Electronic Supplementary Material (ESI) for Digital Discovery.

An Interpretable Machine Learning Framework for Modelling Macromolecular Interaction Mechanisms with Nuclear Magnetic Resonance

Samantha Stuart[1†], Jeffrey Watchorn[2†], Frank X Gu[1,2]*

1. Institute of Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada

2. Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, Ontario, Canada

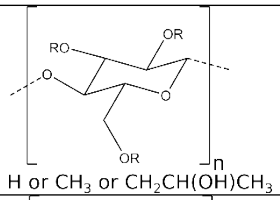
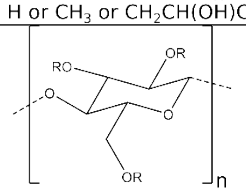
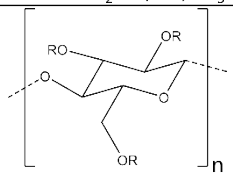
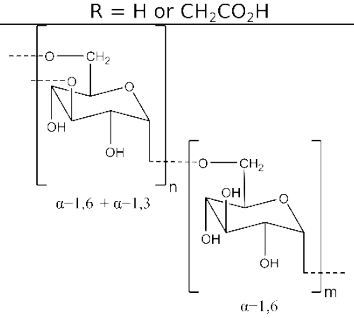
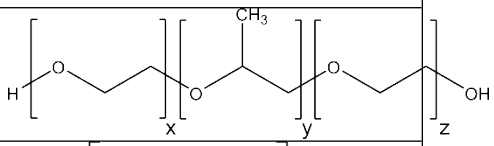
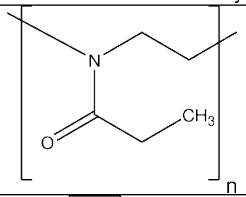
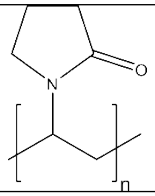
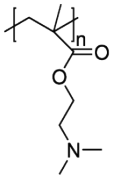
†These authors contributed equally to this work

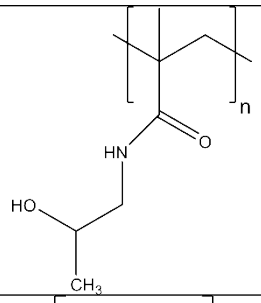
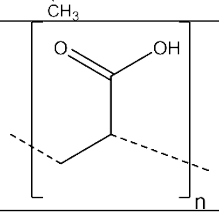
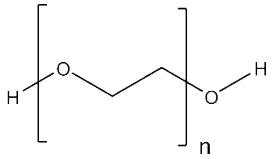
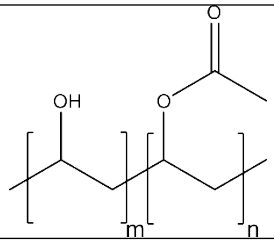
*Corresponding Author

Table of Contents

1. Representative Chemical Structures of Polymer Library

Supplementary Table 1. Representative chemical structures of polymer library dataset.

Polymer Name	Abbreviation	Chemical Structure
Hydroxypropyl methyl cellulose	HPMC	 R = H or CH_3 or $CH_2CH(OH)CH_3$
Hydroxypropyl cellulose	HPC	 R = H or $CH_2CH(OH)CH_3$
Carboxymethyl cellulose	CMC	 R = H or CH_2CO_2H
Dextran from Leuconostoc mesenteroides	DEX	
Poloxamer 407	P407	
Poly(2-ethyl-2-oxazoline)	PEOZ	
Poly(vinylpyrrolidone)	PVP	
Poly((2-dimethylamino)ethyl methacrylate)	PDMAEMA	

Poly-(N-(2-hydroxypropyl)methacrylamide)	PHPMA	
Poly(acrylic) acid	PAA	
Polyethylene glycol	PEG	
Poly(vinyl) alcohol 86-89%	PVA	

2. δ 1H Chemical Shift Feature

2.1. Fingerprint Generation

To incorporate proton cohort information as a macromolecular fingerprint vector, for each proton sample we conducted a hashing workflow as follows: (1) bin the NMR spectrum into 0.1ppm step size intervals (2) identify the chemical shifts of the cohort for a given proton sample, (3) encode the presence or absence of cohort chemical shifts in terms of the binned intervals of the NMR spectrum. After the workflow was conducted for all proton samples in the dataset, binned intervals that contained no examples were dropped.

2.2. Fingerprint Interpretation

Ultimately, including the cohort chemical shift vector provided a contextual polymer fingerprint vector associated with each proton observation. Cohort vector chemical shift intervals, as principal component factor loadings, can be applied to suggest polymer functionalization design suggestions post-hoc for achieving interactive or inert properties. It merits noting, however, that by definition the chemical shift interval of a sample proton is excluded from its cohort fingerprint. Hence cohort interval factor loadings should be interpreted with this inverse relationship in mind. For instance, loadings indicating a cohort shift interval is negatively correlated with interaction may be in reference to the fact the chemical shift of an interactive sample proton was by definition excluded from its cohort vector. Reviewing the true identities, and chemical shifts, of the protons being classified as interactive by the final heuristics should thus be prioritized, and will reveal when this is the case.

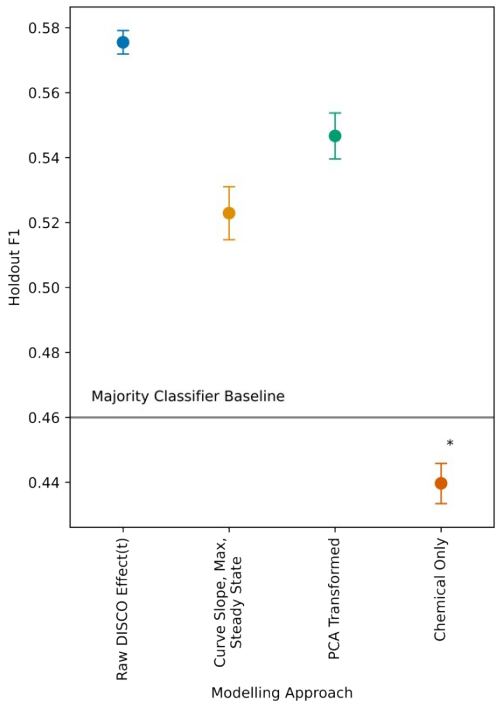
3. Disco Effect Feature

3.1. Benchmarking Study

To compare the different means of incorporating information from proton DISCO Effect(t) buildup curves in modelling, we performed a model pipeline benchmarking exercise with nested cross validation. Data standardization, PCA, and hyperparameter optimization were all fit to training folds, and transformations applied to the validation folds, using scikit-learn Pipeline estimators to mitigate data leakage (1,2).

In the inner nested cross validation fold, the same decision tree hyperparameter grid was used (Supplementary Table 2) with Stratified 5-Fold Grid Search Cross Validation (GridSearchCV estimator in scikit-learn), to automatically construct a cross validated model to the dataset with the highest ROC-AUC. In nested cross validation, holdout sets are iteratively separated from the remainder of the data, creating an "outer fold." We used Leave-One-Out Cross Validation (LOO-CV) to construct the outer folds for nested cross validation, as is common for small datasets (2–5). Each inner fold then comprised the training and validation data, which were used to construct the optimal model to the fold, without seeing the holdout. The best performing models for each inner fold were constructed using the hyperparameter grid in ESI Table 2, where Stratified 5-Fold GridSearchCV identifies the decision tree with highest ROC-AUC from the training and validation data. Similarly, all data standardization and PCA transforms were fit to training folds, and transformed on validation folds. Then, the data transforms and best model were fit to all of the training and validation data, and applied to transform the untrained holdout data for prediction at the optimal classification threshold. The workflow resulted in 99 outer folds, such that all samples were eventually assigned unbiased, untrained predictions. Additionally, given decision trees are sensitive to the random seed used for initialization, we conducted technical replicates of each benchmark at three different random seeds (0, 148, 601). Scikit-learn Pipeline estimators, with an additional ColumnTransformer estimator step isolating the first-pass PCA for CDE computation, were used respectively to compute the benchmarks.

We ultimately benchmarked performance using the macro average Holdout F1 scores (n=3) obtained for the different implementations of the DISCO Effect(t) features. To provide baselines for comparison in benchmarking, we include the performance metrics of a null model (a majority "dummy classifier") reporting all samples as the majority class label (0), as well as a version of the model with only the chemical descriptors, and no DISCO Effect(t) feature. The results of the benchmarking are summarized in Supplementary Figure 1, with the full benchmarking data provided in Supplementary Table 1 and Supplementary Table 2. Fully trained performance metrics of the best descriptive model resulting from each pipeline are also provided. Descriptive models (equivalently fully trained models) were created using the same methodology described in the main manuscript to create the final descriptive model.



Supplementary Fig. 1 Influence of DISCO Effect. Incorporating DISCO Effect as a modelling feature significantly improved model performance in terms of Holdout F1 scores from nested cross validation, compared with exclusively chemical descriptors. Models trained from chemical descriptors only performed worse on average than the null model baseline F1=0.46. Normality of data for statistical testing was verified by Q-Q plot, and equality of variance verified by Bartlett's test. Statistical testing was conducted by One-Way ANOVA (n=3, p=0.000028). Following ANOVA, a post-hoc t-test for multiple comparisons after Bonferroni correction revealed no significant differences between the three DISCO Effect feature implementations (p>0.12), but significantly worse performance by the models trained from chemical descriptors only, relative to all three disco effect features (p<0.032). * indicates a statistically significant difference in Holdout F1 scores from all other modelling approaches respectively (n=3, p<0.05). Error bars indicate standard error of the mean.

The three implementations of DISCO Effect features tested were: the direct inclusion of the mean absolute DISCO Effect(t) datapoints at each saturation time (+7 additional dimensions), including buildup curve attributes (linear region slope (t=1.0 – t=0.25), maximum point, steady state (t=1.75), for +3 additional dimensions), and the Cumulative DISCO Effect computed by retaining the first component following a Linear PCA of the 7-dimensional buildup curve data (+1 dimension). Statistical testing revealed that all three DISCO Effect feature implementations significantly improved holdout F1 scores over the pipeline comprising chemical-only descriptors. Models constructed using chemical descriptors only (i.e. sample proton chemical shift, parent polymer molecular weight, cohort proton chemical shift fingerprint only) ultimately scored worse on average than the majority classifier (F1=0.46) in terms of nested cross validation Holdout F1 (n=3). Prior to statistical testing, normality of data was verified by Q-Q plot, and equality of variance verified by Bartlett's test. Significant F-test results were obtained by one-way ANOVA (n=3, p=0.000028). Following ANOVA, a post-hoc t-test for multiple comparisons after Bonferroni correction revealed the three DISCO Effect feature implementations performed equally, without a statistically significant best performer, in terms of Holdout F1 score (n=3, p<0.05), yet all significantly improved over the chemical-only descriptor pipeline (n=3, p<0.032). The Python packages pingouin (version 0.5.1), and scikit-posthocs (version 0.7.0) were used to conduct the analysis. The code to regenerate this analysis is available in the GitHub for this paper.

Following this benchmarking, we elected to include the Cumulative DISCO Effect as calculated by Linear PCA (retaining first component only) of the mean absolute DISCO Effect(t) buildup curves. We chose this implementation as it added the least additional dimensionality to the modelling dataset, thus facilitating greatest interpretability over the analogous seven column implementation of DISCO Effect(t). However, a slight trend in increased Holdout F1 scores in pipelines applying all seven DISCO Effect(t) saturation times as independent variables suggest that downstream predictive works may improve their out of sample performance by including the whole buildup curve in the feature set, at the expense of facile interpretability.

Supplementary Table 2. DISCO Effect Feature Benchmarking Results. The final model interpreted in the main paper body is indicated in green. The null model baseline predicts all samples to be members of the majority class (0), which results in F1=0.46 as a minimum performance threshold. Best hyperparameters are those returned by GridSearchCV as the parameters that maximized ROC-AUC after Stratified 5-Fold Cross Validation on the full dataset.

DISCO Effect(t) feature approach:	Added dimensionality	Random seed	Best max depth	Best min samples per leaf	Best min samples per split	Fully trained F1	Holdout F1
Full Buildup Curve	7	148	4	5	2	0.818681	0.571892
Full Buildup Curve	7	0	4	7	2	0.817190	0.582719
Full Buildup Curve	7	601	4	20	2	0.728592	0.571892
Curve max, slope, steady state	3	148	4	3	7	0.857963	0.546703
Curve max, slope, steady state	3	0	4	3	7	0.857963	0.523225
Curve max, slope, steady state	3	601	5	3	7	0.898148	0.498734
Cumulative DISCO Effect	**1**	**148**	**5**	**3**	**2**	**0.869737**	**0.551284**
Cumulative DISCO Effect	1	0	5	3	15	0.818681	0.530063
Cumulative DISCO Effect	1	601	4	5	40	0.812025	0.558642
None - Chemical Only	0	148	4	10	30	0.660140	0.446064
None - Chemical Only	0	0	4	10	30	0.660140	0.427314
None - Chemical Only	0	601	4	10	2	0.660140	0.445689
Null Model: Majority Classifier	N/A	N/A	N/A	N/A	N/A	0.46	

Supplementary Table 3 Decision Tree Hyperparameter Grid. Hyperparameter grid used for all Grid Searches performed during benchmarking.

Maximum depth	4, 5, 6, 7, 8, 9, 10
Minimum samples per split	2, 3, 5, 7, 10, 15, 20, 30, 40

Minimum samples per leaf	1, 2, 3, 5, 7, 10, 15, 20

3.2. Cumulative DISCO Effect Calculation by Linear PCA

To provide an optimal implementation of a cumulative sum for buildup curve comparison ("Cumulative DISCO Effect"), we applied linear PCA to the absolute, standardized, saturation transfer buildup curves and retained only the first component. The retained component in the final model had positive factor loadings at all saturation time points (Supplementary Table 3), making it representative of an optimized cumulative sum for buildup curve comparison, and explained 68.8% of the variance in proton saturation transfer buildup curves.

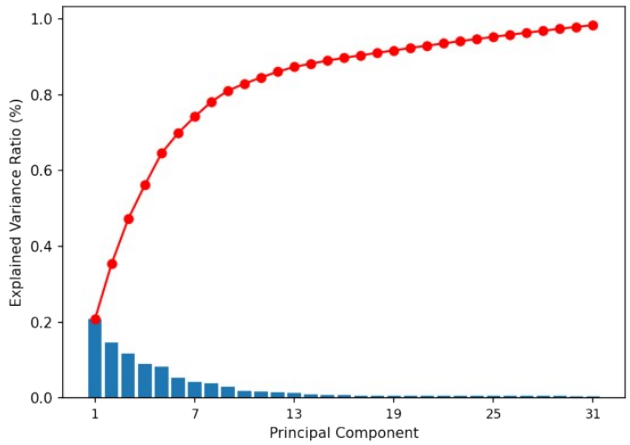
Supplementary Table 4. Cumulative DISCO Effect Feature Creation. Linear PCA factor loadings for the Cumulative DISCO Effect feature.

	t=0.25	t=0.5	t=0.75	t=1.0	t=1.25	t=1.5	t=1.75
PC1	0.391	0.397	0.300	0.320	0.377	0.418	0.425

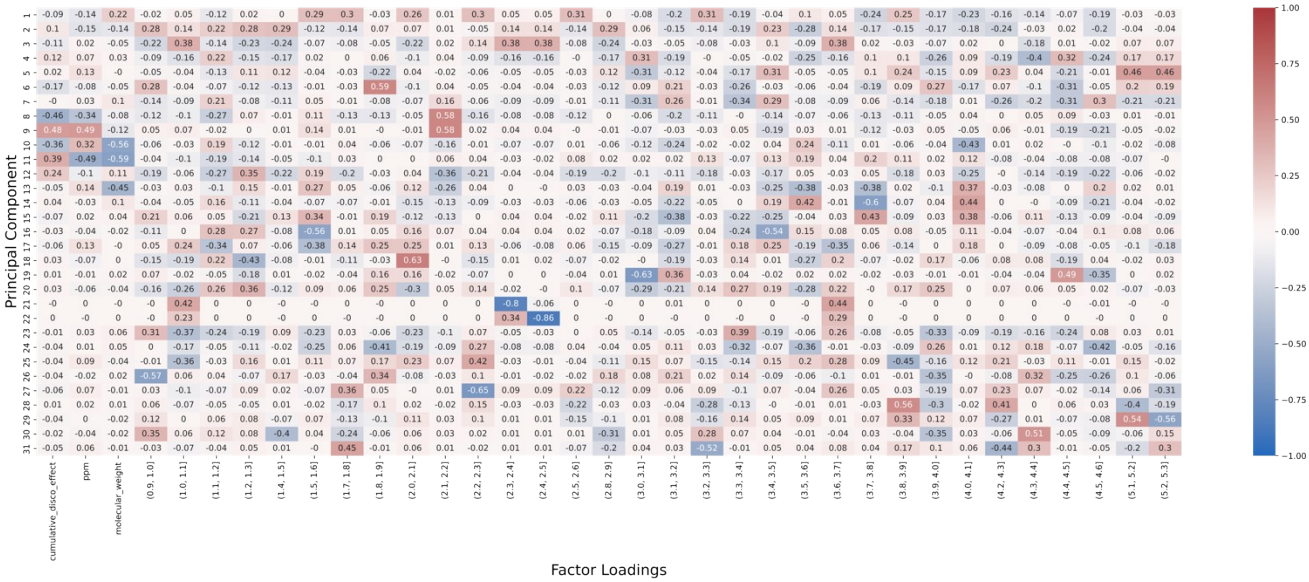
4. Principal Component Analysis

4.1. Full Dataset Principal Component Analysis

To remove linear intercorrelations from the full feature set prior to modelling, the full dataset was standardized, and transformed by Linear PCA. The optimal number of components to retain was selected automatically using Minka's MLE. The scree plot and factor loadings of the Linear PCA performed upon the full feature set are provided. 31 factors were retained, having 98.4% explained variance.



Supplementary Fig. 2 Principal component scree plot. Scree plot describing the variance explained by each component of the full feature set.



Supplementary Fig. 3 Principal Component Factor Loadings. All factor loadings relating the retained 31 principal components for modelling, and original features, are tabulated.

5. Decision Tree Tuning & Regularization

5.1. Hyperparameter Tuning

To constrain splitting within our final model, maximum tree depth, minimum samples per leaf, and minimum samples per split were optimized to the dataset using Stratified 5-Fold Grid Search Cross Validation (1). The grid search evaluated all parameter combinations from the ranges provided (504 candidates per fold) under 5-folds of cross-validation and selected the best tree hyperparameters from the search space in terms of maximum average area under the receiver operating characteristic curve (ROC-AUC). Parameter range inputs to the grid search were selected in accordance with literature in decision tree hyperparameter optimization (6). The exact range used is provided in Supplementary Table 2. The hyperparameters selected for the final model were a maximum tree depth of 5, 3 minimum proton samples per leaf, and no restraint on minimum proton samples per split. Scikit-learn Pipeline estimators were used to prevent data leakage between training and validation splits. By constructing the final model via grid search, the ultimate regions classifying interaction behavior were created without human bias.

5.2. Decision Tree Classification Metrics

Classification metrics are reported in terms of F1 score, the harmonic mean of precision and recall, given the class imbalance in the proton data (7):

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

where TP = number of true positives, FP = number of false positives, and FN = number of false negatives.

6. Detailed Description of Machine-Learned Heuristics

In Supplementary Table 4, a detailed breakdown of each machine learned heuristic, and the decision tree data underlying it is provided.

Supplementary Table 5. Summary of descriptive interaction heuristics. †

ID	Model Heuristic	Protons in Heuristic Scope		Rationale	Rule Plot
H1	PAA's mucoadhesive mechanism is similar at each of its protons. The mechanism is characteristic in the dataset to the molecular weight, CDE, and chemical composition of PAA in tandem.	PAA 450kDa (**2.03**, **1.54**, **1.28**)		PC 12 > 1.7 classified all PAA protons interactive	Fig. 4B
H2	PDMAEMA's chemical shifts and CDE characterize its mechanism. Proton clustering juxtaposed PEOZ and PDMAEMA as chemical and interactive "opposites."	PDMAEMA 10kDa (**4.4, 3.46, 2.05, 1.42**, 1.13, 0.92)		PC 3 <= -3.17 grouped protons from PDMAEMA, high PC3 scores clustered protons from PEOZ	Fig. 4B
H3	Monotonic increase in chemical shift (i.e. electronegativity of neighboring groups) correlated with PDMAEMA interaction.	PDMAEMA 10kDa (**4.4, 3.46, 2.05, 1.42,** 1.13, 0.92)		PC 31 < =-0.06 classified monotonically increasing chemical shifts from H2 as interactive	ESI Fig. 4B
H4	HPC's chemical composition distinguished a subset of 8 protons, across two HPC molecular weights, as similar.	HPC 370kDa (**4.07, 3.77, 3.46**, 3.14, **1.13**) HPC 80kDa (4.07, 3.77, 1.13)		PC 7 > 2.38 groups HPC protons together	ESI Fig. 5B
H5	Tuning molecular weight, without any additional chemical functionalization, unlocked interaction in HPC.	HPC 370kDa (**4.07, 3.77, 3.46**, 3.14, **1.13**) HPC 80kDa (4.07, 3.77, 1.13)		PC11 <=-0.65 classifies interactive H5 protons (370kDa) vs inert (80kDa)	Fig 4D
H6	Protons across polymer species of downfield chemical shifts (avg. ppm $_{9>1.03}$=4.14ppm, avg. ppm $_{9<=1.03}$=3.14ppm), high CDE (avg. CDE $_{9>1.03}$= 3.63, avg. CDE $_{9<=1.03}$= -0.41), and molecular weights in 80-150kDa (avg. $MW_{9>1.03}$=111kDa, avg. $MW_{9<=1.03}$=212.3kDa) show similar propensity for interaction.	PVA 105kDa (**4.08**, 1.58), CMC 131kDa (**4.58**), HPMC 86kDa (**4.48**), HPMC 120kDa (4.48), HPC 80kDa (4.58), DEX 150kDa (5.20)		PC9 > 1.03 groups together seven cross-species protons from the remaining unclassified set, similar protons cluster nearby	Fig 5B
H7	A reduction in CDE (avg. $CDE_{11>0.22=}$ 5.83, avg. $CDE_{11<=0.22}$= 0.704) provided the fine resolution necessary to characterize dominant interactions within a mechanistically similar proton group across species and MW.	PVA 105kDa (**4.08**), CMC 131kDa (**4.58**), HPMC 86kDa (**4.48**)		PC 11<=0.22 narrows the H6 group to classify three interactive protons across polymer species from inert	Fig 5B
H8	Protons across species are bimodally distributed on chemical composition. One distribution contains a secondary interactive proton, which suggests others in its cluster may have propensity for secondary interaction. The second distribution contains purely inert protons.	CMC 131kDa (4.36, 4.25, 4.09, 3.93, **3.76**, 3.58, 3.35, 3.14) CMC 90kDa (4.58, 4.36, 4.25, 4.09, 3.94, 3.76, 3.58, 3.35, 3.14) DEX 150kDa (3.48, 3.72, 3.88, 4.02, 4.22, 5.3) HPC 370kDa (4.58), HPC 80kDa (3.14, 3.46) HPMC 120kDa (1.16, 3.08, 3.38, 3.71, 4.05) HPMC 86kDa (1.16, 3.08, 3.38, 3.71, 4.05)	PDMAEMA 10kDa (1.3, 2.09) PHPMA 40kDa (0.94, 1.16, 1.82, 3.04, 3.19, 3.92) PVP 55 (1.54, 1.78, 2.03, 2.27, 2.51, 3.22, 3.6, 3.89), PVP 1300kDa (1.54, 1.78, 2.03, 2.27, 2.51, 3.22, 3.6, 3.89), P407 13kDa (1.19, 3.47, 3.54, 3.6, 3.76), PEG 2, 10, 20 kDa (3.7) PEOZ 50kDa (1.01, 2.22, 2.32, 2.41, 3.42, 3.62) PVA 105kDa (2.12)	PC15 <= -0.84 classifies the secondary interaction site in CMC131kDa, similar protons cluster nearby	Fig 6B

† **Boldface** indicates a true interactive proton. Underline indicates an "undervalued" inert proton candidate for physical property tuning towards interaction, without additional chemical functionalization.

7. Decision Tree Rule Plots & Interpretations

7.1. Rule 1

Heuristic 1

The first heuristic applied PC12 > 1.7 to classify all three protons in PAA as interactive (Fig 3B). PAA scored highly on PC 12 due to its cohort chemical shifts (1.5, 1.6] and (1.2, 1.3], and as PAA possessed the highest molecular weight in the dataset where interactions result (450kDa). PAA CDE was additionally correlated to the model's classification. In PAA, positive correlation with CDE is likely a function of PAA proton buildup curves each having positive linear slope, a characteristic of interactive protons. All three PAA protons scored similarly on PC12 relative to the other protons, which reflected their nonspecific interaction mechanism of chain interpenetration and entanglement (8). Some research directions derived from Heuristic 1 that merit investigation include molecular weight as a contributing factor to PAA's interaction mechanism, and whether reductions in molecular weight, or functionalization of PAA with more downfield chemical shifts negatively factor loaded to PC12 silences interactions. In this regard, evidence has been reported of linear 450kDa PAA weakening its mucoadhesion by decreasing its molecular weight to 100kDa (9).

The positive factor loadings in PC12 underlying this analysis were, in ascending order: 0.11 molecular weight, 0.19 (1.5, 1.6] chemical shift, 0.24 CDE, and 0.35 for the (1.2, 2.3] chemical shift. Negative factor loadings in PC12, oppositely correlated to PAA mucoadhesion, were: -0.1 ppm, -0.1 (3.0, 3.1], -0.11 (3.1, 3.2], -0.14 (4.3, 4.4], -0.17 (3.4, 3.5], -0.18 (3.8, 3.9], -0.18 (3.2, 3.3], -0.19 (2.5, 2.6], -0.19 (4.4, 4.5], -0.19 (0.9, 1.0], -0.2 (2.8, 2.9], -0.2 (1.7, 1.8], -0.21 (2.2, 2.3], -0.22 (4.5, 4.6], -0.22 (1.4, 1.5], -0.25 (4.0, 4.1], -0.27 (1.1, 1.2], -0.36 (2.1, 2.2].

7.2. Rule 2

Heuristic 2

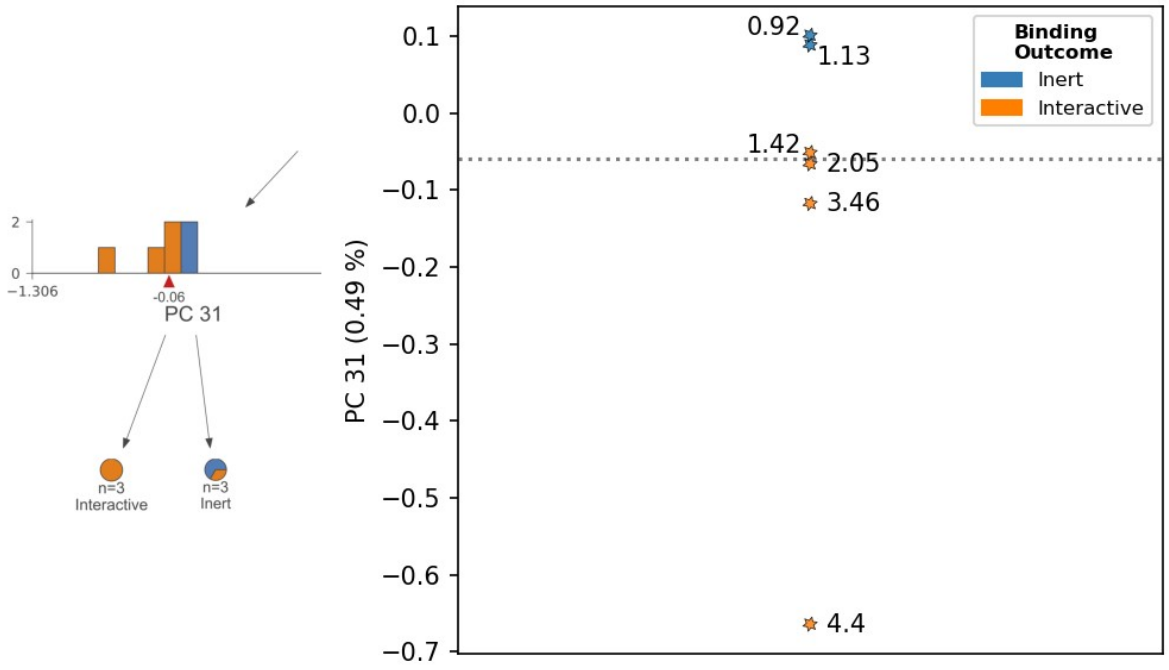
Heuristic 2 was constructed for PDMAEMA protons, which had characteristically low scores on PC3 (<= - 3.171). Figure 4A annotates the decision tree at this juncture. Figure 4B then shows the principal component biplot of protons classified by Heuristics 1 and 2, annotated with their true binding outcomes, and δ_1H chemical shifts. Chemical composition and CDE both amplified the low PC3 scores of PDMAEMA protons. As with PAA, positive linear buildup curve slope contributed to PDMAEMA proton CDE. Juxtaposing PDMAEMA protons with low scores on PC3, a second inert proton cluster from PEOZ scored highly on PC3. With this, Heuristic 2 frames PDMAEMA and PEOZ as "opposites" in both their proton chemical composition and propensity for interaction.

Underlying Heuristic 2, the negative factor loadings in PC3 score (associated with PDMAEMA) were: -0.11 (CDE), -0.14 (1.1, 1.2], -0.18 (4.3, 4.4], -0.22 (2.0, 2.1], -0.22 (0.9, 1.0], -0.23 (1.2, 1.3], -0.24 (1.4, 1.5], -0.24 (2.8, 2.9]. Positive factor loadings on PC3 (associated with PEOZ) were: 0.1 (3.4, 3.5], 0.14 (2.2, 2.3], 0.38 (3.6, 3.7], 0.38 (2.3, 2.4], 0.38 (2.4, 2.5], 0.38 (1.0, 1.1]. In terms of research directions from Heuristic 2, further exploration of PDMAEMA and PEOZ for "opposing interaction behavior triggers" is merited. Specifically, functionalizing PDMAEMA with chemical shifts positively factor loaded on PC3, or PEOZ with chemical shifts negatively loaded to PC3. Recently, the functionalization of PEOZ by methacrylation, and application of PEOZ in combination with Carbopols® both served to trigger its mucoadhesion (10,11).

7.3. Rule 3

Heuristic 3

Heuristic 3 was constructed within the PDMAEMA proton subset identified by Heuristic 2. Interactive PDMAEMA protons were separated from inert using PC31 (PC31<= -0.06), as shown in ESI Fig. 4. All applicable factor loadings in PC31 to this subset were minute (<0.1), except for one PDMAEMA cohort proton group, 0.3 (4.3, 4.4]. Nonetheless, PC31 globally ranked the protons: 0.92 > 1.13 > **1.42** > **2.05** > **3.46** > **4.4**, where monotonic increases in chemical shift downfield characterized interactive from inert sites (interactive bolded). Proton chemical shift is related to the electronic environment of neighboring functional groups, where an increase in chemical shift is a result of proton deshielding. Hence, it is possible the model reflected that a reduction in functional group shielding (for example by the introduction of groups with high electronegativity) increased propensity for interaction in PDMAEMA, as a short chain linear polymer.

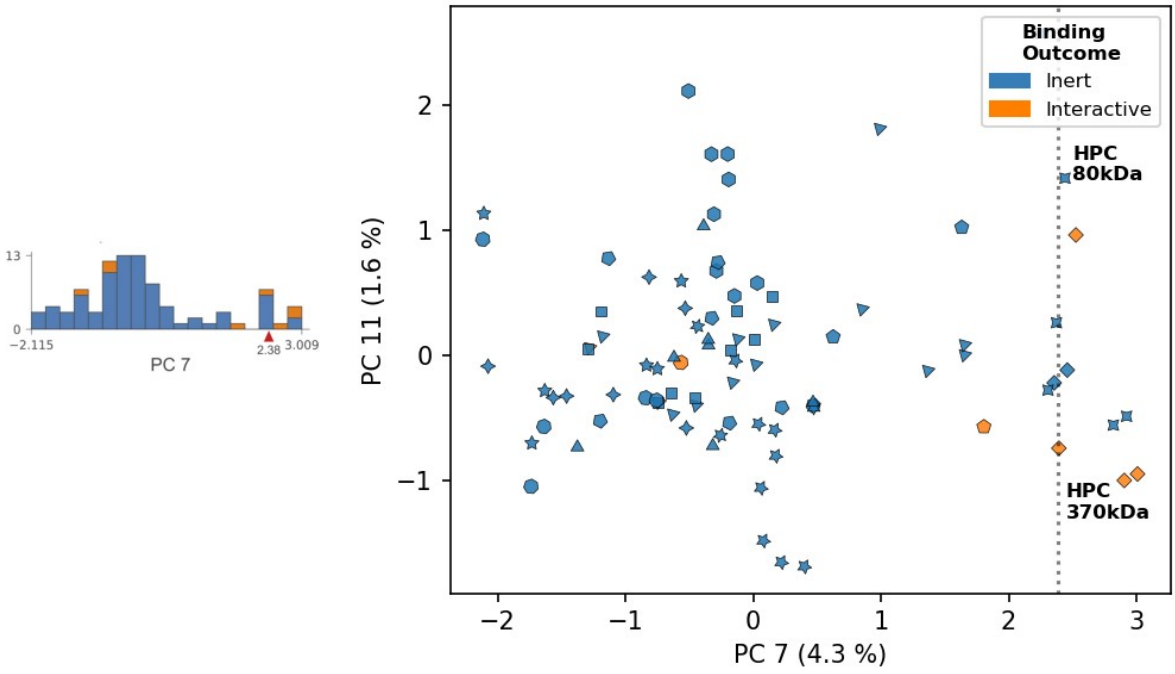


Supplementary Fig 4: Principal component plot of Heuristic 3. A) Decision tree classification node. B) Principal component scores of each proton annotated with ^{1}H Chemical Shift.

7.4. Rule 4

Heuristic 4

The model identified and separated eight HPC protons with mechanistic similarities by maximizing scoring on PC 7 (PC7 > 2.384) in Heuristic 4, depicted in ESI Fig. 5. Three proton chemical shifts from HPC at 80kDa (1.13, 3.77, 4.07ppm), and five from 370kDa (**1.13**, 3.14, **3.46**, **3.77**, **4.07**ppm), were grouped together. The positive PC7 factor loadings underlying these high scores reflected HPC's chemical makeup: 0.1 molecular weight, 0.16 (2.1, 2.2], 0.21 (1.1, 1.2], 0.26 (3.1, 3.2], 0.29 (3.4, 3.5], 0.3 (4.5, 4.6]. In turn, negative PC7 factor loadings were the chemical shifts characteristic of other polymers in the dataset, but absent in HPC's binding region: -0.11 (1.4, 1.5], -0.11 (2.8, 2.9], -0.14 (3.8, 3.9], -0.14 (0.9, 1.0], -0.18 (3.9, 4.0], -0.2 (4.3, 4.4], -0.21 (5.1, 5.2], -0.21 (5.2, 5.3], -0.26 (4.2, 4.3], -0.31 (4.4, 4.5], -0.31 (3.0, 3.1], -0.34 (3.3, 3.4].



Supplementary Fig 5: Principal component plot of Heuristic 4. A) Decision tree classification node. B) Principal component scores of each proton labeled where HPC protons are sectioned to generate Heuristic 5. Marker shapes correspond to each unique polymer sample.

7.5. Rule 5

Heuristic 5

Heuristic 5 revealed that tuning molecular weight, without additional chemical functionalization, unlocked interaction in HPC. The decision rule PC11 <=0.65 cleaved Heuristic 4's proton subset into two groups, primarily by molecular weight. HPC 370kDa achieved stable mucoadhesive interactions at **4.07**, **3.77**, **3.46**, and **1.13** ppm, and remained inert at 3.14ppm. No interactions resulted at any HPC 80kDa molecular weight protons. Additionally, the average CDE of protons below the decision boundary was lower than those above it, and ppm were shifted more downfield (avg. CDE $_{PC11<=0.65}$ =-0.74, avg. CDE $_{PC11>0.65}$ =-0.62), (avg. ppm $_{PC11<=0.65}$ =3.77ppm, avg. ppm $_{PC11>0.65}$ =2.64ppm). The applicable factor loadings and their signs reflected these trends: -0.19 (1.1, 1.2], -0.49 ppm, and -0.59 molecular weight, 0.39 CDE, 0.2 (3.7, 3.8], 0.13 (3.4, 3.5].

7.6. Rule 6

Heuristic 6

Heuristic 6 was the first to draw cross-species comparison, PC9 > 1.03 sectioned off seven protons from various polymers which displayed similar propensities for interaction (Fig. 4B). These protons were (interactive in bold): CMC 131kDa **4.58**ppm, CMC 90kDa 4.58ppm, HPMC 86kDa **4.48**ppm, HPMC 120kDa 4.48ppm, DEX 5.20ppm, PVA: **4.08,** 1.58ppm. Protons scoring above the threshold in Heuristic 6 possessed on average more downfield chemical shifts (avg. ppm $_{9>1.03}$=4.14ppm, avg. ppm $_{9<=1.03}$=3.14ppm), high CDE (avg. CDE $_{9>1.03}$= 3.63, avg. CDE $_{9<=1.03}$= -0.41), and lower molecular weights (avg. $MW_{9>1.03}$=111kDa, avg. $MW_{9<=1.03}$=212.3kDa), than those below it.
Reflecting these trends, the applicable factor loadings on PC9 and their signs underlying this classification were: 0.14 (1.5, 1.6], 0.48 CDE, 0.49 ppm, and 0.58 (2.1, 2.2], and -0.12 molecular weight, -0.12 (3.7, 3.8], -0.19 (3.4, 3.5], -0.19 (4.4, 4.5], and -0.21 (4.5, 4.6]. The high weighting of the (1.5, 1.6] and (2.1, 2.2] cohort groups explain the elevated PC9 scorings of PVA protons relative to the others classified, as PVA was the only material to contain both cohort shifts.

7.7. Rule 7

Heuristic 7

Applying the rule PC11 <= 0.22 in Heuristic 7 perfectly classified three dominant interactions from three distinct polymer species, within the subset identified by Heuristic 6. These were: CMC 131kDa **4.58**ppm, HPMC 86kDa **4.48**ppm, PVA: **4.08**ppm. A reduction in CDE across the decision border (avg. $CDE_{11>0.22=}$ 5.83, avg. $CDE_{11<=0.22}$= 0.704) provided the fine resolution necessary to distinguish dominant interactions within this mechanistically similar proton group, across species and molecular weight. Molecular weights and chemical shifts were similar between the groups, indicating CDE primarily enabled classification (avg. $MW_{11<=0.22}$=107.3kDa, avg. $MW_{11>0.22}$=113.8kDa) (avg. $ppm_{11<=0.22}$=4.38, avg. $ppm_{11>0.22}$=3.96). The applicable factor loadings on PC11 to the proton sample and their signs underlying this classification reflect the observed trends: 0.12 (4.0, 4.1], 0.39 CDE and -0.49 ppm, 0.59 molecular weight.

7.8. Rule 8

Heuristic 8

Heuristic 8 pertains to the remaining unclassified protons in the dataset, which are bimodally distributed in two clusters along PC15 (Fig. 5A). PC15 scores correlated predominantly to cohort chemical shift patterns (ESI, Supplementary Table 4), without loadings above 0.1 in magnitude for CDE, ppm, or molecular weight. Overall, the protons exhibited inert interactions, with the exception of one proton, a secondary interaction from CMC 131kDa at 3.76ppm. The decision rule PC 15<=-0.84 partitioned three protons from the smaller cluster (Fig. 5B), including the interactive site (interactive in bold): CMC 131kDa **3.76**ppm, HPMC 120kDa 3.71ppm, PVP 55kDa 1.54ppm. The three protons segregated by the decision rule demonstrated on average similar chemical shifts (avg. $ppm_{15<=-0.84}$=3.0, avg. $ppm_{15>-0.84}$=3.15), higher CDE (avg. $CDE_{15<=-0.84}$ = 0.40, avg. $CDE_{15>-0.84}$ = -0.43), and lower molecular weight (avg. $MW_{15<=-0.84}$=102, avg. $MW_{15>-0.84}$=217) than the inert proton bulk opposite the decision border. It is possible the model identified that these characteristics correlate to propensity for secondary interaction at these sites in their respective species.

Broader examination of the cluster containing the decision rule revealed recurring proton pairings from the same parent polymer with chemical shifts in the (4.0, 4.1] and (3.7, 3.8] intervals. These pairings were: CMC 131kDa (4.09ppm, **3.76**ppm), CMC 90kDa (4.09ppm, 3.76ppm), DEX 150kDa (4.02ppm, 3.72ppm), HPMC 86kDa (4.05ppm, 3.71ppm), and HPMC 120kDa (4.05ppm, 3.71ppm). P407 at 3.76ppm additionally clustered, without a (4.0, 4.1] shift. The molecular weight gated structure-activity relationship shared by DEX, CMC, HPC, and HPMC identified in discussion of Heuristics 6 & 7 can be expanded to include any influences exerted by, or secondary interactions occurring at (3.7, 3.8] shifts in this regard. Unintuitively, the 1.54ppm shift from PVP 1300kDa, and two sites from PHPMA (1.82ppm, 0.94ppm) additionally clustered in this group associated with secondary interaction.

The PC15 factor loadings and signs underlying Heuristic 8 reflect the observed trends: 0.11 (4.3, 4.4], 0.11 (2.8, 2.9], 0.13 (1.4, 1.5], 0.19 (1.8, 1.9], 0.21 (0.9, 1.0], 0.34 (1.5, 1.6], 0.38 (4.0, 4.1], 0.43 (3.7, 3.8], and -0.12 (2.0, 2.1], -0.13 (2.1, 2.2], -0.13 (4.4, 4.5], -0.2 (3.0, 3.1], -0.21 (1.2, 1.3], -0.22 (3.3, 3.4], -0.25 (3.4, 3.5], and -0.38 (3.1, 3.2].

8. References

1. Wang AYT, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. Chemistry of Materials. 2020;32(12):4954–65.
2. Mudali D, Roerdink JBTM, Teune LK, Leenders KL, Renken RJ. Comparison of decision tree and stepwise regression methods in classification of FDG-PET brain data using SSM/PCA features. In: Proceedings of the 8th International Conference on Advanced Computational Intelligence, ICACI 2016. IEEE; 2016. p. 289–95.
3. Sanchez-Lengeling B, Roch LM, Perea JD, Langner S, Brabec CJ, Aspuru-Guzik A. A Bayesian Approach to Predict Solubility Parameters. Adv Theory Simul [Internet]. 2019 Jan 1 [cited 2021 Dec 8];2(1):1800069. Available from: https://onlinelibrary-wiley-com.myaccess.library.utoronto.ca/doi/full/10.1002/adts.201800069
4. Efron B. The Jackknife, the Bootstrap and Other Resampling Plans. The Jackknife, the Bootstrap and Other Resampling Plans. 1982 Jan;
5. Kwaria RJ, Mondarte EAQ, Tahara H, Chang R, Hayashi T. Data-Driven Prediction of Protein Adsorption on Self-Assembled Monolayers toward Material Screening and Design. ACS Biomater Sci Eng [Internet]. 2020 [cited 2022 Mar 6];6(9):4949–56. Available from: https://dx.doi.org/10.1021/acsbiomaterials.0c01008
6. Mantovani RG, Horváth T, Cerri R, Junior SB, Vanschoren J, de Carvalho ACP de LF. An empirical study on hyperparameter tuning of decision trees. ArXiv [Internet]. 2018 Dec 5 [cited 2022 May 12]; Available from: https://arxiv.org/abs/1812.02207v2
7. Hastie T et. all. The Elements of Statistical Learning [Internet]. Vol. 27, The Mathematical Intelligencer. 2009 [cited 2022 May 15]. 83–85 p. Available from: http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf
8. Watchorn J, Burns D, Stuart S, Gu FX. Investigating the Molecular Mechanism of Protein-Polymer Binding with Direct Saturation Compensated Nuclear Magnetic Resonance. Biomacromolecules. 2022 Oct 14;23(1):67–76.
9. Lam HT, Zupančič O, Laffleur F, Bernkop-Schnürch A. Mucoadhesive properties of polyacrylates: Structure – Function relationship. International Journal of Adhesion and Adhesives. 2021 Jun 1;107:102857.
10. Shan X, Aspinall S, Kaldybekov DB, Buang F, Williams AC, Khutoryanskiy V V. Synthesis and Evaluation of Methacrylated Poly(2-ethyl-2-oxazoline) as a Mucoadhesive Polymer for Nasal Drug Delivery. ACS Appl Polym Mater. 2021 Nov 12;3(11):5882–92.
11. Ruiz-Rubio L, Alonso ML, Pérez-álvarez L, Alonso RM, Vilas JL, Khutoryanskiy V V. Formulation of Carbopol®/poly(2-ethyl-2-oxazoline)s mucoadhesive tablets for buccal delivery of hydrocortisone. Polymers. 2018 Feb 11;10(2):175.