Journal of Pharmaceutical Innovation
 | www.ispe.org/jpi

Research Article

An efficient method-development strategy for quantitative chemical imaging using terahertz pulse spectroscopy

Robert P. Cogdill[1], Steven M. Short[2], Ryanne Forcht[2], Zhenqi Shi[2], Yaochun Shen[3], Philip F. Taday[3], Carl A. Anderson[2], and James K. Drennen III[2]

[1]Duquesne University Center for Pharmaceutical Technology, 410A Mellon Hall, 600 Forbes Avenue, Pittsburgh, PA 15282, USA
[2]Duquesne University Graduate School of Pharmaceutical Sciences, 410A Mellon Hall, 600 Forbes Avenue, Pittsburgh, PA 15282, USA
[3]TeraView, Platinum Building, St John's Innovation Park, Cowley Road, Cambridge CB4 OWS, UK
Corresponding author: Drennen, J.K. (drennen@duq.edu).

The purpose of our research was to investigate efficient procedures for generating multivariate prediction vectors for quantitative chemical analysis of solid dosage forms using terahertz pulse imaging (TPI) reflection spectroscopy. A set of calibration development and validation tablet samples was created following a ternary mixture of anhydrous theophylline, lactose monohydrate, and microcrystalline cellulose (MCC). Spectral images of one side of each tablet were acquired over the range of 8 cm^{-1} to 60 cm^{-1}. Calibration models were generated by partial least-squares (PLS) type II regression of the TPI spectra and by generating a pure-component projection (PCP) basis set using net analyte signal (NAS) processing. Following generation of the calibration vectors, the performance of both methods at predicting the concentration of theophylline, lactose, and MCC was compared using the validation spectra and by generating chemical images from samples with known composition patterns. Sensitivity was observed for the PLS calibration over the range of all constituents for both the calibration and the validation datasets; however, some of the calibration statistics indicate that PLS overfits the spectra. Multicomponent prediction images verified the spatial and composition fidelity of the system. The NAS–PCP calibration procedure yielded accurate linear predictions of theophylline and lactose, whereas the results for MCC prediction were poor. The poor sensitivity for MCC is assumed to be related to the relative lack of phonon absorption bands, which concurs with the characterization of MCC as being semi-crystalline. The results of this study demonstrate the use of TPI reflection spectroscopy and efficient NAS–PCP for the quantitative analysis of crystalline pharmaceutical materials.

Continued on page 64.

Journal of Pharmaceutical Innovation

Introduction

A confluence of analytical technologies and techniques has opened new windows for pharmaceutical scientists and engineers to observe rapidly and non-destructively the physical and chemical characteristics of pharmaceutical drug products and solid dosage units. Spectroscopic technologies, such as near infrared (NIR) and Raman, which were developed largely outside of the pharmaceutical industry, are becoming relatively common solutions for drug development and manufacturing process control (e.g., process analytical technology [1]). More recently, however, pharmaceutical applications have had a key role in the development of novel instrumentation for terahertz (THz) spectroscopy and imaging [2].

The THz band gap is situated in the far-infrared (IR) region of the electromagnetic spectrum, just before microwave radiation, and covers the range from 10 cm^{-1} to 330 cm^{-1}, or 300 GHz to 10 THz. Owing to the low-frequency nature of the radiation, the solid-state optical resonance phenomena detected in the THz range are largely the result of large-mass intramolecular oscillations, intermolecular bond vibrations (e.g., hydrogen bonds), or phonon lattice dynamics in crystalline materials [3-5]. Because many of the polymers used as binders, fillers, coating agents, and packaging materials during solid dosage manufacturing do not exhibit such dynamics, these materials have extremely low THz absorptivity, and hence, are nearly transparent in the THz region. However, active pharmaceutical ingredients, which are typically used in crystalline form, are expected to have intense THz spectral-absorption signatures. Intermolecular and phonon lattice vibrations are strongly related to the composition and geometry of molecular crystals and are highly localized. Hence, spectral features in the THz range can often be verified using *ab initio* calculations, which is one of the reasons why THz spectroscopy is a powerful tool for the verification of crystal form [3-5] and is complementary to X-ray diffraction and Raman spectroscopy. Furthermore, because many packaging materials and excipients are semitransparent, THz radiation can probe relatively deep within drug tablets and directly through some containers [2,6].

Early studies of the far-IR region were performed using traditional dispersive or Fourier transform optics with non-coherent illumination sources. Illumination sources such as glowbars have intensity in the THz region that is only slightly greater than the level of background (e.g., blackbody) radiation. Moreover, a large fraction of light is lost to the optical inefficiencies of traditional spectrometers. Hence, far-IR analyses have, in most cases, been too slow and insensitive for use beyond theoretical investigations. A new form of instrumentation, THz pulse spectrometry (TPS), has absolved these limitations.

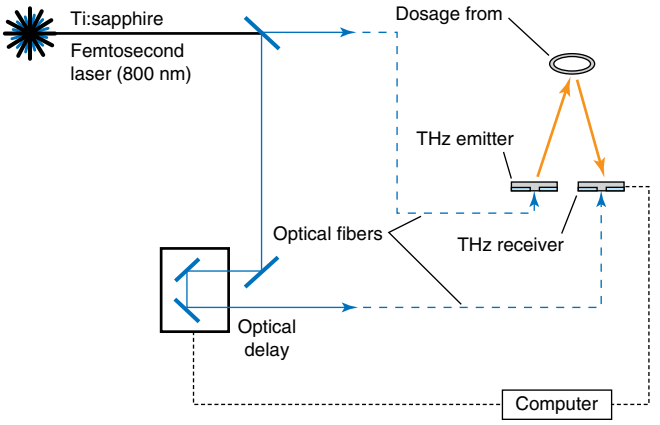

Figure 1. Schematic view of the TeraView coating scan optical path.

Terahertz pulse spectroscopy

Terahertz pulse spectrometers use coherent generation and detection of femtosecond THz pulses, based on the principles of the Auston switch [7], to probe the far-IR region without the need for dispersive or traditional Fourier interferometric optics for wavelength or frequency selection (Figure 1). The TPS measurement process begins with the firing of a pulsed Ti:sapphire laser, producing an ~80-fs pulse of 800-nm radiation (~80-MHz repetition rate). A portion of the NIR pulse (pump beam) is directed onto a voltage-biased gallium arsenide (GaAs) substrate, which serves as the Auston switch emitter, while the remainder of the pulse is directed toward a variable time delay (probe beam). The incident NIR radiation causes an acceleration of charge across the gap, producing an ~500-fs pulse of broadband THz radiation that propagates from the GaAs substrate by the electro-optic effect (Figure 2). Focusing optics direct the pulse onto the sample, where radiation is transmitted, reflected (or scattered), or absorbed. Only reflected and absorbed ra-

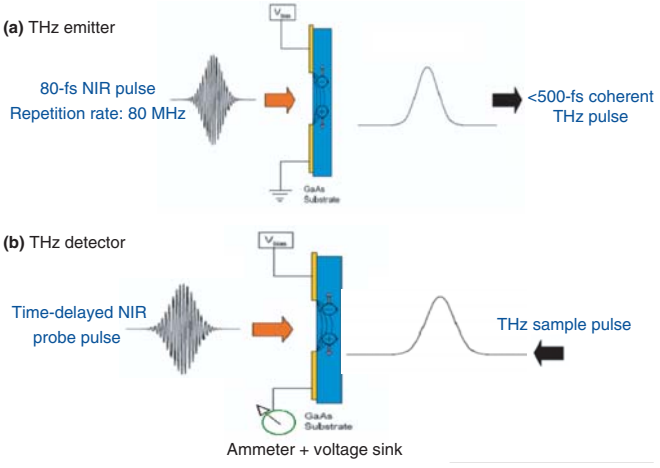

Figure 2. Schematic of an Auston-switch-gated THz emitter and detector. Emitter is shown in **(a)** and detector is shown in **(b)**.

diation will be considered in this study. Subsequently, collection optics direct reflected radiation to a second Auston switch, where it is coincident with the time-delayed probe beam. When the reflected THz radiation and NIR probe beams simultaneously strike the detector switch, a transient current that is proportional to the THz reflection is generated and measured by readout electronics.

Time-domain THz spectra are generated by measuring the intensity of THz reflections over a range of probe beam delays (Figure 3). The features in TPS time-domain spectra correspond to relatively intense reflection events that occur when the incident THz plane wave meets an interface between two materials with different refractive indices. Typically, a positive peak indicates a transition from lower to higher refractive index [8]. If the sample refractive index is known, the time delay of the probe beam can be converted to distance. The time-domain TPS signal can be considered analogous to an ultrasound B-scan, whereby the depth and thickness of interfaces can be estimated with finite accuracy. By raster-scanning across the surface of a sample, a 2D map or image of TPS signals can be generated. Hence, time-domain TPS and THz pulse imaging (TPI) have been used to determine accurately and non-destructively the thickness and uniformity of many solid objects such as tablet coatings and active layers within laminated tablets (e.g., multiple layers of distinct compositions) [9].

Frequency-domain THz spectra are generated from the time-domain signals by fast Fourier transformation (FFT), followed by normalization using a reference spectrum. The reference spectrum for reflection TPS is collected by scanning the surface of a metallic mirror. A ratio of the time-domain FFT power spectra for the sample (E_s) and 100%-reflectance mirror reference (E_r) channels yields a complex conjugate pair for each discrete optical frequency (υ), from which the absorption coefficient (α) and refractive index (n) intensities can be calculated [6] (Equation 1).

$$\sqrt{\varepsilon} \equiv n(\upsilon) + j\,\frac{\alpha(\upsilon)}{4\pi\upsilon} = \frac{1 - E_s(\upsilon)/E_r(\upsilon)}{1 + E_s(\upsilon)/E_r(\upsilon)} \quad \text{(Eqn 1)}$$

Previous research has demonstrated frequency-domain transmission TPS for both qualitative and quantitative analyses [4,5,9]. Intuitively, one can envision how a combination of the 3D spatial measurement capabilities of TPI and the quantitative power of TPS would produce a powerful comprehensive platform for the non-destructive analysis of solid dosage form. At this time, however, the quantitative capabilities of TPS and TPI in the reflection mode remain to be explored. Indeed, to the best of our knowledge, this is the first work to demonstrate the quantitative capabilities of reflection THz-pulsed spectroscopy and imaging.

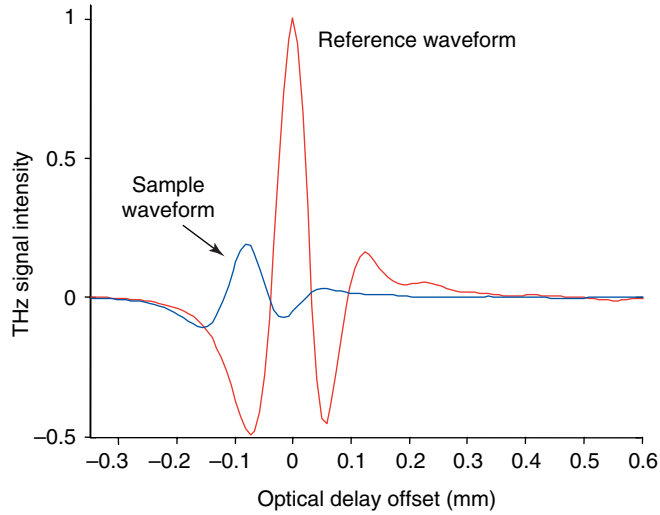

Figure 3. Example THz time-domain intensity scan of a solid compact of theophylline, lactose and MCC, and a 100%-reflectance mirror reference signal. Horizontal axis describes the distance traveled by the optical delay line and is proportional to photon time of flight.

Objectives

The objectives of this work were to:

(i) investigate the quantitative capability of reflection TPS using typical pharmaceutical materials
(ii) demonstrate the combination of TPI and chemometrics for non-destructive 2D chemical mapping
(iii) compare the relative merits of quantitative reflection-TPS calibration for absorption and refractive index spectra
(iv) determine the feasibility of using pure-component projection (PCP) efficient calibration techniques for rapid quantitative TPS-method development.

Experimental

Three-component tablet production

A ternary-mixture design was chosen to provide a moderate chemometric challenge. Anhydrous theophylline powder (lot no. 92577, Knoll, Ludwigshafen, Germany) was chosen as the model compound because its crystal form has absorption features in the THz region. Lactose 316 Fast Flo® NF monohydrate (lot no. 8502113061, Hansen Laboratories, New Berlin, WI, USA) and microcrystalline cellulose (MCC, Avicel® PH 200, lot no. M427C, FMC BioPolymer, Mechanicsburgh, PA, USA) were chosen as compression excipients. The approximate mean particle sizes of the theophylline, lactose, and MCC (reported by documentation from their respective suppliers) were ~90, ~100, and ~180 µm, respectively. All ingredients were used 'as is', with no additional characterization or modification of particle size. In all, 24 design points were selected for tablet production (Figure 4); 16 mixtures were used for THz calibration and eight mixtures were reserved for validation testing. A single tab-

Continued on page 66.

Journal of Pharmaceutical Innovation

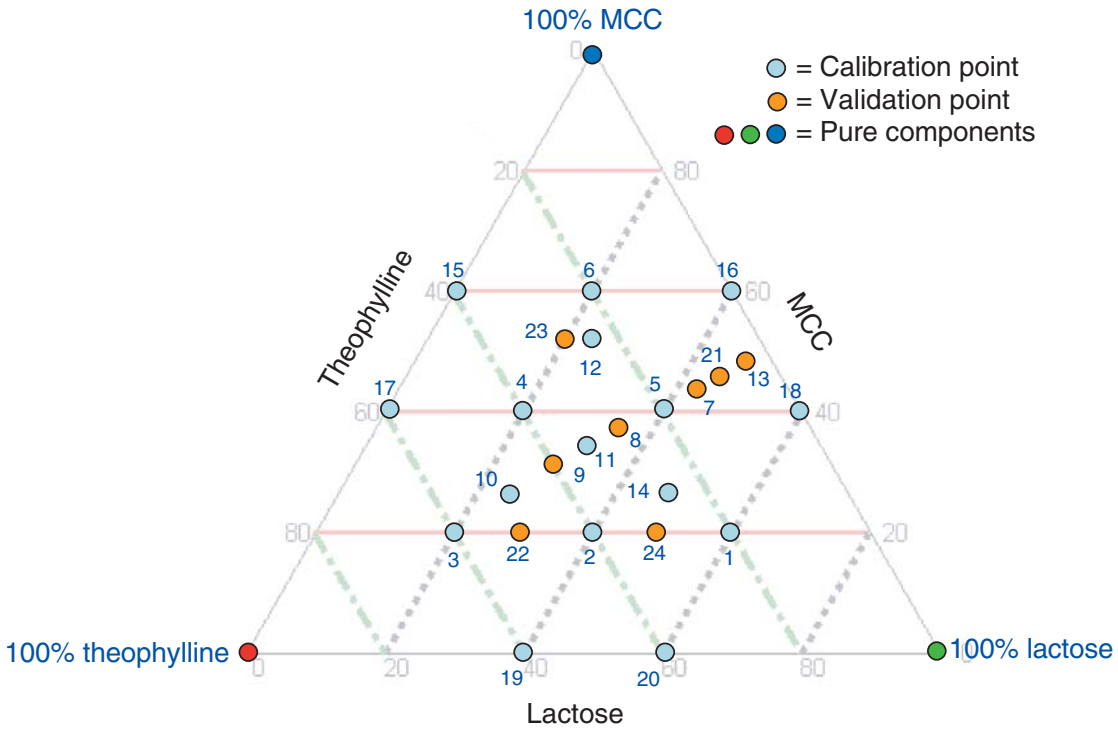

Figure 4. Ternary diagram of the balanced mixture design used for calibration. The validation samples are not balanced and were chosen with emphasis on the range of theophylline tested. Twenty-four design points were selected for tablet production; 16 mixtures were used for THz calibration and eight mixtures were reserved for validation testing.

let was produced for each mixture on a parts-by-weight (w/w) basis, each having a theoretical weight of 1000 g; the complete mixture design is shown in Table 1. Tablet materials were individually dispensed onto weighing paper during weight determination (Data Range, model no. AX504DR, Mettler Toledo, Columbus, OH, USA) and placed directly into 15 × 45-mm glass scintillation vials (Fisher Scientific International, Hampton, NH, USA). After all constituents were dispensed, the nominal mixture concentrations were updated using the observed mass data.

The constituent mixtures were agitated manually and by slowly rotating on a jar mill (US Stoneware, East Palestine, OH, USA). After ~15 min of continuous rotation, an NIR reflectance spectrum was acquired for each mixture by scanning through the bottom of the scintillation vial (FOSS NIR Systems 5000). An *ad hoc* calibration model was created after each blending run, using the acquired spectra and nominal constituent concentrations for all mixtures to estimate blend uniformity. The mixtures were assumed to be effectively homogenous when further mixing failed to yield an increase in the coefficient of determination of the calibration model.

Once mixed, the contents of each scintillation vial were transferred to a 13-mm die for tablet compression. All tablets were compacted using a flat-faced punch with 6000 lb of compression force over a 10-second dwell time (Carver Automatic Tablet Press, model 3887.1SD0A00, Wabash, IN, USA). Three mixture design points (nos 22, 23, and 24) were mixed using sufficient material to create two additional 'split' tablets for verification of dynamic range and resolution of THz chemical imaging. The split tablets were made by inserting a thin rigid divider into the tablet die. With the divider in place, approximately the same volume of material from one of the three design points (22, 23, or 24) was placed on each side of the partition. Immediately before inserting the tablet punch, the divider was gently removed, taking care to avoid disturbing the powder bed. After gently tamping the punch to stabilize the split-sided wad, the tablets were compressed using the procedure described earlier. An additional split tablet was created that consisted of pure lactose and pure MCC.

Data acquisition, instrumentation, and software

Following compaction, a THz scan from one side of each tablet was acquired in reflection mode using a TPI coating scan (TeraView, Cambridge, UK). The flat-face punches used to compress the test tablets created a sharp edge around the circumference of the tablet, which was expected to affect the reflection signature adversely. Hence, a spot with a diameter of ~6 mm near the center of each tablet was scanned (a larger spot was scanned for the split tablets used for image analysis). Depending on the size and geometry of the

tablet area to be scanned, ~750–1250 scans were collected for each side of a tablet, requiring ~10 min of scan time. A single tablet (no. 10) was scanned 10 times to estimate the error of short-term repeatability. The spectral range of interest was ~8–60 cm^{-1}. The scanning parameters were constant for all tablets analyzed.

Conversion of the time-domain TPI waveforms to frequency-domain spectra was handled by batch software provided by TeraView. Both refractive index and optical absorption spectra were calculated and subjected to all subsequent data analyses. Chemometric analyses were performed in the MATLAB programming environment (v7.1, MathWorks, Natick, MA, USA) using the PLS_Toolbox (v3.0, Eigenvector Research, Manson, WA, USA), together with many analysis routines developed in-house (Duquesne University Center for Pharmaceutical Technology, Pittsburgh, PA, USA).

Data analysis

To simplify the data analysis process, only the mean spectrum for each tablet was used for calibration and validation. Using a subset of the thousands of available spectra for each tablet image might have improved the generality of the calibration to a degree, but such objectives were not a focus of this research. Based on the experiences of other researchers, it was anticipated that some pretreatment of the frequency-domain spectra would be required to reduce baseline effects related to physical sample interactions. Typical spectroscopic preprocessing operations such as scatter correction, detrending, and derivatives, in addition to combinations of procedures [10], were tested. The preprocessing method chosen for further method development would be selected based on simplicity and reduction of leave-one-out cross-validation error. The standard errors (*SE*) of calibration (*SEC*), cross-validation (*SECV*), and validation (*SEP*) were calculated according to the following formula (Equation 2):

$$SE = \sqrt{\frac{\Sigma (\hat{y} - y)^2}{n - 1}} \qquad \text{(Eqn 2)}$$

where y, $\hat{y}$ and n are the measured and predicted concentrations and the number of samples for the associated dataset, respectively. In addition to the error statistics, the calibration limit of detection (LOD) was estimated using the standard deviation of prediction (σ) for a series of 10 repeat scans of a single tablet and the International Union of Pure and Applied Chemistry (IUPAC: http://www.iupac.org/dhtml_home.html) formula for LOD [11] (Equation 3):

$$LOD = \frac{k\sigma}{m} \qquad \text{(Eqn 3)}$$

Sample no.	Theophylline (wt/wt)	Lactose (wt/wt)	MCC (wt/wt)
1	0.20	0.59	0.21
2	0.39	0.39	0.21
3	0.60	0.20	0.20
4	0.40	0.20	0.40
5	0.21	0.40	0.40
6	0.20	0.20	0.60
7	0.11	0.45	0.44
8	0.30	0.35	0.35
9	0.40	0.30	0.30
10	0.50	0.25	0.25
11	0.33	0.33	0.34
12	0.25	0.25	0.50
13	0.03	0.49	0.49
14	0.25	0.50	0.25
15	0.40	0.00	0.60
16	0.00	0.40	0.60
17	0.60	0.00	0.40
18	0.00	0.60	0.40
19	0.60	0.40	0.00
20	0.40	0.60	0.00
21	0.05	0.47	0.48
22	0.50	0.30	0.20
23	0.30	0.20	0.50
24	0.30	0.50	0.20

[a]Shaded rows indicate validation samples.

Table 1. Three-component sample mixture plan.[a]

where m is the slope of a univariate classical least-squares fit of the predicted and reference data, and k is the confidence factor (for this work, $k = 3.0$). It is important to note that, for inverse least-squares (ILS) univariate correlation, m is in the numerator of the LOD calculation.

Empirical modeling of the TPS frequency spectra and concentration data was performed using partial least-squares (PLS) type II, whereby the regression coefficient vectors for each of the tablet constituents were estimated simultaneously from the same orthogonal basis set [12]. The ideal number of PLS latent variables for regression was estimated by minimizing the *SEC* and the *SECV*. Between two and six PLS factors were expected to be required for adequate model complexity, based on the number of chemical constituents and physical factors expected to affect the shape of spectra. Whereas high-rank models can be used successfully, calibrations involving significantly more factors than the number of chemical constituents should be approached with caution.

One of the objectives of this research was to investigate the feasibility of applying efficient PCP-method development [13] to reduce the time and expense of future TPS calibra-

Continued on page 68.

Journal of Pharmaceutical Innovation

tion efforts. The analysis of most pharmaceutical solids for process development or control offers a uniquely rich collection of *a priori* information relative to the types of multivariate calibration problem for which most chemometric methods, such as PLS, were developed. The use of chemometric methods based on ILS is largely reflective of their early applications in food, agriculture and petrochemical analysis, in which there might be many constituents that exhibit natural variation over wide ranges [14]. Hence, by necessity, calibration models were often built using hundreds of samples with extremely diverse concentrations.

However, pharmaceutical solids consist of relatively few constituents for which spectral profiles can be acquired easily. Because chemical reactions in drug tablets are typically eschewed, the shape of pure-component spectra in the compacted tablet can be expected to remain extremely similar to the pure spectrum of the raw component. The remaining interacting factors such as matrix hardness and distribution of ingredient particle size affect the spectra in predictable ways. Finally, because the scale of observed composition variation for the analysis of pharmaceutical solids is rather narrow, nonlinearity related to prediction over wide ranges of composition is minimal. Therefore, a linearly additive model of preprocessed optical spectra should be a reasonable assumption. The foundation of efficient spectroscopic calibration is based on the application of *a priori* information such as pure-component spectra, knowledge of relevant constraints, and assumption of a spectral model (i.e., Beer's law) [13].

Two alternative techniques for efficient multivariate calibration using pure-component spectra have been described in the literature: generalized least-squares (GLS) and net analyte signal (NAS). Both techniques are based on a linearly additive model for the observed spectral intensities. Although we can expect the THz frequency-domain reflection spectra to deviate from the linearly additive model to some extent, it is important to realize that the purpose of efficient calibration is to obtain a method that is useful within the expected bounds of the task with as few resources as possible. Hence, until more-complicated models are proven to be of greater value, the simplest methods that can reach our objectives will be applied.

Generalized least-squares techniques have most often been applied in conjunction with ILS methods as a means to attenuate the variance–covariance matrix of predictor variables when confidence is relatively low [13,15-18]. The adjustment is carried out by effectively normalizing the predictor covariance matrix by the inverse of the covariance matrix for known errors such as variance in repeatability or residual predictor variance. By making the assumption that the pure-component spectrum of a tablet component is the maximum-likelihood predictor of that component in a noise-free system, it should follow intuitively that multiplicative attenuation of the pure-component vector by GLS weighting will yield the maximum-likelihood predictor in the presence of structured noise in the predictor variables, given the available training data.

Thus, the GLS–PCP prediction filter [13,19] can be estimated using nothing more than the pure-component spectrum for the component of interest and an estimate of the various spectral noise sources (Equation 4):

$$B_{GLS} = P((T_n'T_n)^{-1} (G^t P)') \quad \text{(Eqn 4)}$$

where G is a matrix of pure-component spectra, and T_n and P are the principal component scores and loadings, respectively, for a matrix of noise spectra; B_{GLS} is the estimated prediction filter. Given sufficient spectral sensitivity, the projection of sample spectra onto a single GLS–PCP prediction vector will yield a univariate output response that is linear with respect to concentration in the component of interest. The GLS–PCP method did not perform well using the data generated from this experiment and will not be discussed further; based on the GLS–PCP formulation, it seems that the algorithm requires extensive characterization of the noise matrix. Otherwise, if the noise matrix is nearly singular, the GLS–PCP results are unstable.

In contrast to GLS–PCP, NAS processing is an additive (or, more intuitively, subtractive) adjustment of the pure-component spectrum for generating a prediction filter. There are (at least) two methods of generating NAS prediction vectors; the first is a generalization of ILS that requires a complete calibration dataset for implementation, and the second involves reducing the rank of the hyperspace spanned by the pure-component vectors. This work is concerned with the second type, whereby the pure-component spectrum for a component of interest is modified by projection into the null space spanned by the pure-component vectors of the other constituents [20] (Equation 5):

$$B_{NAS,k} = [I - G^t_{-k} (G_{-k})^+]G_k \quad \text{(Eqn 5)}$$

where G_k is the pure-component spectrum for the constituent of interest, G_{-k} is the matrix of pure-component spectra for the other components, I is the identity matrix, $^+$ is the Moore–Penrose pseudo-inverse, and $B_{NAS,k}$ is the estimated NAS–PCP prediction filter for the constituent of interest.

Rather than limiting the null-space projection to the pure-component spectra, the significant principal components of interference spectra generated from repeat scans of one of the tablets were added to the projection basis (G_{-k}). The number of principal components included was determined in the same way that the number of PLS factors was chosen. Although this deviates from Lorber's NAS derivation [20], it has the beneficial effect of suppressing spectral noise with structured covariance. It is expected that some analytical signal is lost by including additional interference

components in the G_{-k} matrix; however, such portions of the NAS hyperspace are unreliable for prediction because they cannot be separated from noise signals. The separation of reliable and unreliable analyte signal is the primary advantage of NAS methods. Because few samples were created for this work, repeat spectra from only a single tablet were included in the orthogonalization. For practical applications of NAS–PCP, in which production-scale samples are relatively plentiful, spectral noise matrices can be created to model the effect of many targeted variances (e.g., batch-to-batch, long- and short-term instrument drift) [21].

Once the NAS–PCP filter coefficient matrix was generated, scans of the pure-component material were projected into the basis to produce concentration scores that were useful for normalizing the filter response. Each coefficient vector was offset and scaled such that the response for the corresponding constituent was unity and the mean response for the other constituents was zero. Although this would portend a rather crude means of calibration to an absolute scale, it ensured at least that NAS–PCP predictions would be on a scale similar to the PLS results. Indeed, under ideal conditions, this might prove to be an effective means of establishing a NAS–PCP calibration that is sufficiently accurate on an absolute basis for deployment in certain applications. Finally, because there were no parameters to optimize for NAS–PCP method development, the entire sample dataset can be considered as validation.

Results and discussion

Analysis of spectra

The raw THz refractive index and absorption spectra extracted from the calibration and validation images are shown in Figures 5,6. The dominant feature of both types of spectrum is the intense baseline variation observed. No univariate or multivariate correlation was observed between the tablet compositional factors and the baseline shape. Furthermore, because tablet compression force and powder particle size were not controlled and assumed to be constant, the source of the baseline variation was probably not related to physical or compositional characteristics of the samples. Time-series analysis of the baseline revealed that an abrupt change in baseline occurred between sample sets 1–10 and 11–24. The tablets were created and scanned in two groups (each group being scanned during a period of one or two days), which corresponds with the baseline shift. It is apparent that an intermediate instrument drift was responsible for the aberrant signals. Further research is required to determine the long-term magnitude of instrument variability.

A comparison of the raw refractive-index sample spectra and scaled pure-component spectra illustrates the high level of sensitivity for crystalline lactose and, to a lesser extent, theophylline in the THz region (Figure 5). In contrast to lactose and theophylline, which have visually distinct features in the refractive index spectra, the MCC spectrum appears as a featureless curving baseline. This is not unexpected because many pharmaceutical polymers are mainly transparent in the THz region. Such a lack of phonon resonance features indicates that MCC is more aptly described as amorphous, having insufficient long-range order to sustain lattice vibrations. The raw-sample and pure-com-

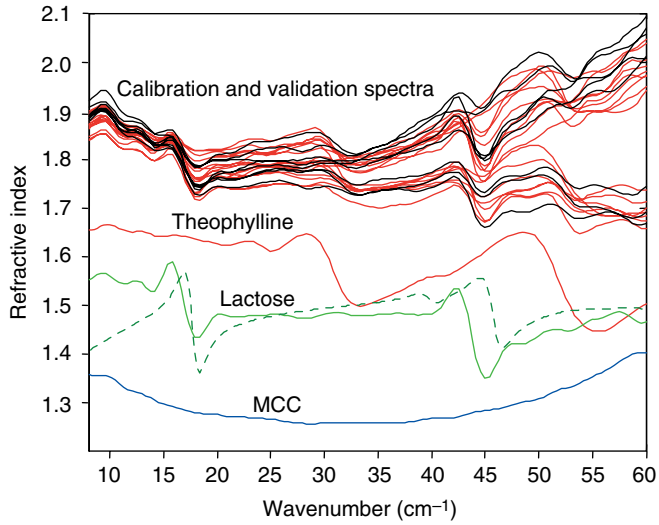

Figure 5. Raw refractive-index reflection spectra for the calibration and validation samples, and the constituent pure-component spectra. A frequency-domain transmission spectrum of lactose has been included for comparison (broken line). All pure-component spectra have been scaled and offset to facilitate plotting.

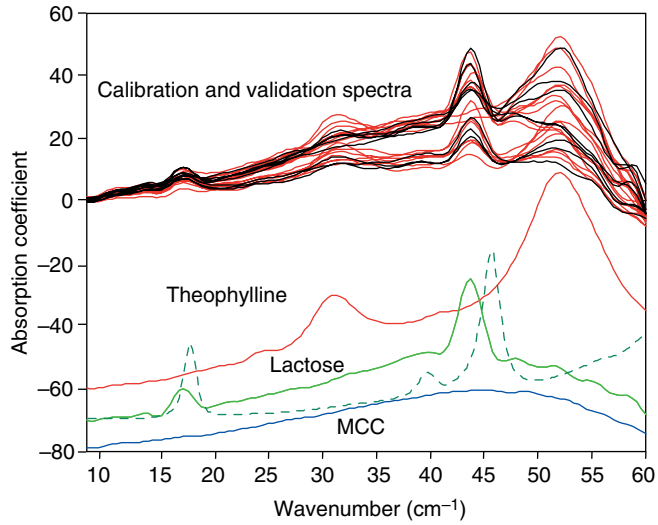

Figure 6. Optical absorption reflection spectra for the calibration and validation samples, and the constituent pure-component spectra. A frequency-domain absorption spectrum of lactose, collected in the transmission mode, has been included for comparison (broken line). All pure-component spectra have been scaled and offset to facilitate plotting.

Continued on page 70.

Journal of Pharmaceutical Innovation

ponent absorption spectra also have many similar features (Figure 6). Indeed, assigning bands observed in the sample spectrum to pure components seems to be more direct for the absorption spectra. As with the refractive index spectra, the MCC absorption spectrum is virtually featureless.

The lactose THz transmission spectrum has been documented in earlier studies [22]; comparison of the lactose reflection and transmission spectra (unrelated samples) illustrated a strong correlation between the measurement modes. Although a slight shift is observed, recent findings of other researchers indicate that spectroscopic phase shifts between TPS transmission and reflection spectra can be attributed to the alignment of the 100%-reflectance mirror reference [23]. Further work is required to determine the root of any differences in frequency-domain spectra between the two measurement modes. Transmission spectra of theophylline and MCC were unavailable at the time of these experiments.

Some similarities can be observed between the refractive index and absorption spectra; qualitatively, the refractive index spectra look similar to a first-derivative profile of the absorption spectra. This correspondence is typical of the anomalous dispersion patterns that are expected to be observed near absorption bands and is described by the classical Kramers–Kronig relation [24-26]. However, it is important to keep in mind that the reflection measurement of compacted tablets presents many significant deviations from the thin-film-type assumptions for which these relations have been developed. At the moment, there is much to be learned regarding the ideal transform between the time and frequency domain and the absolute validity of the refractive index and absorption signals measured by diffuse-reflection TPS of compacted dosage forms [23]. For the objectives of this work, however, it is most important to determine whether sensitive repeatable measurements of composition can be extracted from TPS data.

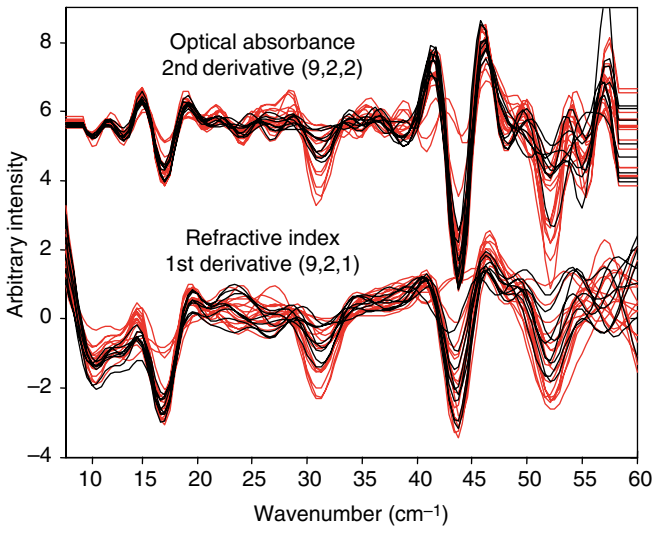

Figure 7. Preprocessed optical absorption and refractive index reflection spectra for the three-component calibration and validation tablets.

The spectral-preprocessing treatment was optimized using visual analysis of the spectra and cross-validation in conjunction with PLS type II regression. Ultimately, Savitzky–Golay 1st-derivative (2nd-order polynomial, 9pt filter) pretreatment [10,27] was selected for the refractive index data, and Savitzky–Golay 2nd-derivative (2nd-order polynomial, 9pt filter) was chosen for the absorption spectra. The spectral derivative operations are apparently effective at mitigating the baseline shift (Figure 7). Smoothing derivatives are also efficient at amplifying useful features in the spectra; in a signal-processing sense, Savitzky–Golay filtering is similar to a midband filter.

It is important to keep in mind that preprocessing operations are not necessary to generate useful calibration models in conjunction with PLS regression because the basis-set estimation process is often effective at isolating features of interest from a background of structured noise. In most cases, calibration models for similar data types (i.e., NIR) estimated without preprocessing require one or more additional latent variables to achieve accurate results. However, the advantage of preprocessing is that operations such as scatter correction and derivatives are invariant filters, whereas approximations of the structured noise using additional latent variables can be over-specific to the noise sources present in the calibration data. In the event that the instrumental drift profile changes, the noise model might no longer apply, whereas invariant preprocessing operations might continue to be effective at mitigating the baseline variation. Thus, whether latent variable ILS or PCP models are used for method development, a finite amount of data preprocessing is necessary to maximize calibration robustness [10].

Partial least-squares calibration

The optimization and performance statistics for the PLS and NAS–PCP calibrations are summarized in Tables 2,3. Five PLS latent variables were used to model both the refractive index and the absorption spectra. The performance statistics for both datasets were acceptable with fewer latent variables (not shown); however, this would be discovered only after validation. The performance results for refractive index and absorption spectra were essentially the same overall. It is crucial during chemometric-method development that the independence of the validation dataset be maintained. Hence, the optimal PLS rank was chosen based on interpretation of the *SECV* scree plot (illustrating the trend of *SECV* as related to the model rank); the PLS rank after which *SECV* failed to improve significantly was selected [28]. Whereas the calibration and cross-validation performance was very good for all constituents, the validation results for

Data type	Optical absorbance					
Method	**PLS**			**NAS–PCP**		
Preprocessing	2nd derivative (9, 2, 2)					
Latent variables	5			3		
Constituent	**Theophylline**	**Lactose**	**MCC**	**Theophylline**	**Lactose**	**MCC**
***SEC* (wt/wt)**	0.024	0.017	0.024	0.046	0.089	0.400
***SECV* (wt/wt)**	0.048	0.029	0.055	NA	NA	NA
***SEP* (wt/wt)**	0.040	0.111	0.108	0.055	0.158	0.331
R^2_{cal}	0.984	0.992	0.984	0.960	0.945	0.477
R^2_{cval}	0.937	0.977	0.918	NA	NA	NA
R^2_{val}	0.959	0.517	0.479	0.926	0.523	0.593
Mass bal.$_{cal}$ (μ, σ)	1.00, 3.82E-16			NA		
Mass bal.$_{val}$ (μ, σ)	1.00, 2.41E-16			NA		
LOD (wt/wt)	0.11	0.14	0.23	0.05	0.01	NA

[a]Abbreviation: NA, not applicable.

Table 2. PLS and NAS–PCP calibration and validation performance statistics using optical absorption spectra.[a]

lactose and MCC were not as good. When prediction error is significantly greater than error of calibration or cross-validation, overfit of the calibration data should be suspected. However, based on the prediction plots (Figure 8), it is reasonable to assume that the poor validation statistics were related to the limited number of available samples and the narrow span of variance in the validation reference chemistry for these constituents. Furthermore, predictions for the same two validation samples were outlying in lactose and MCC, thereby off-setting one another. Because the predicted mass balance for the calibration and validation samples was essentially constant across all samples, one of two conclusions can be made: (i) some material was not accounted for when the tablets were created, causing the concentrations to deviate from the design; or (ii) MCC predictions using the PLS calibration were based on concentration closure, thereby forcing all predicted MCC concentrations to equal the mass balance remaining after prediction of theophylline and lactose. Because the spectral features of MCC are extremely subtle (relative to theophylline and lactose), the latter conclusion seems more probable. However, further results are required to prove whether or not MCC can be independently predicted from TPS reflection spectra.

Limit of detection was calculated independently for each of the constituents, using all of the available predictions (i.e., the calibration and validation predictions were pooled) to calculate the slope variable, and 10 repeat scans of a single tablet (no. 10) to calculate the error of repeatability (σ). Because the slope of the PLS predictions was unity, however, the LOD was essentially three times the error of repeatability. The LOD results for theophylline and lactose were consistent, whereas the LOD for MCC was nearly twice as much. The LOD results

Data type	Refractive index					
Method	**PLS**			**NAS–PCP**		
Preprocessing	1st derivative (9, 2, 1)					
Latent variables	5			3		
Constituent	**Theophylline**	**Lactose**	**MCC**	**Theophylline**	**Lactose**	**MCC**
***SEC* (wt/wt)**	0.022	0.012	0.019	0.045	0.105	0.451
***SECV* (wt/wt)**	0.037	0.036	0.054	NA	NA	NA
***SEP* (wt/wt)**	0.028	0.109	0.116	0.030	0.136	0.636
R^2_{cal}	0.986	0.996	0.990	0.973	0.962	0.125
R^2_{cval}	0.963	0.965	0.923	NA	NA	NA
R^2_{val}	0.981	0.488	0.581	0.974	0.555	0.069
Mass bal.$_{cal}$ (μ, σ)	1.00, 2.2E-16			NA		
Mass bal.$_{val}$ (μ, σ)	1.00, 2.18E-16			NA		
LOD (wt/wt)	0.08	0.08	0.14	0.01	0.05	NA

Table 3. PLS and NAS–PCP calibration and validation performance statistics using refractive index spectra.

Continued on page 72.

Journal of Pharmaceutical Innovation

for all components seem to be acceptable for most applications. Moreover, if additional spectra were added to the model, a better approximation of the noise structure could be developed that would tend to reduce the LOD over time.

The calibration coefficient vectors for all models are shown superimposed with the associated preprocessed pure-component spectra in Figure 9. The PLS regression vectors for both theophylline and lactose were highly correlated to the pure-component spectra with no major uncorrelated peaks. This indicated that there was ample sensitivity for these constituents and that the pure-component sensitivity vectors were essentially mutually orthogonal. The MCC regression vector had little correlation with the MCC pure-component spectrum, which indicated that the MCC calibration was based on closure. Superimposition of the PLS regression vectors demonstrates that the MCC regression vector is a linear combination of the theophylline and lactose vectors (not shown). Furthermore, although the absolute value of the correlation coefficients between pure-component spectra is dependent on the type of preprocessing applied, the correlation coefficients between theophylline and lactose were consistently low (<< 0.5). However, the correlation between either theophylline or lactose and MCC was much higher (>0.75), which also supports the supposition that the MCC regression correlated to linear combinations of the other two components. Finally, because the pure-component sensitivities of theophylline and lactose are essentially orthogonal, these components are extremely amenable to the NAS adjustment; in these cases, the bulk of spectral NAS adjustment is related to the suppression of structured noise factors, leaving a large part of the original NAS intact for prediction.

Three-component prediction images for the split tablets are shown in Figure 10. The lactose–MCC split-tablet image illustrates the spatial resolution in a maximal-contrast situation. Within the split-mixture tablets, the dividing line is less visible. The presence of individual red, green, and blue spots indicates that the scale of spatial resolution is sufficiently fine to detect individual constituent domains. There seem to be some larger-scale patterns of distortion; it

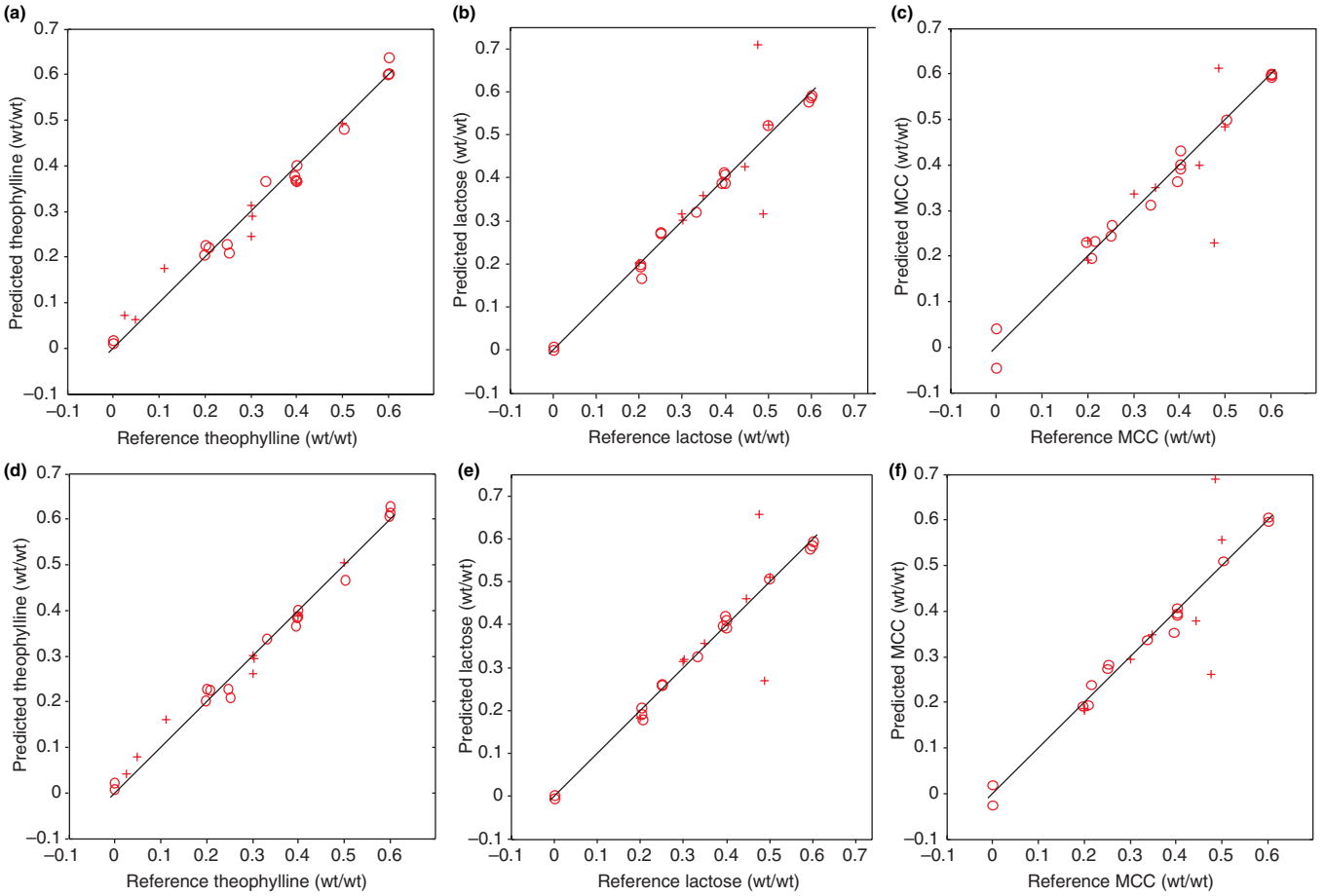

Figure 8. Prediction plots for optical absorption spectra. Optical absorption **(a–c)** and refractive index **(d–f)** spectra of calibration (circles) and validation (crosses) PLS predictions for theophylline (a,d), lactose (b,e), and MCC (c,f). The panels have been scaled to be nearly equal; the solid line in each plot represents the ideal (unity slope) prediction line.

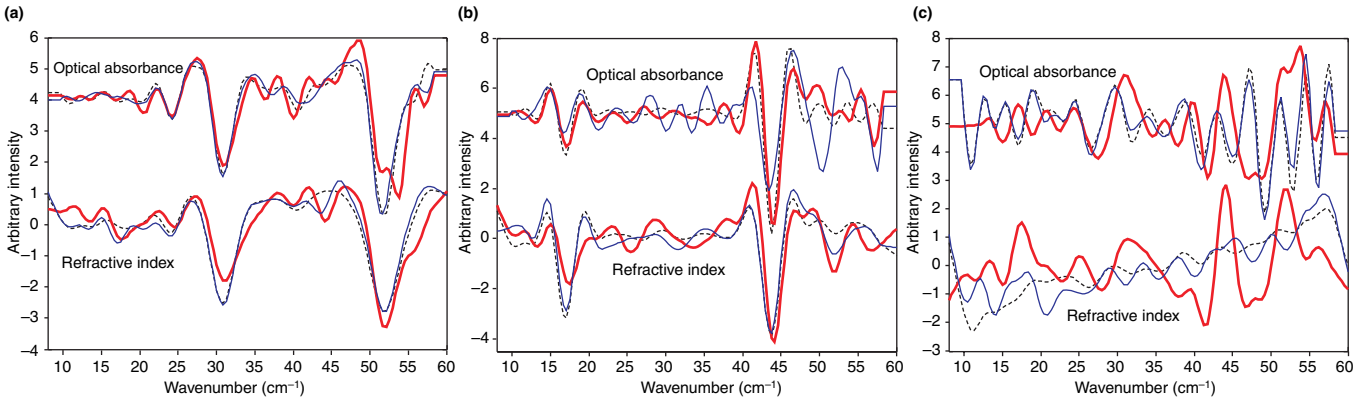

Figure 9. Pure-component and calibration coefficient vectors. PLS regression coefficient vectors (thick red lines), NAS–PCP prediction filters (thin blue lines), and pure-component spectra (broken lines) for optical absorption (upper curves) and refractive index (lower curves) data. Vectors for theophylline **(a)**, lactose **(b)**, and MCC **(c)** are shown. All vectors and spectra have been scaled for plotting.

is unknown whether these artifacts were related to the physical properties of the tablets, the limitations of the algorithms used to generate the image hypercubes from the raster-scan data, or the limitations of the instrumentation. Ultimately, given the present optical configuration, the instrument is limited by the diffraction of THz radiation, which – according to Rayleigh criteria – is limited to no better than half the wavelength of incident radiation. Hence, over the range of 8 cm^{-1} to 60 cm^{-1}, the minimum size range of truly discernable features is ~625–83 μm.

NAS–PCP calibration

In general, the performance of NAS–PCP at predicting theophylline and lactose was somewhat worse than that observed for PLS although the results were very good considering that no regression (or reference chemistry) was involved. Indeed, the linearity of both the PLS and NAS–PCP results is remarkable given the extremely wide range of concentrations sampled for all constituents. However, the NAS–PCP predictions of MCC were completely uncorrelated with MCC concentration (Figure 11). This supports the view that the PLS model for MCC was based on concentration closure rather than on spectroscopic sensitivity. The LOD results for NAS–PCP were uniformly better than those for PLS (without considering MCC). One possible explanation is that, because the NAS–PCP prediction filters are 'smoother', the solution is more 'generalized' and less influenced by small spectral variation; alternatively, the PLS regression vectors might have been overfitted, which tends to reduce generalization. Finally, the NAS–PCP algorithm might have had an inherent advantage because repeatability error is explicitly suppressed during filter estimation, rather than implicitly modeled by regression. Mass-balance predictions were not analyzed for the NAS–PCP calibration because the MCC predictions were of no value.

As would be expected, the NAS–PCP coefficient vectors were highly correlated with the pure-component spectra (Figure 9). The theophylline and lactose vectors were modified only slightly: again, probably because the components were nearly mutually orthogonal to begin with. The MCC prediction filter was completely uncorrelated with the pure-component spectrum. Because the pure-component spectrum was essentially a curving baseline with some minor features, orthogonalization to the other components might have resulted in a linear combination of the converse of each of the vectors. An objective of future research projects will be to determine whether such dissimilarity between the pure-component and NAS–PCP vectors is a reliable indicator that prediction results will be unreliable.

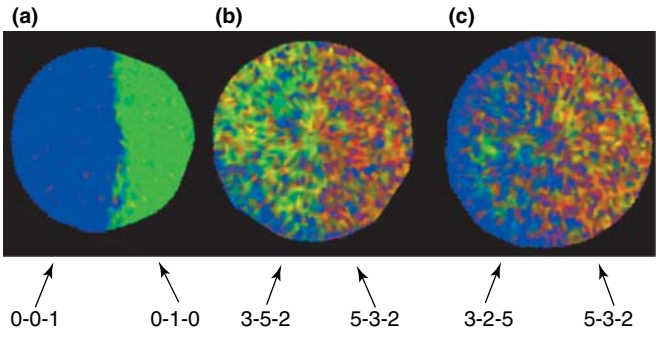

Figure 10. Pseudocolor prediction images of split tablets. Images scaled so that red, green, and blue intensities are proportional to the concentration of theophylline, lactose, and MCC, respectively. Tablet compositions are (left–right): **(a)** (0-0-1:0-1-0), **(b)** (3-5-2:5-3-2), and **(c)** (3-2-5:5-3-2). Off-color domains in the lactose–MCC tablet (a) can be attributed to spectral noise and/or prediction error. Non-primary colored domains (e.g., yellow, orange, and purple) indicate region of uniform mixtures in which domain size for individual constituents is below the diffraction limit or spatial resolution.

Continued on page 74.

Journal of Pharmaceutical Innovation

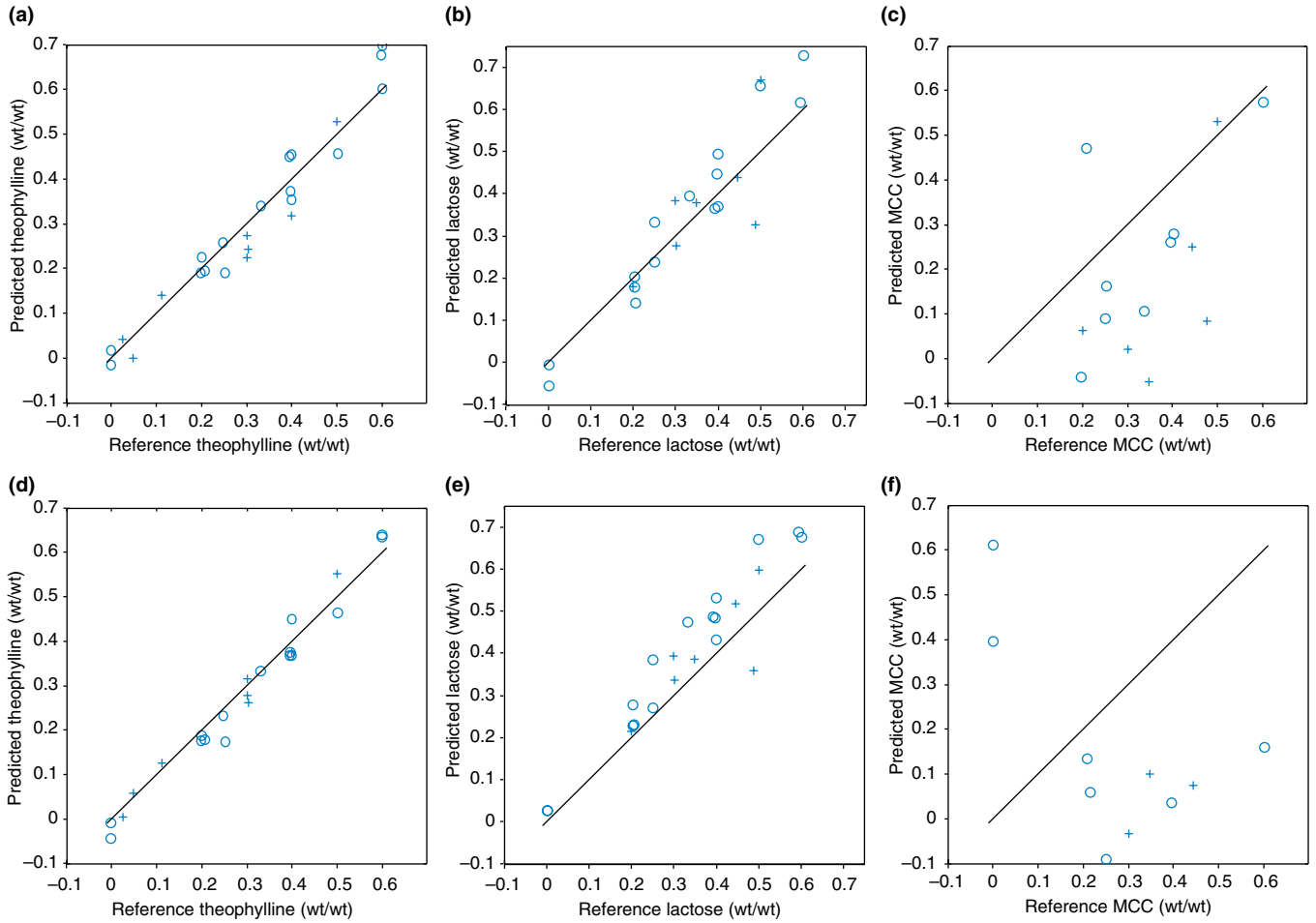

Figure 11. Prediction plots for refractive index spectra. Optical absorption **(a–c)** and refractive index **(d–f)** spectra of calibration (circles) and validation (crosses) NAS–PCP predictions for theophylline (a,d), lactose (b,e), and MCC (c,f). The panels have been scaled to be nearly equal; the solid line in each plot represents the ideal (unity slope) prediction line. The MCC prediction panes contain fewer data points because many of the predicted values were outside the range of feasible composition; the plots have been included to demonstrate that there is a total lack of sensitivity, rather than simple slope or bias error.

Concluding remarks

The results of this study illustrate a procedure by which TPS can be used in conjunction with multiple chemometric tools for the quantitative chemical mapping of crystalline pharmaceutical materials. Although the observed performance of TPS for the quantitative analysis of such common pharmaceutical solids was probably no better than the results that might be obtained using techniques such as NIR or NIR imaging spectroscopy, the objectives of this study were to investigate the analytical capability of TPS at the most basic level. Ultimately, the advantages of TPS (over techniques such as NIR and Raman) will be explored using more-complicated solid-state characterization problems for which there are currently few viable options in terms of analytical instrumentation.

Based on the observations, there is no apparent advantage to using either optical absorption or refractive index for method development. Additional intensive studies are required to determine whether either method of spectral transformation is of greater use than the other. The performance of the NAS–PCP technique at predicting theophylline and lactose, without the need for regression or chemical reference data, is extremely encouraging because, ultimately, PCP methods will greatly reduce the time and expense of method development. Additionally, the results demonstrate that NAS–PCP can be used as a validation tool with which to determine the specificity of a multivariate calibration model. Further research studies will focus on higher-order mixture designs in an effort to break the concentration closure, thereby, resolving more accurately the quantitative limits of TPS for pharmaceutical analysis.

References

1. FDA (2004) PAT – a framework for innovative manufacturing and quality assurance (http://www.fda.gov/cder/guidance/6419fnl.pdf) .
2. Taday, P.F. (2004) Applications of terahertz spectroscopy to pharmaceutical sciences. *Philos. Trans. R. Soc. Lond. A* 362, 351-364.
3. Day, G.M. *et al.* (2006) Understanding the Influence of polymorphism on phonon spectra: lattice dynamics calculations and terahertz spectroscopy of carbamazepine. *J. Phys. Chem. B* 110, 447-456.
4. Strachan, C.J. *et al.* (2004) Using terahertz pulsed spectroscopy to study crystallinity of pharmaceutical materials. *Chem. Phys. Lett.* 290, 20-24.
5. Taday, P.F. *et al.* (2003) Using terahertz pulse spectroscopy to study the crystalline structure of a drug: a case study of the polymorphs of ranitidine hydrochloride. *J. Pharm. Sci.* 92, 831-838.
6. Shen, Y.C. *et al.* (2005) Detection and identification of explosives using terahertz pulsed spectroscopic imaging. *Appl. Phys. Lett.* 86.
7. Neil, G.R. *et al.* (2003) Production of high power femtosecond terahertz radiation. *Nucl. Instrum. Methods Phys. Res. A* 507, 537-540.
8. Woodward, R.M. *et al.* (2002) Terahertz pulse imaging in reflection geometry of human skin cancer and skin tissue. *Phys. Med. Biol.* 47, 3853-3863.
9. Hangyo, M. *et al.* (2005) Terahertz time-domain spectroscopy of solids: a review. *Int. J. Infrared Millimeter Waves* 26, 1661-1690.
10. Cogdill, R.P. *et al.* (2005) Process analytical technology case study, part ii: development and validation of quantitative for tablet API content and hardness. *AAPS PharmSciTech* 6.
11. Long, G.L. and Winefordner, J.D. (1983) Limit of detection: a closer look at the IUPAC definition. *Anal. Chem.* 55, 712A-724A.
12. De Jong, S. (1993) SIMPLS: an alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* 18, 251-263.
13. Cogdill, R.P. and Anderson, C.A. (2005) Efficient spectroscopic calibration using net analyte signal and pure-component projection methods. *J. Near Infrared Spectrosc.* 13, 119-132.
14. Williams, P. and Norris, K., eds (2001) *Near-infrared Technology in the Agricultural and Food Industries* (2nd edn), American Association of Cereal Chemists.
15. Martens, H. and Næs, T. (1989) *Multivariate Calibration* (1st edn), John Wiley & Sons.
16. Næs, T. (1985) Multivariate calibration when the error covariance matrix is structured. *Technometrics* 27, 301-311.
17. Næs, T. (1986) Multivariate calibration by covariance adjustment. *Biometrical J.* 28, 99-107.
18. Næs, T. (1985) Comparison of approaches to multivariate linear calibration. *Biometrical J.* 27, 265-275.
19. Marbach, R. (2002) On Wiener filtering and the physics behind statistical modeling. *J. Biomed. Opt.* 7, 130-147.
20. Lorber, A. (1986) Error propagation and figures of merit for quantification by solving matrix equations. *Anal. Chem.* 58, 1167-1172.
21. Haaland, D.M. and Melgaard, D.K. (2000) New prediction-augmented classical least-squares (PACLS) methods: application to unmodeled interferents. *Appl. Spectrosc.* 54, 1303-1312.
22. Walther, M. *et al.* (2005) Metal-wire terahertz time-domain spectroscopy. *Appl. Phys. Lett.* 87, 261107.
23. Lucarini V. *et al.* (2005) Detection and correction of the misplacement error in terahertz spectroscopy by application of singly subtractive Kramers–Kronig relations. *Phys. Rev. B* 72(12), 125107(6).
24. Mouret G. *et al.* (2006) Anomalous dispersion measurement in terahertz frequency region by photomixing. *Appl. Phys. Lett.* 88(18), 181105(3).
25. Parthasarathy R. *et al.* (2005) Dielectric properties of biological molecules in the terahertz gap. *Appl. Phys. Lett.* 87(11), 113901(3).
26. Yu, B. *et al.* (2004) Torsional vibrational modes of tryptophan studied by terahertz time-domain spectroscopy. *Biophys. J.* 86, 1649-1654.
27. Savitzky, A. and Golay, M.J.E. (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36, 1627-1639.
28. Johnson, R.A. and Wichern, D.W. (1998) *Applied Multivariate Statistical Analysis* (4th edn), Prentice Hall. JPI