Shape-Aware Machine Learning for Transcription Factor Binding Site Recognition

Daria Goptsii
daria.goptsii@mail.mcgill.ca
School of Computer Science, McGill University
Montréal, Canada

Riv Beauchesne
charles.beauchesne@mail.mcgill.ca
School of Computer Science, McGill University
Montréal, Canada

Abstract

Understanding the mechanisms of transcription factors (TFs) binding specificity is critical for gene regulation, as it determines when and where genes are expressed in response to developmental and environmental cues. We developed a machine learning classifier using logistic regression to successfully predict bound TF binding sites from unbound ones. Adding DNA shape features to our model improved significantly the accuracy compared to sequence motif only. We further expanded our hypothesis by adding 9 additional DNA shape features to have a distinct 13 DNA shape features model with sequence motif scoring. This model was the most accurate with a 99% accuracy.

CCS Concepts

• Computing methodologies → Machine learning; **• Applied computing →** *Computational biology*.

Keywords

Transcription factors, DNA shape, motif models, logistic regression, machine learning, genomics

1 Introduction

Understanding how proteins recognize specific regions of DNA is one of the central questions in molecular biology. Transcription factors (TFs), in particular, play a key role in regulating gene expression by binding to specific DNA sequences. While DNA is often depicted as a uniform, idealized double helix, the physical structure of DNA is much more complex and variable. Local variations in DNA shape, including the geometry of base pairs, minor groove width, and helical twist, can influence how TFs recognize and bind to their target sequences.

This structural diversity is critical for gene regulation. TFs must locate their target sequences among millions of possible sites in the genome, and subtle differences in DNA shape can help guide this search. Previous studies have shown that incorporating DNA shape features into predictive models can improve the accuracy of TF binding site identification beyond what is possible with sequence motifs alone [1]. However, the contribution of individual shape features, and how they complement sequence information, remains incomplete.

In this project, we aim to examine whether structural features of DNA can reliably distinguish TF binding sites from unbound regions. By combining experimentally derived binding data with DNA structural information, we seek to determine the relative importance of different features in predicting binding. Specifically, we consider a set of 13 DNA shape features, including six intra–base-pair parameters, six inter–base-pair parameters, and minor groove width (MGW), alongside sequence motif scoring.

Through this analysis, we aim to quantify the extent to which DNA shape enhances the predictive power of computational models of TF binding. Our approach allows us not only to improve classification accuracy but also to gain insight into which structural features most strongly influence TF–DNA recognition. This study provides a framework for integrating sequence and shape information to better understand the determinants of TF binding specificity.

From a machine learning perspective, we cast TF binding prediction as a supervised binary classification problem. Each 100 bp genomic window is encoded as a fixed-dimensional feature vector and labeled as either bound (positive) or unbound (negative). Within this framework, we systematically compare several standard classifiers used in high-dimensional settings, including Logistic Regression, a Multi-Layer Perceptron, and a Linear SVM. For each model, we train and evaluate separate instances on PWM-only features, DNA shape-only features, and combined shape-plus-PWM features. This design allows us to quantify how much predictive signal is carried by sequence motifs versus DNA structure, and to assess whether more flexible models provide any advantage over an interpretable linear baseline in this setting.

2 Feature Extraction

2.1 Positive windows (label = 1)

Positive samples correspond to experimentally validated binding sites for the target transcription factor, in our case `CTCF`. We obtain the genomic coordinates of these sites from
`factorbookMotifPos.txt`
and retain only those entries associated with the selected TF. For each binding event, the midpoint of the reported interval is computed and used as the centre of a window of fixed length $W = 100$ bp. This window spans 50 bp upstream and downstream of the centre, ensuring that potential flanking influences on TF binding are preserved. Before extraction, boundary checks confirm that the window lies entirely within chromosome limits. The corresponding DNA sequence is then retrieved from the loaded genome and converted to uppercase to maintain consistency. Each such window is assigned the positive class label. For computational efficiency during development, the number of positive windows can optionally be restricted by subsampling.

2.2 Negative windows (label = 0)

Negative samples are generated to represent regions where the TF is not expected to bind. These are drawn from ENCODE regulatory regions provided in
`wgEncodeRegTfbsClusteredV3.GM12878.merged.bed`.
Only regions located on chromosomes for which sequence data are available are considered. A region is selected at random, and a starting coordinate is uniformly sampled such that a full 100 bp window fits entirely inside the region. The candidate window is then compared against all positive windows on the same chromosome, and it is discarded if any overlap is detected. This strict non-overlap constraint prevents the model from learning ambiguous or borderline regions. The DNA sequence for each accepted window is extracted in the same manner as for the positives. To maintain class balance, we generate an equal number of negative and positive samples.

2.3 PWM score

Each window is further characterized by a motif-based feature derived from a position weight matrix (PWM) specific to the target transcription factor. The PWM, loaded from
`factorbookMotifPwm.txt`,
contains log-odds values for each nucleotide at every motif position. To compute a binding affinity estimate for a given sequence window, the motif is slid across the entire 100 bp sequence, and a score is computed for each possible alignment. The maximum PWM score across all alignments is retained as a single summary feature. This score provides an interpretable measure of how well the window sequence resembles the canonical TF motif.

2.4 DNA shape features

In addition to sequence information, we incorporate four DNA shape features—minor groove width (MGW), Roll, propeller twist (ProT), and helix twist (HelT)—which capture local structural properties of the DNA backbone. These features are widely used because they reflect how transcription factors interact with the three-dimensional conformation of the double helix. As described by Yang et al. [3], “DNA shape features include minor groove width (MGW), Roll, propeller twist (ProT), and helix twist (HelT) and are predicted with our DNAshape approach”. These four features form the canonical first-order DNA shape representation derived from Monte Carlo simulations [4], and they provide a compact but biologically meaningful structural profile for each nucleotide position.

Following our main hypothesis, we included 9 additional DNA shape features: five intra–base-pair features (Shear, Stretch, Stagger, Buckle, Opening) and four inter–base-pair features (Shift, Slide, Rise, Tilt). These additional features offer a more in-depth view of the DNA shape. They were also used in a study very similar to this one, where the authors found better results when including all of them in their model [1]. We decided to try to replicate their findings.

2.5 Choice of 100 bp window

Although the core CTCF motif is relatively short (approximately 15–20 bp), we use a much larger 100 bp window because transcription factor binding is influenced by the surrounding nucleotide context. Structural and sequence-based signals in the bases adjacent to the motif contribute to TF–DNA recognition, and [3] show “the contributions of DNA shape readout across diverse TF families and its importance in core motif-flanking regions”. Including ±50 bp around the motif therefore allows the model to capture not only the conserved core sequence but also the nearby shape and sequence features that influence binding specificity.

The choice of a 100 bp context is also supported computationally. Whitfield et al. [2] observe that the motif scoring tool POSSUM “computes log-likelihood scores using local nucleotide abundances within a 100-bp window”, and report that “adjusting the size of this window had little effect upon the AUC computed for a given PWM”. Their results justify using 100 bp as a robust window size.

Finally, from a modelling perspective, all training examples must have the same length so that they can be converted into feature vectors of equal dimensionality. Using a fixed 100 bp window ensures that both sequence features and DNA shape profiles can be represented consistently across all positive and negative samples.

2.6 Final dataset

The final dataset contains one row per genomic window, with columns specifying the chromosome, start and end coordinates, the 100 bp DNA sequence, the PWM-derived binding score, and DNA shape feature profiles. Each window is assigned either a positive or negative label depending on whether it overlaps a known binding site. The dataset is exported to CSV format and later used as input for machine learning models designed to predict TF binding.

3 Model Selection

To evaluate how well different learning algorithms capture the relationship between sequence or shape features and transcription factor binding, we explored three model families: Logistic Regression, a Multi-Layer Perceptron (MLP) trained with the Adam optimizer, and a Linear Support Vector Machine (Linear SVM). Each of these models was tested across multiple hyperparameter configurations, but all share the desirable property of being well-suited for high-dimensional binary classification tasks such as TF binding prediction.

To compare the performance of the models, we used a balanced dataset of 10,000 windows (5,000 positive and 5,000 negative) and evaluated performance across several feature subsets. The three feature sets were:

(1) **PWM_only**, containing a single PWM-derived score per window;
(2) **Shape_only**, containing the four DNA shape profiles concatenated into a 400-dimensional vector; and
(3) **Shape_plus_PWM**, a 401-dimensional feature vector combining shape features with the PWM score.

These subsets allowed us to test whether PWM information alone, shape information alone, or a combination of both provides the strongest predictive result.

The results for Logistic Regression, MLP, and Linear SVM on this reduced dataset are summarized in Tables 1, 2, and 3 in the Appendix. Across all models, the Shape-only and Shape+PWM feature sets performed substantially better than PWM-only, highlighting the strong predictive value of DNA shape features. Performance across Logistic Regression, MLP, and Linear SVM was very similar on this reduced dataset when using Shape-only or Shape+PWM features:

all models achieved ROC–AUC values between approximately 0.97 and 0.99. Therefore, we selected Logistic Regression for our further experiments due to its lower complexity in comparison to MLP and Linear SVM model.

4 Logistic Regression with Four DNA Shape Features

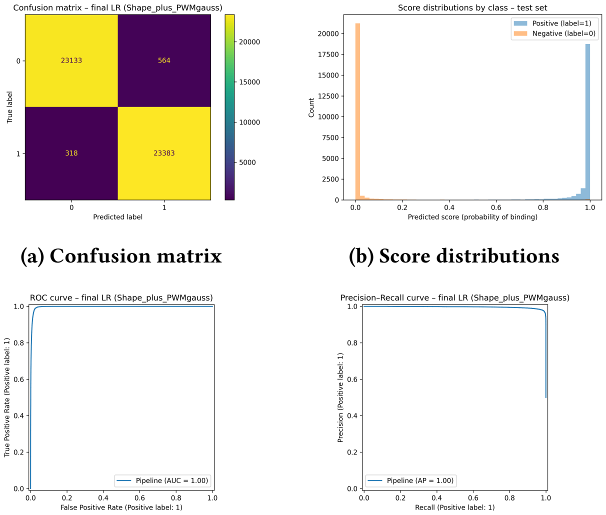
Having selected Logistic Regression as our primary model family, we next evaluated its performance on different feature representations of the CTCF binding dataset. In particular, we considered five feature sets derived from the same windows: *PWM_only*, *PWM_gauss*, *Shape_only*, *Shape_plus_PWM*, and *Shape_plus_PWMgauss*. The PWM_only representation uses a single scalar feature per window, namely the maximum PWM score of the CTCF motif. To obtain a slightly richer but still low-dimensional representation, we constructed *PWM_gauss* by expanding this scalar into a five-dimensional vector using Gaussian basis functions. Concretely, we placed five Gaussian kernels evenly across the observed PWM range and mapped each raw PWM value to its responses under these kernels.

The remaining feature sets rely on four DNA shape profiles (MGW, ProT, Roll, HelT) extracted from bigWig tracks. *Shape_only* concatenates these profiles into a 400-dimensional vector (4 features × 100 bp), while *Shape_plus_PWM* and *Shape_plus_PWMgauss* augment this vector with either the raw PWM score or its Gaussian expansion.

Before training, we applied standard preprocessing across all feature sets: windows containing any non-finite values in DNA shape arrays or PWM scores were removed, and all features were standardized within a `StandardScaler–LogisticRegression` pipeline to ensure well-conditioned optimization. We used a stratified split of the cleaned dataset, holding out 15% as a test set while the remaining 85% was used for training and validation. For each feature representation, we conducted a hyperparameter search over the inverse regularization strength $C \in \{0.01, 0.1, 1.0, 10.0\}$ using 5-fold stratified cross-validation with ROC–AUC as the scoring metric. The best estimator for each feature set was then retrained on the full train+validation partition and finally evaluated on the held-out test set.

The overall comparison across feature sets is summarized in Table 4 in the Appendix. From these results, we selected Logistic Regression with the *Shape_plus_PWMgauss* features, which gave the highest test ROC–AUC and strong cross-validation performance. We therefore retrained this configuration with the best regularization parameter ($C = 0.1$) on the full dataset for use in the final analysis. The resulting dataset statistics, test performance, and hyperparameters for this 4-feature model are reported in Table 5.

On the full dataset (315,985 samples; 405 features), *Shape_plus_PWMgauss* with $C = 0.1$ achieves test Accuracy 0.9814, ROC–AUC 0.9969, and PR–AUC 0.9962. The diagnostic plots in Figure 1 indicate that the final Logistic Regression model using the *Shape_plus_PWMgauss* features separates CTCF-bound and unbound windows with very high accuracy. The confusion matrix (Figure 1) shows only a small number of false positives and false negatives relative to the large number of correct predictions, and the score distributions demonstrate that positive and negative windows receive sharply distinct probability estimates with minimal overlap. The ROC and Precision–Recall curves further confirm that the classifier maintains strong performance across decision thresholds.



(a) Confusion matrix

(b) Score distributions

(c) ROC curve

(d) Precision–Recall curve

Figure 1: Diagnostics for Logistic Regression with 4 DNA shape features + PWM (Gaussian-expanded).

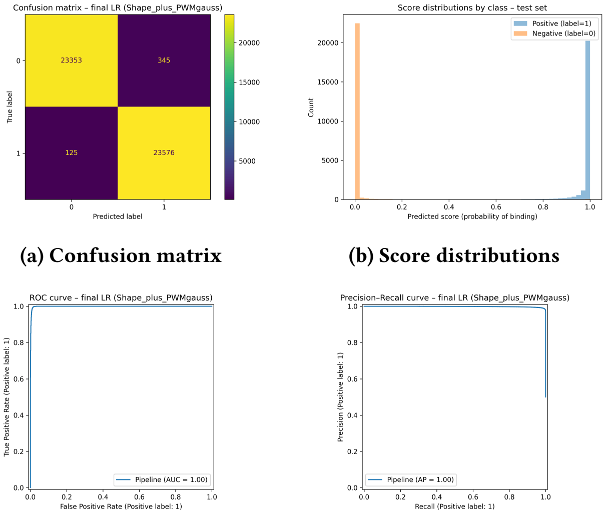
The contribution of each feature family in this 4-feature model (MGW, ProT, Roll, HelT, PWM (Gauss)) was quantified by summing the absolute standardized coefficients of the Logistic Regression model within each group. The corresponding totals appear in the left block of Table 7. HelT, Roll, and ProT provide the strongest influence on the decision boundary, while MGW and the PWM score contribute less overall. This pattern already suggests that DNA shape, and in particular helical and angular parameters, carries a larger share of the predictive signal than the sequence motif alone.

5 Logistic Regression with 13 DNA Shape Features

Considering that the accuracy of the *Shape_plus_PWMgauss* feature set with only four shape features on the CTCF binding dataset was the highest among the 4-feature models, we selected this configuration as the basis for adding further features. In particular, we incorporated nine additional DNA shape features: Rise, Shift, Slide, Tilt, Buckle, Shear, Stretch, Stagger, and Opening. These features were chosen from the same list used in [1], allowing us to compare our results directly with their findings and attempt to reproduce their observations within our own analysis.

Training the expanded Logistic Regression model with all 13 shape features plus PWM (Gaussian-expanded) substantially increased the dimensionality to 1305 features. On 315,991 samples and this 1305-dimensional feature space, the model attains test Accuracy 0.9901, ROC–AUC 0.9987, and PR–AUC 0.9985 with $C = 0.1$. The dataset statistics, test performance, and hyperparameters for this expanded model are reported in Table 6. Compared to the 4-feature model (Table 5), these results show a slight improvement in all metrics, indicating that the additional shape features provide complementary information that further sharpens the decision boundary between bound and unbound windows.

The diagnostic plots in Figure 2 indicate that the final expanded Logistic Regression model using the 13 shape features plus PWM (Gaussian-expanded) separates CTCF-bound and unbound windows with very high accuracy. As in the 4-feature case, the confusion matrix shows only a small number of misclassified examples relative to the total number of test samples, while the score distributions remain sharply separated between the two classes. The ROC and Precision–Recall curves are very close to the ideal top-left and top-right corners, respectively, confirming consistently strong generalization performance.



(a) Confusion matrix

(b) Score distributions

(c) ROC curve

(d) Precision–Recall curve

Figure 2: Diagnostics for Logistic Regression with 13 DNA shape features + PWM (Gaussian-expanded).

Although the running time of the program increased substantially when moving from 4 to 13 shape features (roughly from approximately 2 hours to 4 hours in our setting), we observed a modest but consistent improvement in the accuracy metrics. This trade-off between computational cost and marginal gains in performance is important for practical deployment scenarios, where a smaller subset of features may be preferable if it offers near-identical predictive power at lower cost.

The contribution of each feature family in the 13-feature model was again quantified by summing the absolute standardized coefficients within each group. The right block of Table 7 reports the total $|\beta|$ importance for all 13 shape features and the PWM (Gaussian-expanded). HelT, Slide, and Roll provide the strongest influence on the decision boundary, while Buckle and Shear contribute less overall. Taken together with the 4-feature results, these findings suggest that certain shape dimensions, particularly helix twist and lateral displacements (Slide, Rise), play an especially prominent role in how CTCF recognizes its binding sites.

6 Discussion and Future Directions

In this study, we applied Logistic Regression models to evaluate how well DNA shape features can predict TF binding sites compared to models relying only on sequence motifs. Across all experiments, models based solely on DNA shape consistently outperformed motif-only models, indicating that structural information provides a stronger signal for distinguishing bound from unbound sites. When we combined shape features with motif scores, prediction accuracy improved even further, suggesting that shape and sequence contribute complementary information. The feature-importance analysis supported this conclusion: DNA shape features received higher weights than the motif feature across all models, highlighting their central role in determining binding specificity.

Among the shape features, HelT emerged as one of the strongest and most consistent predictors, implying that variations in helix twist may play a substantial role in how CTCF binds to DNA. Furthermore, the nine additional shape features helped improve the model slightly, but were significantly more time and space consuming than the 4-feature configuration. Although including 13 features was excessive for our purposes, this expanded model allowed us to identify which features contributed most to prediction performance. In the case of CTCF, the five most influential features were HelT, Slide, Roll, MGW, and Rise. Using only these features would likely have been sufficient to achieve strong predictive performance, without the added complexity of the full feature set.

Our results are consistent with those reported in [1], particularly when using the full 13-feature model. This suggests that both approaches reliably capture the underlying determinants of TF binding. Although including all DNA shape features appears useful for estimating the relative contribution of each feature, it also introduces unnecessary complexity. In practice, it represents an overkill for predictive performance, since a smaller subset of features is sufficient to achieve comparable accuracy.

Similar observations have been reported in previous work, where DNA shape was shown to enhance TF-binding predictions beyond what motif models could achieve [3]. These studies also emphasize that proteins often “read” the three-dimensional structure of DNA in addition to the sequence. Together, these results reinforce the idea that TF binding cannot be fully explained by sequence motifs. Instead, local DNA geometry provides an essential context that helps determine whether a site is bound or not. Incorporating shape features into binding prediction models increases their accuracy and provides a more complete and realistic approach to understanding protein–DNA interactions.

Although our results highlight the value of DNA shape features in predicting TF binding, there is still room for further exploration. First, extending this analysis to multiple TFs would help to determine whether the effects found in this study are generalizable or TF-specific. Different proteins may have different preferences, and comparing models across TFs could reveal broader patterns in how DNA shape contributes to binding specificity.

Secondly, integrating dynamic information, such as how DNA shape fluctuates over time or under different cellular conditions, could provide a more complete picture of TF binding behavior. Exploring such temporal and contextual factors could contribute to a deeper understanding of the biophysical principles underlying protein–DNA recognition and could lead to more accurate predictive models.

Finally, future work could examine potential sources of bias and fairness issues in TF binding prediction, such as GC-content biases or differential performance across genomic contexts. Systematic evaluation of these factors would be important if shape-aware TF binding models were to be used in downstream applications such as variant effect prediction or regulatory element annotation.

7 Statement of Contribution

Daria Goptsii implemented DNA shape feature extraction, compared multiple machine learning models using appropriate evaluation metrics, and built and tuned the final Logistic Regression model. Riv Beauchesne implemented the PWM scoring algorithm and genome data parsing, and conducted additional experiments using the extended set of 13 DNA shape features. Both authors contributed to the final report and interpretation of findings.

References

[1] Jinsen Li, Jared M. Sagendorf, Tsu-Pei Chiu, Marco Pasi, Alberto Perez, and Remo Rohs. 2017. Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Research* 45, 22 (2017), 12877–12887. doi:10.1093/nar/gkx1145

[2] Troy W. Whitfield, Jie Wang, Patrick J. Collins, E. Christopher Partridge, Shelley Force Aldred, Nathan D. Trinklein, Richard M. Myers, and Zhiping Weng. 2012. Functional analysis of transcription factor binding sites in human promoters. *Genome Biology* 13, 9 (2012), R50. doi:10.1186/gb-2012-13-9-r50

[3] Lin Yang, Yaron Orenstein, Arttu Jolma, Yimeng Yin, Jussi Taipale, Ron Shamir, and Remo Rohs. 2017. Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Molecular Systems Biology* 13, 2 (2017), 910. doi:10.15252/msb.20167238

[4] Tianyin Zhou, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. 2013. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Research* 41, W1 (2013), W56–W62. doi:10.1093/nar/gkt437

Appendix

To fit the ACM two-column layout, wide tables summarizing model performance and feature-family importance are placed in the Appendix. The complete source code is available on GitHub.[1]

[1] https://github.com/rivbeau/final_project_COMP561

Table 1: Logistic Regression performance across feature sets (10k windows; 3-fold CV).

Feature Set	Model	Test Acc.	Test ROC–AUC	Test PR–AUC	CV ROC–AUC
PWM_only	LogisticRegression	0.7173	0.7973	0.7810	0.7892
Shape_only	LogisticRegression	0.9607	0.9917	0.9872	0.9885
Shape_plus_PWM	LogisticRegression	0.9647	0.9927	0.9884	0.9898

Table 2: MLP (Adam) performance across feature sets (10k windows; 3-fold CV).

Feature Set	Model	Test Acc.	Test ROC–AUC	Test PR–AUC	CV ROC–AUC
PWM_only	MLP_Adam	0.7267	0.7973	0.7810	0.7945
Shape_only	MLP_Adam	0.9713	0.9760	0.9729	0.9905
Shape_plus_PWM	MLP_Adam	0.9693	0.9814	0.9773	0.9913

Table 3: Linear SVM performance across feature sets (10k windows; 3-fold CV).

Feature Set	Model	Test Acc.	Test ROC–AUC	Test PR–AUC	CV ROC–AUC
PWM_only	LinearSVM	0.7167	0.7973	0.7810	0.7892
Shape_only	LinearSVM	0.9600	0.9777	0.9739	0.9889
Shape_plus_PWM	LinearSVM	0.9667	0.9829	0.9795	0.9901

Table 4: Logistic Regression summary across feature sets (CTCF, 100 bp windows; full dataset).

Feature set	**Acc**	**ROC–AUC**	**PR–AUC**	**CV ROC–AUC**	**#Feat**	**Best** *C*
PWM_only	0.754061	0.844289	0.875162	0.842276	1	0.01
PWM_gauss	0.768387	0.844289	0.875162	0.842276	5	0.01
Shape_only	0.972130	0.994667	0.993239	0.994417	400	0.10
Shape_plus_PWM	0.976856	0.996154	0.995350	0.996268	401	0.10
Shape_plus_PWMgauss	0.981392	0.996940	0.996287	0.997219	405	0.10

Table 5: Final expanded Logistic Regression (4 shape features + PWM, Gaussian-expanded).

Category	**Value**
Dataset Statistics	
Total samples used (after cleaning)	315,985
Number of features	405
Train size (85%)	268,587
Test size (15%)	47,398
Test Performance	
Accuracy	0.981392
ROC–AUC	0.996940
PR–AUC	0.996287
Model Hyperparameters	
Regularization parameter *C*	0.1
Penalty	L2
Solver	lbfgs
Max iterations	1000

Table 6: Final expanded Logistic Regression (13 shape features + PWM, Gaussian-expanded).

Category	Value
Dataset Statistics	
Total samples used (after cleaning)	315,991
Number of features	1305
Train size (85%)	268,592
Test size (15%)	47,399
Test Performance	
Accuracy	0.990084
ROC–AUC	0.998730
PR–AUC	0.998487
Model Hyperparameters	
Regularization parameter C	0.1
Penalty	L2
Solver	lbfgs
Max iterations	1000

Table 7: Feature-family importance (sum of absolute standardized coefficients).

Feature Group	Total $\|\beta\|$	Feature Group	Total $\|\beta\|$
4-feature model		*13-feature model*	
MGW	14.01	MGW	18.01
ProT	18.84	ProT	13.05
Roll	18.93	Roll	19.74
HelT	22.15	HelT	27.32
PWM (Gauss)	10.15	Rise	14.96
		Shift	3.58
		Slide	21.73
		Tilt	5.15
		Buckle	2.33
		Shear	3.01
		Stretch	5.48
		Stagger	8.10
		Opening	12.85
		PWM (Gauss)	10.14