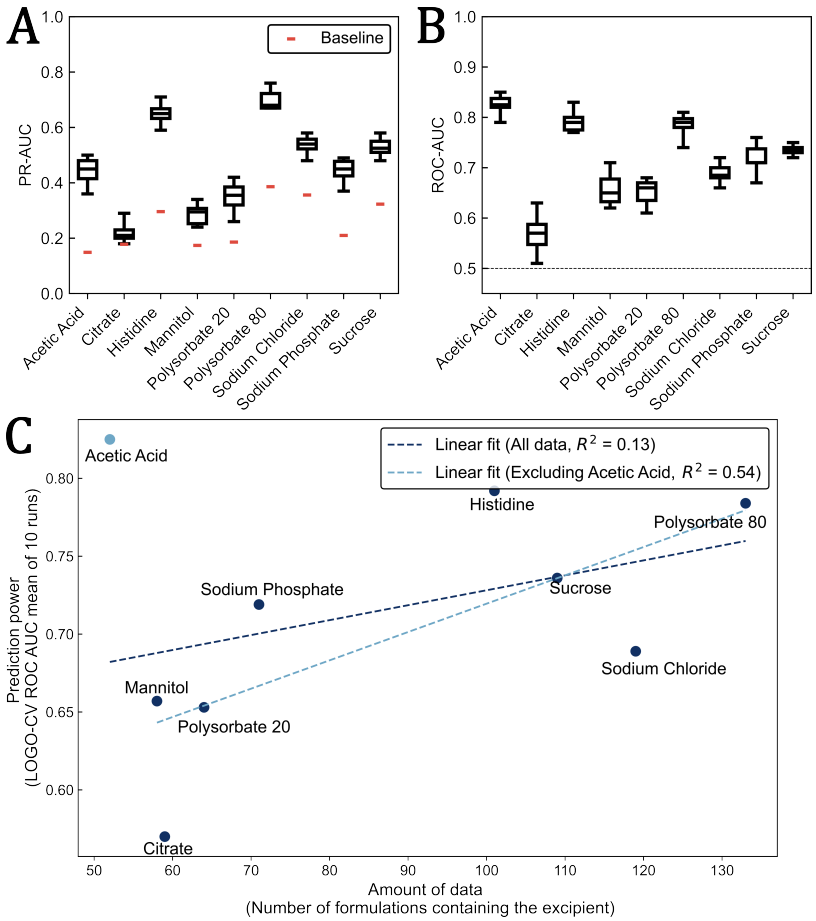


Figure 2. **Summary of predictive power for the nine excipients in ExPreSo.** Cross-validation data against the train set was carried out using leave-one-group-out methodology (see Methods). Data was aggregated over 10 experiment runs, each using a different random seed value for creation of the test set and model parameters. In each boxplot, whiskers indicate min and max, and the central line indicates the median. A) Precision Recall Area Under the Curve (PR-AUC). The baseline (red dash) indicates the precision of a random predictor. B) Receiver Operating Characteristic Area Under the Curve (ROC-AUC). The baseline (dotted line) indicates the performance of a random predictor. C) ROC-AUC Prediction power according to ROC-AUC showed a correlation to the abundance of the excipient in the dataset, with the exception of acetic acid.

While some excipients performed consistently over several runs, such as polysorbate 80 and histidine (see Figure 2), other excipients showed more varied behavior, such as citrate and mannitol. In this small dataset, it is clear that the randomized selection of which formulations were present in the blind test set or train set had a large effect on performance metrics.

The ExPreSo algorithm can be run so that it is either interpretable or fast, based on the chosen subset of input features used. For our "Interpretable" model, we excluded sequence based features such as protein language model embeddings and dipeptide frequency. The remaining protein-based features were derived from 3D modeling, and are human interpretable when examining feature importances. To enable ExPreSo as a standalone predictor, we created a "Fast" model that excludes features derived from the 3D structures, which require slower molecular modeling algorithms. Protein-based features in this Fast algorithm are derived from sequence embeddings and dipeptide frequencies, which can be calculated for a sequence in milliseconds. However, protein embeddings do not have a direct interpretation and dipeptide frequencies are only weakly linked to protein properties, so that the Fast model has low human interpretability when examining feature importances. The performance of the Interpretable and Fast models were comparable (Figure 3).

To address the risk of platform formulations biasing our dataset, we calculated the Jaccard similarity index of formulations of the same company. The Jaccard index compares how similar two sets are, and it is defined as the amount of elements in the sets' intersection divided by the number of elements in the union of the two sets. We noticed that very few of the formulation pairs had a high Jaccard index (see Supplementary Data). This was true for the full dataset, and also when the top companies were considered separately. This suggests that most companies, even though they might take a similar initial formulation base for their experiments, always tailor their final formulation to the particular drug substance.

We went one step further in testing whether our algorithm was company-biased and created a "Protein-Based" version of the algorithm that does not take into account any information related to the manufactured product, such as the company, pH, year of approval, and route of administration. Instead, the input features of this model contained only the surface properties derived from molecular modeling, protein type, protein language model embeddings, and dipeptide frequencies. If the non-protein features were very relevant and thus our dataset heavily biased due to platform formulations, we would expect to see a strong decrease in performance when using only protein based features. Interestingly, we observed only a small decrease in predictive power for the Protein-Based model (Figure 3), showing that the dataset and algorithm are surprisingly resilient to the bias of platform formulations. Combined with the Jaccard index analysis of formulation similarity, this indicates a more tailored excipient selection than expected for the formulations in the dataset. Firstly, it's possible that the preferred excipients by certain companies were strongly suited to the protein scaffold that the company was using for multiple related drug substances. Secondly, during developability screening and downstream processing, there might be a selection of drug candidates that are stable in the preferred platform formulation, as part of a 'platform fit' approach.[3]

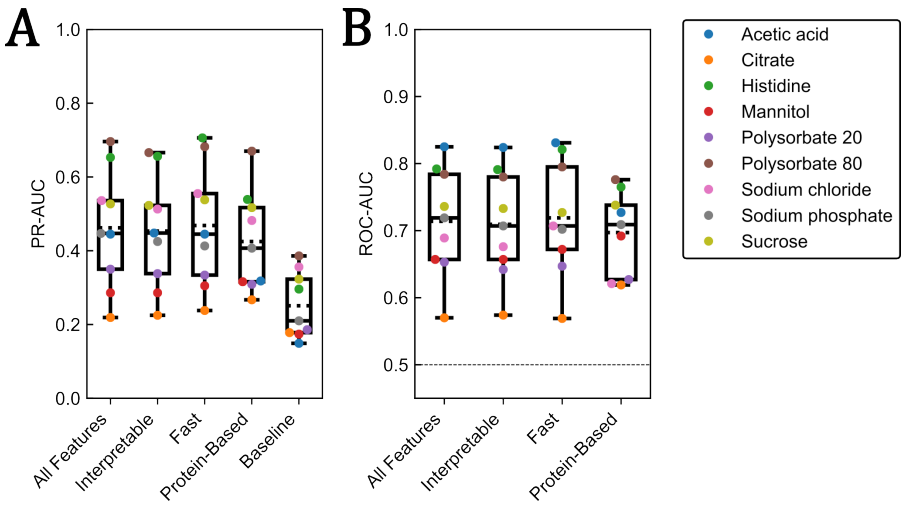


Figure 3. **Fast, interpretable, and protein-only versions of ExPreSo show similar performance.** ExPreSo was validated using different sets of input features (descriptors). The "Interpretable" model lacked sequence-based features such as pLM embeddings. The "Fast" model lacked features derived from molecular modeling. The "Protein-Based" model used only features based on the protein sequence and 3D structure. The AUC values shown here were obtained by LOGO cross-validation (see Methods). Performance metrics shown are Precision Recall Area Under the Curve (PR-AUC), and Receiver Operator Characteristic Area Under the Curve (ROC-AUC). The continuous middle line in the boxplot represents the median, whereas dotted line represents the mean. For the ROC-AUC, the baseline (dotted line) indicates the performance of a random predictor.

Care was taken to ensure ExPreSo showed little overfitting, which is a common problem in small datasets with a high number of input features. In machine-learning, overfitting is typically observed when the performance in validation metrics is higher for a cross-validation test set (where observations are used for both training and testing) than for the hold-out "blind" test set that contains observations upon which the algorithm is never trained. In simpler machine-learning models with a single prediction target, overfitting is typically a qualitative observation, usually visualized by plotting the validation metric (e.g. ROC) for the cross-validation and blind test sets. For ExPreSo this method would be too unwieldy, considering we wanted to simultaneously visualize the performance against nine different prediction targets (excipients), four different model types with different feature inputs (all, interpretable, fast, protein-based), measured with two different validation metrics (ROC-AUC, PR-AUC), and averaged over 10 runs to reduce variability. We therefore developed an "overfitting index", which corresponds to the difference in the performance metric between the cross-validation and blind test validation, normalized to the value of the cross-validation:

$$I = \frac{M_{\mathrm{CV}} - M_{\mathrm{B}}}{M_{\mathrm{CV}}}$$

where I is the overfitting index, M_{CV} is the performance metric from cross-validation, and M_{B} is the performance metric from a blind test set. We use M_{CV} as a baseline because this is much more consistent than M_{B} for small datasets. A higher overfitting index indicates a greater difference between cross-validation datasets, and therefore less predictive power towards new observations. For example, a ROC-AUC of 0.7 after cross-validation and 0.6 for the blind test set yields an overfitting index of (0.7-0.6)/0.7, or approximately 0.14, which can be roughly interpreted as an overestimation of cross-validation ROC-AUC by 14%.

We applied several techniques to reduce overfitting, including the automatic removal of correlated features, feature reduction using principal component analysis (PCA), and limiting all models to use only the 20 most predictive features (see Methods). After implementation of these improvements, all algorithms had minimal overfitting (Figure 4). The difference between LOGO-CV and blind test performance was less than 25% for all excipients and with a median of less than 10%. This gives certainty that the model has similar predictive power for new drug substances.

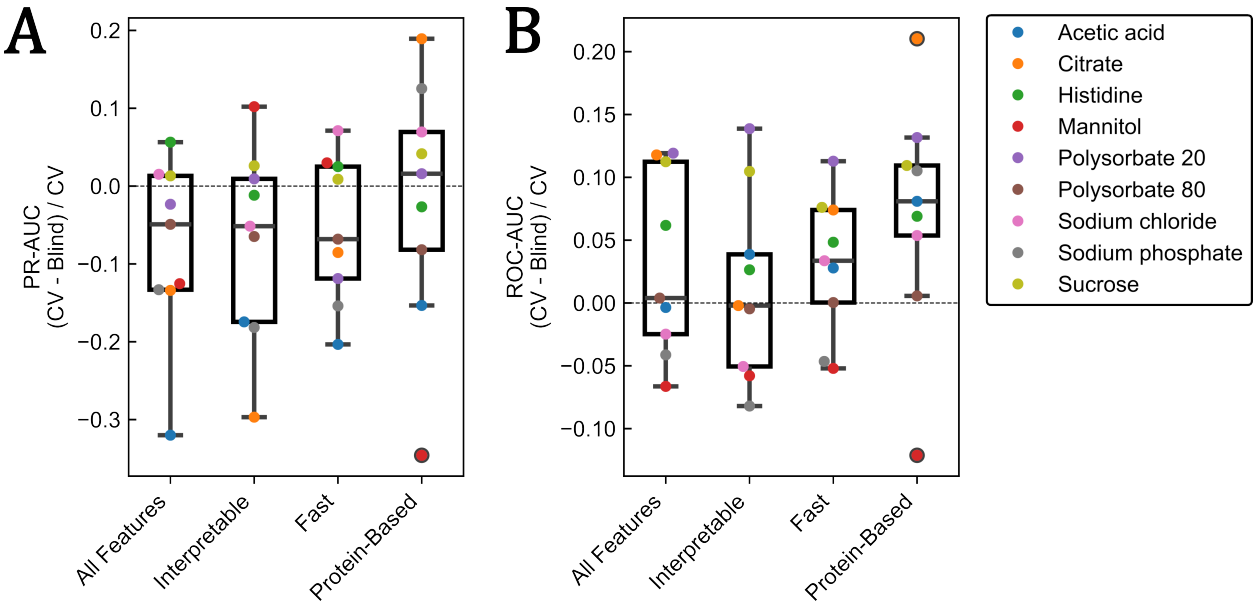


Figure 4. **Minimal overfitting was observed.** Shown is the "overfitting index", defined as the relative difference in performance between the LOGO-CV and the blind test, for all analyzed feature subsets. Higher values indicate higher overfitting. Values above zero (dotted line) indicate better performance for the LOGO-CV validation than the blind test. The middle line in the boxplot represents the median. The whiskers extend until datapoints within 1.5 times the interquartile range.

Examination of the feature importances for the Interpretable model showed that both TPP and protein features were important for prediction (see Supplementary Figure S1). We observed that many of the important features were proxies for protein types, such as length, mass, and percentage of beta sheet residues. However, the top features confirmed that the model is performing as expected, for example, by separating mAbs from non-mAbs. As the dataset grows and when a more stringent clustering of similar proteins can be performed we expect that feature importances would indicate more directly the surface properties that are relevant to excipient selection.

In our analysis acetic acid showed the highest predictability, even though it was the excipient with the least data (Figure 2C). We propose that the presence of acetic acid is highly predictable due to its complete absence in lyophilized formulations, and association with low pH.

Algorithms such as ExPreSo can be used to provide excipient suggestions, in order to guide screening experiments during formulation development. However, it is important that the prediction score from ExPreSo for an excipient is interpreted correctly. Specifically, ExPreSo predicts the likelihood that this drug substance would be formulated with the excipient of interest, if it were in a drug product within this historical dataset with the input target product profile. The output from ExPreSo could then be included as one of the factors considered by a formulation expert, alongside the many other factors

that may guide excipient preselection such as pre-formulation experiments from that drug substance, the available formulations of similar molecules, and the latest scientific literature for the modality, protein family, and individual excipients under consideration.

A possible improvement for ExPreSo would be to instead of having a simple classifier, to create a multi-class predictor. Rather than predicting the presence/absence of a single excipient (e.g. polysorbate 80), the algorithm would predict the assignment to one of three classes (polysorbate 80, polysorbate 20, or no detergent). This method is not trivial in our current dataset since formulations might contain more than one excipient belonging to these groups, and excipients might perform multiple roles, therefore belonging to different groups (for example citrate can be grouped as antioxidant or buffer). As the number of marketed ADCs increases, another improvement would be the addition of features describing the properties of the attached linkers and payloads, not just the protein component as currently implemented. Another minor improvement would be the addition of formulations containing multiple proteins, which are currently excluded for simplicity. However, by far the greatest improvement to algorithms such as ExPreSo would be an increase in dataset size. Because most drug candidates fail at some stage of clinical trials, the dataset of approved drugs used here represents only a small fraction of the formulated therapeutic proteins in existence. Increasing the availability of the data currently held by individual pharmaceutical companies would improve the efficiency of biopharmaceutical formulation as a whole, reduce the costs and risks associated with drug development, and bring great benefits to patients worldwide. Whether private consortia or public initiatives will arise to address this challenge remains to be seen.

When interpreting ExPreSo results to design experiments it is important to consider the underlying assumption that the presence of each excipient is independent from one another. This is of particular importance when algorithms such as ExPreSo are used to predict an entire formulation. For example, the presence of some excipient pairs that share a similar mode of action (e.g. polysorbate 80/polysorbate 20 or sucrose/trehalose) is strongly anticorrelated and it is highly unlikely to find both excipients present in a single formulation. When predicting these excipient pairs, ExPreSo might suggest testing either or both of them. Therefore, we recommend to always use ExPreSo only as a suggested starting point for a design of experiments (DoE) screening experiment, and to analyze the logical sense of each excipient suggested by ExPreSo before starting experiments. Theoretically, ExPreSo-like algorithms could also be applied to support one factor at a time (OFAT) screening approaches, where there already exists a core formulation, and excipients are only modified one at a time. In this case it would be necessary to include the known excipients (both present and absent) as input features.

Another possible use of algorithms of this type would be to assess the potential fit of a drug substance candidate to preexisting platform formulations during the developability process. In this case, the algorithm would predict the excipient likelihood for many candidates and the likelihood for the excipients in the platform formulation could be turned into a metric. This metric can then be used as one of the factors considered during a developability analysis.

Finally, it is important to highlight that ExPreSo currently only predicts excipient presence, not their concentration, and cannot yet replace the careful testing of different concentration combinations for the recommended excipients.

4. Conclusion

To our knowledge, ExPreSo is the first machine learning algorithm developed to suggest excipients for biopharmaceutical formulation development. In a simple approach, we created a group of independent machine learning classifiers. Each algorithm can suggest whether a particular excipient would be present in a stable formulation of the drug substance under consideration. The algorithms had good predicting power with minimal overfitting and can be used to suggest excipients for new drug products entering the market. ExPreSo also proved surprisingly resilient to the influence of platform formulations in the dataset, but did perform slightly better when input features included drug product details such as company. The Fast version of ExPreSo delivers results in seconds. This approach significantly reduces the time required for human-based literature review of formulations for similar drug substances.

ExPreSo helps excipient preselection by using data-derived decision making and thus reducing human bias. In the future, as AI algorithms for excipient suggestion become stronger, it should be possible to reduce the number of excipients screened during formulation development. This will decrease costs to pharmaceutical companies and patients, and reduce the time required for drug product development. As ExPreSo relies on approved drug formulations and improves with increasing data (Figure 2C), the growing number of biologics entering the market (Figure 1D) will considerably expand its dataset in the coming years. This expansion will enhance ExPreSo's predictive accuracy and increase the range of excipients it can effectively predict, thereby boosting its overall utility. Also, after curating a dataset with the concentrations of the excipients in each formulation, future iterations of ExPreSo could also predict excipient concentrations, providing further information to formulation scientists.

5. Acknowledgments

We would like to thank Dmitrij Frishman of Technische Universität München for helpful feedback on the bachelor thesis of Nicholas Lee in the initial stages of the project. We would also like to thank Theodore W. Randolph of University of Colorado Boulder, Andreas Seidl, and Nehil Chaturvedi for helpful comments.

6. References

1. Akers MJ. Excipient-drug interactions in parenteral formulations. *J Pharm Sci*. 2002;91(11):2283-2300. doi:10.1002/jps.10154

2. Warne NW. Development of high concentration protein biopharmaceuticals: The use of platform approaches in formulation development. *Eur J Pharm Biopharm*. 2011;78(2):208-212. doi:10.1016/j.ejpb.2011.03.004

3. Mieczkowski CA. The evolution of commercial antibody formulations. *J Pharm Sci*. 2023;112(7):1801-1810. doi:10.1016/j.xphs.2023.03.026

4. Desai M, Kundu A, Hageman M, Lou H, Boisvert D. Monoclonal antibody and protein therapeutic formulations for subcutaneous delivery: high-concentration, low-volume vs. low-concentration, high-volume. *MAbs*. 2023;15(1). doi:10.1080/19420862.2023.2285277

5. Zarzar J, Khan T, Bhagawati M, Weiche B, Sydow-Andersen J, Alavattam S. High concentration formulation developability approaches and considerations. *MAbs*. 2023;15(1). doi:10.1080/19420862.2023.2211185

6. Jiskoot W, Hawe A, Menzen T, Volkin DB, Crommelin DJA. Ongoing challenges to develop high concentration monoclonal antibody-based formulations for subcutaneous administration: Quo Vadis? *J Pharm Sci*. 2022;111(4):861-867. doi:10.1016/j.xphs.2021.11.008

7. Ghosh I, Gutka H, Krause ME, Clemens R, Kashi RS. A systematic review of commercial high concentration antibody drug products approved in the US: formulation composition, dosage form design and primary packaging considerations. *MAbs*. 2023;15(1). doi:10.1080/19420862.2023.2205540

8. Prašnikar M, Bjelošević Žiberna M, Gosenca Matjaž M, Ahlin Grabnar P. Novel strategies in systemic and local administration of therapeutic monoclonal antibodies. *Int J Pharm*. 2024;667(May):0-2. doi:10.1016/j.ijpharm.2024.124877

9. Holmes D. Buy buy bispecific antibodies. *Nat Rev Drug Discov*. 2011;10(11):798-800. doi:10.1038/nrd3581

10. Spiess C, Zhai Q, Carter PJ. Alternative molecular formats and therapeutic applications for bispecific antibodies. *Mol Immunol*. 2015;67(2):95-106. doi:10.1016/j.molimm.2015.01.003

11. Dumontet C, Reichert JM, Senter PD, Lambert JM, Beck A. Antibody–drug conjugates come of age in oncology. *Nat Rev Drug Discov*. 2023;22(8):641-661. doi:10.1038/s41573-023-00709-2

12. Joubbi S, Micheli A, Milazzo P, et al. Antibody design using deep learning: from sequence and structure design to affinity maturation. *Brief Bioinform*. 2024;25(4). doi:10.1093/bib/bbae307

13. Mieczkowski C, Zhang X, Lee D, et al. Blueprint for antibody biologics developability. *MAbs*. 2023;15(1). doi:10.1080/19420862.2023.2185924

14. Navarro S, Ventura S. Computational methods to predict protein aggregation. *Curr Opin Struct Biol*. 2022;73:102343. doi:10.1016/j.sbi.2022.102343

15. Fernández-Quintero ML, Ljungars A, Waibl F, et al. Assessing developability early in the discovery process for novel biologics. *MAbs*. 2023;15(1). doi:10.1080/19420862.2023.2171248

16. Udugama IA, Lopez PC, Gargalo CL, Li X, Bayer C, Gernaey K V. Digital Twin in biomanufacturing: challenges and opportunities towards its implementation. *Syst Microbiol Biomanufacturing*. 2021;1(3):257-274. doi:10.1007/s43393-021-00024-0

17. Barata TS, Zhang C, Dalby PA, Brocchini S, Zloh M. Identification of protein-excipient interaction hotspots using computational approaches. *Int J Mol Sci*. 2016;17(6). doi:10.3390/ijms17060853

18. Amani A, York P, de Waard H, Anwar J. Molecular dynamics simulation of a polysorbate 80 micelle in water. *Soft Matter*. 2011;7(6):2900-2908. doi:10.1039/C0SM00965B

19. Li J, Wang H, Wang L, Yu D, Zhang X. Stabilization effects of saccharides in protein formulations: A review of sucrose, trehalose, cyclodextrins and dextrans. *Eur J Pharm Sci*. 2024;192(August 2023):106625. doi:10.1016/j.ejps.2023.106625

20. Sudrik CM, Cloutier T, Mody N, Sathish HA, Trout BL. Understanding the role of preferential exclusion of sugars and polyols from native state IgG1 monoclonal antibodies and its effect on aggregation and reversible self-association. *Pharm Res*. 2019;36(8):1-12. doi:10.1007/s11095-019-2642-3

21. Cloutier T, Sudrik C, Mody N, Sathish HA, Trout BL. Molecular computations of preferential interaction coefficients of IgG1 monoclonal antibodies with sorbitol, sucrose, and trehalose and the impact of these excipients on aggregation and viscosity. *Mol Pharm*.

2019;16(8):3657-3664. doi:10.1021/acs.molpharmaceut.9b00545

22. Cloutier TK, Sudrik C, Mody N, Hasige SA, Trout BL. Molecular computations of preferential interactions of proline, arginine.HCl, and NaCl with IgG1 antibodies and their impact on aggregation and viscosity. *MAbs*. 2020;12(1):1-12. doi:10.1080/19420862.2020.1816312

23. Wang Y, Williams HD, Dikicioglu D, Dalby PA. Predictive Model Building for Aggregation Kinetics Based on Molecular Dynamics Simulations of an Antibody Fragment. *Mol Pharm*. Published online 2024. doi:10.1021/acs.molpharmaceut.4c00859

24. Rospiccio M, Arsiccio A, Winter G, Pisano R. The role of cyclodextrins against interface-induced denaturation in pharmaceutical formulations: A molecular dynamics approach. *Mol Pharm*. 2021;18(6):2322-2333. doi:10.1021/acs.molpharmaceut.1c00135

25. Kalayan J, Curtis RA, Warwicker J, Henchman RH. Thermodynamic origin of differential excipient-lysozyme interactions. *Front Mol Biosci*. 2021;8(June):1-13. doi:10.3389/fmolb.2021.689400

26. Calero-Rubio C, Ghosh R, Saluja A, Roberts CJ. Predicting protein-protein interactions of concentrated antibody solutions using dilute solution data and coarse-grained molecular models. *J Pharm Sci*. 2018;107(5):1269-1281. doi:10.1016/j.xphs.2017.12.015

27. Wang J, Olsson S, Wehmeyer C, et al. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent Sci*. 2019;5(5):755-767. doi:10.1021/acscentsci.8b00913

28. Cloutier TK, Sudrik C, Mody N, Sathish HA, Trout BL. Machine learning models of antibody-excipient preferential interactions for use in computational formulation design. *Mol Pharm*. 2020;17(9):3589-3599. doi:10.1021/acs.molpharmaceut.0c00629

29. Jo S, Xu A, Curtis JE, Somani S, Mackerell AD. Characterization of antibody-excipient interactions for rational excipient selection using the site identification by ligand competitive saturation-biologics approach. *Mol Pharm*. 2020;17(11):4323-4333. doi:10.1021/acs.molpharmaceut.0c00775

30. Zhang C, Gossert ST, Williams J, et al. Ranking mAb–excipient interactions in biologics formulations by NMR spectroscopy and computational approaches. *MAbs*. 2023;15(1). doi:10.1080/19420862.2023.2212416

31. Li X, Orr AA, Sajadi MM, et al. Investigating the interaction between excipients and monoclonal antibodies PGT121 and N49P9.6-FR-LS: A comprehensive analysis. *ChemRxiv*. doi:10.26434/chemrxiv-2024-1zv1q

32. Somani S, Jo S, Thirumangalathu R, et al. Toward biotherapeutics formulation composition engineering using site-identification by ligand competitive saturation (SILCS). *J Pharm Sci*. 2021;110(3):1103-1110. doi:10.1016/j.xphs.2020.10.051

33. Zalar M, Svilenov HL, Golovanov AP. Binding of excipients is a poor predictor for aggregation kinetics of biopharmaceutical proteins. *Eur J Pharm Biopharm*. 2020;151(December 2019):127-136. doi:10.1016/j.ejpb.2020.04.002

34. Rao VA, Kim JJ, Patel DS, Rains K, Estoll CR. A comprehensive scientific survey of excipients used in currently marketed, therapeutic biological drug products. *Pharm Res*. 2020;37(10). doi:10.1007/s11095-020-02919-4

35. Strickley RG, Lambert WJ. A review of formulations of commercially available antibodies. *J Pharm Sci*. 2021;110(7):2590-2608.e56. doi:10.1016/j.xphs.2021.03.017

36. Wang SS, Yan Y, Ho K. US FDA-approved therapeutic antibodies with high-concentration

formulation: summaries and perspectives. *Antib Ther*. 2021;4(4):262-273. doi:10.1093/abt/tbab027

37. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

38. Heinzinger M, Elnaggar A, Wang Y, et al. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*. 2019;20(1):1-17. doi:10.1186/s12859-019-3220-8

39. Clifford JN, Høie MH, Deleuran S, Peters B, Nielsen M, Marcatili P. BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Sci*. 2022;31(12):e4497. doi:10.1002/pro.4497

40. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A*. 2021;118(15). doi:10.1073/pnas.2016239118

41. Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell*. 2022;44(10):7112-7127. doi:10.1109/TPAMI.2021.3095381

42. Littmann M, Heinzinger M, Dallago C, Olenyi T, Rost B. Embeddings from deep learning transfer GO annotations beyond homology. *Sci Rep*. 2021;11(1):1-14. doi:10.1038/s41598-020-80786-0

43. Cock PJA, Antao T, Chang JT, et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422-1423. doi:10.1093/bioinformatics/btp163

44. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;63(1):3-42. doi:10.1007/s10994-006-6226-1

45. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830. http://scikit-learn.sourceforge.net.

46. Waight AB, Prihoda D, Shrestha R, et al. A machine learning strategy for the identification of key in silico descriptors and prediction models for IgG monoclonal antibody developability properties. *MAbs*. 2023;15(1). doi:10.1080/19420862.2023.2248671

47. Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Int Res*. 2002;16(1):321–357. doi:10.1613/jair.953

48. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg U V, Bengio S, et al., eds. *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.; 2017:4765-4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

49. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150-3152. doi:10.1093/bioinformatics/bts565

50. Li W, Godzik A. CD-HIT: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22(13):1658-1659. doi:10.1093/bioinformatics/btl158

51. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324