Interpretable Signal Analysis with Knockoffs Enhances Classification of Bacterial Raman Spectra

Charmaine Chia, Matteo Sesia, Chi-Sing Ho, Stefanie S. Jeffrey, Jennifer Dionne,
Emmanuel J. Candès, and Roger T. Howe

***Abstract*—Sophisticated machine learning models are widely applied to signal data because they can detect complex patterns and leverage them effectively to make predictions. However, such models tend to be difficult to interpret, which is particularly concerning for critical biomedical applications, such as the identification of bacterial infections from spectroscopic data. Feature extraction and selection can identify structures in the data that are both informative and non-redundant, leading to simpler and more easily understandable models, without necessarily sacrificing predictive accuracy. In this paper, we present a signal classification method that combines wavelet-based feature extraction with a knockoff filter to control the false discovery rate. We apply the method to Raman spectroscopy data in order to classify bacterial samples. We show that the features thus obtained allow an intuitive logistic regression model to achieve predictive accuracy comparable to that of less understandable alternative approaches.**

***Index Terms*—Bacteria, Classification, Interpretability, Knockoffs, Machine learning, Raman spectroscopy, Signal.**

I. Introduction

New sensor technologies have contributed to the advent of “big data” in biomedicine, of which signal data are an important modality. From one-dimensional electrocardiography and electroencephalography signals from the heart and brain, to two-dimensional tissue images of tumor histology, to three-dimensional magnetic resonance images, these consist of sequential measures of an observable along one or more independent axes such as time, distance, or frequency. Signal data differ from structured forms of data in that the meaning of each independent variable is not as distinctively and intuitively definable. Informative features must be extracted from these raw data using signal processing and machine learning (ML) techniques before useful patterns can be detected and leveraged to make predictions.

While predictive accuracy is usually prioritized in ML, model interpretability is gaining attention. Interpretability is crucial when models inform the decisions of experts and can have serious consequences, such as in applications involving healthcare. Furthermore, when the signal source itself is not well-understood, interpretable models can yield deeper insights and facilitate inferences. Along these lines, the ML framework discussed in [1] proposes three metrics for evaluating models: 1) predictive accuracy (the goodness-of-fit to the underlying data), 2) descriptive accuracy (the fidelity of the interpretation in describing relations learned by the model), and 3) relevancy (the usefulness and comprehensibility of the interpretation to the target audience).

Simpler models (e.g., linear regression, trees, naïve Bayes) are easier to interpret, though often at the expense of predictive accuracy due to their limited flexibility. By contrast, sophisticated models such as neural networks [2–4] can automatically extract predictive features and capture complex relations in the data, but their “black-box” nature makes it difficult to understand their decisions. Various techniques have been proposed to improve the descriptive accuracy of ML models; for example, saliency methods help visualize the activation of individual input features [5], while attribution methods like LIME [6] and SHAP [7] quantify the impact of each feature on the output predictions. However, these *post hoc* techniques are inadequate for developing simpler models.

With regard to relevancy, studies report that people favor explanations that are short, contrast instances with different outcomes, and highlight abnormal causes [8]. In other words, we seek to understand which features are important, and how these affect the outcome. Data scientists often pursue these goals through feature selection, in addition to feature extraction, to ensure that their conclusions are based on relevant and non-redundant predictors. For example, one may want to identify a smaller set of genetic variants linked to disease susceptibility among thousands of possibilities [9], or to identify which specific morphological features from brain electroencephalogram signals can diagnose epilepsy [10].

There exist many variable selection methods [11, 12], but most lack reliable theoretical guarantees, in the sense that their output may include unexpected numbers of false discoveries—unimportant features that are either irrelevant or redundant (see the Appendix for a more precise definition of this concept). Ideally, one would like to avoid false discoveries altogether, since they may lead to misleading interpretations and incorrect

This work was partly supported by NSF grants DMS 1712800 and 1934578, a Math+X grant (Simons Foundation), and by the Stanford School of Engineering’s Catalyst for Collaborative Solutions.

C. Chia (email: charmaine@aetherbio.com) was in the Department of Electrical Engineering at Stanford University, Stanford, CA 94305, coadvised by R. T. Howe and S. S. Jeffrey. She is now with Aether Biomachines, Menlo Park, CA 94025 USA.

M. Sesia (email: sesia@marshall.usc.edu) was in the Department of Statistics at Stanford University, advised by E. Candès. He is now in the Department of Data Sciences and Operations at the University of Southern California, Los Angeles, CA 90089 USA.

C-S. Ho was in the Department of Applied Physics at Stanford University, advised by J. Dionne.

decisions, although a limited number of them can be tolerated. Therefore, a reasonable objective for feature selection is to control the false discovery rate (FDR): the expected proportion of false discoveries [13]. For feature selection within a linear regression setting, the FDR can be provably controlled with the *knockoff filter* [14]. This work was recently extended to a more general non-parametric regression setting [15], which includes the multi-class classification problem considered in this paper. The idea of [15] is to generate *knockoff* features that behave as the actual unimportant ones do, but are known with certainty to be redundant. Important features can then be selected by looking for those that significantly stand out from the knockoffs (see Appendix for a review of this method).

Knockoffs are attracting attention because they give strong statistical guarantees while preserving predictive power. However, most applications have focused on structured data, in which the features are well-defined a priori: single-nucleotide polymorphisms [9, 16–19], virus mutations [20], or demographic/behavioral cancer biomarkers [21], to name some examples. Only few extensions to unstructured data have been reported, namely involving computed tomography (CT) [22], functional magnetic resonance images [23], and economic time series [24]. Thus, the relatively unexplored area of unstructured data provides an interesting use case for knockoffs.

In this paper, we combine feature extraction and selection to obtain a powerful and interpretable signal analysis framework, and demonstrate its utility by applying it to a data set of fast Raman spectroscopy measurements of common bacteria collected at the Stanford Hospital [25]. Raman spectroscopy measures the interaction of laser light with a sample, producing a spectrum where peaks indicate wavelengths at which the light is strongly absorbed by the chemical bonds present therein. Thus, this technique yields an optical fingerprint of the sample. Fast Raman measurements follow the same principle, but their spectra are noisier and more difficult to recognize due to shorter measurement times. Therefore, reliable ML algorithms are useful to automate the recognition of such optical fingerprints. Recently, a convolutional neural network (CNN) was found to be successful at using these data to predict outcomes such as bacterial strain and antibiotic susceptibility [25]. These results are promising, since rapid and culture-free pathogen identification could advance the treatment of bacterial infections and sepsis. At the same time, such high-stakes medical decisions call for more interpretable models that can be easily examined and understood by humans (who, for instance, may wish to know the presence of which chemical bonds drives the machine decision).

Our approach begins with a feature extraction step that transforms the signal data into a more intuitive representation summarizing the presence of localized peaks in the spectra. Then, we apply the knockoff filter to select a subset of features that are likely to be predictive and non-redundant, and finally we use these to fit a simple multinomial logistic regression model that predicts the outcome of interest. Our analysis shows that the proposed method performs similarly to the CNN of [25] in terms of predictive accuracy (sometimes even better), while creating a more compact and interpretable model.

II. DATA SET

We test the proposed framework on data consisting of 60,000 Raman spectra of dried monolayer bacteria and yeast samples taken with fast (one-second) scans, from [25]. Thirty distinct isolates were measured, including multiple isolates of Gram-negative and Gram-positive bacteria, as well as Candida species; 2000 spectra were measured for each isolate, most of which were taken over single cells. The spectra consist of 992 measurement points evenly distributed in the spectral range of 381.98 to 1792.4 cm^{-1}. The measured Raman intensities were normalized to lie between 0 and 1. Further details about these measurements can be found in [25]. The data can be downloaded from https://github.com/csho33/bacteria-ID.

In addition to the Raman spectra (X), three sets of associated outcome labels (Y) are available from this data set:

1) Isolate labels → 30 classes;
2) Empiric antibiotic treatment → 8 classes;
3) Methicillin resistance of Staphylococcus aureus strains → 2 classes

To summarize, the sizes of the data matrices are:

- Raw signal data (X): $60,000 \times 992$;
- Outcome labels (Y): $60,000 \times 1$, except for the 3rd set of labels, which apply only to Staphylococcus aureus strains, giving a $10,000 \times 1$ outcome matrix.

III. METHODS

A. Feature extraction

Feature extraction is the transformation of raw data into a more discriminatory representation for the prediction task. There exist a variety of feature extraction methods for signal data, which can be categorized into four broad families [2].

1) Time/position methods extract characteristic properties from specific windows of measurement points.
2) Frequency methods break signals into their spectral components, giving information complementary to the above; e.g., the Fourier transform [26].
3) Time/position-frequency methods capture both frequency and time/position information in non-linear and non-stationary signals; e.g., the wavelet transform [27].
4) Sparse signal decomposition methods seek sparse data representations in terms of basis sets that are defined empirically; e.g., convolutional dictionary learning [28].

In general, different feature extraction methods may be better suited for different kinds of data, and they should be chosen based on their natural interpretability given the dynamics of the signal source or other relevant prior knowledge [2]. In our application, we opt for a discrete wavelet transform (DWT), which projects the signal X onto a compact orthogonal basis set of wave-like oscillations at different frequencies, beginning and ending with zero amplitude. The basis wavelet is a 24-point Coiflet with five DWT levels. The result of the transform, X', is a set of 1105 features for each of the 60,000 samples, which represent the concatenated approximation and detail coefficients from the five-level wavelet filtering procedure. Starting from the wavelet representation, the original signal can be reconstructed using an Inverse Discrete Wavelet

Transform (IDWT). Our wavelet approach is intuitive here because the data are spectroscopic, with expected peak-like resonance features [29]. In fact, spectroscopic signals are traditionally analyzed by fitting Gaussian or Lagrangian peaks, thus offering a natural interpretive basis for our wavelets.

B. Knockoff generation

We generate knockoffs for both the raw data (X) and the wavelet features (X') following the model-X method in [15], as implemented by the second-order knockoff machines in [20]; see the Appendix for some technical background on knockoffs. The Python code for this task is publicly available from: https://github.com/msesia/knockoff-filter. We apply this algorithm to generate knockoffs that are approximately pairwise exchangeable with the data in terms of their second moments. More precisely, we generate the knockoff features $\tilde{X} \in \mathbb{R}^{n \times p}$ given the original features $X \in \mathbb{R}^{n \times p}$ (the construction for X' is analogous, so we focus on X for simplicity) such that the mean vector and the covariance matrix of $[X, \tilde{X}]$ (the augmented data matrix defined by concatenating the columns of X with those of $\tilde{X}$) match those of $[X, \tilde{X}]_{\text{swap}(j)}$, for any $j \in \{1, \ldots, p\}$. (Above, $\text{swap}(j)$ is the operator that swaps X_j with $\tilde{X}_j$.) Simultaneously, we try to make each element of $\tilde{X}$ as different as possible from the corresponding element of X [15, 20], so that the knockoff filter will be powerful for feature selection [14]. By such construction, the covariance matrices of $[X, \tilde{X}]$ and $[X, \tilde{X}]_{\text{swap}(j)}$, for any j, are approximately equal to:

$$G = \begin{bmatrix} \Sigma & \Sigma - \text{diag}(s) \\ \Sigma - \text{diag}(s) & \Sigma \end{bmatrix}, \qquad (1)$$

where Σ is the covariance matrix of X and the vector s is maximized subject to the constraint that the matrix G be positive semi-definite [14, 15]. We refer to [15] and [20] for further details on knockoff generation. It is worth mentioning here that the method in [20] can accommodate a more general construction (*deep knockoffs*) that also matches higher moments of $[X, \tilde{X}]$ to those of $[X, \tilde{X}]_{\text{swap}(j)}$, which leads to a more robust variable selection procedure in some situations, but seems to make little difference with our data. Therefore, we focus on second-order knockoffs for simplicity.

C. Feature selection with the knockoff filter

Following the generation of knockoffs, the augmented raw and wavelet representation data, $[X, \tilde{X}]$ and $[X', \tilde{X}']$, are separately provided as input to a classifier (after standardizing all columns to have unit variance), which is trained on each of the three sets of Y labels. The number of features available to each model is thus twice that of the original data. The classification model is based on logistic regression with ℓ_1 (lasso) regularization [30], which results in sparse models with several coefficients equal to zero, thus already performing feature selection to some degree. For the 30-class isolate identification and 8-class antibiotic treatment classification tasks, we use a multinomial logistic regression model [31], which outputs a probability distribution across all the classes; the class with the largest estimated probability is taken as the final prediction. For the 2-class methicillin resistance classification, standard (binomial) logistic regression is used.

These models are fitted using the `glmnet` R package [32]. We denote the estimated model coefficients for each task by $\hat{\beta}_1(\lambda), \ldots, \hat{\beta}_{2p}(\lambda)$; the parameter λ controls the strength of the ℓ_1 penalty and is tuned by 10-fold cross-validation. The $\hat{\beta}_j(\lambda)$ and $\hat{\beta}_{j+p}(\lambda)$ coefficients are used to define a score, namely $W_j = |\hat{\beta}_j(\lambda)| - |\hat{\beta}_{j+p}(\lambda)|$, for each of the p original features, as explained further in the Appendix. Feature selection is then performed by selecting variables with $W_j \geq T$, where T is a data-adaptive threshold computed by the knockoff filter [14] to control the FDR below 10%, so that we can expect about 90% of the selected features to be important or redundant [15]. We denote by $\hat{S} \subseteq \{1, \ldots, p\}$, or $\hat{S}'$, the subset of features thus selected from X, or X', respectively.

D. Classification

We compare the predictive performance of the features selected using the knockoffs procedure, $\{X_j\}_{j \in \hat{S}}, \{X'_j\}_{j \in \hat{S}'}$, against that of the full sets of raw and wavelet features, X, X'. For this purpose, we train classification models based on ℓ_1-regularized (multinomial) logistic regression, as before, for all 12 prediction tasks arising from combination of these four input data sets and the three sets of output labels:

$$\begin{bmatrix} X \\ \{X_j\}_{j \in \hat{S}} \\ X' \\ \{X'_j\}_{j \in \hat{S}'} \end{bmatrix} \times \begin{bmatrix} Y_{30-\text{class}} \\ Y_{8-\text{class}} \\ Y_{2-\text{class}} \end{bmatrix}. \qquad (2)$$

Figure 1 summarizes our method. The out-of-sample performance of each model is assessed with 5-fold cross-validation. By comparing the results from these 12 prediction tasks, we can evaluate both (A) the effect of applying feature extraction, and (B) the effect of feature selection via the knockoff filter. Finally, we compare the performance of our models to previous results in [33], which were obtained using a CNN, a support vector machine (SVM), and logistic regression models based on different features.

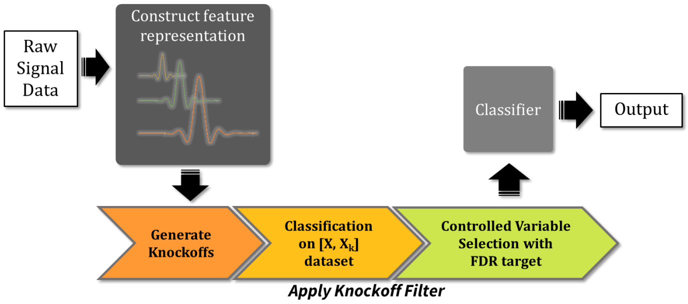
Raw Signal Data
Construct feature representation
Classifier
Output
Generate Knockoffs
Classification on [X, X_k] dataset
Controlled Variable Selection with FDR target
Apply Knockoff Filter

Fig. 1. Analysis framework for bacterial Raman spectra data. First, wavelet features are extracted from the raw data; then, knockoffs are used to select a predictive and non-redundant subset of them; finally, a simple classification model is fitted on the selected features.

IV. RESULTS AND DISCUSSION

Table I summarizes the classification errors obtained by our method for the three prediction tasks (30, 8, and 2

TABLE I
COMPARISON OF MODEL PERFORMANCES ON RAW AND WAVELET FEATURES, BEFORE AND AFTER FEATURE SELECTION.

Input	# of input features	# of nonzero coefficients	Test error (%)	Error s. d. (%)
30 classes				
X	992	457	7.4	0.3
$\{X_j\}_{j\in\hat{S}}$	982	454	7.3	0.2
X'	1105	328	6.7	0.3
$\{X'_j\}_{j\in\hat{S}'}$	152	121	5.3	0.2
8 classes				
X	992	552	5.5	0.2
$\{X_j\}_{j\in\hat{S}}$	913	532	5.5	0.3
X'	1105	436	5.3	0.2
$\{X'_j\}_{j\in\hat{S}'}$	103	93	4.9	0.2
2 classes				
X	992	687	7.2	0.6
$\{X_j\}_{j\in\hat{S}}$	551	526	6.8	0.5
X'	1105	239	6.1	0.7
$\{X'_j\}_{j\in\hat{S}'}$	63	63	5.7	0.5

TABLE II
COMPARISON OF MODEL PERFORMANCES ON RAW-PCA AND KNOCKOFF-FILTERED WAVELET FEATURES.

Input	# of input features	# of nonzero coefficients	Test error (%)	Error s. d. (%)
30 classes				
X_{PCA152}	152	117	6.1	0.3
$\{X'_j\}_{j\in\hat{S}'}$	152	121	5.4	0.2
8 classes				
X_{PCA102}	103	83	5.2	0.3
$\{X'_j\}_{j\in\hat{S}'}$	103	93	4.9	0.2
2 classes				
X_{PCA63}	63	29	6.4	0.8
$\{X'_j\}_{j\in\hat{S}'}$	63	63	5.7	0.5

classes), using each of the four input data sets: X, $\{X_j\}_{j\in\hat{S}}$, X', $\{X'_j\}_{j\in\hat{S}'}$. The rows corresponding to results involving the knockoff filter are shaded and arranged below those corresponding to results obtained without controlled variable selection. The third column counts the number of non-zero coefficients in the final regularized logistic regression model, which is fitted on the input features after tuning the parameter λ by 10-fold cross-validation.

A. Effect of feature extraction

To examine the effect of feature extraction, we compare the classification errors in Table I corresponding to the input data sets X and X' (i.e., the white rows), for each of the three tasks. In each case, we observe a decrease in test error, from 7.4% to 6.7% for the 30-class task, from 5.5% to 5.3% for the 8-class task, and from 7.2% to 6.1% for the 2-class task.

As an additional comparison, Table II reports the predictive performance (within a regularized logistic regression model) of our wavelet features next to that of features obtained by performing a component analysis (PCA) on the raw signal data. (Recall that PCA extracts directions with maximal variance in the data matrix.) To facilitate the comparison, the number of principal components is fixed to match the number of wavelet features selected by the knockoff filter for each classification task. Again, the wavelet features yield lower classification errors, which should not be very surprising given that they have a much more intuitive interpretation for our kind of data.

B. Effect of feature selection

To examine the effect of feature selection with the knockoff filter, we compare the classification errors in adjacent pairs of rows in Table I, for each of the three tasks. In general, we observe a significant improvement in classification error between X' and $\{X'_j\}_{j\in\hat{S}'}$, as the knockoff filter selects a subset of the wavelet features before the final classifier is trained. In particular, the test error decreases from 6.7% to 5.3% for the 30-class task, from 5.3% to 4.9% for the 8-class task, and from 6.1% to 5.7% for the 2-class task. These results are notable given that the number of features input into the classifier is reduced significantly—of the original 1055 wavelet features, we are left with only 152, 103, and 63 features for the respective tasks. It thus seems that, without the upstream feature selection with FDR control performed by the knockoff filter, the classifier may overfit the training data despite its own ℓ_1 regularization, causing poorer performance at test time.

The effect of controlled feature selection is less obvious when this is applied to the raw signal data, X. In this case, the dimensions of the knockoff-filtered data, $\{X_j\}_{j\in\hat{S}}$, are reduced only slightly compared to the original ones, and the resulting changes in classification error are minimal. This suggests that the raw signal features are on average less informative than the wavelet representation features, and cannot be directly "sparsified" without loss of predictive power. This is in contrast with the wavelet features, which are more informative and natural predictors, but are redundant. Therefore, knockoffs can sparsify the latter without sacrificing predictive accuracy (which, on the contrary, is even improved by this process).

C. Comparisons with other classifiers

Table III compares the performance of our framework—see Figure 1—with that of other classifiers applied to the same data. The benchmarks are a convolutional neural network (CNN), a support vector machine (SVM), and logistic regression (LR) without regularization [33]. The CNN is applied directly to the raw signals, while the SVM and LR take the top 20 principal components as input features [33]. We denote our framework as KWLR (knockoff-filtered wavelet logistic regression). Again, performance is assessed through 5-fold cross-validation.

Our proposed framework (KWLR) performs noticeably better for the 30-class task, which is the most difficult one, yielding a prediction error that is almost half that of SVM and LR. The performance of KWLR is worse than that of the CNN for the 8-class and 2-class prediction tasks (there are more examples per class than in the 30-class problem, and clearer distinctions between classes), likely because the

TABLE III
BENCHMARKING OF TEST ERRORS OBTAINED WITH OUR FRAMEWORK TO OTHER MODELS. RESULTS IN WHITE ROWS ARE QUOTED FROM [33].

Input	Classifier	# of input features	# of nonzero coefficients	Test error (%)	Error s. d. (%)
30 classes					
X	CNN	992	992	6.2	0.1
X_{PCA20}	SVM	20	20	11.3	0.2
X_{PCA20}	LR	20	20	10.7	0.2
$\{X'\}_{j\in\hat{S}'}$	KWLR	152	121	5.3	0.2
8 classes					
X	CNN	992	992	1.0	0.1
$\{X'\}_{j\in\hat{S}}$	KWLR	103	93	4.9	0.2
2 classes					
X	CNN	992	992	4.6	0.5
$\{X'\}_{j\in\hat{S}'}$	KWLR	63	63	5.7	0.5

LR classifier can only learn linear mappings from the input features to output classes, whereas the neural network can model complex, non-linear relationships. While beneficial for the easier 8-class and 2-class tasks, the additional flexibility of the CNN leads to overfitting when the problem is more statistically challenging. In contrast, when the data samples are limited or the signals are weak, neural networks and other black-box methods no longer have a clear predictive advantage over simpler and more interpretable models.

D. Visualization of the wavelet features

Figure 2(a) shows an example of a raw Raman signal, which we denote as $X^{(1)}$. From this, we extract wavelet features $X'^{(1)}$ and generate corresponding knockoffs $\tilde{X}'^{(1)}$. Figure 2(b) shows the IDWT projection of $\tilde{X}'^{(1)}$ back into the signal domain. We observe that the knockoff signal preserves some characteristics of the original signal, such as its general shape and noise pattern, but it is clearly distinct.

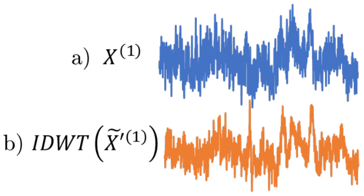

Fig. 2. (a) Raman signal, and (b) knockoff copy of its wavelet representation, projected back into the signal domain.

Figure 3 plots the correlations between the original features (a) and the cross-correlations between the original and knockoff wavelet features (b). The first 100 features or so, corresponding to the lower level DWT coefficients, show the strongest local (among adjacent features) cross-correlations (see insets), while most other features are approximately uncorrelated. The property in (1), which follows from the construction of the knockoffs, implies that Figure 3(b) should look very similar to Figure 3(a), except for the values on the diagonal, which can be lower in (b). This suppression of the diagonal values reflects our attempt to make the knockoffs as different as possible from the real features [15] (this can only be partly achieved for the first 100 wavelets because they have stronger correlations among themselves).

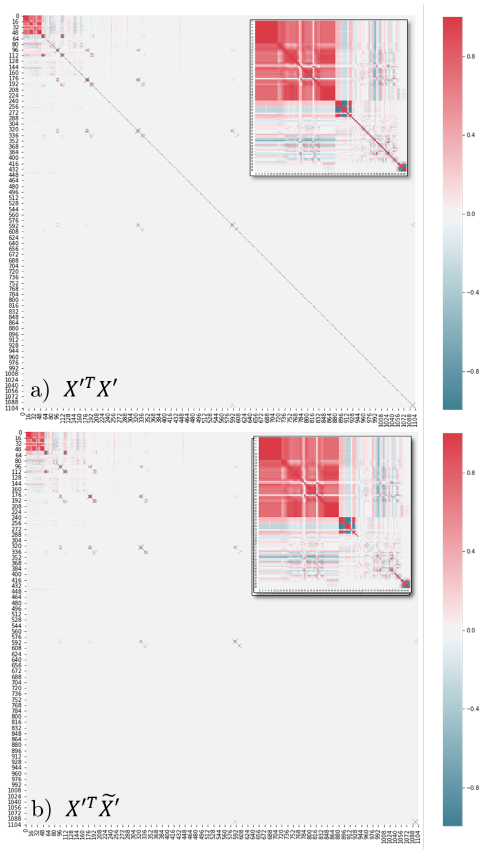

Fig. 3. Correlation map for (a) original wavelet features; (b) original and knockoff wavelet features. The inset highlights the first 100 features.

This second-order construction of knockoffs assumes a multivariate Gaussian approximation for the feature distribution, which may not necessarily be very accurate. In any case, our classification results indicate that the second-order knockoffs are effective in performing controlled feature selection, while retaining power in the selected features. Further improvements in feature selection may be possible if one adopts the more general knockoff construction described in [20], which can model the underlying feature distribution more flexibly.

Figure 4 visualizes a set of knockoff-filtered features in the wavelet and signal domains (the latter is obtained through an IDWT). Most of these wavelets are at lower frequency, as higher frequency wavelets tend to be filtered out. Thus, most noise in the signal domain is removed, while certain peaks are accentuated. Such peaks reveal interpretable structures that are important for the prediction task, in a way that the noisy raw signal cannot directly capture. In particular, we expect peaks in our Raman spectra to indicate distinguishing chemical signatures found in different classes of bacteria.

Examining the higher order (i.e., more spatially localized) wavelet features selected by our method for the 2-class task, we indeed observe that these are consistent with peaks pre-

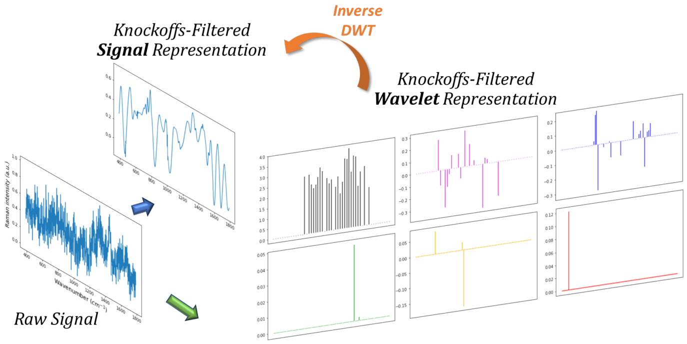

Fig. 4. Visualization in the wavelet and signal domain of features selected by the knockoff filter.

viously identified as being relevant to discriminating between methicillin-resistant (MRSA) and methicillin-sensitive strains (MSSA) of the Staphylococcus aureus bacteria [34]. Recall that we say a feature is "selected" if its fitted coefficient is non-zero after performing knockoff selection and lasso regression. The non-zero detail coefficients from levels 2, 3, 4, and 5 of the wavelet transform for a single MRSA sample are represented in Figure 4 by spikes in the blue, green, yellow, and red plots, respectively. Figure 5 shows the IDWT of these features alongside the spectrum averages from each class. Specifically, it appears that (level 2) wavelets with peaks close to 781 cm^{-1}, 1004 cm^{-1}, 1159 cm^{-1}, and 1523 cm^{-1} were selected. These correspond to the breathing modes for the pyrimidine ring and phenylalanine, as well as the C-C and C=C stretching modes for staphyloxanthin, a carotenoid pigment produced by *S. aureus* that gives it its characteristic golden color. Further, peaks around 1456 cm^{-1} and 1004 cm^{-1} were also selected (level 3 and 4 wavelets), corresponding to the CH_2/CH_3 bending mode and phenylalanine breathing mode. These findings agree with the considerably less noisy Raman microspectroscopy data in [34], which found that the ratios of the 1159 cm^{-1}, 1523 cm^{-1}, and 1456 cm^{-1} peaks to the 1004 cm^{-1} peak were highly predictive of methicillin resistance, and potentially also indicate differences in pigmentation and lipid concentration in *S. aureus* strains. In addition to these known peaks, we also observed that level 5, 4, and 2 wavelets corresponding respectively to peaks around 470 cm^{-1}, 620 cm^{-1}, and 1351 cm^{-1} were important. It is possible that future research will shed light on their chemical origins and allow us to more fully understand the phenotypic differences between methicillin-resistant and methicillin-sensitive *S. aureus* bacteria.

V. Conclusion

This paper demonstrates that domain knowledge-driven feature extraction and controlled variable selection via knockoffs can improve model interpretability for learning tasks involving signal data, relative to more obscure black-box algorithms, and sometimes even increase predictive performance. We have presented an application of this method to a problem of bacterial classification through Raman spectroscopy, although the general framework is also applicable to other kinds of signal data, both within biomedical applications and beyond.

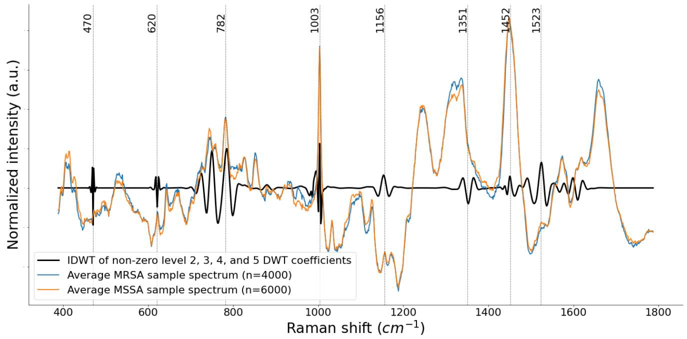

Fig. 5. Visualization in the signal domain of level 2, 3, 4, and 5 wavelet features selected by the knockoff filter for the 2-class (MRSA vs. MSSA) prediction task (black), along with the averaged spectra from each class (color). Many of the selected features have peaks (vertical lines) at wavelengths corresponding to known relevant chemical bonds.

The approach described here also has the advantage of being computationally more affordable to train than many typical black-box ML algorithms, as we utilize simpler models with fewer input features. Even though we focused here on a particularly simple logistic regression classifier, our framework can easily fit more flexible models to the selected features, which may capture nonlinear relations and thus sometimes lead to better performance.

In conclusion, we have suggested a systematic and principled approach to the analysis of signal data, which can facilitate the development of human-interpretable models with good predictive performance. Future research may explore the use of more automated feature extraction models within our framework, such DeepPINK [35], or more complex learning algorithms accompanied by an appropriate quantitative measure of feature importance, e.g., SHAP values [7]. Alternative knockoff generation algorithms such as that in [20] could enhance the robustness of the feature selection step and thus further improve predictive accuracy. Finally, it would be interesting to investigate the impact of the FDR level on the predictive accuracy of our method. In this paper, we have focused on the standard level of 10% for simplicity, and because larger values did not seem to bring much improvement. However, the optimal choice may generally be data-dependent.

References

[1] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. "Definitions, methods, and applications in interpretable machine learning". In: *Proc. Natl. Acad. Sci. U.S.A.* 116.44 (2019), pp. 22071–22080.

[2] S. Krishnan and Y. Athavale. "Trends in biomedical signal feature extraction". In: *Biomed. Signal Proces. Control* 43 (2018), pp. 41–63.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. "A neural probabilistic language model". In: *J. Mach. Learn. Res.* 3.Feb (2003), pp. 1137–1155.

[4] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang. "Deep learning for health informatics". In: *IEEE J. Biomed. Health Inform.* 21.1 (2016), pp. 4–21.

[5] A. Borji. “Saliency prediction in the deep learning era: Successes and limitations”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2019).

[6] M. T. Ribeiro, S. Singh, and C. Guestrin. “”Why should I trust you?” Explaining the predictions of any classifier”. In: *Proc. 22 ACM SIGKDD Int. Conf. Know. Disc. Data Mining*. 2016, pp. 1135–1144.

[7] S. M. Lundberg and S.-I. Lee. “A unified approach to interpreting model predictions”. In: *Adv. Neural. Inf. Proces. Syst.* 2017, pp. 4765–4774.

[8] T. Miller. “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artif. Intell.* 267 (2019), pp. 1–38.

[9] M. Sesia, E. Katsevich, S. Bates, E. Candès, and C. Sabatti. “Multi-resolution localization of causal variants across the genome”. In: *Nat. Comm.* 11.1 (2020), p. 1093.

[10] N. Wang and M. R. Lyu. “Extracting and selecting distinctive EEG features for efficient epileptic seizure prediction”. In: *IEEE J. Biomed. Health Inform.* 19.5 (2014), pp. 1648–1659.

[11] Y. Saeys, I. Inza, and P. Larrañaga. “A review of feature selection techniques in bioinformatics”. In: *Bioinformatics* 23.19 (2007), pp. 2507–2517.

[12] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu. “Feature Selection: A Data Perspective”. In: *ACM Comput. Surv.* 50.6 (2017).

[13] Y. Benjamini and Y. Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *J. R. Stat. Soc. B.* 57 (1995), pp. 289–300.

[14] R. F. Barber and E. J. Candès. “Controlling the false discovery rate via knockoffs”. In: *Ann. Stat.* 43.5 (2015), pp. 2055–2085.

[15] E. Candès, Y. Fan, L. Janson, and J. Lv. “Panning for gold: model-X knockoffs for high-dimensional controlled variable selection”. In: *J. R. Stat. Soc. B.* 80 (2018), pp. 551–577.

[16] E. Katsevich and C. Sabatti. “Multilayer knockoff filter: controlled variable selection at multiple resolutions”. In: *Ann. Appl. Stat.* 13 (2019), pp. 1–33.

[17] M. Sesia, C. Sabatti, and E. Candès. “Gene hunting with hidden Markov model knockoffs”. In: *Biometrika* 106 (2019), pp. 1–18.

[18] A. Shen, H. Fu, K. He, and H. Jiang. “False discovery rate control in cancer biomarker selection using knockoffs”. In: *Cancers* 11.6 (2019), p. 744.

[19] M. Sesia, S. Bates, E. Candès, J. Marchini, and C. Sabatti. “Controlling the false discovery rate in GWAS with population structure”. In: *bioRxiv preprint* (2020). doi: 10.1101/2020.08.04.236703.

[20] Y. Romano, M. Sesia, and E. J. Candès. “Deep knockoffs”. In: *J. Am. Stat. Assoc.* 0.ja (2019), pp. 1–27.

[21] J. R. Gimenez, A. Ghorbani, and J. Zou. “Knockoffs for the mass: new feature importance statistics with false discovery guarantees”. In: *22nd Int. Conf. Artif. Intell. Stat.* 2019, pp. 2125–2133.

[22] X. Li, X. Dong, J. Lian, Y. Zhang, and J. Yu. “Knockoff filter-based feature selection for discrimination of non-small cell lung cancer in CT image”. In: *IET Image Proces.* 13.3 (2018), pp. 543–548.

[23] T.-B. Nguyen, J.-A. Chevalier, and B. Thirion. “ECKO: ensemble of clustered knockoffs for robust multivariate inference on fMRI data”. In: *Intern. Conf. Inform. Proces. Medical Imag.* Ed. by A. C. S. Chung, J. C. Gee, P. A. Yushkevich, and S. Bao. Cham: Springer, 2019, pp. 454–466.

[24] Y. Fan, J. Lv, M. Sharifvaghefi, and Y. Uematsu. “IPAD: stable interpretable forecasting with knockoffs inference”. In: *J. Am. Stat. Assoc.* (2019), pp. 1–13.

[25] C.-S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. Saleh, S. Ermon, and J. Dionne. “Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning”. In: *Nat. Commun.* 10.1 (2019), pp. 1–8.

[26] R. N. Bracewell and R. N. Bracewell. *The Fourier transform and its applications*. Vol. 31999. McGraw-Hill New York, 1986.

[27] S. Mallat. *A wavelet tour of signal processing (2. ed.).* Academic Press, 1999, pp. I–XXIV, 1–637.

[28] C. Garcia-Cardona and B. Wohlberg. “Convolutional dictionary learning: A comparative review and new algorithms”. In: *IEEE Trans. Comput. Imag.* 4.3 (2018), pp. 366–381.

[29] R. L. McCreery. *Raman spectroscopy for chemical analysis*. Vol. 225. John Wiley & Sons, 2005.

[30] R. Tibshirani. “Regression shrinkage and selection via the lasso”. In: *J. R. Stat. Soc. B* 58.1 (1996), pp. 267–288.

[31] G. Tutz, W. Pößnecker, and L. Uhlmann. “Variable selection in general multinomial logit models”. In: *Comput. Stat. Data An.* 82 (2015), pp. 207–222.

[32] J. Friedman, T. Hastie, and R. Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *J. Stat. Softw.* 33.1 (2010), p. 1.

[33] C.-S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. Saleh, S. Ermon, and J. Dionne. “Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning”. In: *arXiv preprint 1901.07666* (2019).

[34] O. D. Ayala, C. A. Wakeman, I. J. Pence, J. A. Gaddy, J. C. Slaughter, E. P. Skaar, and A. Mahadevan-Jansen. “Drug-resistant staphylococcus aureus strains reveal distinct biochemical features with Raman microspectroscopy”. In: *ACS Infect. Dis.* 4.8 (2018), pp. 1197–1210.

[35] Y. Lu, Y. Fan, J. Lv, and W. S. Noble. “DeepPINK: reproducible feature selection in deep neural networks”. In: *Adv. Neural. Inf. Proces. Syst.* 2018, pp. 8676–8686.

APPENDIX

A. Review of the knockoff filter method

For completeness, we give a brief overview of the knockoff framework [15]. We consider n independent pairs of observations $(X^{(i)}, Y^{(i)})$, such that $Y^{(i)}$ depends on its corresponding

p features, $X^{(i)} = (X_1^{(i)}, \dots, X_p^{(i)})$, with $i \in \{1, \dots, n\}$, through some unknown conditional distribution $F_{Y|X}$:

$$Y^{(i)} \mid X_1^{(i)}, \dots, X_p^{(i)} \sim F_{Y|X}.$$

We seek to find the smallest subset of *important* features, $S \subseteq \{1, \dots, p\}$, upon which $F_{Y|X}$ depends; i.e., Y should be independent of $\{X_j\}_{j \notin S}$ conditional on $\{X_j\}_{j \in S}$. We denote by $H_0 = \{1, \dots, p\} \setminus S$ the set of *null* (unimportant) features.

The false discovery rate (FDR) for some $\hat{S} \subseteq \{1, \dots, p\}$ is defined as the expected fraction of null features among it:

$$\text{FDR} = \mathbb{E}\left[\frac{|\hat{S} \cap H_0|}{\max(1, |\hat{S}|)}\right].$$

The goal is to discover as many important features as possible while keeping the FDR below a specified level. This can be achieved by generating, in silico, a *knockoff* copy $\tilde{X}$ of X, which should satisfy the following two properties:

1) Y is independent of $\tilde{X} \mid X$;
2) $[X, \tilde{X}]$ and $[X, \tilde{X}]_{\text{swap}(j)}$ have the same distribution, for any $j \in \{1, \dots, p\}$, where $[X, \tilde{X}]_{\text{swap}(j)}$ is the vector obtained by swapping X_j with $\tilde{X}_j$.

The first condition above simply states that knockoffs are null (this is immediately guaranteed if $\tilde{X}$ is generated before looking at Y). The second condition states that the features in X and $\tilde{X}$ are pairwise exchangeable, which implies that null features have on average the same explanatory power for Y as their corresponding knockoffs. These two properties allow knockoffs to serve as negative controls [15], as explained below. Note that the equality in distribution (second property) is generally difficult to enforce exactly, so we make some approximation and only match the first two moments, following in the footsteps of the previous literature [15, 20].

After augmenting the original feature matrix with the knockoffs, i.e., as $[X, \tilde{X}] \in \mathbb{R}^{n \times 2p}$, a learning model is trained to predict Y, from which feature importance measures $Z = (Z_1, \dots, Z_{2p})$ are then extracted for each of the augmented features. For example, if we adopt a simple regularized logistic regression model, we can define the scores $Z_j = |\hat{\beta}_j(\lambda)|$, where $\hat{\beta}(\lambda)$ denotes the vector of estimated coefficients, and λ is the regularization parameter tuned by cross-validation. Ideally, we would like null features to have Z_j close to zero, although this may not generally be the case in practice; hence the need to calibrate these measures through the knockoffs.

For each $j \in \{1, \dots, p\}$, an importance statistic W_j is defined by contrasting Z_j with Z_{j+p}, i.e., $W_j = Z_j - Z_{j+p}$. Therefore, $W_j > 0$ indicates that X_j appears to be more important than its knockoff copy, which provides evidence against the null hypothesis $j \in H_0$.

By construction, each W_j has equal probability of being positive or negative if $j \in H_0$; more precisely, the signs of null W_j are independent and identically distributed flips of a fair coin [14]. The knockoff filter leverages this property to select features with sufficiently large W_j, according to a data-adaptive threshold that depends on the desired FDR level. The intuition is that the proportion of false discoveries in $\{j : W_j \geq t\}$ can be estimated conservatively by:

$$\widehat{\text{FDP}}(t) = \frac{1 + |\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}|}.$$

In particular, the FDR can be provably controlled [14] below $q \in (0, 1)$ by selecting $\hat{S} = \{j : W_j \geq T\}$, with:

$$T = \min\left\{t : \widehat{\text{FDP}}(t) \leq q\right\}.$$