

REPORT



## A machine learning strategy for the identification of key *in silico* descriptors and prediction models for IgG monoclonal antibody developability properties

Andrew B. Waight<sup>a</sup>, David Prihoda<sup>b</sup>, Rojan Shrestha<sup>a</sup>, Kevin Metcalf<sup>a</sup>, Marc Bailly<sup>a</sup>, Marco Ancona<sup>b</sup>, Talal Widatalla<sup>c</sup>, Zachary Rollins<sup>c</sup>, Alan C Cheng<sup>c</sup>, Danny A. Bitton<sup>b</sup>, and Laurence Fayadat-Dilman<sup>a</sup>

<sup>a</sup>Discovery Biologics, Protein Sciences, Merck & Co., Inc, South San Francisco, CA, USA; <sup>b</sup>Discovery Informatics, MSD Czech Republic s.r.o, Prague, Czech Republic; <sup>c</sup>Computational and Structural Chemistry, Merck & Co., Inc, South San Francisco, CA, USA

### ABSTRACT

Identification of favorable biophysical properties for protein therapeutics as part of developability assessment is a crucial part of the preclinical development process. Successful prediction of such properties and bioassay results from calculated *in silico* features has potential to reduce the time and cost of delivering clinical-grade material to patients, but nevertheless has remained an ongoing challenge to the field. Here, we demonstrate an automated and flexible machine learning workflow designed to compare and identify the most powerful features from computationally derived physiochemical feature sets, generated from popular commercial software packages. We implement this workflow with medium-sized datasets of human and humanized IgG molecules to generate predictive regression models for two key developability endpoints, hydrophobicity and poly-specificity. The most important features discovered through the automated workflow corroborate several previous literature reports, and newly discovered features suggest directions for further research and potential model improvement.

### ARTICLE HISTORY

Received 23 February 2023  
Revised 28 July 2023  
Accepted 11 August 2023

### KEYWORDS

Biophysical; computational; descriptors; developability; IgG1; machine learning



### Introduction


Over the past 25 years, monoclonal antibodies (mAbs) have become one of the fastest growing therapeutic modalities, and today, they are the predominant treatment for several disease areas. As a result of decades of research, advances in the understanding of antibody engineering are opening new avenues to more sophisticated therapeutic molecules, such as antibody drug conjugates, multi-specifics, and other antibody-like therapies designed to have increased efficacy and/or higher tolerability in patient populations. Regardless of the particular antibody modality, a crucial step in the evaluation of lead biomolecules prior to clinical and downstream development is a careful assessment of molecule quality via properties collectively referred to as “developability characteristics.”<sup>1,2</sup>

Developability characteristics generally evaluate biophysical properties such as poly-specificity, thermostability, hydrophobicity, electrostatic properties, self-interaction and aggregation propensity. These common assays are often supplemented with other custom and/or proprietary methodologies. Although the specific assays may vary, the aim of all such developability assessments is to identify well-behaved biomolecules that meet the requirements for the production of clinical-grade material. Given the crucial role of developable characteristics in the selection of lead candidate molecules, it is now routine to attempt identification of developable molecules at the design or early research stage.<sup>3,4</sup> Although the approval rate for biotherapeutic candidates is notoriously low, early elimination of poorly behaved

molecules confers considerable advantages in terms of data quality, time, downstream cost, and overall efficiency. Therefore, assessment of developability characteristics as early as possible is the current preferred workflow in lead antibody molecule selection, with the result being that the field of *in silico* developability assessment for biologics has grown exponentially alongside the clinical advancement of mAbs and related antibody platforms.

Recently, structural biology has been propelled forward by the integration of machine learning and deep learning methods.<sup>5–7</sup> These advancements have been made possible by the accumulation of data repositories of sequence and structural data over several decades. Despite these dramatic improvements in *de novo* structure prediction, prediction of the biophysical properties of protein molecules remains far behind in terms of both prediction accuracy and public data sources. Currently represented datasets of sequence data and Protein Data Bank structures are on the order of  $10^8$  and  $10^6$  (with  $10^3$  antibody structures), respectively.<sup>8–11</sup> In contrast, experimental data of antibody biophysical developability characteristics in the public domain are generally derived from fewer than 500 molecules. Moreover, the difficulty in training generalizable learning algorithms is compounded by the vast antibody sequence space, with an estimated  $> 10^{12}$  possible antibodies in a single repertoire.<sup>13</sup> In addition, the number of physiochemical features that can be computed for protein macromolecules is near infinite, with the result that all the antibody

**CONTACT** Andrew B. Waight  [andrew.waight@merck.com](mailto:andrew.waight@merck.com)  Discovery Biologics, Protein Sciences, Merck & Co., Inc, South San Francisco, CA, USA

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19420862.2023.2248671>

© 2023 Merck and Co. Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

developability datasets currently available to our knowledge consist of many more possible features than observations.

*In silico* features can be sequence-based, including information based on composition and alignment of homologous sequences, or even sequence likelihoods and representations from language models trained on natural B-cell repertoire sequence databases.<sup>12,14–16</sup> Feature information can also be obtained through structure-based calculations, which require a structure first to be generated, either through experimental techniques such as X-ray crystallography or cryogenic electron microscopy (Cryo-EM), but more often through homology modeling or deep learning methods.<sup>5,16–20</sup> With structural information, atomic coordination and geometric calculations can be performed, resulting in features that take solvent accessibility, hydrophobicity, or charge distribution into account. Furthermore, structural modeling allows for energetic calculations based on classical force fields, allowing for approximations of the free energy of the protein folded state. However, although mostly accurate structural prediction of Fab molecules based on homology modeling methods has been available for more than two decades, challenges to the accurate modeling of complementarity-determining region H3 (CDRH3) remain.<sup>16,19,21,22</sup> The result of such inaccuracies in CDRH3 structural prediction is a significant amplification of error in many of the physiochemical calculations that rely on atomic coordinates (e.g., CDR surface patch calculations). Despite the potential error introduced by CDRH3 flexibility, structural features are one of dominant features used in antibody property assessment.<sup>2</sup>

In this study, we calculated and collected protein features generated from three of the most popular software packages and developed a custom machine learning pipeline designed to identify the features which best explain the variance of key developability endpoints. We have applied these feature sets and data science methods in conjugation with recently collected and curated relatively large developability datasets (hydrophobic interaction chromatography (HIC) for 770 IgG molecules; poly-specificity reagent (PSR) assay data for 390 IgG molecules) to identify top performers. We describe an automated and flexible machine learning developability pipeline using commercially available protein descriptors and discuss the most salient features for the selected biophysical property predictions. By identifying the importance of such features collected across the available structural landscape, we hope not only to demonstrate the predictive utility of individual components from each feature set, but also to highlight the calculation methods that appear to be the most promising avenues for further investigation and refinement.

## Results

In this study, we focused on mAb molecules with the pre-developability endpoints of HIC and PSR assays. For each assay dataset, we limited the datapoints to IgG molecules with isotypes or mutations that have no known effects with regards to the assay under study.

As described above, the most attractive use case for predictive models in therapeutic antibody developability

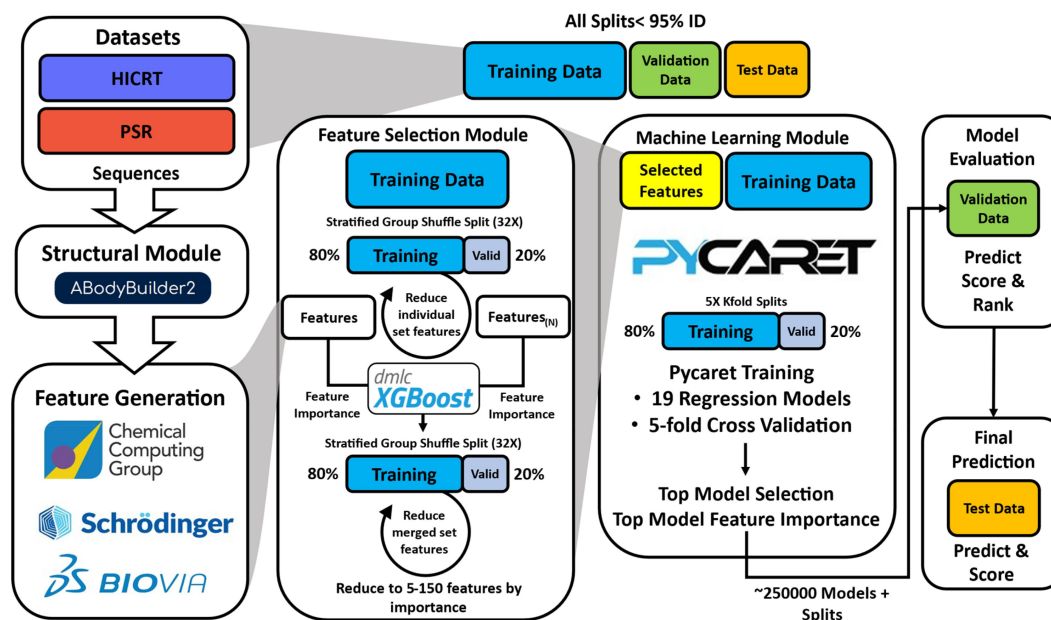
endpoints involves the *a priori* prediction of sequences yet untested by the research group, given the generally unique nature of *de novo* discovered antibodies. However, the training data collected here and collated from individual smaller individual studies pose a non-trivial challenge to the statistical validation of these models through the conventional means of training, validation, and test sets. To begin, the data for each assay are composed of clustered antibody groups, related by sequence identity, of unequal sizes ranging from singlets to over 50 family members. This aspect poses limitations on the use of typical metrics such as cross-validation, which require equal divisions of the training data for evaluation of confidence and works best on data sets where individual members are equidistant from one another (folds have similar distributions). In addition, data collected on developability bioassays show us that the predictive machine learning requirements are mainly for outlier detection of problem molecules. To explain, for both the HIC and PSR assays, the majority of the collected datapoints fall within a normal range, but prediction is needed for the more extreme and less populated examples, leading to a highly skewed dataset. Such skewed data are similarly challenging for machine learning applications which are known to perform best with Gaussian or evenly distributed data. The result of these data limitations is that evaluation of the resulting trained model performance technically measured on a test set generated from a one-time only dataset split and used in a one-time only model prediction call does not accurately reflect the power of the input features, but rather the random chance in the statistical variation of the machine learning pipeline.

We instead propose here a workflow for antibody developability that applies a stringent sequence identity cutoff and a single one-time split of the total data into training, validation, and test sets. The test data are sequestered from all training and validation processes to ensure no data leakage. A large number of models (250K), generated only with the training data, are then evaluated in their ability to predict the validation data, and the top performing features and hyperparameters are chosen and used in a final single prediction of test set for confirmation. The key advantage to this method is highly consistent validation results, which in turn enables unambiguous comparisons of different computational features and methods.

The machine learning strategy described here is designed to accommodate any number of feature sets and produces a ranking of the most important average features, as well as the top scoring models using the available features (Figure 1).

In this study, structural models for all antibodies were generated from sequence using the ABodyBuilder2 command from the ImmuneBuilder package, and sequence and structural-based descriptors were generated using calculations from the popular software suites, Molecular Operating Environment (MOE), Schrödinger Software suite (Bioluminate, Maestro), and Biovia Discovery Studio.<sup>20,23–25</sup>

Using these descriptors, we demonstrate a workflow which allows for the direct comparison of feature sets to evaluate the predictive quality on a specific training/validation/test set of data. The feature selection component of the pipeline makes extensive use of the eXtreme Gradient Boosting algorithm



**Figure 1.** Schematic of machine learning workflow in this study. For each bioassay endpoint in the Datasets category, the data are split into training validation test sets, ensuring low sequence identity, and representative assay variation in the validation set. Fv sequences are modeled using ABodyBuilder2 and features are generated using three popular software packages. Individual or multiple feature sets are trained with XGBoost regression models using 32 grouped splits and reduced individually to X features (X is a hyperparameter 5-150). Multiple reduced feature sets are then combined, resubmitted, and reduced again to X features. The top features selected by XGBoost are submitted with the training set to a PyCaret workflow and trained on 19 regression models. For each X value, the top five models are then tested for prediction and scored on the validation set for ten random seeds. 5000 cycles (250K) models are evaluated and ranked on the validation data. The top performing model on the validation data is confirmed by prediction on the test set.

(XGBoost) in a process of iterations.<sup>26</sup> Individual feature sets are first ranked and selected by feature importance (feature type = gain), before merging and reducing further in a second round.<sup>27</sup> The final selected features from the XGBoost cycles are submitted to the machine learning library PyCaret, which trains and tests multiple regression models.<sup>28</sup> The top scoring algorithms from PyCaret are assessed by prediction on the validation data, which is held separate until after the model has been trained. The search space is sampled through 5000 complete cycles each resulting in 50 separate PyCaret models for a total of 250,000 total machine learning models on each individual or comparison of feature sets. The top-ranked model, features, and number of features from the validation performance are finally confirmed by usage in the prediction of the test set.

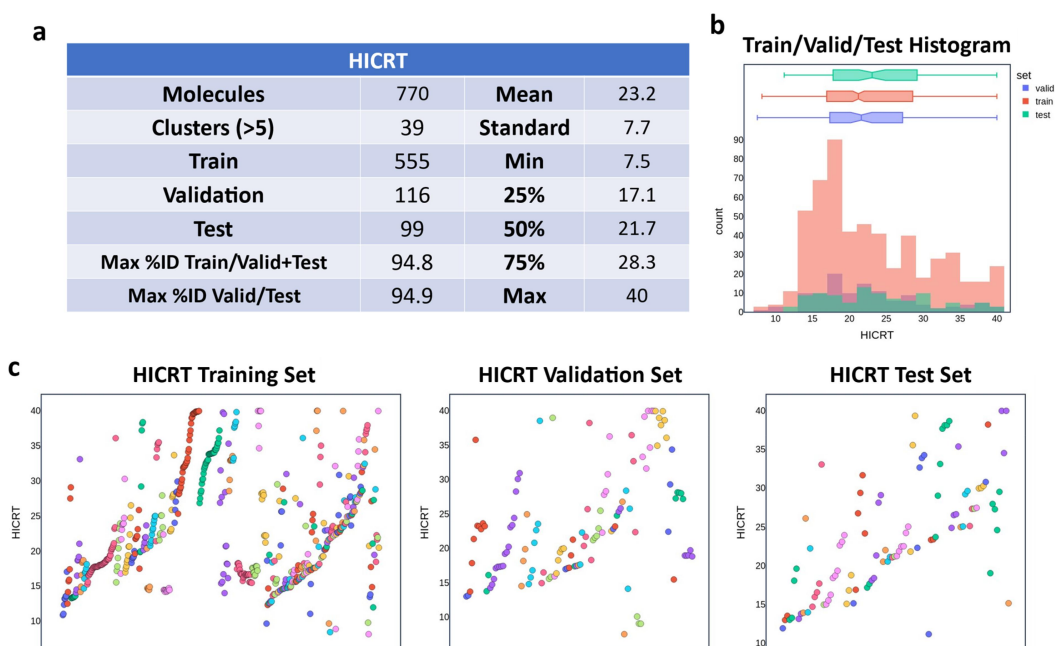
### Processing of collected bioassay data from pipeline projects for machine learning

Data collated from the normal course of therapeutic pipeline projects are naturally grouped, with certain challenging parent sequences or high priority projects often containing a disproportionate number of highly related sequence variants. These aspects already create a challenge for machine learning training, for which diverse datasets are ideal for the model's ability to generalize to new data. For molecules with large family clusters or derivatives, sequence similarity is often the most reliable predictor of biophysical properties, with highly identical molecules such as point mutations tending to behave similarly. Although direct sequence information is not supplied, machine learning methods trained on such data will readily learn from features shared between related sequences

(e.g., pI, molecular weight, and CDR length). Random splitting on such data therefore often results in overfit models with high predictive capacity on validation sets, but poor performance on low identity sequences and therefore unacceptable generalization characteristics. To constrain our prediction models to learning only on the biophysical property calculations, we developed a sequence analysis workflow using identity cutoff criterion in the generation of training and validation datasets used for model training. First, a pairwise mutation matrix is generated, and clusters are assigned based on the identity matrix.<sup>29</sup> Large clusters ( $n > 5$ ) are individually curated for the datasets ensuring as much as possible to contain sequence-related variants with a range of the measured biophysical property in question. The remaining sequences are split using a stratified group method so that the training set comprises approximately 60% of the complete dataset. The remaining 40% is split approximately equally into a validation and test set making sure that family sequences always remain in the same split. A final Clustal distance matrix is calculated to confirm that no two sequences between training validation and test sets exceeds 95% identity over the length of the Fab region.<sup>30</sup> In our experience, careful curation of such data splits to restrict validation and test sets to unseen data, here defined by < 95% identity, is crucial to maintain the generalizability of the final models.

### Hydrophobic interaction chromatography retention time

The hydrophobic interaction chromatography retention time (HIC-RT) used in our study contains 770 IgG datapoints (618 IgG1, 122 IgG4, 30 IgG2) from 39 different clustered projects (Figure 2a). The train, validation and test sets (555, 116 and 99



**Figure 2.** HICRT dataset characteristics and regression performance. (a) Characteristics of the HICRT dataset. (b) Histogram of HICRT values for training validation and test sets. (c) Scatterplot of the training set (left), validation set (middle), and test set (right). Individual points represent molecules in the dataset colored by assigned sequence identity cluster (families).

mAbs, respectively) were separated using a combination of group shuffle split and manual curation to ensure several related clusters spanning a range of HIC-RT values (single and multiple mutations leading to a large change in measurement) as well as individual mAb datapoints representative of the dynamic range of the assay (Figure 2b-c). Fab sequence identity between pairs of training, validation, and test sequences set did not exceed 95%.

The initial process in our workflow explores the best potential features by comparing different combinations (between 5 and 150 total features) using the machine learning library XGBoost. Averaging the XGBoost feature importances for 160K individual models provides a view of the most impactful average features for each dataset.

For the MOE feature set, ASPmax (maximum average surface property), hyd\_strength\_cdr (hydrophobic patch strength for the cdr regions), and hyd\_idx (calculated on sequence by using the Black and Mould hydrophobic index) were the top three features by average importance (Supplementary Figure S1).<sup>31</sup> The best performing model using only MOE features is a K neighbors regressor model ( $R^2 = 0.55$ ,  $\rho = 0.75$  on the test data) with top features identified by sequential feature selection (SFS) of pro\_patch\_cdr\_hyd (area of hydrophobic protein patches near CDRs), HI (hydrophobic imbalance), and ASPmax.<sup>32</sup> The top features of ASPmax, pro\_patch\_cdr\_hyd, HI, and hyd\_idx collected for MOE calculated properties are in agreement with a previously published reports on HIC prediction.<sup>1,33,34</sup>

For the Schrödinger feature set, the most impactful averaged features from 160K individual XGBoost models are CDR\_Hydrophobic\_Patch\_Energy\_gt15 (the sum of residue contributions to strong hydrophobic patches), followed by CDRH3 loop H3\_Aggrescan\_a4v\_pos calculation. AGGRESKAN is a sequence-based algorithm based on

experimental results obtained from amyloid  $\beta$ -peptide and the a4v pos denotes the positive (hydrophobic) values average over a sliding window of 5–11 residues (Supplementary Figure S2).<sup>35,36</sup> The third top feature selected is the H3\_atomic\_contact\_energy, an estimation of CDRH3 desolvation energies based on transferring side chains from n-octanol to water that tracks linearly with hydrophobicity.<sup>37</sup> The best performing model trained on the Schrödinger feature set is a gradient boosting regressor model ( $R^2 = 0.48$ ,  $\rho = 0.69$  on the test data) with top features of CDR\_Hydrophobic\_Patch\_Energy\_gt15 and CDR\_Hydrophobic\_Patch\_Energy, (sum of residue contributions to hydrophobic patches). Also included in important features is the Hydrophobicity\_Hopp\_Woods, which is a sequence-based hydrophobicity score calculated using the Hopp-Woods scale which combines a moving average with six-residue long sliding windows and uses a positive numeric value for all charged residues.<sup>38</sup>

The top XGBoost features from the Discovery Studio feature is Aggr Score (Aggregation Score), equivalent to the SAP calculation using the CHARMM force field, which has been demonstrated to correlate with aggregation (Supplementary Figure S3), followed by ddGsolv (difference between solvation energies of water and condensed (crystalline) phase).<sup>39–41</sup> More positive values of ddGsolv correspond to increased solubility.<sup>42</sup> Solubility score is the third highest aggregate feature from the Discovery Studio feature set, which is calculated as a function of net charge, dipole moment, solvation energy, and the aggregation propensity score.<sup>43</sup> The best performing HICRT model using only Discovery Studio features is a random forest regressor model using ten features ( $R^2 = 0.31$ ,  $\rho = 0.57$  on the test data). The top features remain the same as in the aggregated XGBoost importances, with the addition of dipole moment and developability index which is

a feature calculated from the aggregation propensity score minus the weighted squared total of the charge. The aggregation propensity is a structural feature calculated with the CHARMM force field.<sup>39</sup>

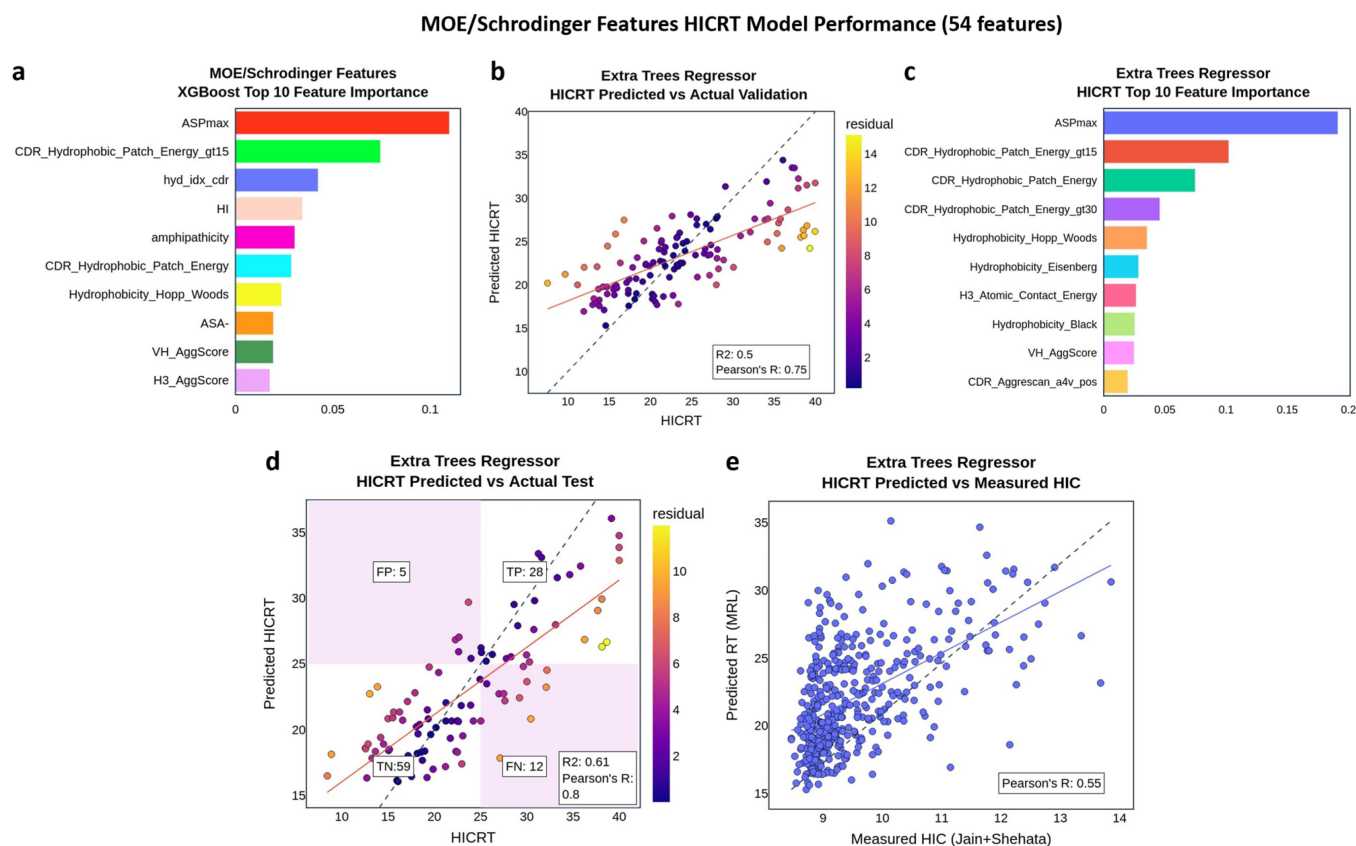
The top models generated with either MOE or Schrödinger features corroborate with each other in that they both rely on similar top features which are a combination of 3D hydrophobic surface patches and sequence-based features based on hydrophobic amino acid scales. However, the overall best predictor of HICRT uses 54 features from both the Schrödinger and MOE feature sets exclusively and outperforms top regression models using all three feature sets (Supplementary Figure S4). This top performing model is an extra trees regressor model ( $R^2 = 0.61$   $\rho = 0.8$  on the test data), and, when used for classification (cutoff of 25 min to delineate high/low HIC behavior), the model correctly identifies 87 of 99 molecules in the test set (28 true positive and 59 true negative) (Figure 3). The most important feature in the top performing extra trees regressor is the ASPmax feature from the MOE feature set, followed by the CDR\_Hydrophobic\_Patch\_Energy\_gt15, and CDR\_Hydrophobic\_Patch\_Energy, and CDR\_Aggrescan\_a4v\_pos, (total CDR calculated Aggrescan values) from the Schrödinger feature set (Figure 3a). Many of the following top features are the sequence-based descriptors using sum of hydropathy scales, including Black and Mould, Hop-Woods, and Eisenberg scales.<sup>44</sup> Other structure-based features of interest are the H3\_Aggscore and VH\_Aggscore (CDRH3 loop and heavy

chain variable region, respectively) from the Schrödinger feature set. AggScore is a 3D structural calculation including hydrophobic, positive, and negative patch calculations, and the structure based patch values are parameterized and smoothed over a sliding window of five residues.<sup>45</sup> AggScore has previously been reported to correlate with mAb HIC profile and clinical-stage mAb therapeutics.<sup>45–48</sup>

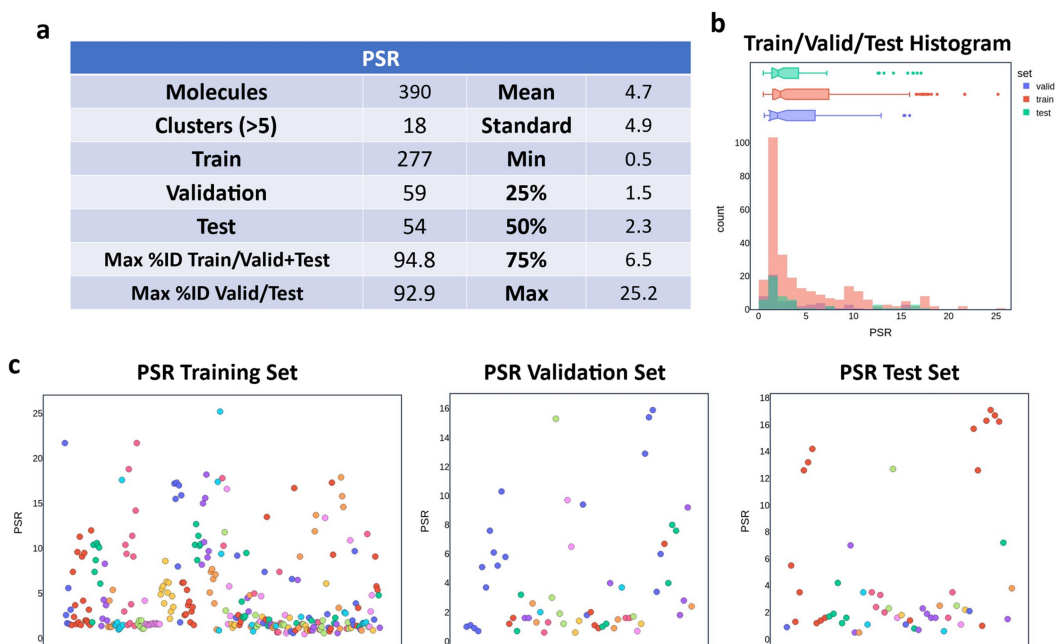
As an additional measurement of the performance of the top model, the extra trees regressor was also applied to prepared sequences from a combined dataset consisting of previously reported HIC values (Figure 3e).<sup>49,50</sup> Following feature generation and removal of sequences with more than 95% ID to our training set, this test set consists of 415 unseen antibody sequences. This combined dataset contains experimental data for both clinically approved and native B-cell derived sequences and is therefore a very suitable test case for our prediction algorithm. Despite differences in the exact HIC assay protocol, there is a decent correlation between the reported values and predicted HICRT from the top performing extra trees regressor (Pearson's  $R = 0.55$ ).

### Poly-specificity reagent binding assay

The PSR data used in the current study consists of 390 IgG molecules (313 IgG1, 48 IgG4 and 29 IgG2) that make up 18 clusters (greater than 5 members) (Figure 4a). As with the HIC-RT dataset, the train, validation, and test sets of 277, 59



**Figure 3.** HICRT dataset – top model regression performance. (a) Top ten XGBoost feature importance by average over 160K models. (b) HICRT top model regression performance on validation data. (c) Top ten feature importance for best performing extra trees regression model. (d) Top HICRT extra trees regression model performance on test set data; shaded areas denote false positive (FP) and false negative (FN) classification performance with HICRT = 25 min as cutoff. (e) Top HICRT extra trees regression model performance on Jain/Shehata HIC data.



**Figure 4.** PSR dataset characteristics and regression performance. (a) Characteristics of the PSR dataset. (b) Histogram of PSR values for training validation and test sets. (c) Scatterplot of the training set (left), validation set (middle) and test set (right). Individual points represent molecules in the dataset colored by assigned sequence identity cluster (families).

and 54 mAbs, respectively, were group shuffle split and then manually curated to ensure several related families with different assay results. Fab identity between pairs of training and validation and test sequences was less than 95% (Figure 4b–c).

Using only MOE features as input to the feature selection module, XGBoost averaged feature importance over 160K models ranked `pro_cdr_net_charge` (CDR net charge) as the most important feature, followed by the `vsurf_ID3`, the hydrophobic interaction energy moment, a feature that describes the vector from the center of mass to the center of the hydrophobic regions for that energy level (Supplementary Figure S5).<sup>51</sup> This feature resembles a dipole moment, and higher values represent a higher concentration of hydrophobic regions on one side or region of the molecule. Additionally, among the top averaged features is the `vsurf_CW1` descriptor, or the capacity factor, which represents the ratio between the hydrophilic regions and the total molecular surface area. The best performing PSR model using only MOE features is an extreme gradient boosting regressor model using 127 features ( $R^2 = 0.32$ ,  $\rho = 0.57$  on the test data). The top features used in this model are `vsa_acid`, the van der Waals total acidic surface area, and `vsurf_ID1` and `vsurf_ID8`, the hydrophobic interaction energy at  $-0.2$  and  $-1.6$  Kcal/mol, respectively. `Vsurf_HB3`, the hydrogen bond donor capacity at  $-1.0$  Kcal/mol, is the fourth top feature and in this context likely corresponds to the electrostatic potential.<sup>52</sup> `FSASA_H` represents the fractional total hydrophobic surface area and is approximately equal in importance to `pro_cdr_net_charge`.

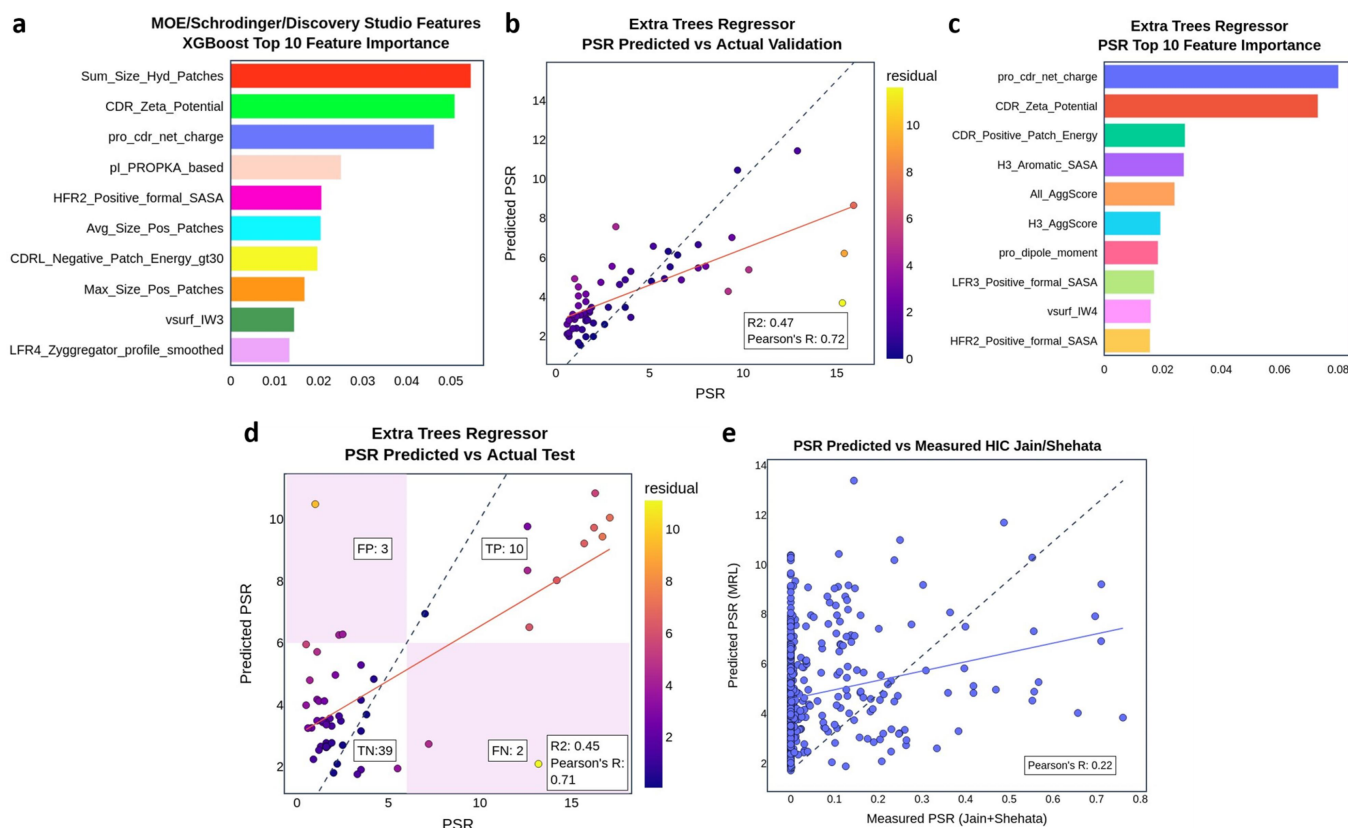
For the Schrödinger feature set, the top averaged XGBoost features from the selection module include `CDR_Zeta_Potential` (calculated zeta potential), `Max_Size_Pos_Patches` (maximum site of positively charged patches, and `Sum_Size_Hyd_Patches` (sum of sizes of hydrophobic patches) (Supplementary Figure S6).

The best performing PSR model using only Schrödinger features is an extra trees regressor model using 79 features ( $R^2 = 0.44$ ,  $\rho = 0.68$  on the test data). The top features in this extra trees regressor model are `CDR_Zeta_Potential`, `CDR_Formal_Charge` (sum of formal charges on the CDR), `CDR_Positive_Patch_Energy` (sum of CDR residue contributions to positively charged patches), `H3_Aromatic_SASA` (sum of surface area of aromatic residues), `CDRH_Zeta_Potential` (zeta potential on the heavy chain), `All_AggScore`, and `H3_AggScore`.

The top aggregate features from XGBoost predictions on the PSR dataset from the Discovery Studio features is the pH of Maximum Stability or pH-dependent relative folding energy, followed by net charge, dipole moment and `ddGsolv`, described above (Supplementary Figure S7).<sup>53,54</sup> The best performing model on the PSR dataset using only Discovery Studio features is a ridge regression model ( $R^2 = 0.24$ ,  $\rho = 0.51$  on the test data) using eight features. The features in descending order of importance, evaluated by SFS are Positive Aggr Score, or sum of the positive surface aggregation propensity (SAP), followed by net charge, dipole moment, and positive QMAP score, which is equivalent to the sum of the positive contributions to the surface charge map (SCM) for each atom also used in the calculation of the SAP technique.<sup>55</sup>

The overall best predictor of PSR from a binary classification standpoint is an extra trees regressor model which uses 91 features from the MOE, Schrödinger, and Discovery Studio feature sets ( $R^2 = 0.45$ ,  $\rho = 0.71$  on the test data) (Figure 5). Using the regression model for classification (cutoff of 6 to delineate good/bad PSR behavior), the model correctly identifies 49 of 54 molecules in the test set (10 true positive and 39 true negative). The most important feature in the top performing extra trees regressor is the `pro_cdr_net_charge` from the

## MOE/Schrodinger/Discovery Studio Features PSR Model Performance (91 features)



**Figure 5.** PSR dataset – top model regression performance. (a) Top ten XGBoost feature importance by average over 160K models. (b) PSR top model regression performance on validation data. (c) Top ten feature importance for best performing extra trees regression model. (d) Top PSR extra trees regression model performance on test set data; shaded areas denote false positive (FP) and false negative (FN) classification performance with PSR = 6 as cutoff. (e) Top PSR extra trees regression model performance on Jain/Shehata PSR data.

MOE feature set followed by CDR\_Zeta\_Potential from the Schrödinger feature set. In confirmation of the analysis using the individual feature sets, these two top features have an outsized impact on the regressor importances. CDR\_Positive\_Patch\_Energy and H3\_Aromatic\_SASA are the third and fourth most important features, followed by All\_AggScore and H3\_AggScore. Unlike the HICRT prediction model, the top performing PSR prediction model trained on in-house data did not correlate well with the reported values from the Jain/Shehata dataset. This is not entirely unexpected as the two PSR methods employed vary significantly, and PSR is known to be sensitive by even relatively minor differences in reagent preparation and assay methods.

The overall best predictor of PSR from a correlative perspective uses 68 features from the MOE and Schrödinger feature sets alone ( $R^2 = 0.5$ ,  $\rho = 0.72$  on the test data) (Supplementary Figure S8). While the regression metrics reflect better performance, applying this model to a classification task results in more misclassified molecules with 46 of 54 molecules correctly classified (8 true positive and 38 true negative). As our use case for predictive models is more often to *a priori* flag challenging mAb molecules over precise accuracy, we have selected the model which uses all three feature sets as our top model for production use.

## Discussion

### Studies in the literature are largely incomparable

Despite the overwhelming need and potential value of accurate *in silico* predictors for biophysical assays and developability properties, the field has been challenged by a lack of high-quality datasets. Typical analyses often require recombinant protein quantities on the order of milligrams, and currently even highly automated expression and purification workflows cannot readily generate datasets needed for naïve deep learning techniques. In addition to compatibility with contemporary datasets, another advantage to machine learning with descriptors calculated from physicochemical properties is that of interpretability, which can aid in guiding the researcher toward future research. However, the possible *in silico* features that can be computed for any given protein are possibly limitless, with the result being that the data available are extremely wide (many features), but not especially deep (many samples) by modern data science standards. For these reasons, most reports in the literature are conducted with a limited group of often related sequences and use only a small group of features, often from a single algorithm or software package. Moreover, datasets are curated differently by individual research groups and often have

incongruent assays, making generalizability of biophysical property predictors an ongoing challenge.

### **Creation and proper curation of dataset splits is important for generalizability**

Here, we have collected and curated HIC-RT and PSR datasets on IgG mAbs larger than any other published datasets known to us.<sup>1,2,49,50</sup> We used these data in conjunction with the calculated features from the most popular software packages to implement several notable technical details, starting with the careful use of < 95% sequence identity between train and testing datasets. In our experience, unless the protein molecules are generated explicitly using sparse sampling, performing simple random train/validation splits results in overfitted models which do not perform well in previously unseen molecule applications and are therefore poorly generalizable. To further illustrate this point, we performed an identical workflow on our HIC-RT data, which consists of train/validation splits derived with clustered versus unclustered data, using a typical stratified shuffle split. In this experiment, despite similar performance on the validation set, performance on the test set dropped significantly ( $R^2 = 0.41$ , Pearson's  $R = 0.65$ ) (data not shown).

### **Predictive models constructed with features calculated on independently produced AbodyBuilder2 scFv structures are comparable to computationally expensive Fab structures**

Initially, our analysis was conducted using modeled Fab structures generated via the native homology modeling workflows particular to each software package MOE, Schrödinger (Bioluminate), and Discovery Studio. For the MOE software, this also includes stochastic titration of 100 conformations and averaging ensembled features as previously described.<sup>56</sup> However, these intrinsic modeling workflows are both computationally expensive and cumbersome (multiple models must be stored), and they also introduce variability into the feature analysis through differences in structural models. The current workflow of using AbodyBuilder2 to first generate single-chain variable fragment (scFv) structures, followed by structure preparation and protonation in each software package prior to feature calculation, is both more computationally efficient and better suited to large scale deployment to discovery end users. Importantly, we found no appreciable differences in predictive model quality by using either the full Fab structure or more computationally expensive structure generation modules (data not shown).

### **Comprehensive screening of machine learning algorithms provides the needed consistency for accurate comparison of features**

We constructed an automatic workflow using XGBoost as the primary method for feature selection and PyCaret to search and train for the best performing model on the

validation set. In our implementation, although there is no data leakage between the training and validation set, the validation set is used exhaustively to select the best model and hyperparameters of feature selection and number of features.

The XGBoost and PyCaret workflows each generate internal splits using only the training data for model generation and the validation data remains segregated from these processes and only accessed in a predictive capacity. We perform the prediction on the validation data 250 thousand times and take the top performing model(s). This exhaustive search is necessary due to the nature of the uneven clustered (and skewed) data collected from pipeline bioassay sources, which gives rise to highly variable performance when using statistical resampling methods such as cross validation. In contrast, because of the exhaustive predictive evaluation on the validation set, the workflow described here produces consistent results.

This consistency overcomes a significant limitation of the data and is crucial for *in silico* feature ranking and comparison. It allows decisions to be made on any available feature set(s) with regards to implementation for in-house machine learning pipelines. The exhaustive nature of the search ensures that the results are reproducible, and feature sets with better performance using this workflow can be implemented with confidence. The caveat to this method is that a separate test set of limited identity (<95%) needs to be generated initially and held entirely separate until the final one-time prediction from the top performing validation models. This necessity even further reduces the available datapoints for model training.

In many cases it is not possible, or efficient, to use many different and disparate software packages to calculate feature sets which give rise to the best predictive models for a given developability assay. We aimed to illustrate in this study how machine learning techniques, when combined with sufficiently powered datasets, can help to identify *in silico* features with the most predictive capacity. Furthermore, this workflow is designed specifically with machine learning operations deployment and extensibility in mind. To explain, as data collection streams increase the size of the datasets for HIC, PSR, and other developability endpoints, this automated workflow is designed to process and rank iterative model development. We expect to see model improvement as dataset sizes (and especially outliers) increase. Of potentially greater importance is the ability of this workflow to accept any number of feature sets and rank the importance and contribution of any descriptor sets in a head-to-head fashion. Advancements in the field of protein descriptors are currently flourishing and include protein language embeddings, surface patch descriptors, or confidence metrics from deep learning modeling.<sup>7,12,15,57-59</sup> The workflow presented here is designed for a facile and robust evaluation of any new calculatable feature ability to predict real therapeutic molecule assay endpoints.

There are also likely limitations and/or possible improvements to the methods outlined here. For example, the reliance on XGBoost for feature selection likely biases the models toward ensembling or bagging algorithms and the brute force search of hyperparameters could likely be optimized by more elegant approaches such as Bayesian optimization. Exploration



of other hyperparameters, such as number of feature selection cycles or methods of selecting feature importance, may also improve our methods.

## Conclusion

Recent advancements in machine learning and artificial intelligence have been greatly facilitated by digitally native datasets of substantial size and breadth to generate models capable of incredible predictive accuracy. Despite breakthroughs in the areas of *de novo* protein design and structure prediction, *in silico* developability prediction of biological therapeutics remains far from being a solved problem. Challenges arise from difficult and underpowered datasets which are expensive and time consuming to produce, but also from an overabundance of possible descriptors and calculatable features. We believe that, by implementing methods such as are demonstrated here, the field can be best positioned to take advantage both of new experimental datasets imminently being collected, as well as the next new breakthrough in feature descriptors, yet to be discovered.

## Materials and methods

### Sequence similarity and clustering

Fab sequences are first concatenated (heavy\_chain, light chain) and a pairwise mutation matrix is generated using the BioPython Align.PairwiseAligner function. Hierarchical clustering with Ward linkage is generated using the Scipy linkage function and returning the flat clusters with a pairwise distance of 40 maximum mutations.<sup>29,60</sup> Final split percent identity is verified by using the Clustal Omega package and generating a full identity matrix with the – percent-id option.<sup>30</sup>

### In silico feature calculations

#### AbodyBuilder2 structure generation

scFv structures were generated for each molecule in the dataset using the AbodyBuilder2 predictor from the Immunebuilder package through the python API and generated with the Aho numbering scheme.<sup>56</sup>

#### MOE

A custom svl script was written to pipeline structure preparation (StructurePreparation, \_LIGX\_Execute) and protonation generation (Protonate\_3D, pH 7.4) before feature calculation. Features were generated by collecting descriptors via the QuaSAR\_descriptorMDB function and the molecular contact surface analysis extension of BioMOE; in total, 269 total MOE features were collected for each molecule.

#### Schrödinger bioluminate

Antibody scFv models were prepared through the protein preparation workflow (prepwizard2\_driver.py) at pH 7.4. Protein descriptors were likewise calculated using the automated protein descriptors script (calc\_protein\_descriptors.py). For each input sequence, Schrödinger Bioluminate calculates 907 descriptors.

### Discovery studio

Protein formulation properties were calculated with custom script from BIOVIA Support using the Protonate (pH 7.4), Calculate Protein Formulation Properties, and Aggregation Scores workflows. Discovery Studio produced 16 individual columns for each molecule using this workflow.

### XGBoost feature selection and pycaret model generation

For each input set of features, the training set was stratified group shuffle split 32 times and separately fit to an XGBoost regression model.<sup>26</sup> The individual feature set importance (importance type = gain) was ranked, collected, and reduced to X total selected features, with X being a critical hyperparameter (X = 5–150, randomly selected). These reduced feature sets were then combined (X \* number of input feature sets) for a subsequent 32 rounds of combined XGBoost regression to result in X total selected features. The top features from the XGBoost rounds were submitted to a PyCaret regression or classification module. PyCaret trains and compares models from the Scikit-learn package (19 different models for regression) and for 10 random seeds, the top 5 scoring models from each training run were tested for predictions on the validation set (50 total models).<sup>61</sup> For each tested set(s) of features, 5000 total end-to-end runs were performed, leading to a total of 250,000 models sampled per output. Top performing models were chosen by average accuracy on the validation set (average of 10 random seeds) and were recapitulated in a separate notebook and evaluated individually on the test set.

### Expression and purification (gene synthesis, transfection, titer estimation, protein a purification)

Transient transfections were done in TubeSpin® bioreactors (TPP Techno Plastic Products AG) using the ExpiCHO Expression System (Thermo Fisher Scientific, Waltham, MA) for the protein production in this study according to the manufacturer's protocol. Briefly, the cells were grown and maintained in ExpiCHO Expression Medium (Thermo Fisher Scientific) and seeded in 10mls of media at  $6 \times 10^6$  cells/mL on the day of transfection. Complexes were formed with 8ug of DNA and 32ul of Expifectamine in OptiPRO™ SFM and incubated for 1 min followed by addition to the cells. The transfected cultures were grown at 37°C, 5% CO<sub>2</sub>, 80% humidity, and 300rpm rotation in a Multitron incubator (Infors HT, Basel, Switzerland) and then shifted to 32°C 24 hr post transfection and were fed with feed and enhancer on days 1 and 5. The cultures were harvested on day 7, the cells were pelleted by centrifugation, and the supernatant was passed through a 0.2 micron filter. The clarified cell culture supernatants were loaded onto Tecan Freedom EVO 200 (Tecan Life Sciences, Männedorf, Switzerland) for antibody purification utilizing miniature columns manufactured by Repligen (Waltham, MA) and packed with MabSelect™ SuRe™ LX (GE Healthcare Life Sciences, Pittsburgh PA). The antibodies were eluted with 20 mM sodium acetate at pH3.5 and immediately neutralized with 0.333 M Tris, 1 M sodium acetate, pH 8.0, and buffer exchanged into 20 mM sodium acetate pH 5.5.

### Hydrophobic chromatography

To determine the hydrophobicity of a given mAb using HIC, 50 µg of sample at 0.5–1 mg/ml were mixed 1/1 (v/v) with a 100 mM sodium phosphate, 2 M ammonium sulfate pH 7.0 buffer solution. Prepared samples are subsequently filtered through a 0.22 µm PVDF membrane prior to loading 60 µl on a Dionex Pro Pac HIC-10 column equilibrated in 100 mM sodium phosphate, 1 M Ammonium Sulfate pH 7.0 (mobile phase A). The samples are eluted using an inverted gradient from mobile phase A to 100 mM sodium phosphate pH 7.0 (mobile phase B). The elution is followed by recording the A280 nm as a function of time, and the data are then exported and analyzed using the Empower software. The retention time of each sample is compared to a reference and is characteristic of the mAb hydrophobicity with longer elution times correlating with higher degree of hydrophobicity.

### Poly-specificity reagent binding assay

The cytosolic preparation is described as SCP (Separated Cytosolic Proteins/Preparation), while the membrane-enriched preparation is labeled as SMP (Solubilized Membrane Proteins/Preparation). The generation of SCPs and SMPs is described by Xu et al.<sup>62</sup> It is important to note, before solubilization (i.e., before detergent addition), the respective cellular fractions are randomly biotinylated; stocks are often at ~ 0.5–1.5 mg/mL of protein. Post-solubilization, the b-SMP stock contains 1% DDM and ~ 0.5–1.5 mg/mL proteins. In reference to developability, binding to non-target PSRs (e.g., generated from parental CHO cells) has been shown to be indicative of compromised binding specificity of antibodies tested. Both SCPs and SMPs can be utilized for this purpose. If the goal is to examine target-specific binding, SMPs derived from transfected mammalian cells are used. This protocol describes the binding assessment method for both SCPs and SMPs, collectively described as PSR throughout the document. PSR binding can be measured with IgG presented on polystyrene beads coated with polyclonal goat anti-human IgG. Polyspecificity is measured using an Attune NxT Flow Cytometer.

### Abbreviations

CDR	complementarity-determining region
Cryo-EM	cryogenic electron microscopy
XGBoost	extreme gradient boosting
HIC	hydrophobic interaction chromatography
HIC-RT	hydrophobic interaction chromatography retention time
MOE	Molecular Operating Environment
mAbs	monoclonal antibodies
PSR	poly-specificity reagent
SFS	sequential feature selection
Fv	fragment variable
SAP	surface aggregation propensity

### Acknowledgments

The authors would like to thank members of the Protein Sciences department in Merck and Co. SSF, Discovery Biologics, for designing, generating, and characterizing molecules used to generate the models used in this publication. We would also like to thank Will Long for assistance in writing and troubleshooting SVL scripts, Jodi Shalusky for providing

Discovery Studio Perl scripts for processing of sequences, and Galen Wo for assistance with Spotfire assisted collection of assay data from disparate sources.

### Disclosure statement

During the execution of this work, all authors were employees of subsidiaries of Merck & Co., Inc., Kenilworth, NJ, USA and stock-holders of Merck & Co., Inc., Kenilworth, NJ, USA.

### Funding

The author(s) reported there is no funding associated with the work featured in this article.

### ORCID

Andrew B. Waight  <http://orcid.org/0000-0002-4110-1452>

### References

- Bailly M, Mieczkowski C, Juan V, Metwally E, Tomazela D, Baker J, Uchida M, Kofman E, Raoufi F, Motlagh S, et al. Predicting antibody developability Profiles through early stage Discovery screening. *MAbs*. 2020;12(1):1743053. doi:10.1080/19420862.2020.1743053.
- Raybould MIJ, Marks C, Krawczyk K, Taddese B, Nowak J, Lewis AP, Bujotzek A, Shi J, Deane CM. Five computational developability guidelines for therapeutic antibody profiling. *Proc Natl Acad Sci USA*. 2019;116(10):4025–30. doi:10.1073/pnas.1810576116.
- Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemp Clin Trials Commun*. 2018;11:156–64. doi:10.1016/j.conctc.2018.08.001.
- Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*. 2004;3(8):711–16. doi:10.1038/nrd1470.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583–89. doi:10.1038/s41586-021-03819-2.
- Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, et al. Accurate prediction of protein structures and interactions using a 3-track neural network. *Sci*. 2021;373(6557):871–76. doi:10.1126/science.abj8754.
- Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Sci*. 2023;379(6637):1123–30. doi:10.1126/science.ade2574.
- Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, Craig PA, Crichton GV, Dalenberg K, Duarte JM, et al. RCSB protein data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res*. 2023;51(D1):D488–508. doi:10.1093/nar/gkac1077.
- UniProt Consortium T, Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49(D1):D480–9. doi:10.1093/nar/gkaa1100.
- Olsen TH, Boyles F, Deane CM. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired

- and paired antibody sequences. *Protein Sci.* 2022;31(1):141–46. doi:10.1002/pro.4205.
11. Schneider C, Raybould MIJ, Deane CM. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Res.* 2022;50(D1):D1368–72. doi:10.1093/nar/gkab1050.
  12. Leem J, Mitchell LS, Farmery JHR, Barton J, Galson JD. Deciphering the language of antibodies using self-supervised learning. *Patterns.* 2022;3(7):100513. doi:10.1016/j.patter.2022.100513.
  13. Khass M, Vale AM, Burrows PD, Schroeder HW. The sequences encoded by immunoglobulin diversity (DH) gene segments play key roles in controlling B-cell development, antigen-binding site diversity, and antibody production. *Immunol Rev.* 2018;284(1):106–19. doi:10.1111/imr.12669.
  14. Prihoda D, Maamary J, Waight A, Juan V, Fayadat-Dilman L, Svozil D, Bitton DA. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *MAbs.* 2022;14:2020203. doi:10.1080/19420862.2021.2020203.
  15. Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. *Bioinforma Adv.* 2022;2(1):vbac046. doi:10.1093/bioadv/vbac046.
  16. Ruffolo JA, Gray JJ. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Biophys J.* 2022;121(3):155a–56. doi:10.1016/j.bpj.2021.11.1942.
  17. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods.* 2022;19(6):679–82. doi:10.1038/s41592-022-01488-1.
  18. Almagro JC, Teplyakov A, Luo J, Sweet RW, Kodangattil S, Hernandez-Guzman F, Gilliland GL. Second antibody modeling assessment (AMA-II). *Proteins Struct Funct Bioinforma.* 2014;82(8):1553–62. doi:10.1002/prot.24567.
  19. Abanades B, Georges G, Bujotzek A, Deane CM. Ablooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics.* 2022;38(7):4.
  20. Abanades B, Wong WK, Boyles F, Georges G, Bujotzek A, Deane CM. ImmuneBuilder: Deep-learning models for predicting the structures of immune proteins. *Commun Biol.* 2023;6:1–8. doi:10.1038/s42003-023-04927-7.
  21. Marks C, Deane CM. Antibody H3 structure prediction. *Comput Struct Biotechnol J.* 2017;15:222–31. doi:10.1016/j.csbj.2017.01.010.
  22. Fernández-Quintero ML, Kokot J, Waibl F, Fischer A-L, Quoika PK, Deane CM, Liedl KR. Challenges in antibody structure prediction. *MAbs.* 2023;15(1):2175319. doi:10.1080/19420862.2023.2175319.
  23. Molecular Operating Environment (MOE), 2022. 02 Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2023.
  24. Schrödinger Release 2022-3: BioLuminate, Schrödinger, LLC, New York, NY, 2021.
  25. I. BIOVIA, Dassault Systèmes, Discovery Studio, 2021, San Diego: Dassault Systèmes, 2021.
  26. Chen T, Guestrin C. Xgboost: A Scalable Tree boosting System [Internet]. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 [cited 2023 Jan 4]. page 785–94. <http://arxiv.org/abs/1603.02754>
  27. Introduction to Boosted trees — xgboost 2.0.0-dev documentation [Internet]. [accessed 2023 Feb 13]: <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>
  28. Moez A. PyCaret: An open source, low-code machine learning library in Python [Internet]. 2020 [cited 2023 Jan 4]; Available from: <https://www.pycaret.org>
  29. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 2009;25(11):1422–23. doi:10.1093/bioinformatics/btp163.
  30. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 2011;7(1):539. doi:10.1038/msb.2011.75.
  31. Black SD, Mould DR. Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Anal Biochem.* 1991;193(1):72–82. doi:10.1016/0003-2697(91)90045-U.
  32. Ferri FJ, Pudil P, Hatef M, Kittler J. Comparative study of techniques for large-scale feature selection\* \*This work was supported by a SERC grant GR/E 97549. The first author was also supported by a FPI grant from the Spanish MEC, PF92 73546684. In: Gelsema E Kanal L editors. Machine intelligence and pattern recognition. North-Holland; 1994. pp. 403–13. [accessed 2023 Jul 6]. <https://www.sciencedirect.com/science/article/pii/B9780444818928500407>.
  33. Jetha A, Thorsteinson N, Jmeian Y, Jeganathan A, Giblin P, Fransson J. Homology modeling and structure-based design improve hydrophobic interaction chromatography behavior of integrin binding antibodies. *MAbs.* 2018;10(6):890–900. doi:10.1080/19420862.2018.1475871.
  34. Salgado JC, Rapaport I, Asenjo JA. Predicting the behaviour of proteins in hydrophobic interaction chromatography. 1: Using the hydrophobic imbalance (HI) to describe their surface amino acid distribution. *J Chromatogr A.* 2006;1107(1–2):110–19. doi:10.1016/j.chroma.2005.12.032.
  35. de Groot NS, Pallarés I, Avilés FX, Vendrell J, Ventura S. Prediction of “hot spots” of aggregation in disease-linked polypeptides. *BMC Struct Biol.* 2005;5(1):18. doi:10.1186/1472-6807-5-18.
  36. Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S. AGGRESAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinform.* 2007;8(1):65. doi:10.1186/1471-2105-8-65.
  37. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol.* 1997;267(3):707–26. doi:10.1006/jmbi.1996.0859.
  38. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci U S A.* 1981;78(6):3824–28. doi:10.1073/pnas.78.6.3824.
  39. Vanommeslaeghe K, MacKerell AD. CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochim Biophys Acta.* 2015;1850(5):861–71. doi:10.1016/j.bbagen.2014.08.004.
  40. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Prediction of aggregation prone regions of therapeutic proteins. *J Phys Chem B.* 2010;114(19):6614–24. doi:10.1021/jp911706q.
  41. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci USA.* 2009;106(29):11937–42. doi:10.1073/pnas.0904191106.
  42. Tjong H, Zhou H-X. Prediction of protein solubility from calculation of transfer free energy. *Biophys J.* 2008;95(6):2601–09. doi:10.1529/biophysj.107.127746.
  43. Spassov VZ, Kemmish H, Yan L. Two physics-based models for pH-dependent calculations of protein solubility. *Protein Sci.* 2022;31(5):e4299. doi:10.1002/pro.4299.
  44. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol.* 1984;179(1):125–42. doi:10.1016/0022-2836(84)90309-7.
  45. Sankar K, Krystek SR Jr, Carl SM, Day T, Maier JKX. AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches. *Proteins Struct Funct Bioinforma.* 2018;86(11):1147–56. doi:10.1002/prot.25594.
  46. Negron C, Fang J, McPherson MJ, Stine WB, McCluskey AJ. Separating clinical antibodies from repertoire antibodies, a path to in silico developability assessment. *MAbs.* 2022;14(1):2080628. doi:10.1080/19420862.2022.2080628.

47. Sankar K, Trainor K, Blazer LL, Adams JJ, Sidhu SS, Day T, Meiering E, Maier JKK. A descriptor set for quantitative structure-property relationship prediction in Biologics. *Mol Inform.* 2022;41(9):2100240. doi:10.1002/minf.202100240.
48. Trainor K, Gingras Z, Shillingford C, Malakian H, Gosselin M, Lipovšek D, Meiering EM. Ensemble modeling and Intracellular aggregation of an engineered immunoglobulin-like Domain. *J Mol Biol.* 2016;428(6):1365–74. doi:10.1016/j.jmb.2016.02.016.
49. Jain T, Sun T, Durand S, Hall A, Houston NR, Nett JH, Sharkey B, Bobrowicz B, Caffry I, Yu Y, et al. Biophysical properties of the clinical-stage antibody landscape. *Proc Natl Acad Sci U S A.* 2017;114(5):944–49. doi:10.1073/pnas.1616408114.
50. Shehata L, Maurer DP, Wec AZ, Lilov A, Champney E, Sun T, Archambault K, Burnina I, Lynaugh H, Zhi X, et al. Affinity maturation enhances antibody specificity but compromises conformational stability. *Cell Rep.* 2019;28(13):3300–8.e4. doi:10.1016/j.celrep.2019.08.056.
51. Cai Z, Zafferani M, Akande OM, Hargrove AE. Quantitative Structure–Activity Relationship (QSAR) study Predicts Small-molecule binding to RNA structure. *J Med Chem.* 2022;65(10):7262–77. doi:10.1021/acs.jmedchem.2c00254.
52. Platts JA. Theoretical prediction of hydrogen bond donor capacity. *Phys Chem Chem Phys.* 2000;2(5):973–80. doi:10.1039/a908853i.
53. Spassov VZ, Yan L. A pH-dependent computational approach to the effect of mutations on protein stability. *J Comput Chem.* 2016;37(29):2573–87. doi:10.1002/jcc.24482.
54. Spassov VZ, Yan L. A fast and accurate computational approach to protein ionization. *Protein Sci.* 2008;17(11):1955–70. doi:10.1110/ps.036335.108.
55. Agrawal NJ, Helk B, Kumar S, Mody N, Sathish HA, Samra HS, Buck PM, Li L, Trout BL. Computational tool for the early screening of monoclonal antibodies for their viscosities. *MAbs.* 2016;8(1):43–48. doi:10.1080/19420862.2015.1099773.
56. Thorsteinson N, Gunn JR, Kelly K, Long W, Labute P. Structure-based charge calculations for predicting isoelectric point, viscosity, clearance, and profiling antibody therapeutics. *MAbs.* 2021;13(1):1981805. doi:10.1080/19420862.2021.1981805.
57. Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM, Correia BE. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods.* 2020;17(2):184–92. doi:10.1038/s41592-019-0666-6.
58. Gainza P, Wehrle S, Van Hall-Beauvais A, Marchand A, Scheck A, Hartevelde Z, Buckley S, Ni D, Tan S, Sverrisson F, et al. De Novo design of protein interactions with learned surface fingerprints. *Nature.* 2023;617(7959):176–84. doi:10.1038/s41586-023-05993-x.
59. Roney JP, Ovchinnikov S. State-of-the-Art estimation of protein model accuracy using AlphaFold. *Phys Rev Lett.* 2022;129(23):238101. doi:10.1103/PhysRevLett.129.238101.
60. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17(3):261–72. doi:10.1038/s41592-019-0686-2.
61. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Cikit-learn: Machine learning in Python. *Mach learn PYTHON. J Mach Learn Res.* 12:2825–30.
62. Xu Y, Roach W, Sun T, Jain T, Prinz B, Yu T-Y, Torrey J, Thomas J, Bobrowicz P, Vasquez M, et al. Addressing polyspecificity of antibodies selected from an in vitro yeast presentation system: a FACS-based, high-throughput selection and analytical tool. *Protein Eng Des Sel.* 2013;26(10):663–70. doi:10.1093/protein/gzt047.