JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

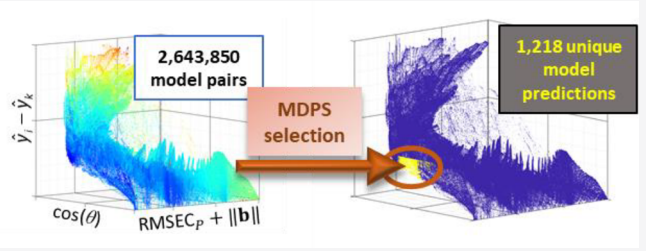
Reliable Model Selection without Reference Values by Utilizing Model Diversity with Prediction Similarity

Robert C. Spiers and John H. Kalivas*

Cite This: J. Chem. Inf. Model. 2021, 61, 2220–2230 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Predictive modeling (calibration or training) with various data formats, such as near-infrared (NIR) spectra and quantitative structure−activity relationship (QSAR) data, provides essential information if a proper model is selected. Similarly, with a general model selection approach, spectral model maintenance (updating) from original modeling conditions to new conditions can be performed for dynamic modeling. Fundamental modeling (partial least-squares (PLS) and others) and maintenance processes (domain adaptation or transfer learning and others) require selection of tuning parameter(s) values to isolate models that can accurately predict new samples or molecules, e.g., number of PLS latent variables to predict analyte concentration. Regardless of the modeling task, model selection is complex and without a reliable protocol. Tuning parameter selection typically depends on only one model quality measure assessing model bias using prediction accuracy. Developed in this paper is a generic model selection process using concepts from consensus modeling and QSAR activity landscapes. It is a consensus filtering approach that prioritizes model diversity (MD) while conserving prediction similarity (PS) fused with a common bias-variance trade-off measure. A significant feature of MDPS is that a cross-validation scheme is not needed because models are selected relative to predicting new samples or molecules, i.e., model selection uses unlabeled samples (without reference values) for active predictions. The versatility and reliability of MDPS model selection is shown using four NIR data sets and a QSAR data set. The study also substantiates the Rashomon effect where there is not one best model tuning parameter value that provides accurate predictions.



INTRODUCTION

Modeling is a universal problem concerned with forming a mathematical relationship (model) between measured variables for a collection of samples and respective reference property quantities. The model is then used to predict future sample content. For example, spectral data such as near-infrared (NIR) is used with analyte concentration values to form a multivariate calibration (also referred to as training) allowing concentration predication of new samples. Another example is to form quantitative structure−activity relationships (QSAR) between molecular descriptors and the activity of interest.

Generating mathematical calibration models for prediction are often performed using partial least-squares (PLS), ridge regression (RR), and other methods.[1,2] These calibration methods require tuning parameter values to be varied (latent variables (LVs) for PLS and ridge parameters for RR) to form a collection of models (regression vectors). From the collection, one model is selected. The PLS LV range is 1 through the mathematical rank k, and the RR ridge parameter η value ranges from 0 to ∞. Thus, model selection is critical for choosing a particular tuning parameter value corresponding to a robust model with an appropriate bias-variance balance (not under- or overfitted). The selected model needs to also deliver accurate predicted property values for new samples (or molecules).

A significant modeling issue is model maintenance. Model corrections become necessary when measurement (*secondary*) conditions (sample (molecule), instrument, and environment matrix effects) for new secondary samples become sufficiently different from the original (*primary*) calibration conditions. Because conditions have changed, the original model does not provide accurate property values for secondary samples.[17]

Model updating is one way to fix this problem. The process usually involves augmenting the original primary calibration set with a small set of secondary samples in order to orient the new model regression vector to an appropriate direction while simultaneously adjusting the model vector magnitude. At least two tuning parameters are required with model updating: one to solve the calibration equation and one to weight the few augmented secondary samples.

Received: December 28, 2020
Published: April 26, 2021

https://doi.org/10.1021/acs.jcim.0c01493
J. Chem. Inf. Model. 2021, 61, 2220–2230

There are a variety of accepted model selection mechanisms for PLS, RR, and others with one tuning parameter. Some of these methods include versions of cross-validation (a common approach),[3,4] permutation,[5] L-curve,[6,7] U-curve,[8−12] sum of ranking differences (SRD),[13,14] H-principle,[15] Mallow's Cp,[16] and others. A few of these methods rely on only one model quality measure such as prediction error (a bias measure), and others combine one or more bias measures with one or more variance or model complexity measure such as the model regression vector magnitude.

At least two model quality measures should be used when selecting only one tuning parameter value for a basic calibration, e.g., PLS or RR where new samples are matrix matched (similar) to the calibration set. One measure should characterize bias and the other variance thereby selecting models with an appropriate bias-variance trade-off. However, in model maintenance processes with multiple tuning parameters, evaluating the bias-variance trade-off for model selection is not as straightforward.[18−20] Specifically, which bias measures to include in the model selection protocol depends on the degree of similarity between the primary and secondary conditions, e.g., the degree of matrix (domain) matching between primary and secondary spectra. These domain differences influencing the ability to model update were reported in other work.[21]

Another difficulty in selecting calibration or maintenance models is the large number of diverse models that can be formed with unique features and yet predict accurately.[22−25] This problem is compounded in model maintenance when two or more tuning parameters are involved. Thus, there are many tuning parameter values that can form applicable models in the effective predictive domain of the new samples. Identifying the one best tuning parameter value(s) is thus not practical. This point is becoming more recognized, and recent work attempts to define this collection of equally accurate predicting models as the point of curvature in a Rashomon curve.[26] The application of the term "Rashomon effect" to modeling was first noted by Leo Breinman.[27] The Rashomon effect basically applies to any situation where a single event can be described in a variety of ways for different reasons that depend on the application. In modeling, there are often a large number of models that can accurately predict a single sample.

Needed is a model selection approach to identify a collection of models providing acceptable bias-variance trade-offs. The approach should function across multiple data sets and calibration and maintenance methods and not depend on the degree of domain differences. Developed and validated in this paper is a novel generic model selection process that uses concepts from consensus modeling[18,28−31] and QSAR activity landscapes to solve the problems.[32−35]

Consensus modeling theory designates using multiple models rather than one for predicting property quantities. For example, consensus modeling expects PLS or RR model vectors at selected respective LVs and η values to maintain diversity and yet generate accurate predictions. Diversity is determined more by vector direction (shape) than magnitude. The goal of model selection established in this paper is to select these diverse models for a robust accurate property prediction. Model predictions are fused to report the final sample property amount.

To accomplish selection of diverse models with accurately predicted property values, tactics are used analogous to those used to form QSAR activity landscape maps for assessing structure−activity relationships between molecular structure and potency (activity) differences.[32−35] Construction of an activity landscape requires computing all possible molecular pairwise comparisons. A landscape map is formed by plotting activity similarities against structure diversity allowing evaluation of the structure−activity relationships.

For the model selection method presented in this work, all possible pairwise model vector comparisons are evaluated for model dissimilarity and sample property prediction similarity. Prediction similarity is used as a substitute for accuracy because if diverse models predict accurately, then the models should also equally predict similar property values. A prediction landscape map is formed by plotting prediction similarities against respective model differences. Because two models can also predict similarly and yet predict inaccurately, a model bias-variance trade-off measure is included with prediction similarities in prediction landscape maps.

The predication landscape map is filtered identifying those tuning parameter values (models) that prioritize model diversity (MD) while conserving prediction similarity (PS) relative to a commonly used bias-variance trade-off measure. Results are presented in this paper evaluating the consistency of the MDPS method over four near-infrared (NIR) data sets to select accurately predicting single and multiple tuning parameter calibration and maintenance models.

A significant feature of MDPS is that because prediction similarities are used to identify models, reference property values (labels) are not needed. Thus, models are selected to only predict new samples, and complex cross-validation data splitting schemes are not needed. The usual cross-validation approach is to select one model from the training set based on an elaborate cross-validation design and then predict the new samples with the one model. With MDPS, models are selected from the collection of models formed by the training set targeting new sample predictions.

The original intent of MDPS was tuning parameter selection for model updating because that is the situation where there is not a strong model selection mechanism. Thus, the paper emphasizes application of MDPS to two (and potentially more) tuning parameter model updating circumstances. The MDPS process is also applied to single tuning parameter RR and PLS methods. The data set focus is NIR spectra, but MDPS is also relevant and important to other modeling situations such as with QSAR data, also examined in this paper or single tuning parameter selection.

METHODS

Original (Primary) Modeling. Multivariate calibration (modeling) requires a collection of m samples of known property reference values signified as an $m \times 1$ $\mathbf{y}$ vector. These calibration samples are measured at w sensors, e.g., NIR spectra measured over w wavelengths, expressed as an $m \times w$ $\mathbf{X}$ matrix.

The $\mathbf{X}$,$\mathbf{y}$ calibration data are used with $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ that is solved for an estimated model regression vector $\hat{\mathbf{b}}$. The PLS and RR methods are used in this paper. Respective tuning parameter values are varied through ranges to create sets of models to select from. The tuning parameter for PLS is the number for the LV range (1, mathematical rank k). For RR, the tuning parameter is the ridge parameter η value that weights an identity matrix augmented to $\mathbf{X}$ and the range $(0,\infty)$ albeit models eventually converge at large η values.

https://doi.org/10.1021/acs.jcim.0c01493
J. Chem. Inf. Model. 2021, 61, 2220−2230

The pool of PLS models to select from is discrete due to the discontinuous nature of LVs, and the RR models are continuously depending on the number of digits in the ridge value past the decimal point. Thus, model selection for PLS and RR represents unique challenges. The novel MDPS is shown to select acceptable PLS and RR models.

Model Updating. While PLS and RR calibrations need only one tuning parameter, model updating methods typically require at least two tuning parameter values. Two model updating methods are used to demonstrate the utility of MDPS to simultaneously select two turning parameters.

The first approach is commonly used and referred to as local mean centering (LMC)[21,36] where the primary (P) and secondary (S) data are mean-centered to respective means in the augmented system given by

$$\begin{pmatrix} \mathbf{y}_P \\ \lambda \mathbf{y}_S \end{pmatrix} = \begin{pmatrix} \mathbf{X}_P \\ \lambda \mathbf{X}_S \end{pmatrix} \mathbf{b}$$

where λ weights the augmented secondary samples with values from the range $(0,\infty)$. Regression vector estimates are obtained by applying PLS to the augmented equation. Thus, LMC requires concurrent selection of two tuning parameter values: the number of LVs and the λ weight.

Because every combination of two tuning parameters must be evaluated with MDPS, the number of total models to select from multiplicatively increases compared to when only one tuning parameter is involved. Confounding LMC model selection is that the degree of domain difference between the primary and secondary data dictates which bias model quality measures to use.[20]

The model updating method termed feature augmentation-2 (FA-2)[21] is similar to LMC, but instead of one model vector, two $\mathbf{b}$ model vectors are formed. The relationship is expressed as

$$\begin{pmatrix} \mathbf{y}_P \\ \mathbf{0} \\ \lambda \mathbf{y}_S \end{pmatrix} = \begin{pmatrix} \mathbf{X}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_P \\ \lambda \mathbf{X}_S & \lambda \mathbf{X}_S \end{pmatrix} \begin{pmatrix} \mathbf{b}_P \\ \mathbf{b}_S \end{pmatrix}$$

where $\mathbf{b}_P$ is the model vector for predicting the primary data, and $\mathbf{b}_S$ symbolizes the model vector oriented to capture secondary features different from primary features. The two respective models are summed to form the final predicting model and assessment by MDPS. Similar to LMC, primary and secondary data are respectively mean-centered, and FA-2 uses two tuning parameters to form a model: PLS LVs and the λ weight. Analogous to LMC, one set of bias-variance model quality measures was not possible for consistent model selection for different data sets relative to respective domain differences.[20]

Model Selection by Model Diversity Prediction Similarity (MDPS). Model selection using the proposed MDPS process requires comparison of all models pairwise for model differences (diversity) and prediction similarities.

Model diversity is determined by using the cosine of the angle formed between the two model vectors being compared and is computed by

$$\cos(\theta)_{i,j} = \hat{\mathbf{b}}_i^{\mathrm{T}} \hat{\mathbf{b}}_j / (\|\hat{\mathbf{b}}_i\| \|\hat{\mathbf{b}}_j\|)$$

where i and j are tuning parameter indexes for two models, and $\|\bullet\|$ designates the Euclidean vector norm (2-norm, L_2 norm). The $\cos(\theta)$ value focuses on vector shape (direction) differences and ranges from 1 for exact similarity to 0 for orthogonal differences. The Euclidian distance between two vectors for magnitude differences is another common vector similarity measure. However, this measure was found to ignore changes attributed to λ in LMC and FA-2.

All prediction similarities used in MDPS are for new samples not part of the calibration or updating set, i.e., the new samples needing to be predicted for property quantity. For consistency between basic (primary) PLS or RR calibrations and the LMC and FA-2 model updating methods, prediction similarities used with MDPS are termed secondary prediction differences (SPD) computed by

$$\mathrm{SPD}_{i,j} = \sum_{n=1}^{s} |\hat{y}_{n,i} - \hat{y}_{n,j}|$$

where $\hat{y}$ signifies PLS, RR, LMC, or FA-2 predicted property values for s new samples. The SPD value quantifies sample prediction differences using prediction values from the tuning parameter specific i and j models. The effect of the number of new s samples used to select models is described in the Results and Discussion section, but generally the size s does not matter. The SPD values are ranged scaled (RS) to vary from 0 to 1 inclusive over all possible model pairs by

$$\mathrm{SPD}_{i,j}^{\mathrm{RS}} = (\mathrm{SPD}_{i,j} - \mathrm{SPD}_{\min}) / (\mathrm{SPD}_{\max} - \mathrm{SPD}_{\min})$$

where min and max are minimum and maximum, respectively.

In order to select models with an appropriate bias-variance trade-off balance, the final prediction similarity for two models includes a weighted U-curve term. Including the U-curve allows removing over- and underfitted models from consideration. Of the many possible U-curves,[8−12] the one used with MDPS is the combination of the range scaled mean regression vector 2-norm ($\overline{\|\hat{\mathbf{b}}\|}_{i,j}^{\mathrm{RS}}$) for models i and j and the corresponding range scaled average root-mean-square error for primary calibration samples ($\overline{\mathrm{RMSEC}}_{P_{i,j}}^{\mathrm{RS}}$). The final composite prediction similarity value for models i and j used in the prediction landscape plot is labeled $C_{i,j}$ computed by

$$C_{i,j} = \mathrm{SPD}_{i,j}^{\mathrm{RS}} + \omega(\overline{\|\hat{\mathbf{b}}\|}^{\mathrm{RS}} + \overline{\mathrm{RMSEC}}_P^{\mathrm{RS}})_{i,j} \quad (1)$$

where ω (≥ 0) weights the bias-variance trade-off determined by the U-curve.

Figure 1 shows an example of an MDPS prediction landscape map obtained by plotting composite prediction similarity $C_{i,j}$ values against the corresponding model diversity $\cos(\theta)_{i,j}$ values for each combination of two models. Figure 1 is for LMC with the Goat data, but it characterizes the situation. Model combinations are colored according to the average RMSE of validation (RMSEV) value for the new samples relative to respective model pairs. Shown in Figure 2 is a three-dimensional plot of Figure 1 where the axes are $\cos(\theta)$, SPD, and the RMSEC U-curve. A U-curve shape can be observed in Figure 2, and selected models tend to be at the bottom of the U-curve.

The contrast between the $C_{i,j}$ axes in (a) with $\omega = 0$ and (b) with $\omega = 0.4$ demonstrates the importance of using a weighted U-curve. Weighting the U-curve in Figure 1b shifts upward those models predicting similarly (small $\mathrm{SPD}_{i,j}$ values) but are the poorly predicting underfitted (red points) and overfitted models (light blue points). Because of these upward shifts in

https://doi.org/10.1021/acs.jcim.0c01493
J. Chem. Inf. Model. 2021, 61, 2220−2230

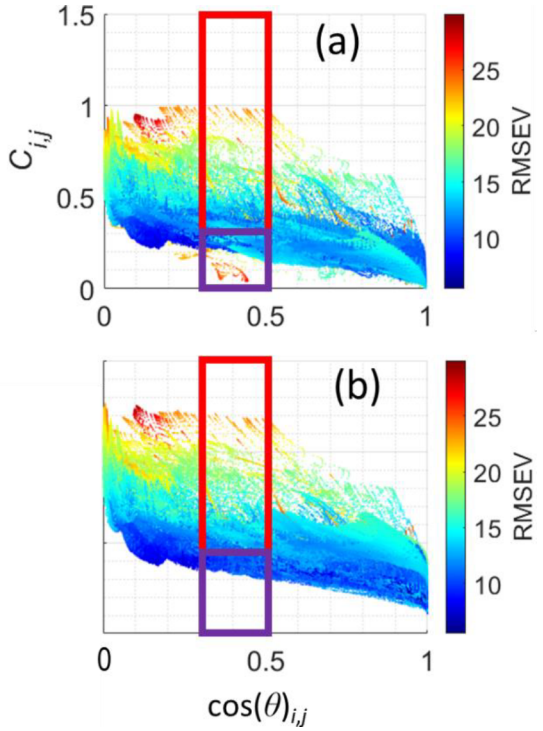


Figure 1. All LMC model pairs for the Goat data situation with 2,300 models creating 2,643,850 combinations. (a) considers $C_{i,j}$ with $\omega = 0$ and (b) has $\omega = 0.4$. Model combinations are colored according to the average RMSEV for respective model pairs. Red boxes encapsulate 316,143 model pairs in the diversity range $0.3 \leq \cos(\theta)_{i,j} \leq 0.5$, and purple boxes contain the 31,614 model pairs selected with the lowest 10% composite prediction similarity $C_{i,j}$ values.

$C_{i,j}$ values, the corresponding models cannot be selected. Figure 2 further shows this effect.

The first step in model selection is to identify all model combinations in a model diversity range (0.3−0.5 for model updating shown by the red rectangle in Figure 1). The 10% subset of model pairs with the lowest composite $C_{i,j}$ values are then selected (purple rectangle). A weighted average of the property predictions from the models in this subset is used for final sample prediction values. Weights determined by the frequency models are selected in the lower 10% subset of model pairs. Specifically,

$$\hat{\mathbf{y}} = \sum_{i=1}^{t} \alpha_i \mathbf{X}_{SU} \hat{\mathbf{b}}_i = \mathbf{X}_{SU} \sum_{i=1}^{t} \alpha_i \hat{\mathbf{b}}_i$$

where α_i denotes the weight associated with the ith model of the t individual models with each α_i determined by the number of times the ith model is selected divided by the total number of models selected in the lower 10%, i.e., the weights sum to 1. Another way to consider the final prediction is that there is one final predicting model $\hat{\mathbf{b}}_f$ formed from a linear combination of the selected models. In this portrayal, $\hat{\mathbf{y}} = \mathbf{X}_{SU}\hat{\mathbf{b}}_f$ where $\hat{\mathbf{b}}_f = \Sigma_{i=1}^{t} \alpha_i \hat{\mathbf{b}}_i$. As a reminder, MDPS selects models for all new samples. Other approaches to forming final predictions for each sample exist including a majority vote of prediction values for a sample, the mode, median, etc. How the model predictions are combined for a final prediction of each sample is up to the user and not part of the MDPS selection. Only the weighted mean is reported here.

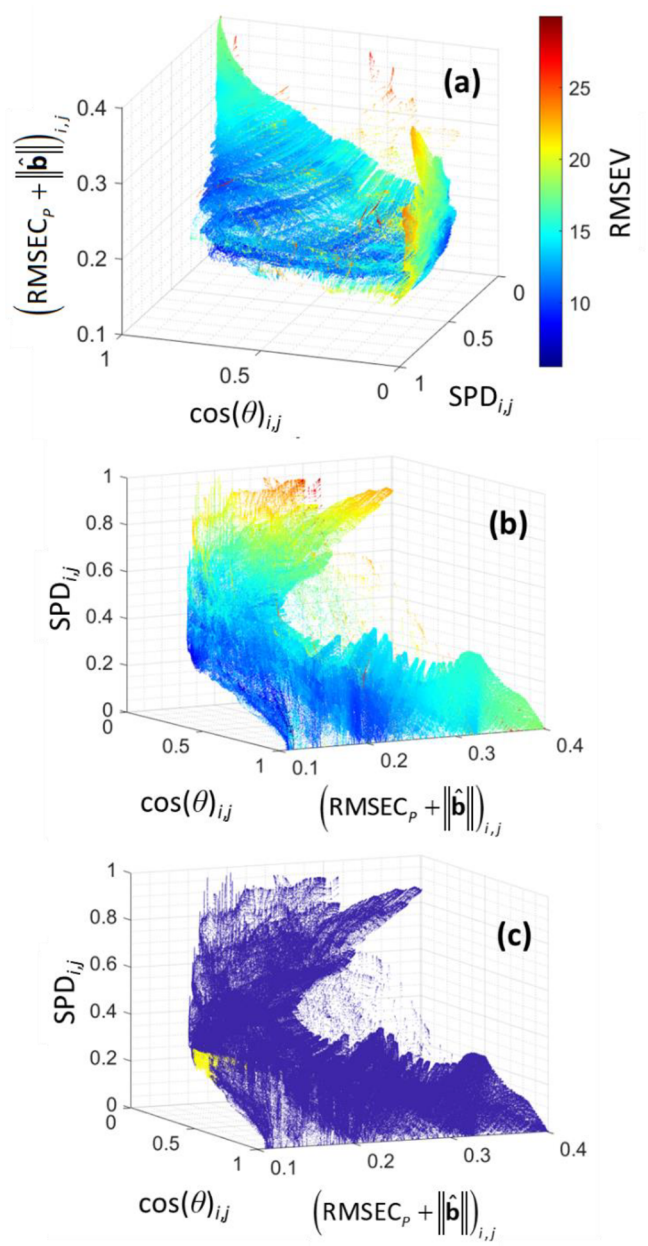


Figure 2. Two views ((a) and (b)) of MPDS terms in eq 1 showing model diversity $\cos(\theta)$ with the composite prediction similarity term separated. Plotted are LMC model pairs for the Goat data set in Figure 1. The RMSEV color bar is the same for (a) and (b), and for (c), purple points are for all model pairs, and yellow points are for those selected in the lower 10% of part b.

For the Goat data set in Figures 1 and 2, there are 2,300 models formed across the LV and λ ranges creating 2,643,850 model combinations. In the model similarity $0.3 \leq \cos(\theta)_{i,j} \leq 0.5$ range, there are 316,143 model pairs leading to 31,614 model pairs remaining in the lower 10% region. Of these pairs, there are 1,218 unique models whose selection frequencies are examined in the Results and Discussion.

In order for MDPS to select good models, there are three adjustable parameters: the $\cos(\theta)$ diversity range, ω for weighting the U-curve, and how many models with the lowest composite prediction similarity values. Thus, it seems selecting model tuning parameter values has been traded for tuning the MDPS process. However, it was found through empirical

https://doi.org/10.1021/acs.jcim.0c01493
J. Chem. Inf. Model. 2021, 61, 2220−2230

observations over many data situations that the values could be fixed relative to the modeling task. For model updating with more than one tuning parameter, the adjustable MDPS parameters were set to $0.3 \leq \cos(\theta)_{i,j} \leq 0.5$, $\omega = 0.4$, and the lowest 10% $C_{i,j}$ values. For the single tuning parameter modeling methods RR and PLS, the MDPS parameters determined best are $0.6 \leq \cos(\theta)_{i,j} \leq 0.8$, $\omega = 1.0$, and the lowest 10% $C_{i,j}$.

EXPERIMENTAL SECTION

Software. All algorithms were developed by the authors using MATLAB R2019a. The MDPS model selection algorithm can be downloaded.[37]

Data Descriptions. Four NIR spectral data sets were used: Corn,[38] Soy,[39] Tablet,[40] and Goat.[41] All 19 calibration and 40 model updating settings were studied. A QSAR data set is also used for three calibration situations.[42]

Corn. Spectra of 80 samples of cornmeal with property values moisture (9.377−10.993%), oil (3.088−3.832%), protein (7.654−9.711%), and starch (62.826−66.472%) were recorded across three instruments: m5, mp5, and mp6. Measured wavelengths ranged from 1100 to 2498 at 2 nm increments for 700 wavelengths. Six updating scenarios are analyzed noted by primary-secondary: m5-mp5, m5-mp6, mp5-m5, mp5-mp6, mp6-m5, and mp6-mp5 for each property for a total of 24 situations.

Soy. Spectra of 60 soy seed samples with property information for moisture (5.9−18.4%), oil (29.0−43.4%), and protein (14.7−22.9%) were measured from 1100 to 2500 nm with 4 nm increments (300 wavelengths in total) on two instruments: R1 and R2. Six updating settings are studied: R1-R2 and R2-R1 each for the three properties.

Tablet. Pharmaceutical tablets were produced and measured in two batches: laboratory (lab) and full production (full) with the active pharmaceutical ingredient (API) Escitolopram. Tablets are subdivided into four types for each batch (types 1−4) based on respective total tablet weights 90, 125, 188, and 250 mg. Because tablet types have different total weights, respective tablet types have different shapes and sizes with tablet thicknesses ranging from 2.9 to 4.3 mm. There are 30 tablets for each batch tablet type making 120 tablets for each batch. The API content for lab is type 1 (4.7432−6.2297), type 2 (6.6456−9.7862%), type 3 (6.6180−9.2715%), and type 4 (6.7182−9.3824%). For the full batch, the API content is type 1 (5.1228−5.8386), type 2 (7.6028−8.40342%), type 3 (7.6400−8.4293%), and type 4 (7.4864−8.4779%). Spectra were measured from 700 to 2500 nm for a total of 404 wavelengths. Primary is always the lab batch, and secondary is always the full batch. Each primary and secondary data set contains two tablet types with one tablet type in each condition being type 1. The other tablet type is one of the remaining three making 9 updating situations: 1&2-1&2, 1&2-1&3, 1&2-1&4, 1&3-1&2, 1&3-1&3, 1&3-1&4, 1&4-1&2, 1&4-1&3, and 1&4-1&4.

Goat. Goat feces samples were analyzed for juniper berry content in 1999 (61 samples with juniper berry 1−50%) and 2002 (48 samples with juniper berry 1−40%). Spectra are measured over 1050 wavelengths from 400 to 2500 nm. Only one updating situation is analyzed: 1999−2002.

QSAR. This data set is a study of carbonic anhydrase (CA) inhibitors and consists of 142 compounds assayed for inhibition of three CA isoenzymes: CAI, CAII, and CAIV. The log of respective inhibition values was modeled using 63 molecular descriptors.

Tuning Parameter Values. For PLS calibration, the number of LVs ranged from 1 through the mathematical rank of the calibration set in the respective $\mathbf{X}$ spectral matrix used with PLS. There are 50 η tuning parameter values for RR formed exponentially decreasing between the highest and lowest singular values of each corresponding $\mathbf{X}$ spectral matrix used with PLS.

For updating by LMC and FA-2, the PLS LVs ranged from 1 through the mathematical rank of each primary $\mathbf{X}_P$ spectral matrix. There are 50 λ tuning parameter values formed exponentially decreasing in the range between the highest and lowest singular values of each corresponding $\mathbf{X}_P$.

Data Splitting for Validation. Random data splitting is used to distribute samples between calibration and validation sets 100 times.[43] Models are selected for each split, and mean RMSE of validation (V) and corresponding R^2 values across the 100 random data splits are assessed. The R^2 values for a data split are obtained from plotting validation set predicted property values against reference values.

For PLS and RR calibrations, each of the 100 random data set splits is divided with 80% for calibration (model formation) and the remaining 20% for validation. Models are selected by MDPS using the 20% validation samples for eq 1 without respective property values. The LMC and FA-2 data set divisions are listed in Table 1 for the number of samples into primary (PRI), augmented secondary calibration (CALS), and secondary validation (VALS) sets.

Table 1. Updating Data Set Specific Sample Divisions for Primary (PRI), Augmented Secondary Calibration (CALS), and Secondary Validation (VALS)

data	PRI	CALS	VALS
Corn	40	5	20
Soy	30	5	15
Tablet	60	6	24
Goat	61	5	20

Model Selection Benchmarks. *MDPS Selection for Model Updating by LMC and FA-2.* Three baseline model prediction errors are needed for comparison to prediction errors from the two model updating methods. These baselines are primary predicting secondary (PPS), secondary predicting secondary (SPS), and small secondary predicting secondary (SSPS). The same secondary validation sets listed in Table 1 are used for each of these three baselines and the MDPS selected LMS and FA-2 models. Shown in figures are PPS, SPS, and SSPS model RMSEV and R^2 boxplot trends at minima and the first two quartiles.

The PPS models come from PLS calibrations of primary samples in Table 1. The PPS RMSEV values for VALS samples should ideally show the necessity for model updating.

Because the SPS RMSEV baseline goal is to show the RMSEV quality that is achievable with an expensive full secondary calibration, all samples in each secondary data set are randomly split 100 times with 60% of the samples being used for calibration and 40% for validation. A viable model updating method should ideally be reasonable relative to SPS RMSEV and R^2 values.

The point of the third baseline SSPS RMSEV values is to verify that a PLS calibration model based on only the small

https://doi.org/10.1021/acs.jcim.0c01493
J. Chem. Inf. Model. 2021, 61, 2220−2230

augmented secondary calibration set CALS in Table 1 is not feasible and the primary data set is needed.

PLS and RR Model Selection. A variety of benchmark model selection methods can be used to establish baselines for comparison to MDPS. For this study, three U-curves are used where model selection at the minima of each U-curve[8−12] is the most straightforward and least judgmental for comparison to MDPS. A U-curve selects the model best minimizing the model 2-norm (or other measure of model complexity but model 2-norm is used here) simultaneously with a prediction error measure.

The three U-curves are compared to MDPS. In one instance, a RMSEC U-curve (U_C) for each random 80% is formed across the tuning parameter range and then used to select a model. The selected model is used to predict the remaining 20% validation set to generate the corresponding RMSEV and R^2 values. This process is repeated for each of the 100 random splits. The other situation performs additional 100 random inner two-way data splits on each of the 80% calibration sets with 60% for forming PLS or RR models and respective RMSEC U-curves on each inner random split. The remaining 40% of the inner split is used to form respective RMSE of cross-validation (CV) U-curves. A model (tuning parameter value) is selected for each of the outer 100 random splits from the minimum of mean RMSECV U-curve across the 100 inner splits (U_{CV}). A tuning parameter value is also selected for each of the outer 100 random splits from the minimum of the average RMSEC and RMSECV U-curves (U_M) across the 100 inner splits. These selected PLS or RR tuning parameter values are the models used from the outer 80% calibration set to predict the 20% validation samples to form final RMSEV and R^2 values. As a reminder, models are only selected by MDPS from the 80% and are based on prediction similarity for the 20% without the property reference values. These MDPS models are then used to compute final RMSEV and R^2 values to compare with respective U-curve selected models.

RESULTS AND DISCUSSION

For each calibration and updating process, minimum and quartile (first and median) boxplots of RMSEV and corresponding R^2 values are used to assess the quality of the models selected over the 100 random data splits. However, with tuning parameter-based calibration and updating methods, there are often an excessive number of under- and overfitted models generated depending on the ranges of tuning parameter values. These subsets of models predict nearly the same, and when compiling model quality measures such as RMSEV values, modeling algorithms can be misrepresented by quartile boxplots. Therefore, it became necessary to identify respective tuning parameter transition regions where variations between models are important (active bias-variance trade-off zones). Beyond these transition zones lie converged over- and underfitted models that need to be removed before quartile boxplots.

Presented in the Supporting Information (SI) is information on how a tuning parameter active bias-variance trade-off zone is identified. Also included are boxplots showing the misrepresentation when tuning parameters are allowed to range into under- and overfitted regions.

Results shown and discussed as follows are based on models selected from the active bias-variance trade-off zones in order to fairly compare respective boxplots. However, it is important for the reader to note that model selection by MDPS, in practice, does not require prior determination of the tuning parameter active bias-variance trade-off zone. Boxplots in the SI demonstrate that identifying convergence areas is unnecessary for MDPS. This statement parallels other attempts to remove under- and overfitted models from the model selection pool[18,19] as well as defining the Rashomon curve in recent work on model selection.[26]

Model Selection for Multiple Tuning Parameters. Model updating methods need to be compared to baseline methods PPS, SPS, and SSPS. These comparisons are used to validate the efficiency and necessity of applying an updating method. For MDPS with LMC and FA-2, the model diversity thresholds of 0.3−0.5 in conjunction with the ω weight 0.4 were used. Shown in Figure 3 for four spectral data set situations are boxplots of corresponding MDPS selected models (including sample-wise) and SPS model trends. Plotted in the SI in Figure S5 are larger versions of these boxplots with the PPS and SSPS trends included.

The boxplots reveal that the MDPS selected models predict at or less than respective first quartiles and perform comparably to the first quartile SPS models. The SPS models have substantially more calibration samples, while the updating methods use five to six secondary samples (Table 1). The selected models generally outperform both SSPS and PPS. Thus, the secondary set can be considered uniquely different from the primary conditions, and the secondary set has enough variance that a small calibration set cannot be used to form an effective model. Results also show that LMC and FA-2 are equally effective for these data sets. The boxplots demonstrate that MDPS can be used to select models for one or more samples.

In addition to forming MDPS boxplots using all s samples in a data split to calculate $SPD_{i,j}$ values in eq 1, a sample-wise study was performed where models were selected for each secondary sample. Sample-wise LMC and FA-2 boxplots shown in Figure 3 demonstrate that MDPS can be used to select models for one sample or a collection of new samples with no real difference in prediction quality.

Models selected by MDPS are found most frequently in ideal tuning parameter combination regions. Part of the reason for this fact is that MDPS selects models specifically to predict only the new samples, and thus, MDPS allows a slight overfit advantage. Histograms and RMSEV heatmaps in Figure 4 characterize the situation for LMC using the Goat data set over all 100 data splits. As a reminder, Figures 1 and 2 represent one of the Goat data splits where 2,300 models are formed producing 2,643,850 model combinations. There are 31,614 model pairs selected by MDPS composed of 1,218 unique models. The dark blue regions in the top heatmap of Figure 4 correspond to the strongest models with the lowest RMSEV values. The bottom heatmap shows that these models are most frequently selected by MDPS. Complementing these heat maps is Figure 5 with the selected regression vectors color coded to selection frequency for one of the 100 data splits.

Additional boxplots are shown in Figure S6 of the SI for four other data sets. Regression vectors and histograms of selected models are displayed in Figure S7 for another data set further demonstrating that the most frequently selected models are those with lower RMSEV values and are characterized by a range in shape and magnitude.

It has been suggested that to characterize the calibration transfer quality for a method, the relationship of the new samples to the updated model space should be assessed with a

https://doi.org/10.1021/acs.jcim.0c01493
J. Chem. Inf. Model. 2021, 61, 2220−2230

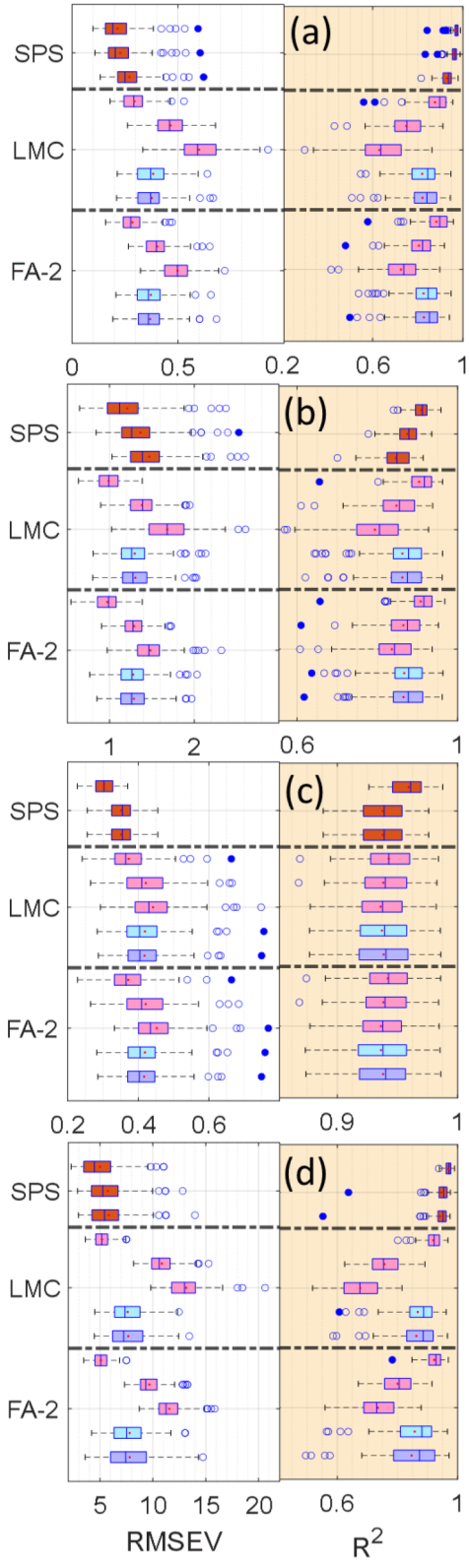


Figure 3. Model updating boxplots of RMSEV and R^2 values for the SPS baseline and MDPS selected models across four NIR data sets. Groups of three boxes with one color correspond to minima, first quartiles, and medians of each category. Blue boxes represent MDPS selected models, and purple boxes are the sample-wise selected models. Data sets are (a) Corn mp5-m5 starch, (b) Soy R2-R1 moisture, (c) Tablet 1&4-1&3, and (d) Goat.

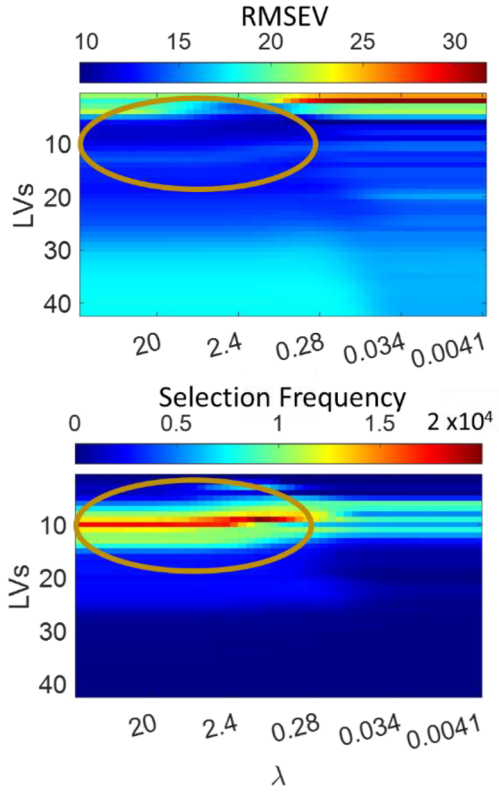


Figure 4. Top heatmap shows mean RMSEV values across the 100 data splits of Goat using LMC. The bottom heatmap shows the MDPS model selection frequency. Circled areas show that models with lower RMSEV values are matched to models most frequently selected.

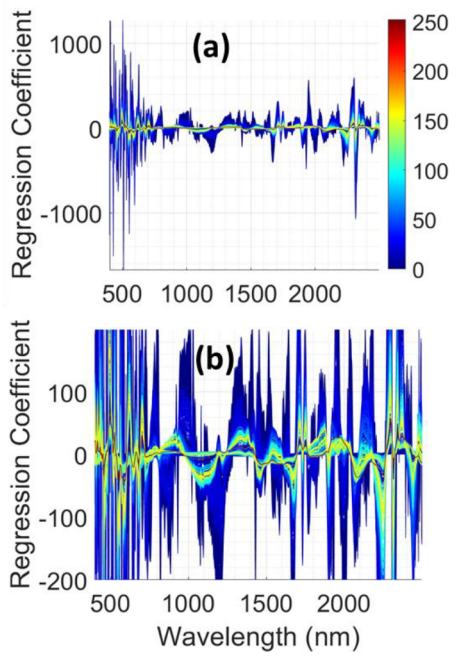


Figure 5. Goat (a) LMC model vectors relative to MDPS selection frequency (color bar) for one of the 100 data splits and (b) a zoomed-in image of part a.

prediction outlier diagnostic.[44] However, such studies were not performed.

https://doi.org/10.1021/acs.jcim.0c01493
J. Chem. Inf. Model. 2021, 61, 2220−2230

Model Selection for Single Tuning Parameter. Various MDPS values for the degree of diversity and U-curve weight in eq 1 were evaluated, and three are presented demonstrating respective effects. These values are C1 ($0.3 \leq \cos(\theta) \leq 0.5$, $\omega = 0.4$), C2 ($0.3 \leq \cos(\theta) \leq 0.5$, $\omega = 1.0$), and C3 ($0.6 \leq \cos(\theta) \leq 0.8$, $\omega = 1.0$). Figures 6 and 7 (and Figures S8–S13

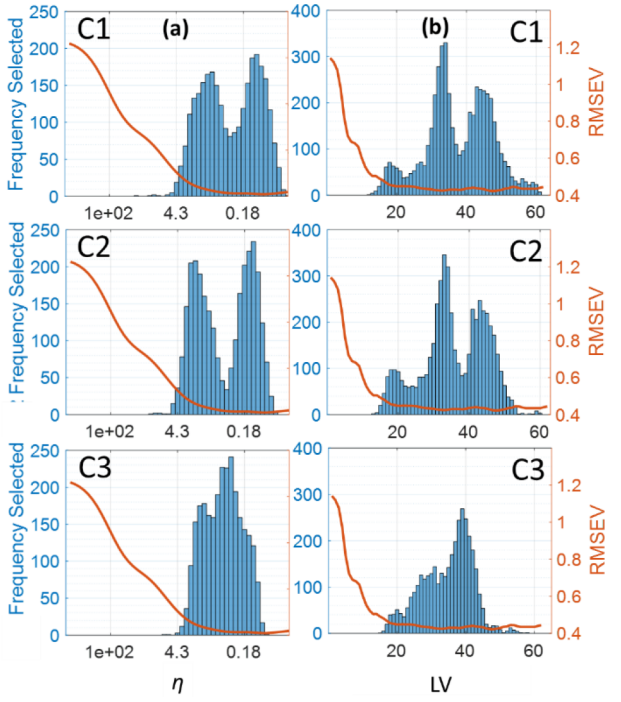


Figure 6. QSAR CAIV mean RMSEV (orange) and MDPS (a) RR and (b) model selection frequencies (blue) using QSAR CAIV data varying the $\cos(\theta)$ and ω thresholds associated with C1 ($0.3 \leq \cos(\theta) \leq 0.5$, $\omega = 0.4$), C2 ($0.3 \leq \cos(\theta) \leq 0.5$, $\omega = 1.0$), and C3 ($0.6 \leq \cos(\theta) \leq 0.8$, $\omega = 1.0$).

in the SI for other data sets) show that C3 is overall the best compromise. Regardless of the data set, when C3 is used, Figure 7 and the corresponding boxplots in the SI (Figures S9, S11, and S13) show that MDPS selects models with either equivalent or lower RMSEV values and corresponding R^2 values compared to the three standard U-curves shown in respective boxplots. The following discussion clarifies these observations.

Plotted in Figure 6 (and Figures S8, S10, and S12) are histograms of the models selected over the 100 data splits and the mean RMSEV curves. The RMSEV curve is the prediction error for the new samples, and the reader is reminded that these samples are not used to form calibration models. These samples are also not used to select models based on a cross-validation but are only used by MDPS to select models to predict these specific samples. In Figure 6, the MDPS selected models using C3 show little improvement in RMSEV and R^2 values compared to C1 and C2. However, this observation is relative to the actual underlying shape of the RMSEV curve. In Figure 6a, the RMSEV plots have a slight natural U-curve behavior. Shown in Figures S8 and S10 are two other data set situations with similar results to the QSAR CAIV data set. The U shaped RMSEV is more pronounced in Figure S10. Thus, the boxplots show fairly equivalent results for models selected by MPDS (using C1, C2, or C3) and U-curves due to correlated U shapes.

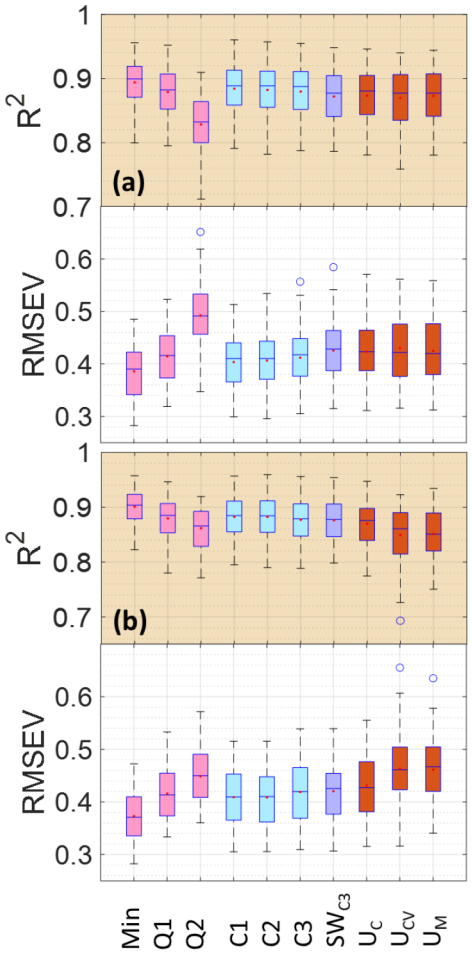


Figure 7. Corn mp5 oil RMSEV and R^2 boxplots for (a) RR and (b) PLS models. Pink boxes are the minima and first two quartiles of all models evaluated (min, Q1 and Q2). Blue boxes correspond to MDPS selected models in Figure 3 histograms using C1, C2, and C3. The purple box is the MDPS model selection sample-wise with C3 (SW_{C3}). Orange boxes are the respective U-curve selected model using outer calibration sets (U_C), mean inner CV (U_{CV}), and the mean inner calibration and CV (U_M).

Contradicting this trend is the data set in Figures S12 and S13 where due to the L shaped RMSEV curve, model diversity C3 is needed to obtain acceptable models and predictions. Specifically, the RMSEV plots resemble the standard L shaped RMSEC plot for the calibration even though the samples used to compute the RMSEV values are not used to form calibration models. To compensate for the lack of U shape, less model diversity is needed, or else underfitted models are included as shown in Figure S12. Thus, C3 is recommended when selecting single tuning parameter values because this diversity requirement is effective in whether the RMSEV is U or L shaped.

By increasing the ω weight in C1 to that used in C2, some of the C1 selected underfitted models are not selected. In both situations, selection is bimodal for the RR histograms and trimodal for PLS. Decreasing the model diversity threshold fused with C2 to that used in C3 creates a unimodal model selection distribution better correlated to those models with lower RMSEV values.

Improvement over the U-curves by MDPS is possible because MDPS selects models to directly predict new samples.

https://doi.org/10.1021/acs.jcim.0c01493
J. Chem. Inf. Model. 2021, 61, 2220−2230

As noted with the two tuning parameter situation, part of the improvement by MDPS is because MDPS holds a small overfitting advantage relative to the new samples being predicted, i.e., MDPS selects models targeting to predict only the new samples. The histograms characterize this slight overfit feature of MDPS. Selecting a model from a calibration set based on some form of cross-validation may not perform as well depending on the degree new samples are spanned by the calibration set. The usual cross-validation approach requires multiple data splits of the training set to select one model across the span of tuning parameter values. This one model is then used to predict the new samples. Conversely, an ensemble of models is used for prediction with MDPS providing a data fusion benefit.

A sample-wise study similar to that used for LMC and FA-2 was performed for RR and PLS. Sample-wise RR and PLS boxplots shown in Figures 7, S9, S11, and S13 demonstrate that MDPS can also be used to select RR or PLS models for one sample or a collection of new samples without substantive differences in prediction quality.

It has been previously reported that numerous optimized model regression vectors can be formed with different magnitudes and shapes to accurately predict a data set.[22−25] It was also suggested that the better predicting models with different tuning parameter values, e.g., different number of PLS LVs or RR ridge parameter values, will cluster together in a score plot of regression vector models over a full range of tuning parameter values.[45,46] The under- and overfitted models should tend to form respective multiple clusters.

This analysis process was termed regression model comparison plot (REMOCOP). Shown in Figure 8 is a principal component analysis (PCA) score plot of the combined set of full tuning parameter ranged RR and PLS regression vectors with C3 for all 100 QSAR CAIV data set splits. Identified are the selected, suitable but not selected, and under- and overfitted models showing respective clusters. To identify approximate under- and overfitted models, an empirical threshold was used for a U-curve (model 2-norm and RMSEV). Models with U-curve values above 0.35 were labeled over- and underfitted, and models below were deemed suitable if not selected by MDPS. With this threshold, the score plot reveals one cluster for the selected models and multiple clusters of under- and overfitted models. Similar score plots were obtained using C1 and C2 with clusters formed relative to histogram distributions in Figure 6. The score plot further demonstrates the Rashomon effect that numerous regression models with different shapes and magnitudes can be formed to effectively predict samples. A PCA plot for another data set is shown in Figure S14 revealing similar trends.

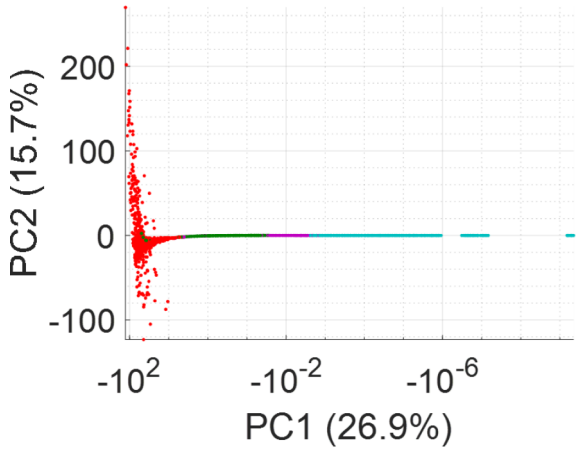


Figure 8. Score plot of RR and PLS models across tuning parameter ranges for all 100 QSAR CAIV data set splits in Figure 6. Point colors are (green) for RR and PLS MDPS selected models, (red) overfitted, (cyan) underfitted, and (magenta) suitable nonselected models. The PC1 axis is on a logarithmic scale.

In the REMOCOP work, it was suggested that score values can be modified to weight the calibration basis vectors (loading vectors) to form other models in the regions of the better models. As previously noted, the final MDPS predicting model $\hat{\mathbf{b}}_f$ is also a linear combination of models, and in essence, it can also be considered a linear combination of basis vectors. More definitely, it has been shown[47,48] that model vectors solving $\mathbf{y} = \mathbf{Xb}$ can be expressed as $\hat{\mathbf{b}} = \mathbf{v}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a $k \times 1$ weight vector of the $\mathbf{V}$ basis vectors from the singular value decomposition of $\mathbf{X} = \mathbf{USV}^T$, and k denotes the rank of $\mathbf{X}$. Each regression vector is distinguished by a specific linear combination of the k basis column vectors in $\mathbf{V}$. Thus, MDPS can be considered a model selection approach forming a final predicting model that is one of the many linear combinations of the basis vectors that can effectively predict samples.

Summarizing, the results presented in Figures 6−8 and in the SI for other data sets document that MDPS can select accurately predicting models for multivariate calibration based on one tuning parameter as well as two tuning parameters. Models can be selected for each new sample to be predicted as well as a set of new samples, assuming the new samples are not outliers to the calibration set. Additionally, results further verify that many models with different shapes and magnitudes can accurately predict new samples. This observation agrees with the Rashomon effect previously noted.

MDPS Selection Time and Memory. A brief comparison of the time needed to select RR and PLS models by MDPS and the three standard U-curves was made. When cross-validation is used, the time-consuming part is the training and model selection. The exact time needed depends on the cross-validation design and number of samples. Once a model is selected, prediction of a new sample is immediate (only the inner product of two vectors is needed). Because cross-validation is not used with MDPS, forming model vectors is rapid, and the time-consuming part is model selection because all possible pairs of models are evaluated. Thus, the exact time depends on the number of models formed and ultimately compared relative to the number of samples to be predicted. Once models are selected, prediction is also essentially instantaneous because sample predictions were already formed in order to select models. Values reported in Table 2 are the total time needed to train and predict one set of new samples using PLS and RR. Another consideration when evaluating the times is that the time depends on the algorithm language, configuration, and computer architecture. As expected, the MDPS time is substantially less than a cross-validation approach to model selection. Model selection for LMC and FA-2 has similar MDPS times.

Memory requirements for MDPS are related to the storage of the collection of regression vectors formed by the modeling method. Additional memory would be needed to store the two data points (model diversity and prediction similarity) for each possible model pair, generally not a large requirement. As with time, the actual memory size will depend on the number of

https://doi.org/10.1021/acs.jcim.0c01493
J. Chem. Inf. Model. 2021, 61, 2220−2230

Table 2. Representative Times to Select Existing RR and PLS Models to Predict a Particular Set of Samples

data set	method	MDPS time (s)	U_C time (s)	U_{CV} and U_M time (s)
Corn mp5 oil	PLS	0.0039	0.0003	3.4123
Corn mp5 moisture	PLS	0.0038	0.0002	3.7582
Soy R1 moisture	PLS	0.0024	0.0001	1.3372
Corn m5 moisture	PLS	0.0036	0.0001	3.4409
QSAR CAI	PLS	0.0023	0.0002	2.1478
QSAR CAIV	PLS	0.0025	0.0001	2.0018
Corn mp5 oil	RR	0.0043	0.0011	96.5706
Corn mp5 moisture	RR	0.0034	0.0004	95.5161
Soy R1 moisture	RR	0.0029	0.0009	21.5554
Corn m5 moisture	RR	0.0033	0.0003	97.8993
QSAR CAI	RR	0.0053	0.0013	1.7735
QSAR CAIV	RR	0.0025	0.0001	1.7624

models and regression coefficients, algorithm language, configuration, and computer architecture. For model updating, models are generated dynamically for the new samples, and for the basic primary calibration by RR or PLS, models can be saved to a hard drive for future use.

CONCLUSIONS

Harnessing model diversity and prediction similarity measures for MDPS is an effective method of consensus model selection achieving prediction errors at, below, or marginally different than baseline expectations for the data sets studied. Because models are selected to predict new samples without using a cross-validation process, the MDPS also maintains a small overfitting advantage. In the one tuning parameter case, MDPS selected models outperformed standard traditional U-curves because models are selected to directly predict the validation samples and a collection of models is used to predict new samples instead of just one model.

As just noted, prediction similarities are used to identify models, and hence, reference property values (labels) are not needed. Thus, models are selected tuned to predict new samples not part of a calibration set. Current work in our laboratory involves using MDPS to select updated models where secondary property reference values are not used to form models as with LMC and FA-2.

It has been shown that many dissimilar models can effectively predict a sample,[22−25] and the MDPS approach leverages this fact to select a collection of dissimilar models accurately predicting samples. However, in essence, the MDPS selected models can be considered forming one final predicting model that is one of the many models obtainable from linear combinations of the basis vectors.[45−48]

Because single tuning parameter calibration methods have less diversity between the better models than multiple tuning parameter based methods, it is best to increase the cosine and U-curve thresholds in MPDS relative to model updating methods. The increase allows MDPS to select accurately predicting models with an appropriate bias-variance trade-off.

The MDPS method was applied to linear regression methods based on tuning parameters using spectra and QSAR data sets. The approach is general and should be useful to other tuning parameter based linear regression approaches involving other data sets. The usefulness of MDPS to nonlinear based methods such as support vector machines is yet to be determined. Current work in our laboratory is evaluating MDPS for model selection relative to three tuning parameters.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c01493.

Discussion on importance of removing converged models when forming quartile-based boxplots for fair comparisons, process for automatic identification of active bias-variance trade-off regions for boxplots, and additional graphics evaluating MDPS for PLS, RR, LMC, and FA-2 (PDF)

AUTHOR INFORMATION

Corresponding Author

John H. Kalivas − *Department of Chemistry, Idaho State University, Pocatello, Idaho 83209, United States;* orcid.org/0000-0001-7056-976X; Email: kalijohn@isu.edu

Author

Robert C. Spiers − *Department of Chemistry, Idaho State University, Pocatello, Idaho 83209, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c01493

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Nos. CHE-1506417 (cofunded by CDS&E) and CHE-1904166 (cofunded by CDS&E and the Office of Investigative and Forensic Sciences in the National Institute of Justice) and is gratefully acknowledged by the authors.

REFERENCES

(1) Næs, T.; Isaksson, T.; Fearn, T.; Davies, T. *A User Friendly Guide to Multivariate Calibration and Classification*; NIR Publications: Chichester, UK, 2002; DOI: 10.1255/978-1-906715-25-0.

(2) Kalivas, J. H. Interrelationships of Multivariate Regression Methods Using Eigenvector Basis Sets. *J. Chemom.* **1999**, *13*, 111−132.

(3) Shao, J. Linear Model Selection by Cross-Validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486−494.

(4) Golub, G. H.; Heath, H.; Wahba, G. Generalized cross Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics* **1979**, *21*, 215−223.

(5) Faber, N.; Rajkó, R. How to Avoid Over-fitting in Multivariate Calibration—The Conventional Validation Approach and An Alternative. *Anal. Chim. Acta* **2007**, *595*, 98−106.

(6) Hansen, P. C. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*; SIAM: Philadelphia, PA, 1998; DOI: 10.1137/1.9780898719697.

(7) Forrester, J. B.; Kalivas, J. H. Ridge Regression Optimization Using a Harmonious Approach. *J. Chemom.* **2004**, *18*, 372−384.

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

(9) 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.
(10) 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.
(11) 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.
(12) 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.
(13) Héberger, K. Sum of Ranking Differences Compares Methods or Models Fairly. *TrAC, Trends Anal. Chem.* **2010**, *29*, 101−109.
(14) Kalivas, J. H.; Héberger, K.; Andries, E. Sum of Ranking Differences (SRD) to Ensemble Multivariate Calibration Model Merits for Tuning Parameter Selection and Comparing Calibration Methods. *Anal. Chim. Acta* **2015**, *869*, 21−33.
(15) Höskuldsson, A. The H-Principle in Modeling with Applications to Chemometrics. *Chemom. Intell. Lab. Syst.* **1992**, *14*, 139−153.
(16) Myers, R. H. *Classical and Modern Regression with Applications*, 2nd ed.; Duxbury: Pacific Grove, 1990.
(17) Brown, S. D. Transfer of Multivariate Calibration Models. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, 2nd ed.; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2020; Vol. 3, pp 359−391, DOI: 10.1016/B978-0-12-409547-2.00644-2.
(18) Shahbazikhah, P.; Kalivas, J. H. A Consensus Modeling Approach to Update a Spectroscopic Calibration. *Chemom. Intell. Lab. Syst.* **2013**, *120*, 142−153.
(19) Tencate, A.; Kalivas, J. H.; White, A. Fusion Strategies for Selecting Multiple Tuning Parameters for Multivariate Calibration and Other Penalty Based Processes: A Model Updating Application for Pharmaceutical Analysis. *Anal. Chim. Acta* **2016**, *921*, 28−37.
(20) Gurung, A.; Kalivas, J. H. Model Selection Challenges with Application to Multivariate Calibration Updating Methods. *J. Chemom.* **2020**, *34*, e3245.
(21) Andries, E.; Kalivas, J. H.; Gurung, A. Sample and Feature Augmentation Strategies for Calibration Updating. *J. Chemom.* **2019**, *33*, e3080.
(22) Brown, C. D. Discordance Between Net Analyte Signal Theory and Practical Multivariate Calibration. *Anal. Chem.* **2004**, *76*, 4364−4373.
(23) Brown, C. D.; Green, R. L. Critical Factors Limiting the Interpretation of Regression Vectors in Multivariate Calibration. *TrAC, Trends Anal. Chem.* **2009**, *28*, 506−514.
(24) Kunz, M. R.; Ottaway, J.; Kalivas, J. H.; Andries, E. Impact of Standardization Sample Design on Tikhonov Regularization Variants for Spectroscopic Calibration Maintenance and Transfer. *J. Chemom.* **2010**, *24*, 218−229.
(25) Kalivas, J. H.; Ferré, J.; Tencate, A. J. Selectivity-relaxed Classical and Inverse Least Squares Calibration and Selectivity Measures with a Unified Selectivity Coefficient. *J. Chemom.* **2017**, *31*, No. e2925.
(26) Semenova, L.; Rudin, C.; Parr, R. A Study in Rashomon Curves and Volumes: A New Perspective on Generalization and Model Simplicity in Machine Learning. *arXiv:1908.01755v2 [cs.LG]*. https://arxiv.org/abs/1908.01755 (accessed 2021).
(27) Breiman, L. Statistical Modeling: The Two Cultures. *Stat. Sci.* **2001**, *16*, 199−231.
(28) Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. Decision Forests: Combining the Predictions of Multiple Independent Decision Tree Models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 525−531.
(29) Van Rhee, A. M. Use of Recursion Forests in the Sequential Screening Process: Consensus Selection by Multiple Recursion Trees. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 941−948.
(30) Xu, L.; Jiang, J. H.; Zhou, Y. P.; Wu, H. L.; Shen, G. L.; Yu, R. Q. Stacked Regression for Model Combination and Fast Spectral Interval Selection in Multivariate Calibration. *Chemom. Intell. Lab. Syst.* **2007**, *87*, 226−230.
(31) Ni, W.; Brown, S. D.; Man, R. Stacked Partial Least Squares Regression Analysis for Spectral Calibration and Prediction. *J. Chemom.* **2009**, *23*, 505−517.
(32) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure-Activity Landscapes. *Drug Discovery Today* **2009**, *14*, 698−705.
(33) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.
(34) Miyao, T.; Funatsu, K.; Bajorath, J. Three-Dimensional Activity Landscape Models of Different Design and Their Application to Compound Mapping and Potency Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 993−1004.
(35) Yongye, A. B.; Byler, K.; Santos, R.; Martínez-Mayorga, K.; Maggiora, G. M.; Medina-Franco, J. L. Consensus Models of Activity Landscapes with Multiple Chemical Conformer, and Property Representations. *J. Chem. Inf. Model.* **2011**, *51*, 1259−1270.
(36) Kalivas, J. H.; Siano, G.; Andries, E.; Goicoechea, H. Calibration Maintenance and Transfer Using Tikhonov Regularization Approaches. *Appl. Spectrosc.* **2009**, *63*, 800−809.
(37) MDPS software link. https://www.isu.edu/chem/faculty/staffdirectoryentries/kalivas-john.html (accessed 2021-04-22).
(38) Wise, B. M.; Gallagher, N. B. *Eigenvector Research*; Manson, WA. http://www.eigenvector.com/data/index.htm (accessed 2021-04-22).
(39) Bouveresse, E.; Hartmann, C.; Massart, D. L.; Last, I. R.; Prebble, P. A. Standardization of Near-Infrared Spectrometric Instruments. *Anal. Chem.* **1996**, *68*, 982−990.
(40) Dyrby, M.; Engelsen, S.; Nørgaard, L.; Bruhn, M. Chemometric Quantitation of the Active Substance (Containing C≡N) in a Pharmaceutical Tablet Using Near-Infrared (NIR) Transmittance and NIR FT-Raman Spectra. *Appl. Spectrosc.* **2002**, *56*, 579−585.
(41) Walker, J. W.; Campbell, E. S.; Lupton, C. J.; Taylor, C. A.; Waldron, D. F.; Landau, S. Y. Effects of Breed, Sex, and Age on the Variation and Ability of Fecal Near-Infrared Reflectance Spectra to Predict the Composition of Goat Diets. *J. Anim. Sci.* **2007**, *85*, 518−526.
(42) Mattioni, B. E.; Jurs, P. C. Development of Quantitative Structure-Activity Relationship and Classification Models for a Set of Carbonic Anhydrase Inhibitors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 94−102.
(43) Xu, Q. S.; Liang, Y. Z. Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1−11.
(44) Guenard, R. D.; Wehlburg, C. M.; Pell, R. J.; Haaland, D. M. Importance of Prediction Outlier Diagnostics in Determining a Successful Inter-Vendor Multivariate Calibration Model Transfer. *Appl. Spectrosc.* **2007**, *61*, 747−754.
(45) Geladi, P.; Swerts, J.; Lindgren, F. Multiwavelength Microscopic Image Analysis of a Piece of Painted Chinaware: Classification and Regression. *Chemom. Intell. Lab. Syst.* **1994**, *24*, 145−167.
(46) Geladi, P. The Regression Model Comparison Plot (REMOCOP). In *Frontiers in Analytical Spectroscopy*; Andrews, D. L., Davies, A. M. C., Eds.; Proceedings of "Spectroscopy Across the Spectrum IV", Norwich, July 11−14 1994, Royal Society of Chemistry: 1995; pp 225−236.
(47) Kalivas, J. H. Interrelationships of Multivariate Regression Methods Using Eigenvector Basis Vectors. *J. Chemom.* **1999**, *13*, 111−132.
(48) Kalivas, J. H. Basis Sets for Multivariate Regression. *Anal. Chim. Acta* **2001**, *428*, 31−40.

https://doi.org/10.1021/acs.jcim.0c01493
J. Chem. Inf. Model. 2021, 61, 2220−2230