Molecular Therapy
Methods & Clinical Development
Original Article

Machine Learning Enables Accurate Prediction of Asparagine Deamidation Probability and Rate

Jared A. Delmar,[1] Jihong Wang,[1] Seo Woo Choi,[2] Jason A. Martins,[2] and John P. Mikhail[2]

[1]Analytical Sciences, Biopharmaceutical Development, AstraZeneca, One MedImmune Way, Gaithersburg, MD 20878, USA; [2]David H. Koch School of Chemical Engineering Practice, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

The spontaneous conversion of asparagine residues to aspartic acid or iso-aspartic acid, via deamidation, is a major pathway of protein degradation and is often seriously disruptive to biological systems. Deamidation has been shown to negatively affect both *in vitro* stability and *in vivo* biological function of diverse classes of proteins. During protein therapeutics development, deamidation liabilities that are overlooked necessitate expensive and time-consuming remediation strategies, sometimes leading to termination of the project. In this paper, we apply machine learning to a large (n = 776) liquid chromatography-tandem mass spectrometry (LC-MS/MS) dataset of monoclonal antibody peptides to create computational models for the post-translational modification asparagine deamidation, using the random decision forest method. We show that our categorical model predicts antibody deamidation with nearly 5% increased accuracy and 0.2 MCC over the best currently available models. Surprisingly, our model also paces or outperforms advanced and conventional models on an independent non-antibody dataset. In addition to deamidation probability, we are able to accurately predict deamidation rate (R^2 = 0.963 and Q2 = 0.822), a capability with no peer in current models. This method should enable significant improvement in protein candidate selection, especially in biopharmaceutical development, and can be applied with similar accuracy to enzymes, monoclonal antibodies, next-generation formats, vaccine component antigens, and gene therapy vectors such as adeno-associated virus.

INTRODUCTION

Therapeutic proteins are an important and growing class of drugs that includes peptides, such as insulin; cytokines, like erythropoietin; monoclonal antibodies (mAbs), which are among the most successful cancer therapies; next-generation formats, such as antibody-drug conjugates, bispecific antibodies, and fusion proteins; as well as vaccine components and gene therapy vectors. While small molecules comprise the largest class of new drug approvals, nearly 30% of US Food and Drug Administration (FDA) approvals in 2018 were protein based, up from 26% in 2017. As of the writing of this paper, half of new drugs approved by the FDA in 2019 represent biologics.[1]

Therapeutic proteins offer new mechanisms of action, higher target specificity, lower toxicity, and longer-acting pharmacokinetics, compared to small molecule drugs.[2–4] However, the development of therapeutic proteins poses additional challenges. Not only must the drug be effective, but it must also be "developable," a concept that encompasses many characteristics including high yield and homogeneity from cell culture, high purity drug substance after purification processing, low viscosity and high stability at the high concentrations necessary for drug product, high stability at *in vitro* long-term storage conditions and *in vivo* after administration, high target specificity, and, for antibodies, unimpaired neonatal Fc receptor (FcRn) binding.[2,5] Nearly all of the factors that make a protein drug developable are derived from the amino acid sequence, including site-specific post-translational modifications (PTMs).[6]

In particular, the spontaneous non-enzymatic conversion of asparagine to aspartic acid or iso-aspartic acid via deamidation is a major pathway of protein degradation and is often seriously disruptive to biological systems.[7–9] Deamidation has been shown to negatively affect both *in vitro* stability and *in vivo* biological function of diverse classes of proteins. Deamidation has been reported as a critical quality attribute in many monoclonal antibodies due to its impact on biological activity.[10–13] In one humanized monoclonal immunoglobulin G1 (IgG1) antibody drug, an asparagine in the heavy-chain complementarity determining region 2 (CDR2) loop was found to deamidate *in vivo*, which greatly decreased the drug's efficacy.[14] In another case, heavy-chain CDR deamidation resulted in an almost complete loss of potency and binding activity of a therapeutic monoclonal antibody.[15] In adeno-associated virus, an emerging new vector for gene therapy, extensive capsid deamidation has been observed that impacts transduction and correlates to a loss of vector activity.[16] Deamidation of asparagine residues can also significantly affect immunogenicity and efficacy of protein-based vaccines. Specifically, progress

Received 6 September 2019; accepted 16 September 2019;
https://doi.org/10.1016/j.omtm.2019.09.008.

Correspondence: Jared A. Delmar, Analytical Sciences, Biopharmaceutical Development, AstraZeneca, One MedImmune Way, Gaithersburg, MD 20878, USA.
E-mail: delmarj@medimmune.com

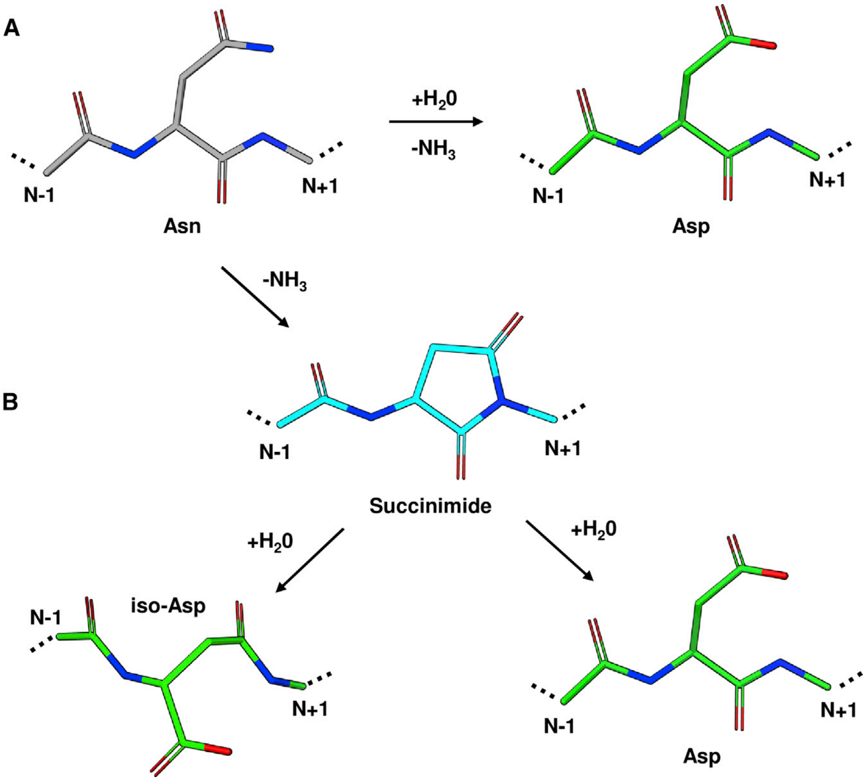


Figure 1. Asparagine Deamidation Reaction

(A and B) Spontaneous degradation of asparagine can occur by (A) direct hydrolysis of the side chain to aspartic acid or (B) via a succinimide intermediate, produced by a nucleophilic attack of the side chain carbonyl by the following (N+1) residue backbone nitrogen, producing either iso-aspartic acid or aspartic acid. Residues are rendered as sticks with Asn, Asp, and iso-Asp, and succinimide carbons colored gray, green, and cyan, respectively (O, red; N, blue).

to develop next-generation anthrax vaccines has been halted by vaccine instability resulting from asparagine deamidation in anthrax protective antigen.[17–21] Even in the nontherapeutic enzyme glucoamylase, used commercially to produce sweeteners and ethanol, asparagine deamidation causes a decrease in enzyme activity and change in thermodynamic stability.[22–24]

Prediction of deamidation liability as early as possible in protein drug development is important because many more candidate drugs are proposed than can be tested. For example, typical antibody generation results in hundreds of candidates, which far exceeds the capacity of a drug development organization.[2,25] Development of a therapeutic protein is so costly in both money and time that, after an initial assessment for screening, only a single candidate is moved forward in most cases.[16,26,27] Sequence liabilities that are not dealt with as early as possible necessitate more expensive and time-consuming remediation strategies later in development[26] and could lead to termination of the project.

Computational tools already exist to facilitate drug candidate screening by the identification of sequence liabilities.[6,28–45] In the case of asparagine deamidation,[36,40–45] currently available tools suffer from several limitations: they provide only a binary (yes, high risk to deamidate; or no, low risk to deamidate) prediction,[36,40,42,43] require an experimental crystal structure,[42–45] or are applicable only to antibody asparagines.[36,40] All offer no[36,40,42] or low accuracy[41,43–45] predictions of deamidation rate. Oversimplified models tend to overestimate the number of deamidation sites greatly, which leads to over-engineering and rejecting good drug candidates too early in development. On the other hand, these models may also overlook asparagines for which deamidation is rarely observed (such as NK or NW sites), which can lead to costly and ineffective drug development.

In this paper, we apply machine learning to a large (n = 776) liquid chromatography-tandem mass spectrometry (LC-MS/MS) dataset of monoclonal antibody peptides to create accurate random decision forest models for the PTM asparagine deamidation.[46] We show that our categorical model predicts antibody deamidation likelihood with nearly 5% increased accuracy and 0.2 MCC over the best currently available models. Surprisingly, our model also paces or outperforms advanced and conventional models on an independent non-antibody dataset, including enzyme, antigen, and viral capsid deamidation sites. In addition to deamidation probability, we are able to accurately predict deamidation rate (R^2 = 0.963 and Q^2 = 0.822), a capability with no peer in current models. We provide evidence that our method can be applied with equal accuracy to predict the likelihood and rate of site-specific asparagine deamidation in any protein of interest.

RESULTS AND DISCUSSION

Feature Selection

Spontaneous deamidation of asparagine to aspartic acid or iso-aspartic acid proceeds by one of two reaction mechanisms (Figure 1). At neutral to basic pH, the most common route is by a nucleophilic attack of the asparagine side chain by the backbone nitrogen of the following (N+1) residue, forming the cyclic succinimide intermediate. Hydrolysis at one of two carbonyls of the succinimide intermediate results in either aspartic acid or iso-aspartic acid. Below pH 5, direct hydrolysis of the asparagine side chain amide to aspartic acid is the dominant reaction.[8,9]

Both mechanisms have been shown to rely on both the primary and three-dimensional (3D) structure, with the residue immediately following the asparagine residue (N+1) having the largest

Table 1. Predictors for Asparagine Deamidation Machine Learning Model

Structural / Chemical Category	Parameter
Primary sequence	pentapeptide deamidation half-life (pphl, days)
	categorical N+1 residue
Backbone orientation	backbone dihedral Phi (φ,°)
	backbone dihedral Psi (Ψ,°)
	nucleophilic C-N attack distance (Å)
Side-chain orientation	side-chain dihedral chi1 ($\chi_{1,}$°)
	side-chain torsion chi2 ($\chi_{2,}$°)
Solvent accessibility	percent solvent accessibility (PSA, %)
	solvent accessible surface area (SASA, Å^2)
Hydrogen bonding	hydrogen bonds to side chain (#)
	Asn local secondary structure (Sheet)
	Asn local secondary structure (Loop)
Machine-learning parameter (for regression model only)	categorical model probability output (%)

12 total parameters were used to inform the categorical machine-learning model to predict deamidation likelihood, comprising 6 general categories. For the regression model to prediction deamidation rate, the output of the categorical model was included as an additional predictor.

effect.[8,9,47–51] The amino side residue preceding the asparagine (N−1) was shown to have little to no effect on deamidation rate.[50,51] Steric hindrance, conformational space, and electrostatic effects introduced by the N+1 residue may all affect the ability of the side chain and/or backbone to align and form the cyclic intermediate.[9] As both reaction mechanisms require hydrolysis to form the final aspartic acid or iso-aspartic acid product, availability of water molecules, or a proton donor, and solvent exposure may directly influence the rate of deamidation.[9] Finally, hydrogen bonding to the side chain or backbone may stabilize asparagine and prevent degradation to aspartic acid.

Taken together, these observations compiled from literature informed 12 total parameters for asparagine deamidation likelihood (Table 1), which our machine-learning models would use to predict deamidation. The N+1 residue was taken into account as both a categorical variable and as the experimental half-life of a synthetic pentapeptide (pentapeptide half-life, pphl) containing the same N−1 and N+1 sequence, measured by Robinson et al.[51] Half-lives were not reported by Robinson et al. for pentapeptides with asparagine in the N+1 position, likely because it is difficult to distinguish between deamidation in the N and N+1 position in this case. Thus, when N+1 = N, we used an average pphl of 5.7 days.

The ability of the side chain and backbone to align to form the cyclic succinimide intermediate was taken into account by the backbone dihedral angles (phi and psi), asparagine side-chain dihedrals (chi1 and chi2), and distance between the side chain carbonyl and backbone nitrogen (nucleophilic attack distance).[42] Solvent accessibility was expressed as a percent of the total residue area (percent solvent accessibility [PSA]), as well as area in Å^2 (solvent accessible surface area [SASA]). Hydrogen bonding to the asparagine side chain was predicted if a potential donor-acceptor pair was found within 3 Å and counted as the total number of predicted bonds (up to 4). Hydrogen bonding to the backbone was accounted for by secondary structure (either sheet or loop). We do not need a third variable for helical secondary structure, because if the local secondary structure is neither a sheet nor loop, an α helix can be assumed.

Because the predictive tool we developed is most valuable during early development or candidate selection of a deamidation-liable protein, when little to no experimental data is available, we relied only on the primary sequence of the proteins to train our models. Two of the deamidation predictors we chose (N+1 residue and pphl) could be gleaned from the primary sequence directly. For the 9 parameters that could not, the structure of each protein was generated by homology modeling and predictors were extracted from the predicted 3D structure. To predict deamidate rate, we used the same 12 predictors for the regression model, with an additional parameter for the output of the first classification model, representing the predicted likelihood of deamidation.

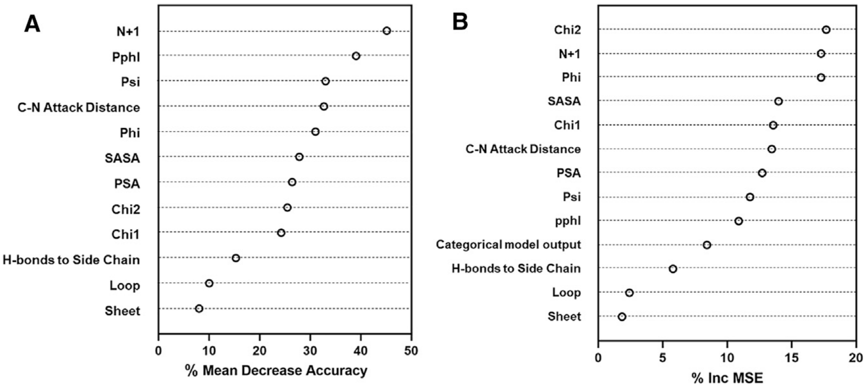


Figure 2. Categorical and Regression Models Predictor Ranking

(A) Importance of each parameter in the categorical model for predicting deamidation probability was measured by the mean decrease in out-of-bag accuracy when that parameter was excluded from the model. (B) Importance of each parameter in the regression model for predicting deamidation half-life was measured by the mean increase in the out-of-bag percent mean squared error (MSE) when that parameter was excluded from the model.

Table 2. Confusion Matrix for Predictions Made by the Categorical Machine Learning Model for Predicting Deamidation Liability on the Training Dataset

Prediction →		
Experiment ↓	Positive	Negative
Positive	137	0
Negative	0	639

Training and Validation Dataset Construction

It is expected that at least 80% of the effort involved in training a machine-learning model is dedicated to data processing and cleaning.[52] Machine learning requires a large amount of data with a high degree of standardization. Thus, we performed a large side-by-side forced degradation study for 34 in-house IgG molecules, comprising a total of 608 asparagine sites that include 304 unique sites. We accelerated deamidation by incubating each molecule at 40°C and pH 8.0 for up to 4 weeks and measured the deamidation abundance at 0-, 2-, and 4-week time points for each site by LC-MS/MS tryptic peptide mapping. The deamidation half-life ($t_{1/2}$) of each asparagine was calculated by least-squares fit of the data to an exponential decay function (Figure S1).

This experimental $t_{1/2}$. was used to train the regression model for half-life prediction. For training a classification machine-learning model to predict the probability of observing deamidation, we set the threshold at 1.0% deamidation per month or $t_{1/2}$ ≈276 weeks. For example, if we observed only a 0.9% increase in deamidation after 4 weeks under stress conditions, we would train the classification model that this site did not deamidate (class "no"). If we measured a 1.1% increase in deamidation after 4 weeks at stress conditions for an asparagine site, this site was reported as class "yes," or deamidated.

Each asparagine located in the variable region of the antibodies in our forced degradation study was included in the training set. Because our training set included six antibody heavy-chain formats[53–56] and two light-chain formats (Figure S2), we also included one copy of each unique constant region asparagine in the training data.

Table 3. Statistics for Predictions Made by the Categorical Machine Learning Model for Predicting Deamidation Liability on the Training Dataset

Statistic	Categorical Model
Accuracy	100.0%
MCC	1.000
Precision	100.0%
Recall	100.0%
Specificity	0.0%
Negative predictive value	0.0%
Miss rate	0.0%
Fallout	0.0%
False discovery rate	100.0%
False omission rate	100.0%

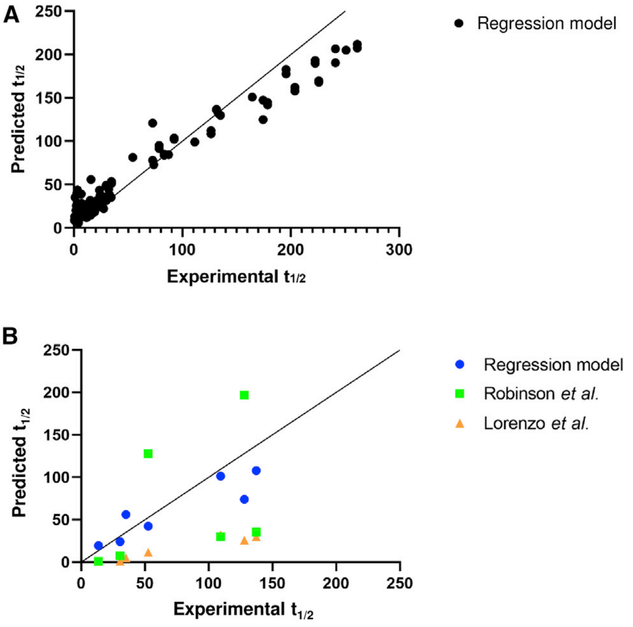


Figure 3. Regression Machine Learning Model for Predicting Deamidation Rate

(A) Predicted half-life ($t_{1/2}$, weeks) was plotted versus the experimental measured $t_{1/2}$ for the training dataset. Individual asparagines are plotted as black circles for and the solid black line indicates where predicted $t_{1/2}$ = experimental $t_{1/2}$. Our regression model predicted the training set with $R^2 = 0.963$. (B) Predicted half-life ($t_{1/2}$, weeks) by our regression model (blue circles), the Robinson et al. model (green squares), and the Lorenzo et al. model (orange triangles) were plotted versus experimentally measured $t_{1/2}$ (weeks) for the validation set. The solid black line indicates where predicted $t_{1/2}$ = experimental $t_{1/2}$. While our regression model predicted the independent validation set with Q2 = 0.822, both the predictions by Robinson et al. and Lorenzo et al. resulted in $Q^2 < 0$.

Initial models trained on our in-house dataset made accurate predictions overall (data not shown). However, they performed relatively poorly on asparagines located in IgG complementarity determining regions (CDRs)—probably because, out of the 304 unique asparagines in our initial dataset, only 25 unique CDR asparagines were found to deamidate. To give our models more examples of CDR deamidation to learn from, we expanded the training dataset to include 33 additional clinical-stage IgG1s (and 39 additional CDR deamidation sites) with deamidation data published by Lu et al.[57] (Table S3). The stress condition used by Lu et al.[57] (40°C and pH 8.5) was similar to our own, and their data was incorporated into the training sets for both our categorical and regression models. Only the final models that including training data from Lu et al. are evaluated here.

To validate our models, we constructed an independent validation dataset with data from 12 additional in-house IgG molecules that were

Table 4. Confusion Matrix for Predictions Made by the Categorical Machine Learning Model for Predicting Deamidation Liability on the Independent Validation Dataset

Prediction →		
Experiment ↓	Positive	Negative
Positive	17	9
Negative	4	165

not contained in the training dataset, stressed at identical conditions to the training set molecules. The observed deamidation frequency in the validation set was consistent with that of the training set (Figure S3). Predictors and LC-MS/MS data for these molecules were added to an independent validation set. This independent validation dataset was supplemented with non-antibody data from both the validation set published by Jia et al.,[42] containing *Aspergillus awamori* glucoamylase,[22–24] anthrax antigen,[17–19,21] and human angionenin RNase,[45] and recent capsid viral protein 3 (VP3) deamidation data published by Giles et al.[16] from adeno-associated virus 8 (AAV8), an emerging vector for gene therapy (Table S3).

Machine-Learning Models for Predicting Deamidation Likelihood and Rate

Both the classification model and regression model were random forest models built in RStudio using the randomForest[46] and caret[58] libraries. The number of trees and number of parameters tried at each split were optimized by hand to minimize the out-of-bag error estimate. Because the output of the classification model is a probability that an asparagine belongs to class "yes," or will deamidate, the probability threshold at which we interpret the prediction as "yes" or "no" was also optimized after model building to maximize the accuracy.

Table 5. Statistics for Predictions Made by the Categorical Machine Learning Model for Predicting Deamidation Liability on the Independent Validation Set

Statistic	Categorical Model
Accuracy	93.3%
MCC	0.691
Precision	81.0%
Recall	65.4%
Specificity	97.6%
Negative predictive value	94.8%
Miss rate	34.6%
Fallout	19.0%
False discovery rate	2.4%
False omission rate	5.2%

Notably, on the independent validation set containing non-antibody proteins our model was able to achieve 93.3% accuracy and a Matthews correlation coefficient (MCC) of 0.691.

Table 6. Confusion Matrix for Predictions Made by our Categorical Model on the Non-mAb Independent Validation Subset

Prediction →		
Experiment ↓	Positive	Negative
Positive	6	1
Negative	4	69

Statistics for the fit to the training set were calculated for both the classification and regression models. Notably, the classification model was able to achieve 100% accuracy on the training set, using 12 parameters to determine whether each of 776 asparagines would deamidate with no mistakes made. The regression model was able to predict $t_{1/2}$ for the 137 deamidated asparagines, 88 of which are unique, in the training set with an R^2 of 0.963. The regression model used the same 12 predictors as the classification model, as well as the prediction output from the classification model, for a total of 13 parameters (Table 1).

The top two predictors of deamidation liability, measured by the mean decrease in out-of-bag accuracy when that parameter is excluded from the categorical model, were the N+1 categorical variable and the pphl (Figure 2A). This is consistent with the literature and it is well accepted that the N+1 residue has the greatest effect on the deamidation liability of all studied parameters.[8,9,47–51] Even a conventional one-parameter method using only the N+1 residue is competitive with advanced techniques (Tables 11 and 16). The next three most important parameters were related to the backbone alignment (psi and phi dihedral angles and nucleophilic attack distance), followed by solvent accessibility (SASA and PSA), side-chain alignment (chi1 and chi2 dihedral angles), and hydrogen bonding (side-chain hydrogen bonds and secondary structure). Similarly, Jia et al.[42] found that tracking hydrogen bonding, secondary structure in particular, did not improve their asparagine deamidation prediction.

To predict $t_{1/2}$, the side-chain orientation was among the most important variables, measured by the increase in the percentage of mean squared error when that parameter is excluded from the regression model (chi2 and chi1 were 1st and 5th most important, respectively). Chi2 is the closest angle to the carbonyl involved in succinimide formation and Jia et al.[42] also observed that chi2 was more important than chi1 in their model.[42] Similar to the categorical model, the N+1 variable was among the best predictors, followed by the backbone dihedral angle phi (Figure 2B). Interestingly, pphl, the second-most important predictor for the categorical model, was only the 9th most important predictor

Table 7. Confusion Matrix for Predictions Made by the Conventional NG/NN/NS Model on the Non-mAb Independent Validation Subset

Prediction →		
Experiment ↓	Positive	Negative
Positive	7	0
Negative	9	64

Table 8. Confusion Matrix for Predictions Made by the Lorenzo et al. Model on the Non-mAb Independent Validation Subset

Prediction →		
Experiment ↓	Positive	Negative
Positive	6	1
Negative	4	69

Table 10. Confusion Matrix for Predictions Made by the Jia et al. Model on the Non-mAb Independent Validation Subset

Prediction →		
Experiment ↓	Positive	Negative
Positive	5	2
Negative	2	71

for the regression model. It is possible that deamidation rate is strongly influenced by structural effects absent from the Robinson et al.[47] pentapeptides. Indeed, we have observed in both our training and validation sets that $t_{1/2}$ defies the hierarchy set forth by pphl. As in the categorical model, parameters that captured hydrogen bonding to the side chain and secondary structure were the least useful in predicting deamidation rate.

To test whether the models were capable of accurately predicting deamidation likelihood and rate, both classification and regression models were applied to the validation set comprised of 12 unseen antibodies held from the training set and 4 non-mAb proteins (Tables 2, 3, 4, and 5; Figure 3). Surprisingly, the categorical model performed with high accuracy for both mAb and non-mAb proteins, predicting the validation set with more than 93% accuracy and a Matthews correlation coefficient (MCC) of 0.691 even though it had never seen a non-mAb deamidation site before (Table 5). Because deamidation data for the non-mAbs came from literature and were either not quantitative or not collected under the same conditions as our in-house data, we did not validate the regression model on these molecules. On the validation set molecules with in-house LC-MS/MS data, the regression model was successful, predicting $t_{1/2}$ with Q2 = 0.822 (Figure 3B).

Out of the 195 total asparagines in our validation set, the categorical model made 13 mistakes. It tended toward the conservative side, underpredicting deamidation in 9 cases and only overpredicting 4 asparagines that were not experimentally observed to deamidate. Interestingly, 5 of the 9 underpredicted sites (almost 40% of our total error) came from one molecule: the AAV8 capsid protein VP3.[16] Further, the 5 sites that our model mispredicted were significantly less deamidated than the other 5 deamidation sites observed in VP3, all of which our model correctly predicted (Tables 17 and 20). It was shown by Giles et al.[16] that the 5 low abundance deamidation sites did not respond significantly to incubation at 70°C or pH 10 or changes to the purification process. Thus, it is possible that deamidation would remain unchanged in our milder forced degradation conditions (40°C and pH 8), and we should instead consider these sites as non-liable. Nevertheless, our model outperformed both conventional (Table 18) and advanced (Table 19) predictions of deamidation for AAV8 capsid protein VP3 asparagines.

Table 9. Confusion Matrix for Predictions Made by the Robinson et al. Model on the Non-mAb Independent Validation Subset

Prediction →		
Experiment ↓	Positive	Negative
Positive	7	0
Negative	18	55

Taken together, the high accuracy at which our models were able to predict deamidation in the diverse proteins in our validation set indicates that these models may be generally applied to predict deamidation in any protein of interest.

Comparison with Advanced and Conventional Models

To evaluate the relative performance of our models, we have applied to our validation set as many currently available predictions of deamidation from the literature as possible. These advanced tools include another machine learning model by Jia et al.,[42] a tree-based approach by Yan et al.,[40] and empirical calculations by Robinson et al.[43] and Lorenzo et al.[41]. In addition, we compared all of these approaches to a conventional one-parameter method based on the primary sequence alone. For the conventional method (named NG/NN/NS here), if an asparagine is followed by glycine, asparagine, or serine (N+1 = G, N, or S), then it is considered as liable to deamidate. All deamidation sites in our validation set were NG, NN, or NS motifs,

Table 11. Statistical Comparison of Predictions Made by Our Categorical Model and Other Models on the Independent Non-mAb Validation Subset

Statistic	Categorical Model	NG/NN/NS	Lorenzo et al.[41]	Robinson et al.[43]	Jia et al.[42]
Accuracy	93.8%	88.8%	93.8%	77.5%	95.0%
MCC	0.686	0.619	0.686	0.459	0.687
Precision	60.0%	43.8%	60.0%	28.0%	71.4%
Recall	85.7%	100.0%	85.7%	100.0%	71.4%
Specificity	94.5%	87.7%	94.5%	75.3%	97.3%
Negative predictive value	98.6%	100.0%	98.6%	100.0%	97.3%
Miss rate	14.3%	0.0%	14.3%	0.0%	28.6%
Fallout	40.0%	56.3%	40.0%	72.0%	28.6%
False discovery rate	5.5%	12.3%	5.5%	24.7%	2.7%
False omission rate	1.4%	0.0%	1.4%	0.0%	2.7%

Statistics were calculated for predictions made by all models on the non-mAb validation subset, which was nearly identical to the validation set used by Jia et al.[42]

Table 12. Confusion Matrix for Predictions Made by Our Categorical Model on the mAb-Only Independent Validation Set

Prediction →		
Experiment ↓	Positive	Negative
Positive	6	3
Negative	0	59

Table 14. Confusion Matrix for Predictions Made by the Yan et al. Model on the mAb-Only Independent Validation Set

Prediction →		
Experiment ↓	Positive	Negative
Positive	5	4
Negative	7	52

so the data was particularly amenable to and not biased against this conventional method. Statistical comparison of the conventional model and our categorical model are shown in Table S1. Conversely, our training set contained many non-NG, -NN, and -NS sites (Figure S3) and was poorly predicted by the conventional method (Table S2).

Because the methods by Robinson et al.[43] and Jia et al.[42] require crystal structures, these models could be applied to only a subset of our validation set, which is nearly identical to the validation set used by Jia et al. for their method and comprised of only non-mAbs.[42] We have removed N395 of *A. awamori* glucoamylase (PDB: 3GLY) from this validation subset as Chen et al.[22] showed that this asparagine is N-glycosylated. Of note, all sites with N+1 = N and N+1 = Q are also missing from the non-mAb validation subset and the model by Jia et al.[42] does not provide predictions for them. Finally, two deamidation sites in anthrax protective antigen (N466 and N537) were corrected to match the observations of Verma et al.[19] The individual confusion matrices are shown in Tables 6, 7, 8, 9, and 10, and a statistical comparison of our method performance on this validation subset to the conventional method and those of Lorenzo et al.,[41] Robinson et al.,[43] and Jia et al.[42] is shown in Table 11. Jia et al.[42] had the highest accuracy on this non-mAb validation set, with one less mistake than our model or that of Lorenzo et al.[41] However, their model also performed last in several categories. Our model and that of Lorenzo et al.[41] had the second-best overall accuracy of 93.8%, were not the worst performers in any category, and had nearly identical MCC to that of the Jia et al.[42] model (Table 11).

The tree-based model proposed by Yan et al.[40] is only applicable to IgG mAbs. Thus, in order to compare our methods, we created another subset of our validation set including only IgG mAbs. On this mAb-only independent validation subset, predictions made by our model (Table 12) were compared against those made by the models of Yan et al.[40] (Table 13), Lorenzo et al.[41] (Table 14), and the conventional one-parameter NG/NN/NS model (Table 15). Again, our categorical model was not the worst performer in any statistic and had the best MCC (0.796) and accuracy (95.6%) at predicting the mAb-only dataset (Table 16).

Unfortunately, we were not able to make a significant comparison to the tree-based method proposed by Sydow et al.[36] Their method is restricted to only the CDR of antibodies.

As of the writing of this paper, we were only able to find two methods in the literature for the prediction of deamidation half-life: by Robinson et al.[43] and Lorenzo et al.[41] Deamidation half-life is both temperature and pH dependent and each model is specific to one condition. Specifically, the Robinson et al.[43] model predicts $t_{1/2}$ for proteins at 37°C and pH 7.4 and Lorenzo et al.[41] at slightly basic pH and up to 40°C. In our experience, these conditions are similar enough to our own (40°C and pH 8.0) to make a direct comparison.

Out of the 26 unique deamidation sites in our validation set, only 7 had available $t_{1/2}$ measurements, all of which were calculated from LC-MS/MS data collected in-house. Thus, we applied each model to these 7 sites for comparison of predictive accuracy. Both the Robinson et al.[43] and Lorenzo et al.[41] models predicted values of $t_{1/2}$ that disagreed enough with the experimental values to produce a Q2 < 0. Our model achieved a Q^2 of 0.822 (Figure 3B). Both the Robinson et al.[43] and Lorenzo et al.[41] methods rely heavily on the pphl as the basis for their half-life prediction, while our model ranked pphl as one of the least useful parameters for prediction $t_{1/2}$ (Figure 2B), which might help to explain the discrepancy in results. While the Lorenzo et al. model tended to underpredict deamidation rate in this independent validation set, both the Lorenzo et al. and NG/NN/NS models overpredicted the number of liable sites in the AAV8 capsid protein VP3 (Tables 17, 18, 19, 20).

Conclusions

We have constructed both a categorical model for predicting whether or not an asparagine is liable for deamidation, and a regression model for determining the rate at which a predicted site deamidates. Both

Table 13. Confusion Matrix for Predictions Made by the Conventional NG/NN/NS Model on the mAb-Only Independent Validation Set

Prediction →		
Experiment ↓	Positive	Negative
Positive	9	0
Negative	9	50

Table 15. Confusion Matrix for Predictions Made by the Lorenzo et al. Model on the mAb-Only Independent Validation Set

Prediction →		
Experiment ↓	Positive	Negative
Positive	6	3
Negative	3	56

Table 16. Statistical Comparison of Predictions Made by Our Categorical Model and Other Models on the Independent mAb-Only Validation Subset

Statistic	Categorical Model	NG/NN/NS	Yan et al.[40]	Lorenzo et al.[41]
Accuracy	95.6%	86.8%	83.8%	91.2%
MCC	0.796	0.651	0.388	0.616
Precision	100.0%	50.0%	41.7%	66.7%
Recall	66.7%	100.0%	55.6%	66.7%
Specificity	100.0%	84.7%	88.1%	94.9%
Negative predictive value	95.2%	100.0%	92.9%	94.9%
Miss rate	33.3%	0.0%	44.4%	33.3%
Fallout	0.0%	50.0%	58.3%	33.3%
False discovery rate	0.0%	15.3%	11.9%	5.1%
False omission rate	4.8%	0.0%	7.1%	5.1%

outperform or pace currently available models based on predictions made on independent validation sets.

Although both models were trained on only mAb deamidation data, we found that they applied with similar accuracy to non-mAb molecules in our validation set, including enzyme, antigen, and viral capsid deamidation sites. In contrast to other methods, ours do not require crystallographic 3D coordinates and are not protein class specific. Rather, the structural information used by our models to predict deamidation is drawn from homology models. Thus, they are applicable to any protein for which a similar protein's structure is available in the PDB.

It is our hope that with more data and increasingly accurate and interpretable models, a fundamental understanding of protein degradation, including deamidation, will be attained, leading to more and better protein-based therapies.

MATERIALS AND METHODS

3D Model Building and Parameter Extraction

For AstraZeneca in-house molecules, full-length homology models were built using Schrödinger BioLuminate.[59] Briefly, the most similar crystal structure from the PDB, by sequence, was first identified by basic local alignment search tool (BLAST).[60] This structure and an in-house constant region template were used as scaffolds for the full-length structure. The Protein Preparation Wizard tool was used to add hydrogens, assign bond orders, remove solvent molecules, optimize H-bond assignments, and perform restrained energy minimization. Molecules from the study by Lu et al.[57] were modeled similarly; however, only the Fv structure was generated. Predictors of asparagine deamidation were extracted from the 3D homology models within Schrödinger via python script.

Table 17. Confusion Matrix for Predictions Made by our Categorical Model on the AAV8 Capsid Protein VP3

Prediction →		
Experiment ↓	Positive	Negative
Positive	5	5
Negative	0	37

Table 18. Confusion Matrix for Predictions Made by the Conventional NG/NN/NS Model on the AAV8 Capsid Protein VP3

Prediction →		
Experiment ↓	Positive	Negative
Positive	9	1
Negative	8	29

Generation of Deamidated IgGs

For IgG deamidation data generated in-house, samples at 10 mg/mL in 50 mM Tris pH 8.0 were incubated at 40°C for 2-week and 4-week time points. Reactants were stored at −80°C prior to analysis by LC-MS/MS.

LC-MS/MS Tryptic Peptide Mapping

20 μL samples at 5 μg/μL were denatured by adding 200 μL of 8 M guanidine, 130 mM Tris, 1 mM EDTA, pH 7.6 denaturing buffer. The samples were then reduced by the addition of 2 μL of 500 mM dithiothreitol. After incubation at 37°C for 30 min, samples were alkylated by the addition of 5 μL of 500 mM iodoacetamide and incubated at ambient temperature for 30 min in the dark. The reduced and alkylated samples were buffer exchanged into a solution containing 2 M urea and 100 mM Tris at pH 8.0 using an Amicon spin filter (EMD Millipore, Billerica, MA, USA; molecular weight cut-off of 10 kDa); 5 μg of trypsin was then added to the spin filter and incubated at 37°C for 4 h. The digested samples were collected from the spin filters, and the digestion was quenched with trifluoroacetic acid.

Peptides produced by enzymatic digestion were eluted on an Acquity Ultra Performance liquid chromatography system (Waters, Milford, MA, USA) equipped with an ethylene bridged hybrid C18 reversed-phase column (1.7 μm, 2.1 mm, 150 mm) using a gradient of 0%–60% acetonitrile at a flow rate of 0.2 mL/min (total elution time of 76 min). Peptides separated on the column were identified by a UV detector and analyzed using an Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific). Peak identification was based on both the exact monoisotopic mass and the tandem mass spectrum of the

Table 19. Confusion Matrix for Predictions Made by the Lorenzo et al. Model on the AAV8 Capsid Protein VP3

Prediction →		
Experiment ↓	Positive	Negative
Positive	9	1
Negative	9	28

Table 20. Comparison of Predictions Made by our Categorical Model and the NG/NN/NS Model on Selected Residues of the AAV8 Capsid Protein

Residue	N+1	Avg % Deamidation Giles et al.[16]	Categorical Model	NG/NN/NS	Lorenzo et al.[41]
N254	N	9%	no	yes	yes
N255	H	ND	no	no	yes
N263	G	99%	yes	yes	yes
N304	N	ND	no	yes	yes
N305	N	8%	no	yes	yes
N337	N	ND	No	yes	yes
N384	N	ND	no	yes	yes
N385	G	88%	yes	yes	yes
N410	N	3%	no	yes	yes
N459	T	7%	no	no	no
N498	N	ND	no	yes	yes
N499	N	17%	yes	yes	yes
N500	S	ND	no	yes	yes
N514	G	84%	yes	yes	yes
N517	S	4%	no	yes	yes
N540	G	79%	yes	yes	yes
N599	S	ND	no	yes	yes
N611	R	ND	no	no	no
N670	S	ND	no	yes	yes
N693	S	ND	no	yes	yes

Individual residue level predictions are shown for each model on a subset of residues from the AAV8 protein. The 5 mispredicted sites by our model were significantly less deamidated than the other 5 deamidation sites observed in the AAV8 capsid, measured by Giles et al.[16]

target ion. Deamidation quantitation was based on peak areas from the extracted ion chromatography of corresponding ions.

In most cases in our collected deamidation data for the training and validation sets, sequencing information by MS/MS could distinguish between deamidation on neighboring asparagines in the same tryptic peptide. However, for two NN sites in the validation set, MS/MS data could not distinguish between the N and N+1 residues. Thus, in these cases, the $t_{1/2}$ was a combined measurement for both sites in the peptide. Half-life predictions made by the regression model for these two sites were also combined prior to analysis.

Random Forest Machine Learning Model Construction

Both the classification model and regression model were random forest models built in RStudio using the randomForest[46] and caret[58] libraries. The number of trees and number of parameters tried at each split were optimized by manually tuned to minimize the out-of-bag error estimate.

For the classification model, 500 trees were generated with 3 variables tried at each split, producing an out-of-bag error estimate of 4.25% on the training set. The probability threshold at which we interpret the prediction as "yes" or "no" was also optimized to 53% after model building. Confusion matrices and variable importance plots were generated using caret and random Forest libraries, respectively.

The regression model was trained only on the subset of training data containing deamidation sites quantified by LC-MS/MS, including our in-house data and that of Lu et al.[57] In this case, 500 trees were generated with 4 variables tried at each split. The out-of-bag predictions explained 63.5% of the variance of the training set. R^2 and Q^2 were calculated and variable importance plots were generated using caret and randomForest libraries, respectively.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omtm.2019.09.008.

AUTHOR CONTRIBUTIONS

J.A.D. and J.W. conceptualized and designed the experiments and curated and interpreted the data; J.A.D., S.W.C., J.A.M., and J.P.M. performed the experiments and analyzed the data; J.A.D. wrote the manuscript.

CONFLICTS OF INTEREST

This work was supported by the global biologics R&D arm of AstraZeneca. Authors J.A.D. and J.W. are employees of AstraZeneca and have stock and/or stock interests or options in AstraZeneca.

ACKNOWLEDGMENTS

The authors would like to thank Sydney Clark, Arun Parupudi, Chen Qian, and Weichen Xu for their help with sample preparation; Ben Niu, Chen Qian, and Anthony Shannon for help with data collection; David Arancibia, Trina Do, Lauren Johnson, Nicholas Knoepfle, and Mohamed Ndiaye for help compiling samples for the training set; Hege Beard, Christopher Negron, and Kevin Halligan for their help with homology model building and parameter extraction; Xiaoyu Chen, Bojana Popovic, and Bruck Taddese for helpful discussion and manuscript review; Fiona Cusdin and Christopher Lloyd for help with sample generation for the validation set; and Gail Wasserman, Xiangyang Wang, Xiaoyu Chen, Milton Axley, Thomas Blacklock, Ray Field, T. Alan Hatton, and the 2018 MIT Practice School Cohort for their continual support of the project.

REFERENCES

1. US Food and Drug Administration (2019). Drugs@fda: FDA approved drug products, https://www.accessdata.fda.gov/scripts/cder/daf.
2. Jarasch, A., Koll, H., Regula, J.T., Bader, M., Papadimitriou, A., and Kettenberger, H. (2015). Developability assessment during the selection of novel therapeutic antibodies. J. Pharm. Sci. *104*, 1885–1898.
3. Elvin, J.G., Couston, R.G., and van der Walle, C.F. (2013). Therapeutic antibodies: market considerations, disease targets and bioprocessing. Int. J. Pharm. *440*, 83–98.
4. Carter, P.J. (2006). Potent antibody therapeutics by design. Nat. Rev. Immunol. *6*, 343–357.

5. Kohli, N., Jain, N., Geddie, M.L., Razlog, M., Xu, L., and Lugovskoy, A.A. (2015). A novel screening method to assess developability of antibody-like molecules. MAbs *7*, 752–758.

6. Raybould, M.I.J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A.P., Bujotzek, A., Shi, J., and Deane, C.M. (2019). Five computational developability guidelines for therapeutic antibody profiling. Proc. Natl. Acad. Sci. USA *116*, 4025–4030.

7. Robinson, N.E., and Robinson, A.B. (2001). Molecular clocks. Proc. Natl. Acad. Sci. USA *98*, 944–949.

8. Pace, A.L., Wong, R.L., Zhang, Y.T., Kao, Y.H., and Wang, Y.J. (2013). Asparagine deamidation dependence on buffer type, pH, and temperature. J. Pharm. Sci. *102*, 1712–1723.

9. Catak, S., Monard, G., Aviyente, V., and Ruiz-López, M.F. (2009). Deamidation of asparagine residues: direct hydrolysis versus succinimide-mediated deamidation mechanisms. J. Phys. Chem. A *113*, 1111–1120.

10. Haberger, M., Bomans, K., Diepold, K., Hook, M., Gassner, J., Schlothauer, T., Zwick, A., Spick, C., Kepert, J.F., Hienz, B., et al. (2014). Assessment of chemical modifications of sites in the CDRs of recombinant antibodies: Susceptibility vs. functionality of critical quality attributes. MAbs *6*, 327–339.

11. Harris, R.J., Kabakoff, B., Macchi, F.D., Shen, F.J., Kwong, M., Andya, J.D., Shire, S.J., Bjork, N., Totpal, K., and Chen, A.B. (2001). Identification of multiple sources of charge heterogeneity in a recombinant antibody. J. Chromatogr. B Biomed. Sci. Appl. *752*, 233–245.

12. Diepold, K., Bomans, K., Wiedmann, M., Zimmermann, B., Petzold, A., Schlothauer, T., Mueller, R., Moritz, B., Stracke, J.O., Mølhøj, M., et al. (2012). Simultaneous assessment of Asp isomerization and Asn deamidation in recombinant antibodies by LC-MS following incubation at elevated temperatures. PLoS ONE *7*, e30295.

13. Vlasak, J., Bussat, M.C., Wang, S., Wagner-Rousset, E., Schaefer, M., Klinguer-Hamour, C., Kirchmeier, M., Corvaïa, N., Ionescu, R., and Beck, A. (2009). Identification and characterization of asparagine deamidation in the light chain CDR1 of a humanized IgG1 antibody. Anal. Biochem. *392*, 145–154.

14. Huang, L., Lu, J., Wroblewski, V.J., Beals, J.M., and Riggin, R.M. (2005). In vivo deamidation characterization of monoclonal antibody by LC/MS/MS. Anal. Chem. *77*, 1432–1439.

15. Yan, B., Steen, S., Hambly, D., Valliere-Douglass, J., Vanden Bos, T., Smallwood, S., Yates, Z., Arroll, T., Han, Y., Gadgil, H., et al. (2009). Succinimide formation at Asn 55 in the complementarity determining region of a recombinant monoclonal antibody IgG1 heavy chain. J. Pharm. Sci. *98*, 3509–3521.

16. Giles, A.R., Sims, J.J., Turner, K.B., Govindasamy, L., Alvira, M.R., Lock, M., and Wilson, J.M. (2018). Deamidation of amino acids on the surface of adeno-associated virus capsids leads to charge heterogeneity and altered vector function. Mol. Ther. *26*, 2848–2862.

17. Verma, A., Ngundi, M.M., and Burns, D.L. (2016). Mechanistic analysis of the effect of deamidation on the immunogenicity of anthrax protective antigen. Clin. Vaccine Immunol. *23*, 396–402.

18. Verma, A., and Burns, D.L. (2018). Improving the stability of recombinant anthrax protective antigen vaccine. Vaccine *36*, 6379–6382.

19. Verma, A., McNichol, B., Domínguez-Castillo, R.I., Amador-Molina, J.C., Arciniega, J.L., Reiter, K., Meade, B.D., Ngundi, M.M., Stibitz, S., and Burns, D.L. (2013). Use of site-directed mutagenesis to model the effects of spontaneous deamidation on the immunogenicity of Bacillus anthracis protective antigen. Infect. Immun. *81*, 278–284.

20. Baillie, L.W. (2009). Is new always better than old?: The development of human vaccines for anthrax. Hum. Vaccin. *5*, 806–816.

21. D'Souza, A.J., Mar, K.D., Huang, J., Majumdar, S., Ford, B.M., Dyas, B., Ulrich, R.G., and Sullivan, V.J. (2013). Rapid deamidation of recombinant protective antigen when adsorbed on aluminum hydroxide gel correlates with reduced potency of vaccine. J. Pharm. Sci. *102*, 454–461.

22. Chen, H.M., Ford, C., and Reilly, P.J. (1994). Substitution of asparagine residues in Aspergillus awamori glucoamylase by site-directed mutagenesis to eliminate N-glycosylation and inactivation by deamidation. Biochem. J. *301*, 275–281.

23. Sierks, M.R., Ford, C., Reilly, P.J., and Svensson, B. (1993). Functional roles and subsite locations of Leu177, Trp178 and Asn182 of Aspergillus awamori glucoamylase determined by site-directed mutagenesis. Protein Eng. *6*, 75–79.

24. Bakir, U., Coutinho, P.M., Sullivan, P.A., Ford, C., and Reilly, P.J. (1993). Cassette mutagenesis of Aspergillus awamori glucoamylase near its general acid residue to probe its catalytic and pH properties. Protein Eng. *6*, 939–946.

25. Lavoisier, A., and Schlaeppi, J.M. (2015). Early developability screen of therapeutic antibody candidates using Taylor dispersion analysis and UV area imaging detection. MAbs *7*, 77–83.

26. Yang, X., Xu, W., Dukleska, S., Benchaar, S., Mengisen, S., Antochshuk, V., Cheung, J., Mann, L., Babadjanova, Z., Rowand, J., et al. (2013). Developability studies before initiation of process development: improving manufacturability of monoclonal antibodies. MAbs *5*, 787–794.

27. Xu, Y., Wang, D., Mason, B., Rossomando, T., Li, N., Liu, D., Cheung, J.K., Xu, W., Raghava, S., Katiyar, A., et al. (2019). Structure, heterogeneity and developability assessment of therapeutic antibodies. MAbs *11*, 239–264.

28. Yu, J., Shi, S., Zhang, F., Chen, G., and Cao, M. (2019). Predgly: Predicting lysine glycation sites for homo sapiens based on xgboost feature optimization. Bioinformatics *35*, 2749–2756.

29. Islam, M.M., Saha, S., Rahman, M.M., Shatabda, S., Farid, D.M., and Dehzangi, A. (2018). iProtGly-SS: Identifying protein glycation sites using sequence and structure based features. Proteins *86*, 777–789.

30. Ju, Z., Sun, J., Li, Y., and Wang, L. (2017). Predicting lysine glycation sites using bi-profile bayes feature extraction. Comput. Biol. Chem. *71*, 98–103.

31. Xu, Y., Li, L., Ding, J., Wu, L.Y., Mai, G., and Zhou, F. (2017). Gly-PseAAC: Identifying protein lysine glycation through sequences. Gene *602*, 1–7.

32. Reddy, H.M., Sharma, A., Dehzangi, A., Shigemizu, D., Chandra, A.A., and Tsunoda, T. (2019). GlyStruct: glycation prediction using structural properties of amino acid residues. BMC Bioinformatics *19* (*Suppl 13*), 547.

33. Akmal, M.A., Rasool, N., and Khan, Y.D. (2017). Prediction of N-linked glycosylation sites using position relative features and statistical moments. PLoS ONE *12*, e0181966.

34. Li, F., Li, C., Revote, J., Zhang, Y., Webb, G.I., Li, J., Song, J., and Lithgow, T. (2016). GlycoMine^{struct}: a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features. Sci. Rep. *6*, 34595.

35. Chuang, G.Y., Boyington, J.C., Joyce, M.G., Zhu, J., Nabel, G.J., Kwong, P.D., and Georgiev, I. (2012). Computational prediction of N-linked glycosylation incorporating structural properties and patterns. Bioinformatics *28*, 2249–2255.

36. Sydow, J.F., Lipsmeier, F., Larraillet, V., Hilger, M., Mautz, B., Mølhøj, M., Kuentzer, J., Klostermann, S., Schoch, J., Voelger, H.R., et al. (2014). Structure-based prediction of asparagine and aspartate degradation sites in antibody variable regions. PLoS ONE *9*, e100736.

37. Aledo, J.C., Cantón, F.R., and Veredas, F.J. (2017). A machine learning approach for predicting methionine oxidation sites. BMC Bioinformatics *18*, 430.

38. Sankar, K., Hoi, K.H., Yin, Y., Ramachandran, P., Andersen, N., Hilderbrand, A., McDonald, P., Spiess, C., and Zhang, Q. (2018). Prediction of methionine oxidation risk in monoclonal antibodies using a machine learning method. MAbs *10*, 1281–1290.

39. Agrawal, N.J., Dykstra, A., Yang, J., Yue, H., Nguyen, X., Kolvenbach, C., and Angell, N. (2018). Prediction of the hydrogen peroxide-induced methionine oxidation propensity in monoclonal antibodies. J. Pharm. Sci. *107*, 1282–1289.

40. Yan, Q., Huang, M., Lewis, M.J., and Hu, P. (2018). Structure based prediction of asparagine deamidation propensity in monoclonal antibodies. MAbs *10*, 901–912.

41. Lorenzo, J.R., Alonso, L.G., and Sánchez, I.E. (2015). Prediction of spontaneous protein deamidation from sequence-derived secondary structure and intrinsic disorder. PLoS ONE *10*, e0145186.

42. Jia, L., and Sun, Y. (2017). Protein asparagine deamidation prediction based on structures with machine learning methods. PLoS ONE *12*, e0181347.

43. Robinson, N.E., and Robinson, A.B. (2001). Prediction of protein deamidation rates from primary and three-dimensional structure. Proc. Natl. Acad. Sci. USA *98*, 4367–4372.

44. Robinson, N.E. (2002). Protein deamidation. Proc. Natl. Acad. Sci. USA *99*, 5283–5288.

45. Robinson, N.E., and Robinson, A.B. (2001). Deamidation of human proteins. Proc. Natl. Acad. Sci. USA *98*, 12409–12413.

46. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

47. Robinson, N.E., Robinson, A.B., and Merrifield, R.B. (2001). Mass spectrometric evaluation of synthetic peptides as primary structure models for peptide and protein deamidation. J. Pept. Res. *57*, 483–493.

48. Capasso, S. (2000). Estimation of the deamidation rate of asparagine side chains. J. Pept. Res. *55*, 224–229.

49. Capasso, S., Mazzarella, L., Sica, F., Zagari, A., and Salvadori, S. (1993). Kinetics and mechanism of succinimide ring formation in the deamidation process of asparagine residues. J. Chem. Soc., Perkin Trans. *2*, 679–682.

50. Tyler-Cross, R., and Schirch, V. (1991). Effects of amino acid sequence, buffers, and ionic strength on the rate and mechanism of deamidation of asparagine residues in small peptides. J. Biol. Chem. *266*, 22549–22556.

51. Robinson, N.E., Robinson, Z.W., Robinson, B.R., Robinson, A.L., Robinson, J.A., Robinson, M.L., and Robinson, A.B. (2004). Structure-dependent nonenzymatic deamidation of glutaminyl and asparaginyl pentapeptides. J. Pept. Res. *63*, 426–436.

52. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. *18*, 463–477.

53. Dall'Acqua, W.F., Kiener, P.A., and Wu, H. (2006). Properties of human IgG1s engineered for enhanced binding to the neonatal Fc receptor (FcRn). J. Biol. Chem. *281*, 23514–23524.

54. Silva, J.P., Vetterlein, O., Jose, J., Peters, S., and Kirby, H. (2015). The S228P mutation prevents in vivo and in vitro IgG4 Fab-arm exchange as demonstrated using a combination of novel quantitative immunoassays and physiological matrix preparation. J. Biol. Chem. *290*, 5462–5469.

55. Oganesyan, V., Gao, C., Shirinian, L., Wu, H., and Dall'Acqua, W.F. (2008). Structural characterization of a human Fc fragment engineered for lack of effector functions. Acta Crystallogr. D Biol. Crystallogr. *64*, 700–704.

56. Dimasi, N., Fleming, R., Zhong, H., Bezabeh, B., Kinneer, K., Christie, R.J., Fazenbaker, C., Wu, H., and Gao, C. (2017). Efficient preparation of site-specific antibody-drug conjugates using cysteine insertion. Mol. Pharm. *14*, 1501–1516.

57. Lu, X., Nobrega, R.P., Lynaugh, H., Jain, T., Barlow, K., Boland, T., Sivasubramanian, A., Vásquez, M., and Xu, Y. (2019). Deamidation and isomerization liability analysis of 131 clinical-stage antibodies. MAbs *11*, 45–57.

58. Kuhn, M. (2008). Building predictive models in r using the caret package. J. Stat. Soft. *28*, 1–26.

59. Zhu, K., Day, T., Warshaviak, D., Murrett, C., Friesner, R., and Pearlman, D. (2014). Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. Proteins *82*, 1646–1655.

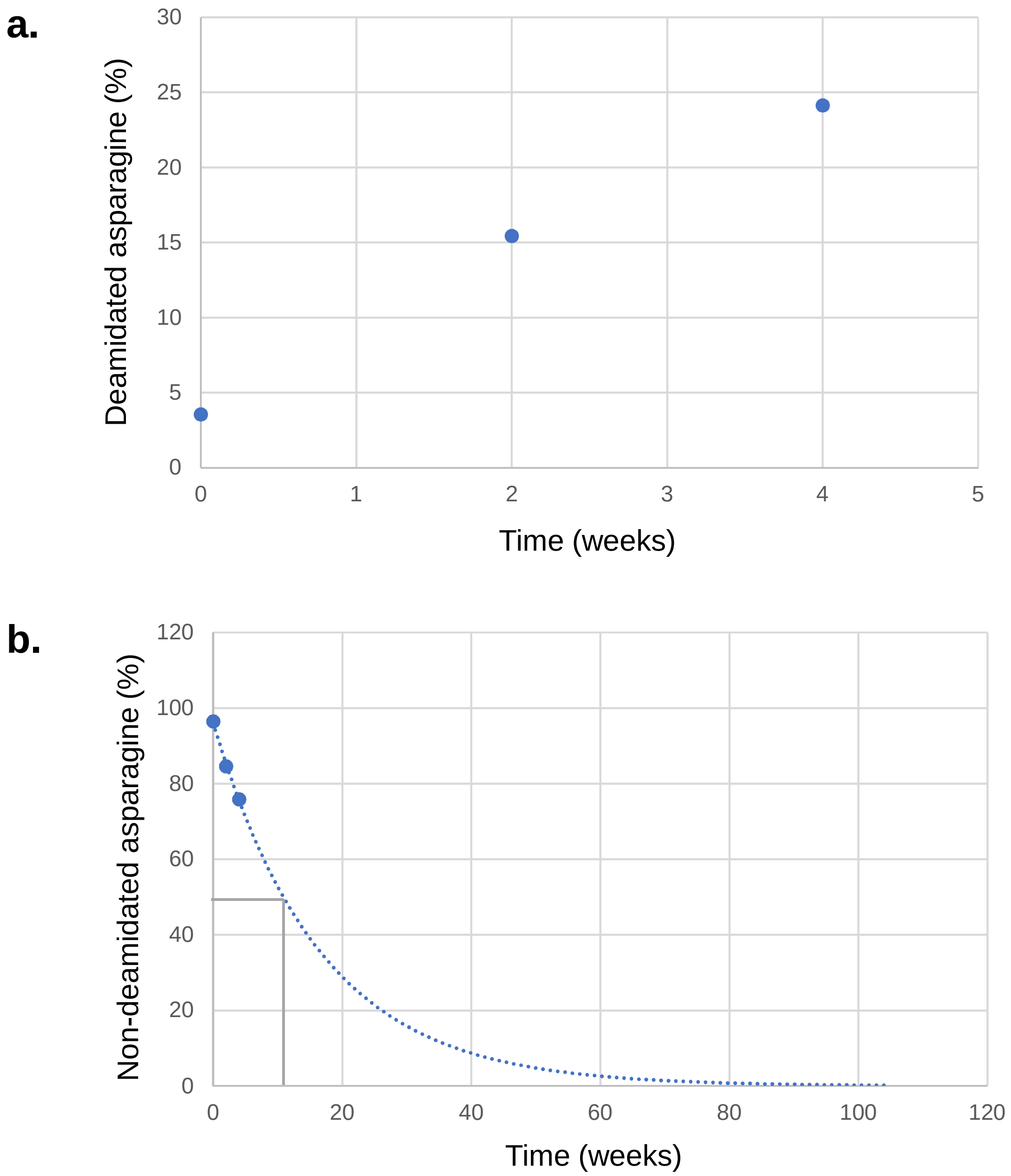
60. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

OMTM, Volume *15*

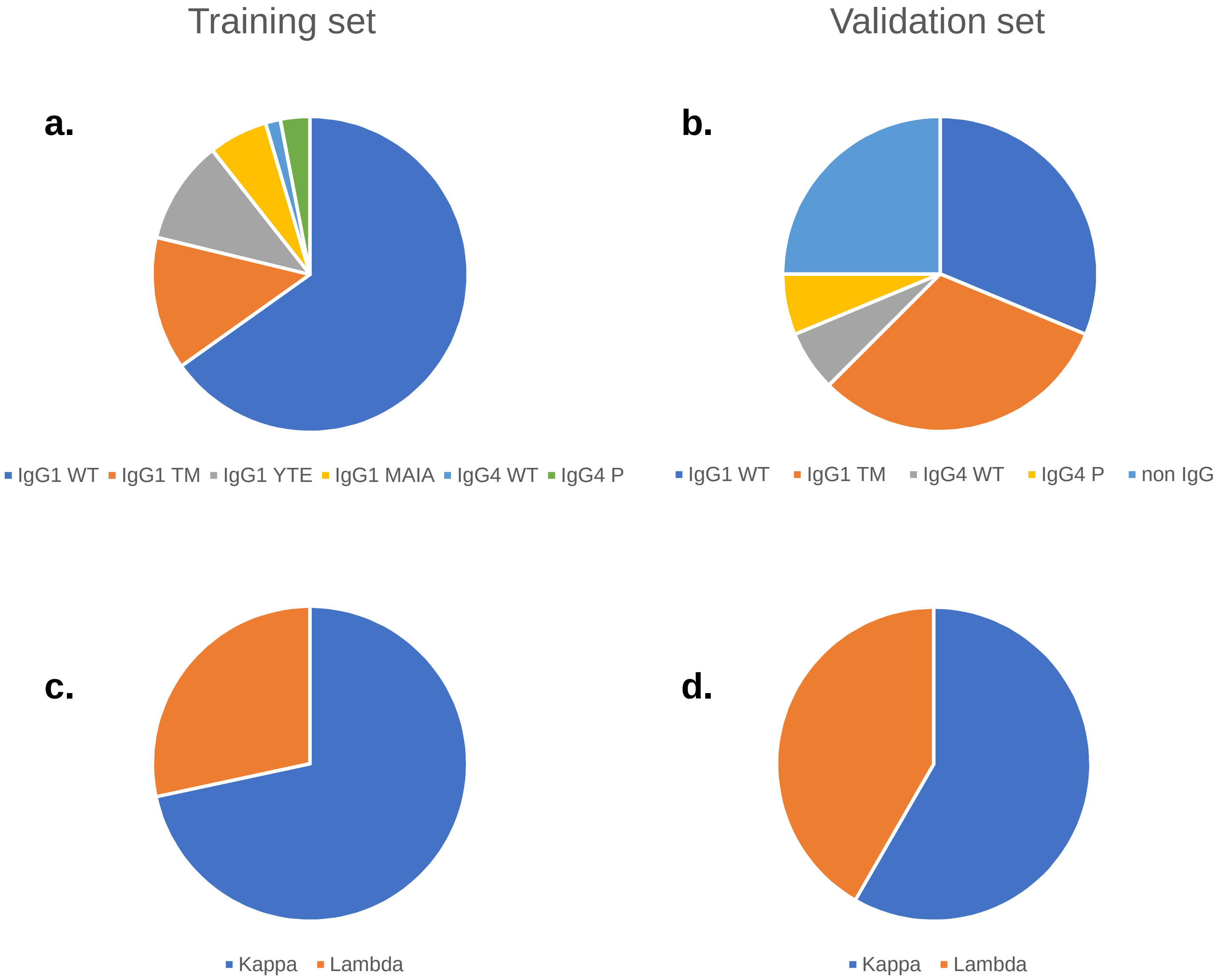
Supplemental Information

Machine Learning Enables Accurate Prediction of Asparagine Deamidation Probability and Rate

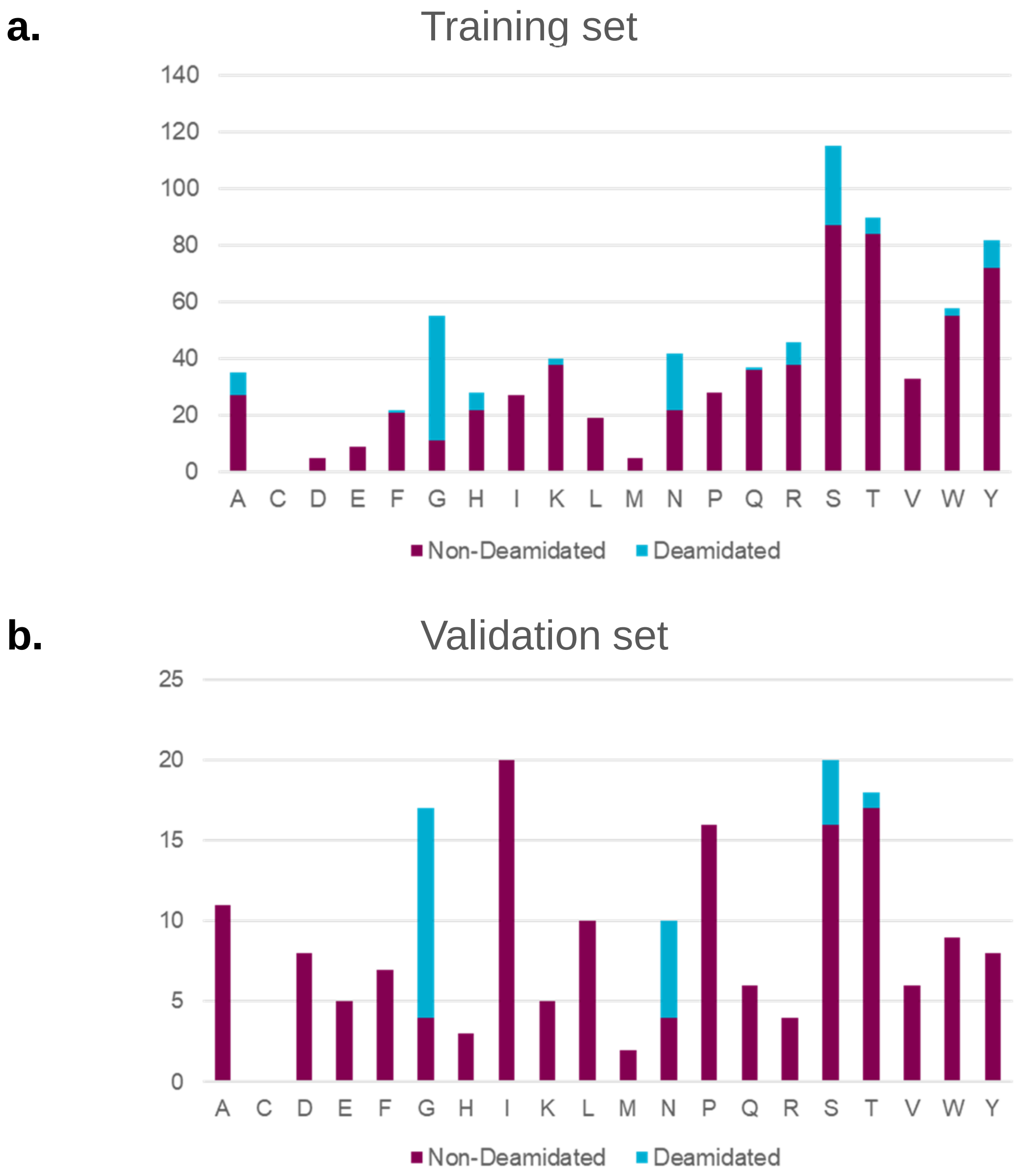
Jared A. Delmar, Jihong Wang, Seo Woo Choi, Jason A. Martins, and John P. Mikhail



Supplemental Figure 1. Calculation of site-specific asparagine deamidation half-life from from LC-MS/MS deamidation abundance. (a) Deamidation abundance for each asparagine in our training set molecules were measured by LC-MS/MS as the sum of aspartic acid and iso-aspartic acid products after 0, 2, and 4 week timepoints at stress conditions (blue dots). (b) The deamidation half-life of each site was calculated by a least squares fit to the abundance of non-deamidated asparagines versus time in weeks (blue dotted line). The half-life ($t_{1/2}$) is the time in weeks for deamidation to reach 50% (in this case 11.5 weeks, indicated by grey lines).



Supplemental Figure 2. Training and validation data set distribution. Distribution of IgG formats and non-IgG formats in (a) the training set and (b) the validation set. There are a total of 64 IgG1s, with 6 unique heavy chain formats, and 3 IgG4s, with 2 unique heavy chain formats, in the training set. The validation set contains 10 IgG1s and 2 IgG4s, with 2 unique heavy chain formats each, and 4 non-mAb proteins. Among IgGs, the light chain constant region format distribution is shown for (c) the training set and (d) the validation set.



Supplemental Figure 3. Distribution of deamidation frequency in training and validation sets. The number of asparagines is plotted versus the N+1 residue for (a) the training data set and (b) the independent validation set. In each case, the number of non-deamidated asparagines observed is colored maroon and the number of deamidated sites is colored cyan. For the training set, the fraction of deamidated sites where N+1 = G, N, or S, was 80%, 48%, and 24%, respectively; whereas in the validation set, we observed 76%, 40%, and 20%, respectively.

a.

Prediction → Experiment ↓	Positive	Negative
Positive	17	9
Negative	5	164

b.

Prediction → Experiment ↓	Positive	Negative
Positive	25	1
Negative	26	143

c.

Statistic	Categorical model	NG/NN/NS
Accuracy	92.8%	86.2%
MCC	0.671	0.625
Precision	77.3%	49.0%
Recall	65.4%	96.2%
Specificity	97.0%	84.6%
Negative Predictive Value	94.8%	99.3%
Miss Rate	34.6%	3.8%
Fallout	22.7%	51.0%
False Discovery Rate	3.0%	15.4%
False Omission Rate	5.2%	0.7%

Supplemental Table 1. Comparison of predictions made by the categorical model and the simple (NG/NN/NS) model on the independent validation set. (a) Confusion matrix for our categorical model; (b) confusion matrix for the NG/NN/NS model; and (c) statistics calculated for both the categorical and NG/NN/NS models.

a.

Prediction → Experiment ↓	Positive	Negative
Positive	137	0
Negative	0	639

b.

Prediction → Experiment ↓	Positive	Negative
Positive	92	45
Negative	120	519

c.

Statistic	Categorical model	NG/NN/NS
Accuracy	100.0%	43.4%
MCC	1.000	0.672
Precision	100.0%	81.2%
Recall	100.0%	92.0%
Specificity	0.0%	32.8%
Negative Predictive Value	0.0%	56.6%
Miss Rate	0.0%	18.8%
Fallout	0.0%	8.0%
False Discovery Rate	100.0%	78.7%
False Omission Rate	100.0%	41.4%

Supplemental Table 2. Comparison of predictions made by the categorical model and the conventional (NG/NN/NS) model on the training set. (a) Confusion matrix for our categorical model; (b) confusion matrix for the NG/NN/NS model; and (c) statistics calculated for both the categorical and NG/NN/NS models.

a.

Training set	mAbs		non-mAbs	
	In-house	Lu *et al.*	Jia *et al.*	Giles *et al.*
All asparagines; Deamidated / Total	98 / 608	39 / 168	0/0	0/0
Unique asparagines; Deamidated / Total	49 / 304	39 / 168	0/0	0/0

b.

Validation set	mAbs		non-mAbs	
	In-house	Lu *et al.*	Jia *et al.*	Giles *et al.*
All asparagines; Deamidated / Total	9 / 68	0 / 0	7 / 80	10 / 47
Unique asparagines; Deamidated / Total	9 / 68	0 / 0	7 / 80	10 / 47

c.

Non-mAb validation subset	mAbs		non-mAbs	
	In-house	Lu *et al.*	Jia *et al.*	Giles *et al.*
All asparagines; Deamidated / Total	0 / 0	0 / 0	7 / 80	0 / 0
Unique asparagines; Deamidated / Total	0 / 0	0 / 0	7 / 80	0 / 0

d.

mAb-only validation subset	mAbs		non-mAbs	
	In-house	Lu *et al.*	Jia *et al.*	Giles *et al.*
All asparagines; Deamidated / Total	9 / 68	0 / 0	0 / 0	0 / 0
Unique asparagines; Deamidated / Total	9 / 68	0 / 0	0 / 0	0 / 0

Supplemental Table 3. Data sources and description. Number of total, deamidated, and unique asparagines for (a) complete training data set, (b) complete validation data set, (c) non-mAb validation data subset, (d) mAb-only validation data subset. Non-unique asparagines in the training set mAbs have a nearly identical site on the opposite heavy or light chain, as the full IgG homology model was generate for in-house molecules. Our regression model was trained and validated only on the deamidated data within each set.