Fitting Rank Order Data in the Age of Context

Kevin Dick & James R. Green
Department of Systems and Computer Engineering
Carleton University
Ottawa, Canada
kevin.dick@carleton.ca; jrgreen@sce.carleton.ca

***Abstract*—Rank order data are pervasive in science and in our daily lived experience. With the advent of high performance computing and the commensurate increase in available data, the opportunity to capture the overall distribution of values by means of nonparametric curve fitting enables the identification of exceptional points in large datasets. With a rank order structure, these distributions may exhibit a "knee" delineating a threshold between exceptional points and those of the baseline. Given an accurate characterization of the distribution of prediction scores, including careful identification of the knee, we have previously shown that predictive performance can be significantly improved by leveraging this "context". This paper examines the nonparametric characterization of such distributions. Locally weighted regression (LOESS) is a widely used nonparametric approach to curve fitting. Here, we revisit the assumptions behind the selection of kernel functions for nonparametric curve fitting of biological and biomedical data exhibiting rare or exceptional instances. We propose a new linear asymmetric kernel function and compare it to the commonly used tricube kernel used in LOESS. We evaluate its ability to fit rank order data in the domain of protein-protein interaction prediction. The proposed linear kernel significantly improved predictive performance ($p < 0.001$) of two state-of-the-art predictors and promises to be widely applicable in related machine learning pipelines and nonparametric regression tasks.**

Keywords—Nonparametric Regression, Kernel Functions, Machine Learning

I. Introduction

Ranking elements is a process ubiquitous in science and in our daily human experience; we attribute value to elements based on their relation to others. When sorting data, we add a ranking structure wherein the pairwise relationship of any two elements in the set is such that the first element is *ranked higher than*, *ranked lower than*, or *ranked equal* to the second. For example, we might rank patients by a test result, rank pharmaceuticals by tumor suppression metrics, or rank putative protein-protein interactions (PPIs) by a prediction score. While evidently useful to any application looking to identify the best subset of elements among others (*i.e.* top-*k*), with the advent of high performance computing and the commensurate increase in available data, data-driven methods leveraging the context in the structure of rank order data are emergent [1].

Historically, predictive studies across the sciences were limited in scope due to the algorithmic complexity prohibiting scaling to comprehensive analyses [2]. The field of protein-protein interaction (PPI) prediction is one example domain wherein only modestly sized interaction sub-networks could be practically predicted due to the computational challenge of predicting all pairwise putative PPIs in a proteome. Furthermore, many biomedical and bioinformatic applications manifest extreme class imbalance further exacerbating the challenge of delineating the rare positive class from the majority negatives such as correctly diagnosing rare diseases within a population [3] or predicting microRNAs within genomes [4].

To identify these exceptional PPIs, we appraise the value of the interaction between protein i and protein j, P_iP_j, by considering it in the context of all the pairwise interactions between protein i and all proteins in the proteome, $P_i *$ as well as $P_j *$; we thus define "context" as the comprehensive appraisal of the value of one element with respect to all others in a set. The conventional treatment of rank order data should thereby adapt to accommodate these emergent opportunities.

With recent algorithmic advances and the increase in computational resources, the comprehensive prediction of complete interactomes is now possible [5]–[7]. That is, we are now able to assign a score to all possible pairs of proteins, corresponding to the likelihood that they will physically interact *in vivo*. Ranking these putative interactions in decreasing order by their likelihood, we expect those among the top-*k* are positive interactions, while those ranked below *k* are negative. Only a rare few of these predicted interactions are expected to constitute true positive interaction among the negatives (*e.g.* previously estimated to be 1:100 [8]). To identify these few exceptional protein pairs requires the determination of a suitable decision threshold to delineate positives from negatives while avoiding an overwhelming number of false positives.

Emergent data-driven approaches leveraging the context of comprehensive sets of all possible predictions enable the refinement of the decision boundary. For PPI prediction, the Reciprocal Perspective (RP) [1] framework seeks to determine local decision boundaries by identifying exceptionally high-ranking protein pairs, relative to the baseline score observed among all pairs. The distribution of all scores is first estimated using a nonparametric robust locally weighted regression (LOESS) procedure [9] on the rank order prediction data. The knee of the resulting distribution is then identified along with other metrics. This method has led to significant improvement in PPI classification performance [1], however the conventional assumptions of the LOESS method may not necessarily be well suited for these rank order data. More specifically, the rarity of positive samples makes fitting the knee of the curve challenging, since traditional symmetric kernel functions do not sufficiently weight these desirable data points.

978-1-5386-6709-5/18/$31.00 ©2018 IEEE

Since William S. Cleveland first introduced the LOESS procedure, it has been used extensively in biomedical literature such as in identifying differentially expressed transcripts from RNA-seq [10] and model calibration for estimating the probability of the occurrence of a binary outcome in medical applications [11]. Here, we propose new asymmetric kernel functions for the application of LOESS to rank order data. This work is inspired from our recent work leveraging context to improve PPI prediction performance.

II. METHODS

The RP method for PPI prediction (RP-PPI) is a data-driven method that identifies exceptional pairs scoring above the baseline floor for each protein in the pair of interest using the One-to-All Curve. These two-dimensional curves plot the rank order distribution of the comprehensively predicted pairwise scores for a protein with the complete proteome and exhibit a characteristic knee delineating the high and low scoring pairs for the one protein. This knee holds particular importance for a given protein as it can be leveraged to differentiate between those high- and low-scoring PPIs on a per-protein basis. Accurate identification of the knee enables us to consider context-based features in addition to a globally defined decision threshold when predicting whether or not a PPI is positive or negative. For example, should no PPIs involving a given protein score higher than the global threshold, this implies that the protein is not predicted to interact with *any* proteins, which contradicts the biological purpose of a protein. To resolve this contradiction, we can examine those PPIs which tend to score higher than the baseline of a given protein, however this requires that we delineate a point indicative of that baseline. The RP method uses the maximum value of the second derivative to identify this point which requires a continuous form for these discrete PPI data. A single low-order polynomial cannot adequately describe the entire curve and we might be tempted to find some other parametric family of curves to fit the data. Ultimately, this would be unwieldly and require more effort than the relatively simple LOESS method. We thus use it to fit a smooth curve to this rank order data to identify this knee.

A. Overview of LOESS

The motivating principal of local regression is that the regression function $g(x)$ at a predictor x can be locally approximated by the value of a function in some specified parametric class (*e.g.* polynomial) by fitting a regression surface to the data points within a specifiable neighborhood of x. LOESS uses weighted least squares to fit a low-order polynomial (usually degree 1 or 2) centered in the neighborhood with a radius parameterized by *window span,* α, which controls the amount of smoothing. Usually, each data point in the window is weighted by a continuous kernel whose value decreases with distance from the window center. The kernel, $K(\cdot)$, is a non-negative, integrable function, usually satisfying two constraints, normalization and symmetry:

$$\int_{-1}^{1} K(u)du = 1 \quad (1)$$

$$K(u) = K(-u), \forall u \quad (2)$$

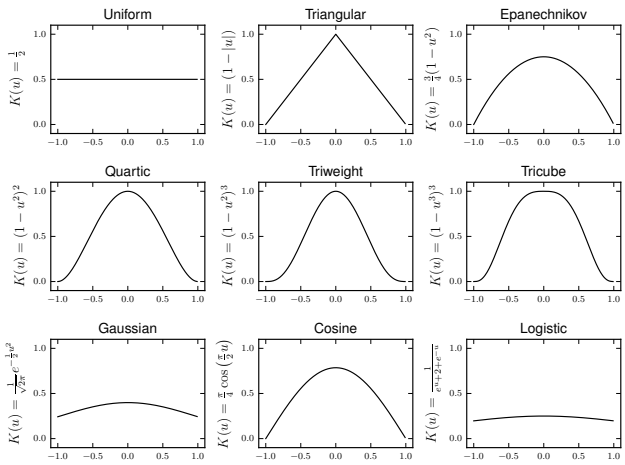

Fig. 1. Commonly used Kernel Functions in Nonparametric Statistics.

A number of commonly used kernel functions are depicted in Fig. 1, each useful to a variety of applications; traditionally the tricube kernel is used as the default.

B. Revisiting Kernel Assumptions for Rank Order Data

When fitting any two arbitrary variables with LOESS, a curve summarizing the relationship is produced. Without any prior knowledge of the underlying structure of the data, any one of the kernel functions in Fig. 1 can be selected to weight the influence of each point in each window; typically, the tricube is used. Rank order data has a monotonic property which, when plotted against rank, introduces a special relationship between the ordinate and abscissa. Numerically, the positive natural value of the abscissa of each datum with respect to its two neighbours is proportionally increasing or decreasing. For example, the relationship between the 1st, 2nd, and 3rd ranked data share the same relationship as between the 41st, 42nd, and 43rd. Intuitively, we might consider applying the uniform kernel to fit these data given this relationship, however this would treat the fit of every point with equivalent importance and does not account for the inherent class imbalance of these data.

Context-based prediction methods for binary class problems attribute lower rank values to the rare positive class and higher ranked values to the majority negative class with the characteristic knee of the curve used to delineate the two classes [1]. A trade-off between window width (which controls smoothing) and sensitivity (the ability to identify low-rank exceptional points) makes the identification of the knee challenging. Only a tiny fraction of points are expected to be above the knee and any substantial window width will include far more baseline samples than exceptional samples. Using a very small window, however, results in an overly noisy curve from which it is difficult to identify the actual knee. Moreover, these high scoring and low ranked values are weakly weighted using a traditional symmetric kernel function, such as the tricube, and the resulting fit does not sufficiently capture the trend exhibited by the minority positive class.

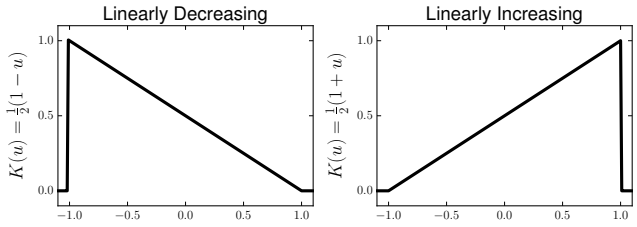

Fig. 2. Proposed Asymmetric Linear Kernel Functions. For data ranked in decreasing order, the linearly decreasing kernel is most appropriate whereas the linearly increasing kernel is suited to data ranked in increasing order.

We here propose relaxing the assumption that the kernel function be symmetric. This relaxation then permits the adaptation of the previously defined kernel functions into a monotonic form more appropriately suited when considering rank ordered data (ranked in either increasing or decreasing order). Fig. 2 depicts the proposed linear kernels over the range $u = [-1,1]$ which conserves (1) and relaxes (2):

$$K(u) = \begin{cases} {}^{1}/_{2}\,(1 \pm u) & \text{if } u \in [-1,1] \\ 0 & \text{if } u \notin [-1,1] \end{cases} \qquad (3)$$

C. Comprehensive Interactome Prediction

To explore the suitability of our proposed asymmetric linear kernel when used with LOESS within RP-PPI, we assemble several datasets of rank order data from five species using two PPI predictors: the Protein-protein Interaction Prediction Engine (PIPE) [6] and Scoring PRotein INTeraction (SPRINT) method [7]. For each organism with n proteins, the comprehensive set of predictions comprises the triangular number, ${}^{1}/_{2}\,(n^2 + n)$, for all pairwise putative interactions. Each organism's proteome and interactome size are summarized in TABLE I.

D. Hyperparameter Tuning: Window Span, α

Randomly selecting 1% of proteins from each dataset, we qualitatively determined the quality of fit for window span parameter α by varying the value and visualizing the result over the coarse-grained ranges $\alpha \in [0.05, 0.10, \ldots 0.40, 0.45]$, followed by a fine-grained range $\alpha \in [0.01, 0.02, \ldots, 0.19, 0.20]$ using both the linear and tricube kernel. For a random sample of 100 of these proteins, we appraised the one-to-all curve and denoted the point we perceived as the "knee" as the ground truth. For each curve, the value of α and kernel (linear vs. tricube) producing the nearest estimate to this perceived knee was recorded. These results were visualized using one-to-all curves.

TABLE I. ORGANISM DATASET SIZES

Organism	Proteome Size	Interactome Size	Positive Training Samples*
H. sapiens	20,160	203,222,880	13,937
S. cerevisiae	6,717	22,562,403	74,588
A. thaliana	16,886	142,576,941	2,956
C. elegans	6,443	20,759,346	7,840
M. musculus	17,759	157,699,920	2,772

* Obtained from [12] as in [1].

E. Experimental Design

Following the experimental design of [1], this work compares the predictive performance resulting from RP features extracted using the proposed linear kernel versus the traditional tricube kernel applied to the comprehensive datasets of PIPE and SPRINT for each of the five organisms (Fig. 3). The RP method was used to compute 15 context-based features for each PPI from the estimated distribution baselines and knees, thereby producing two independent feature matrices for comparison; one for each kernel.

To evaluate the improvements resulting from the linear kernel, the set of positive PPIs and an equivalently sized random subsample of negative PPIs (sampled without replacement) were assembled to train and evaluate a Random Forest classification model using five-fold cross-validation. Prevalence-corrected precision-recall curves (PRC) and the area under the curve (AUC) metric were used to summarize model performance since the class imbalance in our test data is not necessarily reflective of the actual degree of imbalance expected when the classifier is applied to a complete proteome. Finally, to quantify the statistical difference between the linear and tricube kernels, we used bootstrap testing over 1,000 iterations and summarized the resulting curves using the AUC metric over the PRCs: AUPRC.

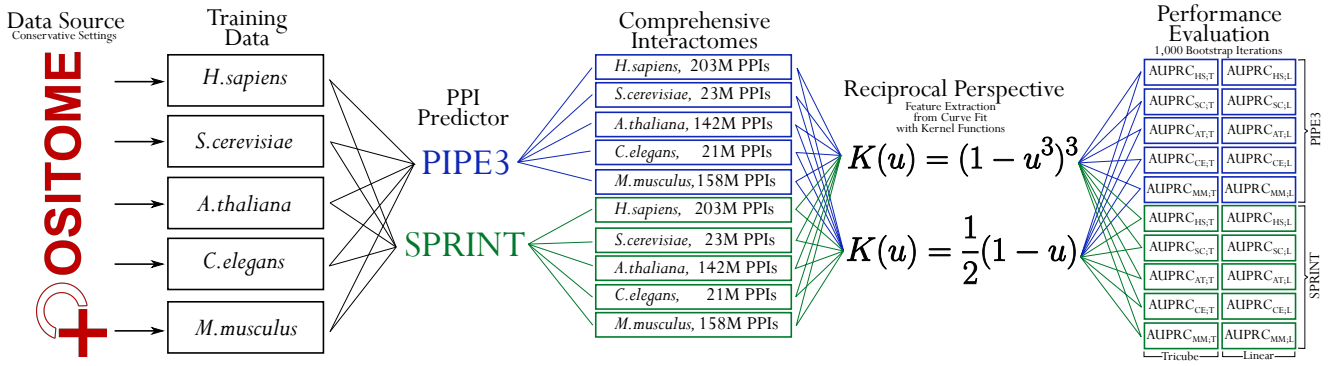

Fig. 3. Overview of the Processing Pipeline and Experimental Design. The training data (sequences and pairs) for the five test organims are acquired using the Positome web service. The data were used to generate the comprehensive interactome of predictions using both PIPE3 and SPRINT. The RP framework was then used to extract context-based features from these datasets. In one condition the tricube kernel was used for the curve-fitting procecure and the proposed linear kernel for the other. The resulting performance of the RP predictions were summarized and compared using the average AUPRC from 1,000 bootstrap iterations.

III. Results & Discussion

The opportunity to now comprehensively predict pairwise relationships heralds the "Age of Context"; through computational advances, the value of any one pair can be appraised in the context of all possible pairings. The one-to-all curves exemplify how rank order data may be leveraged to identify exceptional points within massive datasets. These curves, however, cannot easily be described through classical parametric models and thus the nonparametric LOESS method is used to identify the knee of each curve. Challenges arise due to the rarity of exceptional points, namely the smoothing-sensitivity trade-off. A large window span produces a fit conforming well to the original data; however it is insensitive to the exceptional points. Conversely, a small window span produces a model sufficiently sensitive to the exceptional data; however, this comes at the cost of noise and low degree of smoothing. To address this trade-off, we propose the use of an asymmetric class of kernels to more suitably weight the exceptional points during the curve fitting procedure and evaluate the linearly decreasing variant with a series of qualitative and quantitative experiments.

A. A New Class of Kernels

With the relaxation of the symmetry constraint (2), this work proposed two new asymmetric kernels (Fig. 2; Eqn. 3) of a new class of kernel functions. These functions are strictly asymmetrical however the relaxation of (2) introduces a generalized class of *not necessarily symmetric* kernel functions wherein regions of $K(\cdot)$ respect (2) while others do not. Assuming some creative liberties, we mildly abuse conventional notation to describe this new class using:

$$K(u) \neq K(-u), \forall u \qquad (4)$$

where $\neq$ denotes "not necessarily equal to". Exploring the utility of the kernels in this class is left to future work however, given the implications of our findings on rank ordered data, additional investigations are warranted.

B. Qualitative Improvement of Fit

Comparing both kernel functions through visualization and qualitative appraisal of fit, we note the distinctive trade-off between noise and closeness of fit. When using a very low α (*e.g.* $\alpha < 0.05$) the LOESS curves often successfully capture the rare high scoring interactions and overall trend of the curve; however, many locally fit segments are noisy and poorly conform to the original data. Here, the tricube kernel regularly succeeded in estimating the knee of the curve, whereas the linear kernel estimated a point among the rare low ranked data of the curve corresponding to higher variation in score. This is expected as the variation within this high scoring segment is much greater than that in the baseline of these curves. These noisy locally fit segments therefore produce erratic estimates, a definitive indication of insufficient smoothing.

As α increases and a greater proportion of data are used in each local window, the rare low ranking data fail to be captured in the local fit. The high density of points occurring in the baseline "pull down" the fit and the estimated knee is correspondingly biased higher in rank, effectively missing the high scoring region of the curve.

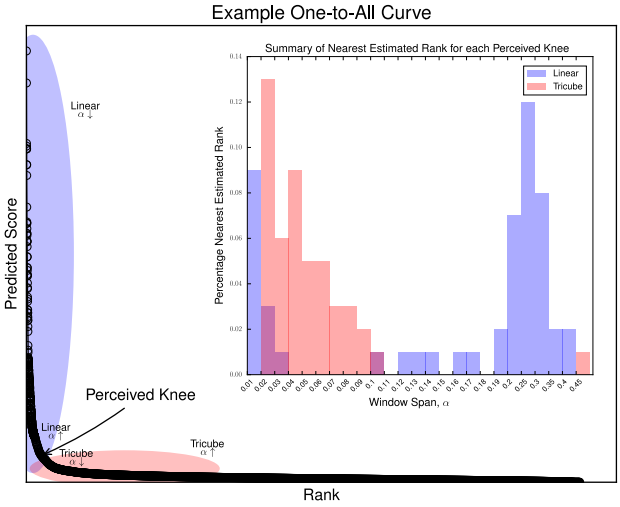

Fig. 4. Summary of LOESS Estimation of the Perceived Knee. The example one-to-all curve highlights the regions where the linear and tricube kernels typically estimate to be the knee of the curve for various values of α. The inset figure illustrates the frequency at which each value of α produces an estimated knee location that best agrees with the ground truth for each kernel type.

Fig. 4 summarizes the α value that produces an estimated knee that most closely agrees with the perceived (ground truth) knee. While this experiment contains a subjective component, it provides preliminary evidence that the linear kernel is more sensitive to the rare high scoring proteins, thereby successfully capturing the rare trend and estimating the knee of the curve.

We also note that no one value of α or choice of kernel appears to be uniquely superior for the estimation of the knee of the one-to-all curves (Fig. 4; inset). Notably, for lower values of α the linear kernel captures the low rank data points as anticipated. The asymmetric kernel attributes larger weight to these data and the resulting fit is thereby more sensitive to the sparse density of low rank data. However, for small windows, the linear kernel tends to estimate the knee well above the perceived knee due to the noise resulting from small window size and high variance in the score (the ordinate). While the typical one-to-all curve have a comparable number of high scoring data, a subset of proteins have considerably fewer and the linear kernel with $\alpha < 0.05$ alone can capture these (since the tricube kernel dampens their influence in the fit). The sensitivity of the proposed linear kernel to these data suggest that it should be preferred when nonparametrically fitting rank order data exhibiting class imbalance. Given the complementarity of window span for each kernel, future work may examine combining the estimate from a linear kernel with larger α with that of the tricube kernel using a smaller α.

C. Quantitative Improvement in Predictive Performance

To determine whether the linear kernel's baseline estimation leads to improved predictive performance in RP-PPI, we compared the AUPRC of each kernel using each method over the five test organisms (Table II). An improvement in predictive performance was observed in *H. sapiens*,

TABLE II. SUMMARY OF CLASSIFICATION PERFORMANCE FOLLOWING 1,000 BOOSTRAP ITERATIONS FOR EACH TEST CONDITION ($\mu \pm SE$)

Organism	Kernel	PIPE AUPRC	SPRINT AUPRC
H. sapiens	Tricube	0.478 ± 0.001	0.500 ± 0.001
	Linear	0.633 ± 0.002	0.696 ± 0.014
S. cerevisiae	Tricube	0.336 ± 0.001	0.294 ± 0.001
	Linear	0.345 ± 0.001	0.322 ± 0.001
A. thaliana	Tricube	0.517 ± 0.009	0.443 ± 0.009
	Linear	0.416 ± 0.008	0.287 ± 0.008
C. elegans	Tricube	0.425 ± 0.002	0.370 ± 0.002
	Linear	0.441 ± 0.002	0.395 ± 0.002
M. musculus	Tricube	0.497 ± 0.004	0.491 ± 0.005
	Linear	0.425 ± 0.003	0.405 ± 0.004

S. cerevisiae, and *C. elegans* and decline in performance in *A. thaliana* and *M. musculus*. These findings are consistent across the two methods used. Considering the null hypothesis (H_0: no significant difference in summary statistics between the tricube and linear kernels) we computed p-values using Welch's unequal variances t-test and found the observed differences in AUPRC to be significant at the $p < 0.001$ level, for the six improved conditions. The increase in predictive performance is likely the result of the improved sensitivity of the linear kernel to appropriately fit the low-ranking data and thereby capture the trend exhibited by the rare positive class. The three organisms with improved performance also had the largest number of available training samples, and therefore are likely to be more robust against noise resulting from a more sensitive method; the incorrect estimation of the knee of a single protein would be less damaging to overall performance.

Conversely, the reduction in predictive performance is likely the result of insufficient training samples as well as noise in the curve fitting process. The sensitivity of the linear kernel to deviations in the data risks biasing the estimated knee to a rank lower than that of the actual knee. While this sensitivity is desired on average over the entire dataset, for smaller training dataset sizes, the biased estimates are much more damaging to the overall performance. Leveraging a larger training dataset (*e.g.* using the *permissive* settings from the Positome web service [12]) we expect gains in performance comparable to the three organisms with sufficiently large training datasets.

In summary, these findings corroborate the hypothesis that the asymmetric linear kernel is more sensitive to extraordinary data with rank order structure and can lead to statistically significant improvement in predictive performance when applied to a sufficiently sized training dataset.

D. Future Work

This work can easily be expanded by defining related asymmetric kernel functions of the *not necessarily symmetric* class and fitting curves to the rank order data of related problems. The RP meta-method is theoretically applicable to any machine learning problem exhibiting extreme class imbalance where the signal used to differentiate between classes does not easily separate the classes using a single global decision function. Investigating the distribution of the rank order data of comprehensive predictions and accurately estimating the knee of these curves promises to lead to similar gains in performance as exemplified in this work.

Furthermore, investigating the fit of the log-transform of rank order data shows promise, given the extremity of rare positive class. The log-transform initially smooths the variance of the ordinate, which should lead to increasingly accurate estimates of the knee of these data.

Finally, in an effort to accelerate these prediction pipelines, subsampling the comprehensive set of predictions to estimate the baseline, in favor of computationally expensive nonparametric LOESS methods, promises to accelerate pipelines reliant upon these comprehensively predicted datasets.

IV. CONCLUSION

With the commensurate increase in available bioinformatic and biomedical data, novel methods leveraging newly available context in rank order data have emerged. To address the challenges in identifying the characteristic knee in these data, we, here, propose the use of asymmetric kernel functions when fitting a smooth LOESS fit to estimate this point. Both the tricube and linear kernel functions correctly estimate the perceived knee when using small and large window spans, respectively. Evidence for improved sensitivity of fit were demonstrated for the asymmetric LOESS kernels introduced here. Finally, significant improvement in AUPRC ($p < 0.001$) were achieved in the three of five organisms with the most robust training datasets.

REFERENCES

[1] K. Dick and J. R. Green, "Reciprocal Perspective for Improved Protein-Protein Interaction Prediction," *Sci. Rep.*, vol. 8, no. 1, p. 11694, Dec. 2018.

[2] O. Y. Al-Jarrah, S. Muhaidat, G. K. Karagiannidis, and K. Taha, "Efficient Machine Learning for Big Data: A Review," *Big Data Res.*, vol. 2, no. 3, pp. 87–93, Sep. 2015.

[3] D.-C. Li, C.-W. Liu, and S. C. Hu, "A learning method for the class imbalance problem with medical data sets," *Comput. Biol. Med.*, vol. 40, no. 5, pp. 509–518, May 2010.

[4] R. J. Peace, K. K. Biggar, K. B. Storey, and J. R. Green, "A framework for improving microRNA prediction in non-human genomes," *Nucleic Acids Res.*, vol. 43, no. 20, p. gkv698, Jul. 2015.

[5] S. Pitre *et al.*, "PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs," *BMC Bioinformatics*, vol. 7, no. 1, p. 365, 2006.

[6] A. Schoenrock, F. Dehne, J. R. Green, A. Golshani, and S. Pitre, "MP-PIPE," in *Proceedings of the international conference on Supercomputing - ICS '11*, 2011, p. 327.

[7] Y. Li and L. Ilie, "SPRINT: ultrafast protein-protein interaction prediction of the entire human interactome," *BMC Bioinformatics*, vol. 18, no. 1, p. 485, Dec. 2017.

[8] S. Pitre *et al.*, "Short Co-occurring Polypeptide Regions Can Predict Global Protein Interaction Maps," *Sci. Rep.*, vol. 2, pp. 686–93, Jan. 2012.

[9] W. S. Cleveland, "Robust Locally Weighted Regression and Smoothing Scatterplots," *J. Am. Stat. Assoc.*, vol. 74, no. 368, pp. 829–836, 1979.

[10] P. Glaus, A. Honkela, and M. Rattray, "Identifying differentially expressed transcripts from RNA-seq data with biological variation," *Bioinformatics*, vol. 28, no. 13, pp. 1721–1728, Jul. 2012.

[11] P. C. Austin and E. W. Steyerberg, "Graphical assessment of internal and external calibration of logistic regression models by using LOESS smoothers," *Stat. Med.*, vol. 33, no. 3, pp. 517–535, Feb. 2014.

[12] K. Dick, F. Dehne, A. Golshani, and J. R. Green, "Positome: A method for improving protein-protein interaction quality and prediction accuracy," in *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2017, pp. 1–8.