# Physicochemical Responsive Integrated Similarity Measure (PRISM) for a Comprehensive Quantitative Perspective of Sample Similarity Dynamically Assessed with NIR Spectra

Robert C. Spiers, Callan Norby, John H. Kalivas*

Department of Chemistry, Idaho State University, Pocatello, Idaho 83209, USA

**ABSTRACT:** Determining sample similarity underlies many foundational principles in analytical chemistry. For example, calibration models are unsuitable to predict outliers. Calibration transfer methods assume a moderate degree of sample and measurement dissimilarities between a calibration set and target prediction samples. Classification approaches link target sample similarities to groups of similar class samples. Although similarity is ubiquitous in analytical chemistry and everyday life, quantifying sample similarity is without a straightforward solution, especially when target domain samples are unlabeled and the only known features are measurables such as spectra (the focus of this paper). The process proposed to assess sample similarity integrates spectral similarity information with contextual considerations between source analyte contents, model, and analyte predictions. This hybrid approach named physicochemical responsive integrated similarity measure (PRISM) amplifies hidden-but-essential physicochemical properties encoded within respective spectra. PRISM is tested on four near-infrared (NIR) datasets for four diverse application areas to show efficacy. These applications are assessment of prediction reliability and model updating for model generalizability, outlier detection, and basic matrix matching evaluation. Discussion is provided on adapting PRISM to classification problems. Results indicate that PRISM collects large amounts of similarity information and effectively integrates it to produce a quantitative similarity evaluation between a target sample and a source domain. The approach is also useful for biological samples with additional physiochemical variations. While PRISM is dynamically tested on NIR data, parts of PRISM were previously applied to other data types and PRISM should be applicable to other measurement systems perturbed by matrix effects.

Essential to analytical chemistry (and other disciplines) is the ability to characterize the similarity between samples. For example, a key issue is ascertaining if a model formed for a certain analyte using a source calibration (training) set, e.g., a partial least squares (PLS) model, can be used to predict target domain samples for the same analyte. In other words, are any target samples outliers to the source domain samples? Unfortunately, the answer depends on how similarity is defined and an abundance of similarity measures exist.[1-8] Specifically, each similarity measure is typically used independently to provide its respective partial view of the intricate underlying similarity structure between samples. Additionally, many measures necessitate optimization[5-7] that can be impractical when using computers for automated decision-making capabilities such as outlier detection.[9] Thus, despite the intuitive nature defining similarity by object (sample) closeness relative to the degree of agreement between respective features, it is difficult to remediate this entrenched human similarity notion with its mathematical realization. Some theorists believe that it is unclear to state object "A is similar to object B" and it is only significant to state "A is similar to B with respect to C."[2,3] In other words, A is only similar to B qualified to the measurable merit(s) used.

The focus of this paper is determining spectral similarity and a comprehensive autonomous similarity measure should be applicable to all spectroscopic situations requiring similarity assessment. A common ensemble process to ascertain an overall similarity between two sample spectra is to combine the Euclidean distance (magnitude difference) with the cosine of angle (shape difference).[10] This binary approach requires optimizing a weighting scheme balancing the two measures. An alternative approach combines Mahalanobis distance and Q-residual as used in the popular soft independent modelling by class analogy (SIMCA) classification algorithm.[8] Both SIMCA measures depend on a principal component analysis (PCA) of the source domain spectra and it is debatable as to how to optimize the number of PCs.[11,12] Another approach assesses similarity by comparing respective dataset covariance shapes, magnitudes, and centroid locations to distinguish structure differences.[6,7]

However, influencing each similarity merit is that certain sample properties are not always strongly responding in spectra and hence, large changes in the prediction property of interest, e.g., analyte amount, may go unnoticed in spectra. Equally, small sample-wise changes in less important sample frameworks could impact spectral structure and the similarity value. These sample-wise matrix effects stem from the degree of inter- and intra-molecular interactions between sample species that further depend on the nature and strengths of respective associations relative to species amounts.[13,14] Other sample and measurement conditions, e.g., pH, temperature and instrument are also considered part of the full matrix effects in this paper. Sample and measurement conditions (except instrument) are grouped under the term *physico*chemical causes of matrix effects. In biological systems, *physio*chemical effects based on additional interactions between physiological and chemical processes are part of the matrix effects. Thus, while a dual spectral similarity view, e.g., SIMCA, is an improvement over a singular perspective, a more complete measure is needed.

Developed in this paper is a new similarity measure named physicochemical responsive integrated similarity measure (PRISM) composed of many similarity measures. These selected similarity measures (and other others can be included) are fused to form the PRISM value. The proposed integrated PRISM unites multiple similarity assessments to assess complementary information achieving a thorough view of sample matrix similarity. This approach reduces the effects of contradictory results (local anomalies) and increases the reliability of the PRISM value. Physicochemical is used to name PRISM because such effects typically dominate spectral differences but it is understood that PRISM evaluates other sources of matrix effect differences. Presented in the Supporting Information (SI) is a mathematical framework illustrating the underlying sample matrix effects and why spectra visually appearing similar are not necessarily similar. Unique to PRISM is that it focuses on using spectra as measurable abstractions of the underlying orders of matrix effect differences between source and target domains. Analogous to the sorting of electromagnetic frequencies by a conventional prism, PRISM discriminates source-target sample combinations according to their PRISM value.

It has been stated that similarity measures should be directly related to the property of interest with the association ascertained by using information about processes responsible for the property.[3] However, this information is rarely known for the target domain. In particular, for quantitative analysis, many properties may be available for the calibration set, but only measured target spectra are known. Compounding proper spectral similarity assessment for quantitative analysis purposes is, as previously noted, the amount of analyte and other matrix effecting species. Only evaluating spectral similarity may not be enough for quantitative analysis. To overcome this barrier, PRISM uses sample matrix effects to strengthen key spectral features by emphasizing analyte content dissimilarities between source and target domain samples using a source calibration model to direct the similarity characterization.[15] Presented in the SI is a mathematical framework realizing this goal.

Because PRISM can be used to assess similarity for both classification and quantitative analysis, similarity is henceforth grouped into spectral contributions denoted by $\text{PRISM}_X$ and those guided by quantitative source models, and consequently, source analyte values, termed $\text{PRISM}_y$. In this framework, measurements from the $\text{PRISM}_X$ measures are complementary to the $\text{PRISM}_y$ measures. Other work has found it useful to combine complementary information by balancing the exploration of the chemical space with exploitation of the prediction model.[16,17] If classification is the situation, then only $\text{PRISM}_X$ is applicable.

The underlying fundamental analytical chemistry question being addressed in this paper is how to determine if source domain-based models (calibration or classifier) are generalizable to target domain samples. PRISM is applied to four crucial analytical chemistry areas using near-infrared (NIR) spectra to demonstrate its versatility. These applications involve quantitative analysis using PLS models to assess prediction reliability, calibration transfer complexity for model generalizability, outlier detection, and matrix matching assessment. The importance of these four fundamental areas is described next. Classification is another important area of analytical chemistry and is a facsimile to outlier detection and matrix matching. Some brief comments on adapting PRISM for classification are provided at the end of Results and Discussion. While PRISM is robustly tested on NIR data, it is applicable to other measurement systems affected by matrix effects. This attribute is also discussed at the end of the Results and Discussion.

## PRISM APPLICATIONS

Because the paper focus is towards quantitative analysis, a brief overview of multivariate calibration and prediction is provided in the SI if the reader is not familiar.

**Prediction Reliability.** The calibration set applicability domain (AD), a term commonly used in quantitative structure activity relationship (QSAR) modeling, identifies where target samples are accurately predicted with confidence by the source model.[18-26] In order for new target samples from the deployment domain to best fit into the AD, target samples must be independent and identically distributed (iid) matched to the source calibration set matrix effects. The similarity measure used to define AD sets the perspective to determine the source model generalizability to the target domain. An effective similarity criterion should be consistent with the axiom that target sample similarity is correlated to prediction accuracy[23,24] and hence PRISM values should be correlated with prediction error. However, it is impractical to expect the correlation to be strongly linear since prediction error is not inherently exactly related to similarity. In particular, a target sample dissimilar to a calibration set can be accurately predicted, but it should not be the case that a target sample deemed similar be poorly predicted.

In the presented prediction reliability application, we strive to assess whether minor, moderate, or substantial prediction errors are expected for target samples relative to the source domain calibration set. This evaluation is performed by comparing target sample PRISM values against the source predicting target (SPT) prediction errors ($\left|y-\hat{y}\right|$) and analyzing the correlation over all the datasets.

Related to the AD, it has been found that prediction error does not depend on the machine-learning method, but on the similarity to the training samples.[23-25] Target samples with the highest similarity and most neighbors were best predicted especially with a narrow training set compared to a diverse set.[22] Thus, if the target domain is well matched to the source domain, then all calibration (machine-learning) methods essentially predict with equal accuracy. The effect of domain sample size on AD was recently studied. In this case, the goal was to assess "hard overlap" versus "soft overlap" scoring mechanisms to ascertain if a machine learning algorithm is actually maintaining superior performance.[26] Such studies were not performed with PRISM.

**Model Update.** If a target sample is deemed a nonmember of the AD, recourse is available by calibration transfer using transfer learning methods.[27,28] These techniques adapt the calibration model to account for the novel target domain conditions, either by spectral preprocessing or direct model adaptation by model updating.[29,30] Model updating reorients the model in direction and magnitude to accurately predict source and target samples. In terms of eq. S5 in the SI, model updating attempts to alter the model $\hat{\mathbf{b}}$ such that the non-analyte part of $\mathbf{u}$ matched to source and target sample matrix effect differences s goes to zero. Many model updating algorithms exist ranging from expensive requiring some target domain sample analyte reference values,[11,32] to cheap requiring no target sample reference values.[29,31,32]

The model updating application assesses updating difficulty relative to model generalizability, an assessment problem that to date, lacks any potential solution.[28] The model updating methods local mean centering (LMC)[32] and null augmented regression eigenvalue (NARE)[29,32] are used to evaluate PRISM and the reader is referred to the respective references for details on the methods. The primary difference is that LMC requires a few target reference samples and NARE does not. It is expected that LMC with target reference values will handle a greater degree of PRISM dissimilarity between source and target domains. Because model updating updates from a source calibration set to a collection of target samples, not just one target sample, then for this work, mean PRISM values between source and target datasets are compared to mean analyte prediction accuracy for LMC and NARE processes across the random data splits. Ranges for the number of PLS LVs and respective LMC and NARE weighting parameters generate thousands of models necessitating model selection for final prediction (see Experimental for dataset dependent LV and weight ranges). The autonomous model diversity and prediction similarity (MDPS) approach[29,30] is used to select models and the corresponding validation errors, reported as root mean square error (RMSEV), and $R^2$ statistics are reported. Also presented are LMC and NARE first quartiles of all generated models to characterize the efficacy of LMC and NARE relative to PRISM values.

**Outlier Detection.** The third application is outlier detection[33] where it is determined if a target sample is within the source AD. Outlier detection is analogous to rigorous one-class classification[34] but classification was not studied. To evaluate the outlier detection ability of PRISM, a sample is removed from a calibration set to act as a target sample and PRISM compares this target sample to each available source calibration set. The process is repeated until every calibration sample has been tested as a target sample. Two methods determining accuracy are evaluated: checking whether target samples are identified as outliers to calibration sets other than the source set of origin and if PRISM-deemed outliers have degraded prediction errors compared to errors obtained using the calibration set of origin. This type of outlier detection should not be associated with outlier 'cleaning' of a sample set. In that situation, analyte reference values are known and outlier detection measures such as studentized residuals can be used. A process similar to PRISM was developed and applied for this case[33] and PRISM could be slightly modified to work with this analytical chemistry problem as well. Shown in the SI is an outlier detection example.

**Matrix Match.** This application situation occurs when multiple source calibration sets are present and a target sample must be classified (matrix matched) into the most similar set before prediction.[7] This classification-then-regression protocol, sometimes termed bucket of models, is especially suitable when a source set is naturally grouped into subsets with constrained intra-group matrix effect variances. To evaluate PRISM for this matrix matching task, samples are removed from calibration sets to act as target samples and PRISM compares these target samples to all source sets determining which set they are most similar to. Two methods for determining accuracy are evaluated: checking whether target samples are matched back to the sets of origin and if prediction accuracies for the PRISM-selected similar sets correspond to the lowest sample prediction errors. Although these two assessment methods often produce equivalent results, depending on the degree of intra- and inter-matrix variances, the selected set with the most accurate prediction may not be sample origin set. This setup is analogous to multi-class classification.[35]

## PRISM

This work assumes that a sample spectrum is a linear sum of its component parts where the primary parts are chemical constituents, amounts, and respective pure spectra. Perturbing these pure spectra are sample matrix effects. Described in the SI is the mathematical framework. Basic concerns with all similarity assessments are which similarity merit to use, PC optimization protocol if pertinent, and the integration method for multiple similarity measures. These items for PRISM are now discussed.

**Overview**. PRISM is a consensus method and like all consensus approaches, similarity measures fused to form PRISM should accurately characterize the situation more times than not. Thus, an extensive analysis was performed to identify effective similarity measures to assure proper consensus assessments. This analysis resulted in 13 spectral measures for $\mathrm{PRISM}_X$ and 10 model-based measures for $\mathrm{PRISM}_{y.}$ Similarity equations are presented in the SI. Other measures can be supplemented by the user to the selected measures.

Five of the 13 $\mathrm{PRISM}_X$ measures are PCA based. A novel aspect of PRISM is that unlike SIMCA and other PC based methods, the number of PCs is not optimized. Recent work has shown strong accuracy is obtainable by using non-optimized windows of PCs integrated to form a consensus perspective.[33,35-38] As noted in these references, PC windows reduce the chance of anomalous similarity value for a particular target sample, i.e., less likely to obtain false positives or negatives.

PLS is the calibration modeling method for analyte prediction. Because the 10 $\mathrm{PRISM}_y$ measures are not as robust to different numbers of PLS latent variables (LVs) as the $\mathrm{PRISM}_X$ measures are to PCs, LV optimization is required. Latent variable selection is well studied and many methods to select an optimal model exist such as cross-validation, U-curve, or model diversity and prediction similarity.[30,39] The U-curve is used here for its speed, bias/variance balanced decision, self-directing, and it has been thoroughly documented to work well.[39-43] To form a U-curve in the $\mathrm{PRISM}_y$ algorithm, 100 random splits of the source data is used with 80% of the samples going to the calibration set and 20% for the validation set. Respective mean RMSEC, RMSECV, model 2-norm, and jaggedness values are range-scaled and summed to form a final mean U-curve. The number of LVs selected resides at the U-curve minimum.

**Algorithm.** To obtain the similarity between a target domain sample and a source domain sample set, PRISM uses differences ($\Delta$) in similarity measures between target samples and each source sample one at a time across all similarity measures. Comparing to each source sample allows PRISM to maximize capturing similarity information embedded in the source domain covariance structure relative to target domain samples. Differences have been used in other studies to expand and better detail the sought information for other purposes.[15,44-47]

As an example, let the Mahalanobis distance (MD) be the spectral similarity measure symbolized by $f$ and $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively designate spectra for a sample in the source domain $D$ and a target sample. The corresponding MDs are denoted $f_1$ and $f_2$. The similarity difference used in PRISM$_X$ is then $\Delta f = (f_1 - f_2)$. Specifically, source domain spectra $\mathbf{X}_D$ form the source covariance space. The usual MD assess the distance from the centroid of $\mathbf{X}_D$ to $\mathbf{x}_1$ and $\mathbf{x}_2$, but instead, $\Delta f$ uses the difference between the MDs of $\mathbf{x}_2$ and $\mathbf{x}_1$ relative to $\mathbf{X}_D$. Contained in the SI is a detailed mathematical description of $\Delta f$. All similarity measures are generalized to this $\Delta f$ form in the SI.

The view of sample-wise difference is that if $\mathbf{x}_2$ is similar to $\mathbf{X}_D$, then it should appropriately respond when probed with similarity measurements. By design, source sample $\mathbf{x}_1$ is similar to $\mathbf{X}_D$ and hence, acts as a similarity benchmark value for comparison with $\mathbf{x}_2$. If $\mathbf{x}_2$ and $\mathbf{x}_1$ are similar to $\mathbf{X}_D$ by a small $\Delta f$, then $\mathbf{x}_2$ is likely matched to $\mathbf{X}_D$. The greater $\Delta f$ is, the more likely $\mathbf{x}_2$ is not similar $\mathbf{x}_1$ and hence, not similar to $\mathbf{X}_D$.

The $\Delta f$ value can be positive or negative. Two approaches are taken when averaging $\Delta f$ values to form composite similarity values. One is referred to as signed where the signs are kept in the averaging and the others is unsigned, where absolute values of $\Delta f$ are averaged. By using signed and unsigned $\Delta f$s, the PRISM$_X$ and PRISM$_y$ measures are correspondingly expanded from 13 to 26 and 10 to 18 values. Signed and unsigned equations are further discussed in the SI.

The collection of sample-wise $\Delta f$ similarity perspectives provide comprehensive PRISM$_X$ and PRISM$_y$ similarity views. For example, a single target sample compared to a source domain set of 80 samples generates 16,320 sample-wise $\Delta f$ similarity measures for the target sample. Reported in the SI is this example showing the calculation. However, these target sample measurements are not useful unless normalized to a reference similarity value to remove dataset dependency. An effective normalization process is to standardize to the similarity variance (structure) of the source domain set. This standardization is performed by sequentially removing each source sample and treating it as a pseudo-target sample to obtain the corresponding sample-wise PRISM$_X$ and PRISM$_y$ $\Delta f$ similarity values. Thus, it is important to note that $\mathbf{x}_2$ can be one of two samples. One is a target domain sample as described earlier and the other is a source domain sample removed to act as a pseudo-target sample. For this second situation, $D$ is now $D$1 indicating one source domain sample has been removed and the source sample used for $\mathbf{x}_1$ are sequentially the remaining $D$1 samples not being used as the pseudo-target $\mathbf{x}_2$. Continuing with the 80 sample source example, 16,116 $\Delta f$ values for each pseudo-target source sample are obtained. In totality, 1,305,600 $\Delta f$ similarity measures are evaluated to determine the final target sample PRISM value.

The $\Delta f$ values for a target sample (over 1.3 million for the example) need to be distilled into a single number to quantify the overall PRISM similarity of that target sample to the source domain. For each pseudo-target sample, the average of the sample-wise PC $\Delta f$ values is taken over the PC windows of each PC based similarity measurement producing one mean $\Delta f$ value for each PC based method. Equally, respective average sample-wise $\Delta f$ values are obtained for each of the other similarity measures. The pseudo-target sample now has 26 PRISM$_X$ and 18 PRISM$_y$ mean measures. The process is repeated for each source sample acting as a pseudo-target sample and then for the actual target sample. Values are assembled into corresponding PRISM$_X$ and PRISM$_y$ similarity arrays of respective sizes 26 x $m$+1 and 18 x $m$+1 where $m$ is the number of source domain samples and the +1 for the target sample. Similarity measures (each row) are normalized to unit length to remove magnitude discrepancies between similarity measures.

The sum of ranking differences (SRD)[48] process with the SRD rank target set to 'max' is used separately on the PRISM$_X$ and PRISM$_y$ arrays to combine all of the respective similarity measures forming two similarity values for each source sample and the target sample labled $s_X$ and $s_y$. SRD is used due to its effective fusion mechanism capable of detecting similarity discrepancies. Alternatively, a sum fusion rule or other fusion process could be used to produce similar results. The $s_X$ and $s_y$ similarity values for the source samples are then Z scored characterizing the source similarity distribution providing final PRISM$_X$ and PRISM$_y$ Z scores for each source sample. Using the source Z distribution mean and standard deviation, the target sample is Z scored to form its PRISM$_X$ and PRISM$_y$ values. These two values for the target sample are the final spectral and model perspectives of target similarity to the source domain. As such, they are not independent similarity merits but rather complementary measures of the same hypothesis. Under this framework, the two Z scores can be combined for one PRISM value using Stouffer's Z score method computed by

$$\mathrm{PRISM} = \sqrt{2}\left[\left(\mathrm{PRISM}_X + \mathrm{PRISM}_y\right)/2\right] \qquad (1)$$

that is similar to Fisher's method of combining p-values.[49] The process is repeated for each target domain sample. Detailed in the SI are the specific steps for a target sample.

The large number of target sample similarity measures have now been combined and standardized to the source distribution such that they are interpretable as a Z score distance relative to the source domain. Because the SRD ranking target is 'max', each target sample will almost always have lower Z scores than the source domain samples and therefore the target PRISM Z score will often be negative. For example, a target PRISM Z score of -5 indicates a very high degree of dissimilarity between the target sample and the source domain, whereas a PRISM Z score of -1 is most likely within the source distribution.

It may be possible to alter the stages just described to provide additional benefits to PRISM. For example, changing the fusion rule or training a machine learning algorithm to weight each similarity measure according to its importance to overall sample similarity. The number of possible changes is too large to study for the purpose of this work.

## EXPERIMENTAL

**Datasets**. *Corn*. This dataset is used to evaluate PRISM for all four applications. Corn contains 80 cornmeal samples measured on three near-infrared (NIR) instruments (m5, mp5, and mp6) from 1100-2500 nm at 2 nm intervals with reference values moisture, oil, protein, and starch.[50] For prediction reliability, outlier detection, and matrix matching, the same sample is removed from each instrument leaving 79 for the instrument-based source domains. The process is repeated until each sample has been removed. For model updating, 100 random splits are used and mean results are reported. To avoid the same samples being selected for both the source and target instrument sample sets, then for each instrument updating situation per random split, 40 source instrument samples are selected for the calibration set and theses samples are removed from target instrument consideration; 5 target instrument labeled samples from the remining 40 target samples are used for LMC; and 15 unlabeled target samples from the remaining 35 target samples for NARE. These 15 samples are used for model updating prediction by LMC and NARE.

*Temperature*. The temperature dataset is used to evaluate PRISM for prediction reliability and matrix matching. Sixteen mixtures containing known amounts of ethanol, water, and 2-propanol were measured on a NIR instrument from 590-1091 nm at 1-nm intervals at 30°, 40°, 50°, 60°, and 70° C.[51] The PRISM scores are evaluated by removing a sample from each temperature dataset and characterizing its spectrum at one temperature against each of the other five temperatures. The process is repeated until each sample has been removed and assessed by PRISM. This data set is also used to substantiate the PRISM matrix matching mathematics described in the SI for a spectral situation with known primary matrix effects perturbing the PLS model orthogonality.

*Melamine*. This dataset is used to evaluate PRISM for model updating. Melamine-formaldehyde samples were measured for their polymerization turbidity temperature point (the analyte) indicating polymer length. Temperature values were converted to Kelvin making positive values. Four different batch polymerization recipes were measured (562, 568, 861, and 862)[31,52] and turbidity point values for all four recipes are bimodal. For this study, each recipe was split into two calibration sets: one with analyte turbidity temperatures from 267-283K and the other from 293-316K denoted '*low*' and '*high*', respectively. Tabulated in Table S3 of the SI are mean temperatures and standard deviations for these eight calibrations sets that form 64 updating settings. Updating situations are labeled by whether the matrix effects and/or analyte amounts are different between the source and target set. The first grouping is termed *calibration* if target samples are from the same source recipe with an equivalent analyte range. The second category, *domain adaptation*, involves the same analyte range but different recipes. If the same recipe but a different analyte range is used, then this case is labeled *easy transfer learning*. The last group, *hard transfer learning*, simultaneously deals with different recipes and analyte ranges.

The original dataset includes 8,127 samples with two NIR spectral ranges from 5458-62542 cm$^{-1}$ and 65957-69752 cm$^{-1}$ unequally divided between recipe and analyte categories. To create equal division of samples and improve computational time, each of the eight source sets (recipe/low or high) were reduced once to 100 randomly selected samples for respective source sets. Similar to the corn dataset, 100 random splits are used and mean results are reported. For each updating situation per random split, 70 samples are selected from the 100 source domain samples for the calibration set and theses samples are removed from target domain consideration when the target domain is same as the source set; 5 labeled target domain samples from the remaining 30 source samples or from the 100-target domain set and are used for LMC; and 20 unlabeled target samples from the remaining 25 source samples or 95 target samples for NARE. These 25 samples are used for model updating prediction by LMC and NARE.

*Mango*. This dataset is used to evaluate PRISM for model updating. The dataset consists of 11,691 NIR spectra measured from 742-990 nm of mango samples with corresponding dry matter amount. The mangos originate from Australia during the 2015-2018 growing seasons and vary in region, cultivar, ripeness, and NIR sensor temperature.[53] The 11,691 samples were downsized to a 10,243 training set and this set was used to select source samples from. The target dataset consists of 501 mango samples originating from Brazil in the 2020 growing season made up of two cultivars included in the source collection and varying ripeness. Spectra were recorded from 684-990 nm on a different instrument than the source.[54] Source wavelengths were trimmed to match the target. For computational efficiency, the source and target were sorted reduced to every 50th sample for the source set, totaling 205 samples, and every third sample for the target set, totaling 167 samples. No effort was made to equally represent source and target domain variances such as cultivar, ripeness, etc. Similar to the corn and melamine datasets, 100 random splits are used and mean results are reported. For each prediction situation per random split, 150 samples are selected from the 205 Australia source domain samples for the calibration set and theses samples are removed from target domain consideration when the target domain is Australia; 10 labeled target domain samples from the remaining 55 Australia samples or 167 Brazil samples are used for LMC; and 30 unlabeled target samples from the remaining 45 Australia samples or 157 Brazil samples for NARE. These 30 samples are used for model updating prediction by LMC and NARE. The target analyte distribution has a small shift from the source. Shown in the SI is Fig. S11 with the two distributions overlayed.

**LMC and NARE Tuning Parameter Ranges.** Initially, the weight value range is 1000 times the singular value range of the source dataset incremented in the one's place. The initial LV range is from 1 LV to sufficiently overdetermined, e.g., 99% rule. Before model selection by MDPS, weight and LV values are dynamically resized per dataset to remove model quality zones of convergence.[29,30]

**Software**. Algorithms were developed by the authors using MATLAB 2022a and the Parallel Computing Toolbox. The PRISM algorithm can be downloaded.[55] Access to the Parallel Computing Toolbox reduces computational time but is not necessary to run the code.

## RESULTS and DISCUSSION

Underlying all the results in this paper is the assumption that samples similar to a source set will be predicted accurately by that set, and dissimilar samples will be predicted poorly. Because prediction reliability, outlier detection, and matrix matching are related, just different goals, outlier detection and matrix matching results are shown and discussed in the SI. Also shown in the SI are enlargements of all the Figures.

**Prediction Reliability**. Plotted in Fig. 1 are the source predicting target (SPT) prediction errors ($|y - \hat{y}|$) against PRISM values for are all 9 corn dataset combinations across the four analytes making 36 situations. The left plot shows all samples acting as targets and the corresponding 36 mean PRISM values are presented on the right. The general trend over all the scenarios is prediction error increasing with more negative PRISM Z scores for greater dissimilarity. The relationship is more linear relative to each analyte. The slope of an analyte specific line characterizes the sensitivity of the prediction error to matrix effects, i.e., how much error to expect per unit change in PRISM.

Recalling that more negative PRISM Z scores indicate a greater degree of dissimilarity from the source set, Fig. 1 shows those samples and sets identified as similar by PRISM are predicted well. Likewise, samples identified dissimilar are typically predicted poorly. As expected, when calibration samples and target samples are measured on the same instrument, Z scores are closest to 0 and all samples are accurately predicted.

Interestingly, two primary instrument comparison groups appear. The grouping on the left in both Fig. 1 plots involves instrument m5 as the source or target instrument respectively compared to instrument mp5 or mp6 indicating m5 is dissimilar to these instruments. Conversely, when instruments mp5 or mp6 are compared to each other, PRISM Z scores specify these instruments as similar. PRISM shows that these instruments can predict samples measured on the other instrument.

Of the four analytes, oil is well predicted regardless of the Z score value. Noticeable is that four of the oil comparison situations involving instrument m5 as the source or target are still predicted well even though greater Z score values indicate unreliable predictions are expected. It is difficult to ascertain exactly why but a possible reason is that respective instrument source models are naturally orthogonal to the oil matrix effect differences. Provided in the SI is further discussion including Fig. S14 inferring model orthogonality as a reasonable explanation. Even though these samples are predicted accurately with the larger Z scores, the usual solution would be to update the model to the new target instrument. The Model Update section shows that oil is still accurately predicted.

Displayed in Fig. 2 are respective PRISM$_X$ and PRISM$_y$ values combined to form the PRISM values in Fig. 1. Revealed is that PRISM$_y$ is responsible for the two groupings observed in Fig. 1. The PRISM$_y$ values produce the differentiation because PRISM$_y$ is designed to detect minor variations that can affect predictive accuracy but do not largely affect spectral shape.

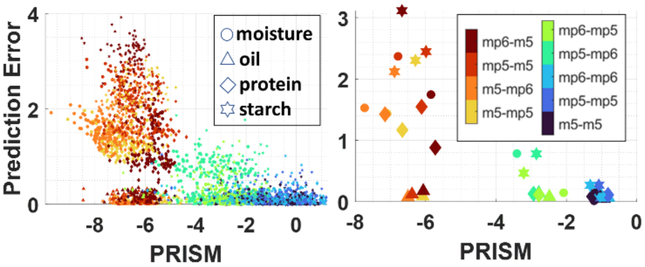


**Fig. 1.** SPT prediction errors against PRISM Z scores for the 36 corn situations color coded to 9 dataset combinations with data point shapes representing the analyte. Shown are all samples (left) and mean PRISM Z scores (right).

Presented in Fig. 3 are SPT prediction errors against PRISM values for all 25 temperature dataset combinations across the three analytes making 75 situations. Trends regarding the correlation of predicted error to PRISM are similar to that as the corn data in Fig. 1. Prediction errors increase as PRISM Z scores become more negative with greater temperature difference disclosing greater dissimilarity between source and target samples. Again, PRISM Z scores effectively indicate at approximately -1 predictions are assuredly accurate compared to anticipating when prediction errors will begin to degrade. This PRISM Z score is where calibration and prediction at the same temperature transitions to mild temperature matrix effect shifts. These temperature results are similar to those previously obtained[56] using SRD rankings in an outlier detection set-up.[33]

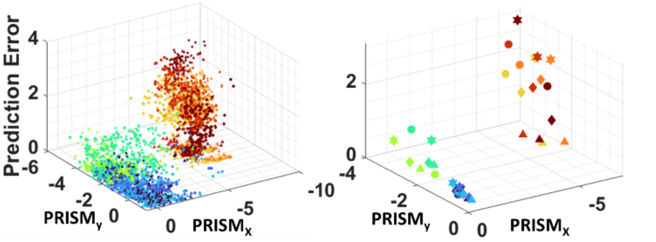


**Fig. 2.** SPT prediction error against PRISM$_X$ and PRISM$_y$ Z scores for the corn data as in Fig. 1.

Summarizing, PRISM Z scores identify when analyte predictions are highly likely to be accurate versus when predictions are more unlikely to be accurate. In this framework, PRISM can be thought as determining if a new sample is an outlier to the calibration set. Lastly, it is worth mentioning again that related to the AD, it has been found that prediction error does not depend on the machine-learning method, but on the similarity of the target domain to the calibration source samples.[21-14] Specifically, if the prediction reliability is deemed satisfactory, say by PLS as used in this study, then all calibration (machine-learning) methods will essentially predict with equal accuracy.

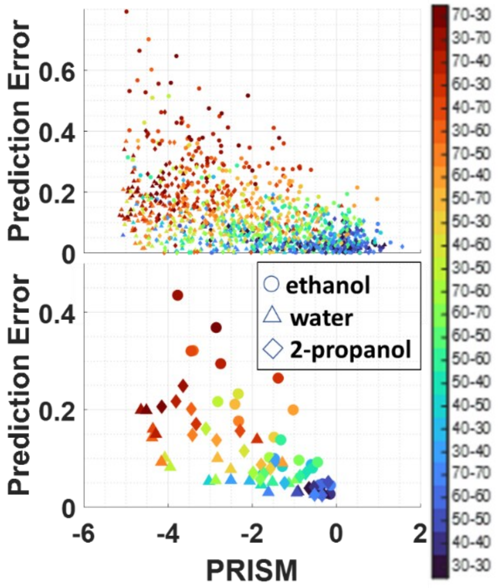


**Fig. 3.** SPT prediction errors against PRISM Z scores for the 75 temperature situations colored to the 25 dataset comparisons with data point shapes symbolizing analytes. Shown are all samples (top) and mean PRISM Z scores (bottom).

**Model Update.** The corn dataset is often used for model updating. Plotted in Fig. 4 are results for no model updating (source predicting target (SPT)), LMC, and NARE for all four analytes and the nine possible combinations of source and target instrument updating. Shown for each situation are mean results across the 100 random data splits for the MDPS[29,30] selected models. First quartile results for all models formed are presented in the SI. The SPT results in Fig. 4 and SI are as expected where prediction errors degrade as the source and target domains become more dissimilar indicated by the more negative PRISM Z scores. As in Figs. 1 and 2, oil is again the exception. The best SPT results are generally when the source and target domains are the same instrument. The SPT $R^2$ values also degrade as the source and target domains become less similar. However, NARE and LMC trends differ for most of the analytes. At more negative PRISM Z scores, NARE does not perform as well as LMC that updates with only a few labeled samples. This observation persists with the first quartile results in the SI. Since LMC consistently produces accurate results compared to NARE, it is generally recommended. However, since the collection of reference values from the target domain can be unreasonable for some applications, NARE can be similarly effective when the analyte domains are matched. Lastly, models selected by MDPS generally provide similar prediction errors as the first quartile (see SI), indicating that MDPS continues to select acceptable models when dataset similarity decreases.

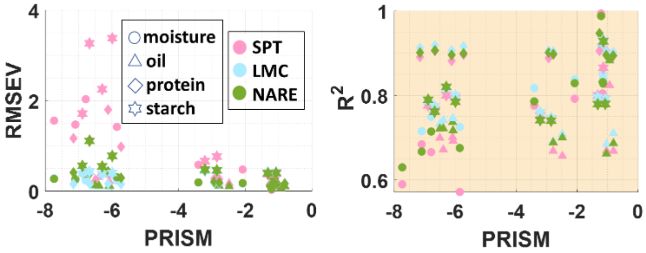


**Fig. 4.** Mean corn RMSEV and R2 values against PRISM Z scores for the 9 updating situations using model updating by LMC and NARE and no updating with SPT. All models selected by MDPS.

As noted earlier, the analyte distributions between source and target domains need to be matched for NARE to be successful. In corn, this is always the case, but the melamine dataset was reconfigured to construct unmatched analyte distributions with no analyte overlap between the high and low domains. Plotted in Fig. 5 are the results for these contrived model updating situations described in the Experimental section. The prominent observation is that LMC is able to maintain low prediction errors for all for situations including the instances where the analyte distributions are not matched (easy and hard transfer learning). Similar to the corn dataset, the best SPT results are generally when the source and target domains are the same (the calibration situation). When the analyte distributions are matched for domain adaptation cases, NARE performs just as well as LMC. However, when the analyte distributions are mismatched as with the hard and easy transfer situations, then NARE often underperforms LMC and SPT independent of Z scores. The particular matrix effect difference in NARE hard and easy transfer causes the models to be orthogonal not only to the matrix effect differences, but also the net analyte difference between the two sets and effectively eliminates any possibility for model extrapolations. Thus, NARE is highly effective relative to Z scores if source and target domains have equivalent analyte distributions as in the calibration and domain adaptation situations. First quartile results in the SI substantiate the results in Fig. 5.

Mango model updating results presented in Fig. S13 of the SI corroborate NARE's difficulty observed in Fig. 5 due to the source and target (Australia and Brazil) analyte distribution mismatch shown in Fig. S11. The LMC results are accurate regardless. The PRISM Z scores plotted in Fig. S12 agree with other datasets where greater negative PRISM Z scores indicate am upper-bound on prediction reliability.

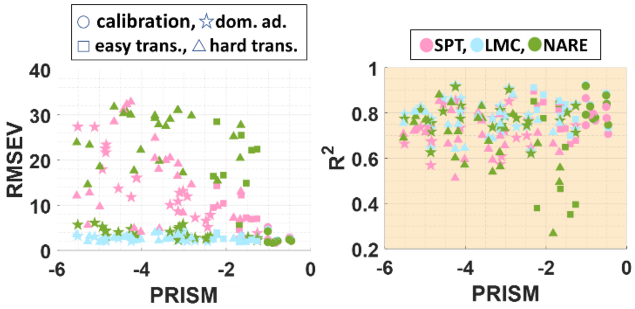


**Fig. 5.** Mean melamine RMSEV and R2 values against PRISM Z scores for the 64 updating situations using model updating by LMC and NARE and no updating with SPT. Data point shapes indicate the model updating categories calibration, domain adaptation, easy transfer learning, and hard transfer learning defined in the Melamine section of the Experimental section. All models are selected by MDPS.

From the results presented in Figs. 5, S12 and S13 it presently seems infeasible for PRISM to consistently identify when the analyte distributions are not matched. An alternative interpretation of PRISM Z scores is that analyte and matrix effect similarities are simultaneously expressed. That is, any given PRISM Z score is degenerate as to whether it only conveys analyte information, only matrix effect content, or partially both facets. If target sample analyte amounts were approximately known, then these values could be decoupled, but with entirely unlabeled target data, this task is difficult if not impossible for the easy and hard transfer cases. Nevertheless, even without a prior information, PRISM, as studied here, weakly correlates to prediction error for the hard and easy transfer situations and is more correlated in the calibration and domain adaptation circumstances.

**Other Data Types and PRISM as a Classification Method.** The dynamics of PRISM has been demonstrated with NIR data. Detailed in this section are other data types previously used with several $\text{PRISM}_X$ and $\text{PRISM}_y$ measures. The section concludes discussing PRISM for one- and multi-class classification.

Many $\text{PRISM}_X$ measures have been used for food and beverage characterization. These include using mid-infrared, NIR, UV, and VIS full spectra and thermogravimetric data for beer product authentication,[35] ICP-MS for authenticating fava beans from Santorini,[57] mid-infrared to detect adulteration of strawberry puree with other fruits,[36] contamination of clams with heavy metals,[57] authentication of meat as either turkey or chicken,[57] and microplastic identification.[58] A collection of 13 chemical measurements were used to classify three wine cultivars frown in Italy.[35] Lastly, $\text{PRISM}_X$ measures were applied to restore defaced serial numbers using lock-in infrared thermography.[37,38] The original $\text{PRISM}_y$ measures were recently developed and applied to NMR data for matrix matching samples to particular calibration sets.[15] Because $\text{PRISM}_X$ and $\text{PRISM}_y$ similarity measures successfully served their purposes with numerous other data types and matrix effects dominate most measurements, it is reasonable to assume that PRISM will work the same with these and other data types as it did with NIR data.

Outlier detection and matrix matching described in the SI are respectively analogues to one- and multi-class classification. Outlier detection and one-class classification both seek to identify whether a sample belongs to one particular group or not. Matrix matching and multi-class classification both identify which source set a target sample is most similar to. The PRISM Z score can be easily altered to become one- and multi-class classification algorithms, since the only change is that the $PRISM_y$ measures are not used. The $PRISM_X$ in the classification context would likely be especially useful due to the interpretability of the Z scores and ease of thresholding for rigorous one-class modeling (without any non-class samples to optimize).

## CONCLUSIONS

Composite PRISM Z scores balance two complementary perspectives, $PRISM_X$ and $PRISM_y$, to fully characterize the similarity of two domains. The $PRISM_X$ view focuses on exploring spectral matrix effected domain differences and $PRISM_y$ probes the domain differences using a model vector directing the comparison towards analyte differences. Presented in the SI is a theoretical framework to understand PRISM.

Because PRISM is based on a statistical Z score, it is data set independent and interpretable Z score thresholds can be set. The PRISM Z score can be considered a similarity ranking relative to the mean source domain matrix effect. Samples with PRISM Z scores within $1\sigma$ of the source Z score mean 0 can be considered similar to the source (calibration set) and predictable, i.e., prediction accuracy is essentially guaranteed by the source model and the sample is not considered an outlier to the source sample set. As sample PRISM Z scores deviate further from 0, these samples are more likely less similar to the source domain and hence, predicted with less reliability. A PRISM Z score around -3 yields moderate potential inaccuracy and a PRISM Z score beyond -5 indicates a high likelihood of inaccurate predictions.

Related to assessing the level of difference between source and target domains for model generalizability is assessing when model updating is needed. In multiple datasets it was shown that SPT degrades when the target sample PRISM Z scores indicate greater dissimilarity, but LMC with only a few target reference values is consistently accurate throughout the PRISM Z score range. NARE is shown to be a strong updating method if the analyte distributions are equivalent. Otherwise, inaccurate predictions are likely. The following guidelines naturally arise from these observations. If the PRISM Z score is less negative than -2 (within $2\sigma$ of 0), then LMC and NARE are often equivalent to SPT and SPT would be favored because of its interpretability. If the PRISM Z score is more negative than -2 and target samples are expected to have an equivalent analyte distribution as the source samples, then NARE can be used for accurate target sample analyte predictions. Otherwise, if the analyte distributions are not matched, then LMC can be used.

The only parameter requiring optimization with PRISM is the number of PLS LVs. By using the U-curve approach, PRISM is essentially an autonomous similarity measure. It is data set invariant and adaptable to adding and/or removing similarity measures to the user's preference. Because PRISM holistically characterizes similarity from an abundance of similarity perspectives, there does not appear to be a need to identify an optimal weighting scheme for each measure. Thus, the user of PRISM needs no advanced chemometrics knowledge.

## ASSOCIATED CONTENT

### Supporting Information

Supporting information contains the underlying mathematics of PRISM, some further experimental details, and additional results with graphics and discussion.

## AUTHOR INFORMATION

### Corresponding Author

* Email: johnkalivas1@isu.edu

## ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. CHE-1904166 (co-funded by CDS&E and the Office of Investigative and Forensic Sciences in the National Institute of Justice) and is gratefully acknowledged by the authors. The authors thank Puneet Mishra for providing the mango data.

## REFERENCES

(1) Todeschini, R.; Ballabio, D.; Consonni, V.; Grisoni, F. A New Concept of Higher-Order Similarity and the Role of Distance/Similarity Measures in Local Classification Methods. *Chemom. Intell. Lab. Syst.*, **2016**, *157*, 50-57.

(2) Goodman, N. Seven Structures on Similarity in Problems and Projects. 437-447. Bobbs-Merri;, New York, **1972**.

(3) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity- A Review. *QSAR Comb. Sci.*, **2003**, *22*, 1006-1026.

(4) Todeschini, R.; Ballabio, D.; Consonni, V. Distances and Similarity Measures in Chemometrics and Chemoinformatics. In Encyclopedia of Analytical Chemistry, R.A. Meyers (Ed.), 2020.

(5) Brereton, R.G.; Lloyd, G.R. Re-evaluating the Role of the Mahalanobis Distance Measure. *J. Chemom.*, **2016***, 30*, 134-143.

(6) Joun-Rimbaus, D.; Massart, D.L.; Saby C.A.; Puel, C. Determination of the Representativity Between Two Multidimensional Data Sets by a Comparison of Their Structure. *Chemom. Intell. Lab. Syst.*, **1998**, *40*, 129-144.

(7) Joun-Rimbaus, D.; Massart, D.L.; Saby C.A.; Puel, C. Characterization of the Representivity of Selected Sets of Samples in Multivariate Calibration and Pattern Recognition. *Anal. Chim. Acta.*, **1997**, *350*, 149-161.

(8) Wold S.; Søstrøm M. SIMCA. A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In Chemometrics: Theory and Publications, B.R. Kowalski (Ed.), American Chemical Society: Washington DC, USA, 1977; 243–282.

(9) Samuel, A.Z.; Mukojima, R.; Horii, S.; Ando, M.; Egashira, S.; Nakashima, T.; Iwatsuki, M.; Takyeama, H. On Selecting a Suitable Spectral Matching Method for Automated Analytical Applications of Raman Spectroscopy. *ACS Omega*, **2021**, *6*, 2060-2065.

(10) Gurung, A.; Kalivas, J.H. Model Selection Challenges with Application to Multivariate Calibration Updating Methods. *J. Chemom.*, **2020**, *34*.e3245.

(11) Valle, S.; Weihua L.; Qin, S.J. Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. *Ind. Eng. Chem. Res.,* **1999***, 38*, 4389-4401.

(12) Zwick, W.R.; Velicer, W.F. Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin,* **1986***, 99*, 432-442.

(13) Miller, C.E. Chemical Principles of Near-Infrared Technology. In Near-Infrared Technology in the Agriculture and Food Industries 2nd ed., P. Williams and K. Norris (Eds.), 2001.

(14) Williams, P.C. Implementation of Near-Infrared Technology. In Near-Infrared Technology in the Agriculture and Food Industries 2nd ed., P. Williams and K. Norris (Eds.), 2001.

(15) Lemos, T.; Emerson, R.M.; Kalivas, J.H. Identifying Chemical, Physical, and Instrumental Matrix Matched Samples by Leveraging Spectral Model Regression Vectors. *Anal. Chem.*, **2020**, *92*, 815-823.

(16) Desai, B.; Dixon, K.; Farrant, E.; Feng, Q.; Gibson, K.R.; van Hoorn, W.P.; Mills, J.; Morgan, T.; Parry, D.M.; Ramjee, M.K.; Selway, C.N.; Traver, G.J.; Whitlock, G.; Wright, A.G. Rapid Discovery of a Novel Seris of Abl Kinase Inhibitors by Application of an Integrated Microfluidic Synthesis and Screening Platform. *J. Med. Chem.*, **2013**, *56*, 3033-3047.

(17) Reker, D.; Schneider, P.; Schneider, G. Multi-objective Active Machine Learning Rapidly Improves Structure-Activity Models and Reveals New Protein-Protein Interaction Inhibitors. *Chem. Sci.*, **2016**, *7*, 3919-39-27

(18) Toplak, M.; Močnik, R.; Polajnar, M.; Bosnić, Z.; Carlsson, L.; Hasselgren, C.; Demšar, J.; Boyer, S.; Zupan, B.; Stålring, J. Assessment of Machine Learning Reliability Methods for Quantifying the Applicability Domain of QSAR Regression Models. *J. Chem. Inf. Model.*, **2014**, *54*, 431-441.

(19) Sheridan, R.P.; Hunt, P.; Culberson, J.C. Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model.,* **2006**, *46*, 180-192.

(20) De, P.; Kar, S.; Ambure, P.; Roy, K. Prediction reliability of QSAR models; An Overview of Various Validation Tools. *Arch. Toxicol.*, **2022**, *96*, 1279-1295.

(21) Bosnić, Z.; Kononenko, I. Comparison of Approaches for Estimating Reliability of Individual Regression Predictions. *Data Knowl. Eng.*, **2008**, *67*, 504-516.

(22) Bosnić, Z.; Kononenko, I. An Overview of Advances in Reliability Estimation of Individual Predictions in Machine Learning. *Intell. Data Anal.*, **2009**, *13*, 385-401.

(23) Sheridan, R.P.; Feuston, B.P.; Maiorov, V.N.; Kearsley, S.K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1912-1928.

(24) Liu, R.; Glover, K.P.; Feasel, M.G.; Wallqvist, A. General Approach to Estimate Error Bars for Quantitative Structure-Activity Relationship Predictions of Molecular Activity. *J. Chem. Info. Model.*, **2018**, *58*, 1561-1575.

(25) Liu, R.; Wang, H.; Glover, K.P.; Feasel, M.G.; Wallqvist, A. Dissecting Machine-Learning Prediction Models of Molecular Activity: Is an Applicability Domain Needed for Quantitative Structure -Activity Relationship Models Based on Deep Neural Networks? *J. Chem. Info. Model.*, **2019**, *59*, 117-126.

(26) Su, M.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Tapping on the Black Box: How is the Scoring Power of a Machine-Learning Scoring Function Dependent on the Training Set? *J. Chem. Info. Model.*, **2020**, *60*, 1122-1136.

(27) Kouw, W.M.; Loog, M. An Introduction to Domain Adaptation and Transfer Learning. **2018**, arXiv:1812.11806 [cs.LG] https://doi.org/10.48550/arXiv.1812.11806

(28) Nikzad-Langerodi, R.; Andries, E. A Chemometrician's Guide to Transfer Learning. *J. Chemom.*, **2021**, *35*:e3373.

(29) Spiers, R.C.; Kalivas, J.H. Calibration Model Updating to Novel Sample and Measurement Conditions without Reference Values. *Anal. Chem.*, **2021**, *93*, 9688-9696.

(30) Spiers, R.C.; Kalivas, J.H. Reliable Model Selection without Reference Values by Utilizing Model Diversity with Prediction Similarity. *J. Chem. Inf. Model.,* **2021**, *61*, 2220-2230.

(31) Nikzad-Langerodi, R.; Zellinger, W.; Lughofer, E.; Saminger-Platz, S. Anal. Chem. **2018**, *90*, 6693−6701.

(32) Andries, E.; Kalivas, J.H.; Gurung, A. Sample and Feature Augmentation Strategies for Calibration Updating. *J. Chemom.*, **2019**, *33*:e3080.

(33) Brownfield, B.; Kalivas, J.H. Consensus Outlier Detection Using Sum of Ranking Differences of Common and New Outlier Measures Without Tuning Parameter Selections. *Anal. Chem.*, **2017**, *89*, 5087-5094.

(34) Rodionova, O. Y.; Oliveri, P.; Pomerantsev, A.L. Rigorous and Compliant Approaches to One-class Classification. *Chemom. Intell. Lab. Syst.*, **2016**, *159*, 89−96.

(35) Brownfield, B.; Lemos, T.; Kalivas, J.H. Consensus Classification Using Non-Optimized Classifiers. *Anal. Chem0.*, **2018**, *90*, 4429-4437.

(36) Lemos, T.; Kalivas, J.H. Self-Optimized One-Class Classification Using Sum of Ranking Differences Combined with a Receiver Operator Characteristic Curve. *Anal. Chem.*, **2020***, 92*, 5354-5361.

(37) Unobe, I.; Lau, L.; Kalivas, J.; Rodriguez, R.; Sorensen, A. Restoration of Defaced Serial Numbers Using Lock-in Infrared Thermography (Part I). *J. Spectr. Imaging*, **2019**, *8*, Article ID a19.

(38) Unobe, I.; Lau, L.; Kalivas, J.; Rodriguez, R.; Sorensen, A. Restoration of Defaced Serial Numbers Using Lock-in Infrared Thermography (Part II). *J. Spectr. Imaging*, **2019**, *8*, Article ID a20.

(39) Kalivas, J.H.; Green, R.L. Pareto Optimal Multivariate Calibration for Spectroscopic Data. *Appl. Spectrosc.,* **2001***, 55*, 1645-1652.

(40) Green, R.L.; Kalivas, J.H. Graphical Diagnostics for Regression Model Determination with Consideration of the Bias/Variance Trade-off. *Chemom. Intell. Lab. Syst*., **2002**, *60*, 173−188.

(41) Gowen, A.A.; Downey, G.; Esquerre, C.; O'Donnell, C.P. Preventing Over-fitting in PLS Calibration Models of Near-Infrared (NIR) Spectroscopy Data Using Regression Coefficients. *J. Chemom.*, **2011**, *25*, 375−381.

(42) Kalivas, J.H.; Palmer,J. Characterizing Multivariate Calibration Tradeoffs (Bias, Variance, Selectivity, and Sensitivity) to Select Model Tuning Parameters. *J. Chemom.*, **2014**, *28*, 347−357.

(43) Takahama, S.; Dillner, A.M. Model Selection for Partial Least Squares Calibration and Implications for Analysis of Atmospheric Organic Aerosol Samples with Mid-Infrared Spectroscopy. *J. Chemom.*, **2015**, *29*, 659−668.

(44) Sheridan, R.P.; Hunt, P.; Culberson, J.C. Molecular Transformation as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 180-192.

(45) Dossetter, A.G.; Griffen, J.; Leach, A. Matched Molecular Pair Analysis in Drug Discovery. *Drug Discovery. Today*, **2013**, *18*, 724-731.

(46) Kramer, C.; Fuchs, J.E.; Whitebread, S.; Gedek, P.; Liedl, K.R. Matched Molecular Pair Analysis: Significance and Impact of Experimental Uncertainty. *J. Med. Chem.*, **2014**, *57*, 3786-3802.

(47) Tynes, M.; Gao, W.; Burrill, D.J.; Batista, E.R.; Perez, D.; Yang, P.; Lubbers, N. Pairwise Difference Regression: A Machine Learning Meta-algorithm for Improved Prediction and Uncertainty Quantification in Chemical Search. *J. Chem. Inf. Model.*, **2021**, *61*, 3846-3857.

(48) Héberger, K. Sum of Ranking Differences Compares Methods or Models Fairly. *Trends Analyt. Chem.,* **2010***, 29*, 101-109.

(49) Heard, N.A.; Rubin-Delanchy, P. Choosing Between Methods of Combining *p*-Values. *Biometrika,* **2018***, 105*, 239-246.

(50) Wise, B. M.; Gallagher, N. B. Eigenvector Research, Manson, WA. http://www.eigenvector.com/data/index.htm

(51) Wülfert, F.; Kok, W.T.; Smilde, A.K. Influence of Temperature on Vibrational Spectra and Consequences for the Predictive Ability of Multivariate Models. *Anal. Chem.*, **1998**, *70*, 1761-1767.

(52) Nikzad-Langerodi, R.; Zellinger, W.; Lughofer, E.; Saminger-Platz, S. Domain-Invariant Partial-Least-Squares Regression. *Anal. Chem.,* **2018***, 99*, 6693-6701.

(53) Anderson, N.T.; Walsh, K.B.; Flynn, J.R.; Walsh, J.P. Achieving Robustness Across Season, Location and Cultivar for a NIRS Model for Intact Mango Fruit Dry Matter Content. II. Local PLS and Nonlinear Models. *Postharvest Biol. Technol.*, **2021***, 171,* 111358.

(54) Mishra, P.; Passos, D. Deep Chemometrics: Validation and Transfer of a Global Deep Near-Infrared Fruit Model to Use It on a New Portable Instrument. *J. Chemom.,* **2021**, *35*:e3367.

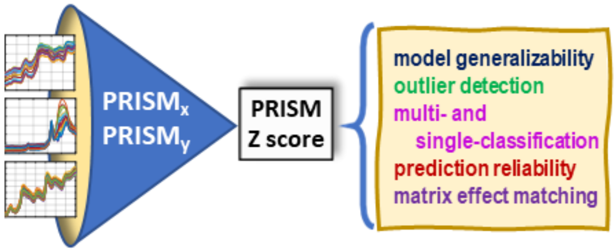
(55) https://www.isu.edu/chem/faculty/staffdirectoryentries/kalivas-john.html, accessed July 2023.

(56) Kalivas, J.H.; Brownfield, B.; Karki, B.J. Sample-Wise Spectral Multivariate Calibration Desensitized to New Artifacts Relative to the Calibration Data Using a Residual Penalty. *J. Chenom.* **2017**, *31*:e2873.

(57) Kalivas, J.H; Lemos, T. Automatic Food and Beverage Authentication and Adulteration Detection by Classification Hybrid Fusion. *J. Chemom.*, **2023**, *37*:e3371.

(58) Chabuka, B.K.; Kalivas, J.H. Application of Hybrid Fusion Classification Process for Identification of Microplastics Based on Fourier Transform Infrared Spectroscopy. *Appl. Spectrosc.*, **2020**, *74*, 1167-1183.

FOR TOC ONLY


$PRISM_x$
$PRISM_y$
PRISM
Z score
model generalizability
outlier detection
multi- and
single-classification
prediction reliability
matrix effect matching