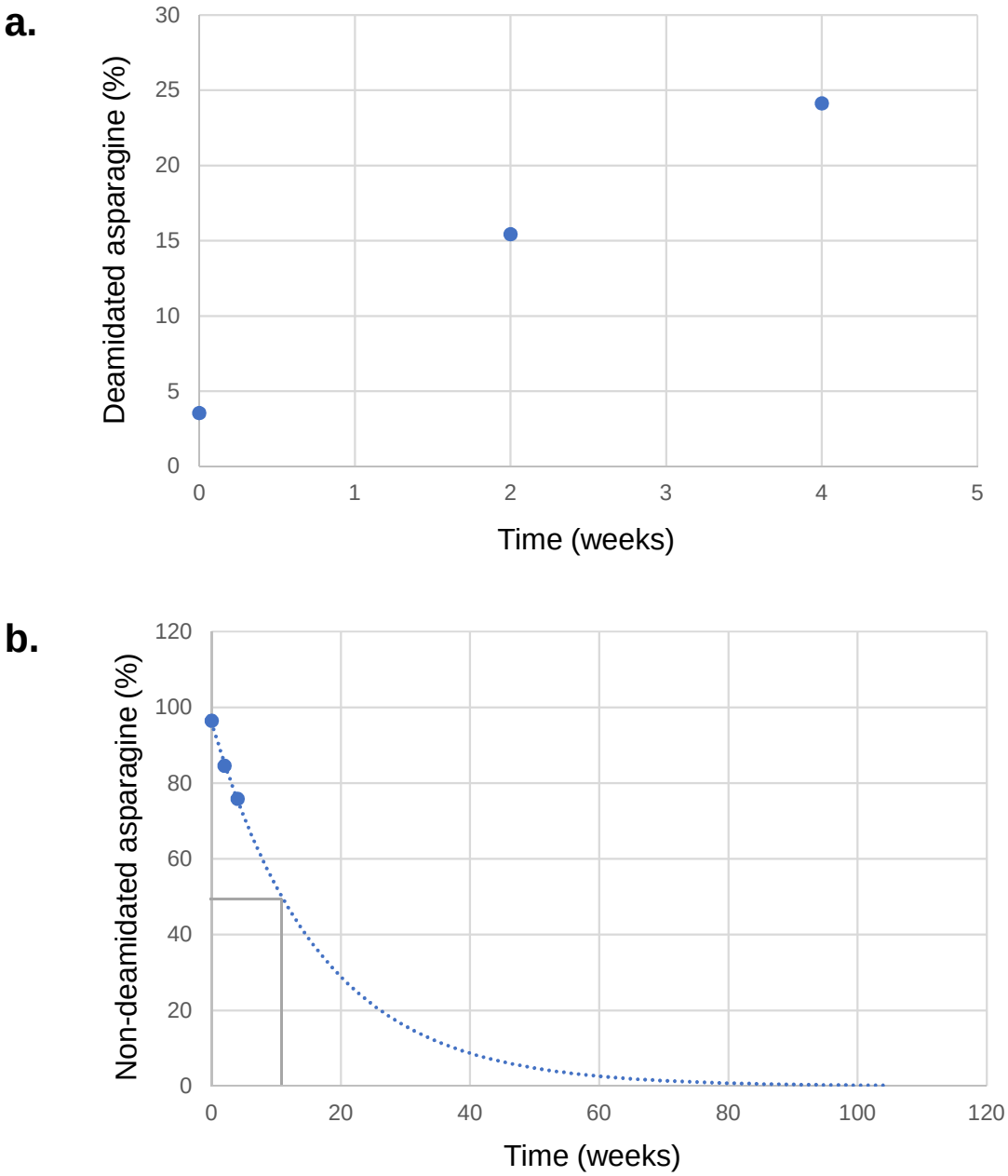
OMTM, Volume *15*

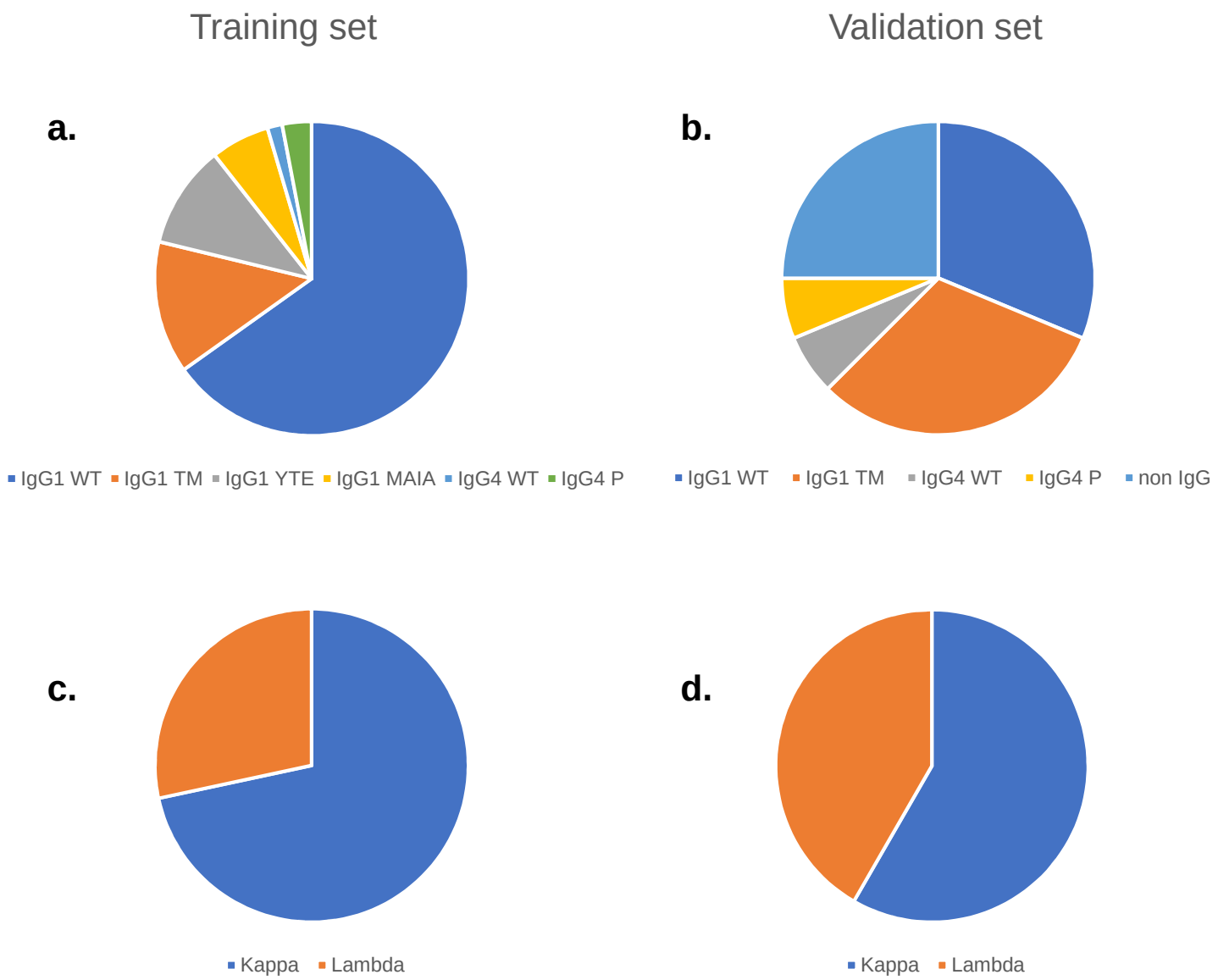
Supplemental Information

Machine Learning Enables Accurate Prediction of Asparagine Deamidation Probability and Rate

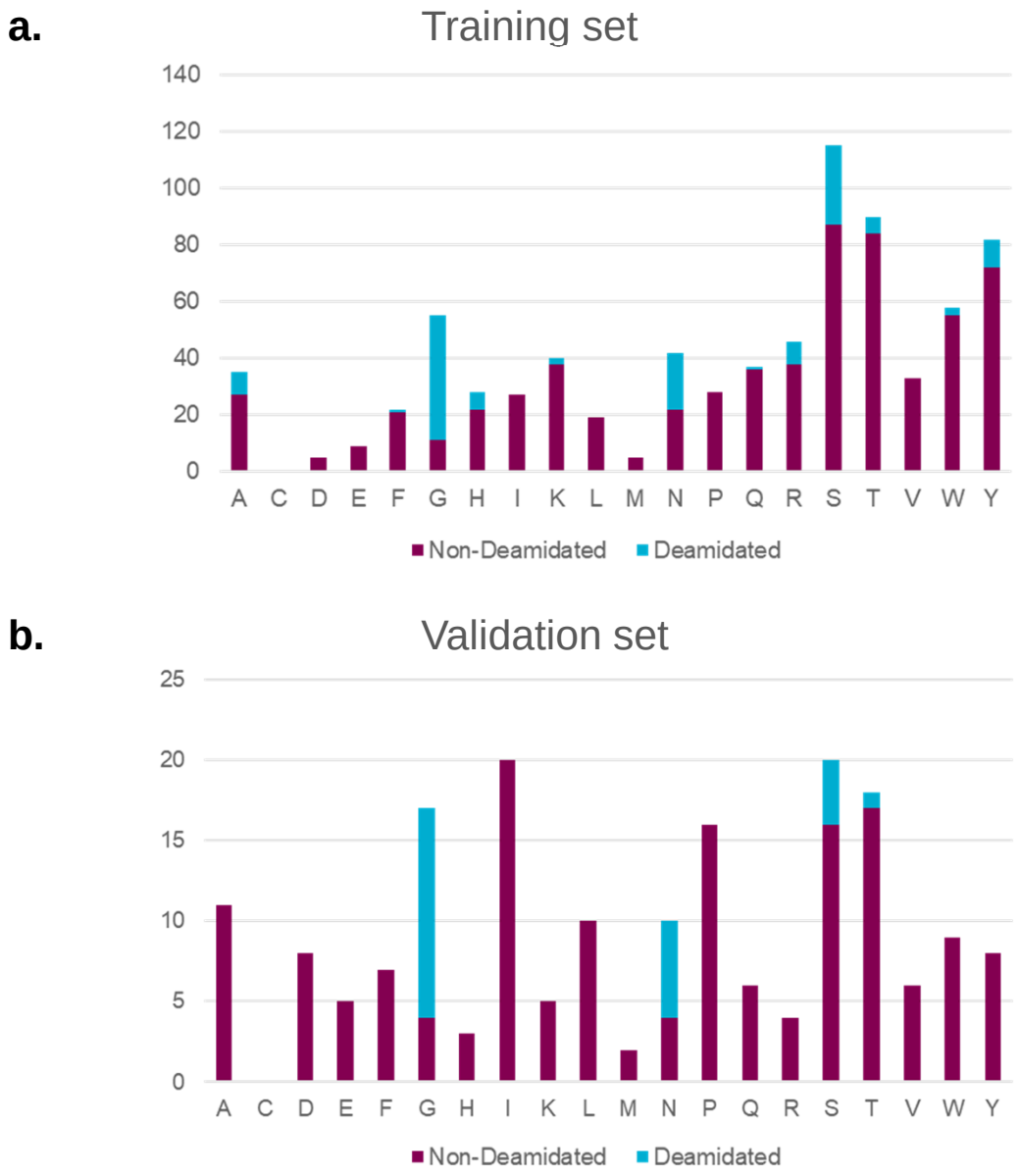
Jared A. Delmar, Jihong Wang, Seo Woo Choi, Jason A. Martins, and John P. Mikhail



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.