Automatic Cocrystal Detection by Raman Spectral Deconvolution-Based Novelty Analysis

Automatic Cocrystal Detection by Raman Spectral Deconvolution Based Novelty Analysis

Mehrdad Yaghoobi,[*,†] Tudor Grecu,[*,‡] Stephanie Brookes,[*,¶] and Colin J. Campbell[*,§]

†*School of Engineering, University of Edinburgh, United Kingdom*

‡*Solvias, Römerpark 2, 4303 Kaiseraugst, Switzerland*

¶*New Modalities and Parenterals Development, Pharmaceutical Technology and Development, Global Operations, AstraZeneca, Macclesfield, UK*

§*School of Chemistry, University of Edinburgh, United Kingdom*

E-mail: M.Yaghoobi-Vaighan@ed.ac.uk; Tudor.Grecu@solvias.com; Stephanie.Brookes@astrazeneca.com; Colin.Campbell@ed.ac.uk

Abstract

Cocrystals are important molecular adducts that have many advantages as a means of modifying the physicochemical properties of active pharmaceutical ingredients, including taste masking and improved solubility, bio-availability and stability. As a result, the discovery of new cocrystals is of great interest to commercial drug discovery programs. Time consuming manual analysis of the large volumes of data that emerges from large scale cocrystal screening programs of up to 1000s of preparations poses a challenge. Raman spectroscopy has been shown to discriminate between cocrystals and physical mixtures and is easy to automate allowing rapid screening of large numbers of potential cocrystals, but the spectral features that encode the information are often subtle (e.g. slight changes in peak positions or intensities). We have employed an automated signal processing routine based on a sparse decomposition algorithm to speed up

the data processing steps while maintaining the accuracy of a trained spectroscopist. We used our algorithm to screen 31 potential cocrystal preparations and found that through the use of a computationally generated threshold, we could achieve a clear classification of cocrystals and physical mixtures in less than a minute, compared to several hours manually.

Introduction

Cocrystals have been defined as multiple component crystals in which all components are solid and neutral under ambient conditions, thus differentiating themselves from solvates or salts.[1,2] In pharmaceutical cocrystals, active pharmaceutical ingredients (APIs) interact via non-covalent bonds with molecules from the Generally Recognised as Safe List, (European Medicine Agency, 2015) forming a bespoke material with potentially superior physical properties compared to the pure starting molecules. The molecules that interact with the API to form cocrystals are called cocrystal formers (CCFs). Pharmaceutical cocrystals have been proven to improve solubility, dissolution, stability, bioavailability, mechanical properties and taste masking, as well as protection and extension of intellectual property.[3–5] In 2015, the European Medicines Agency (EMA) was the first scientific body to regard cocrystals as an alternative to salts for providing appropriate solid state properties.[6] Due to the salt-cocrystal continuum[7] and their conceptual similarities, the EMA recommends that similar principles should be applied for salt and cocrystal safety and efficacy documentation from a regulatory point of view. The FDA perspective on pharmaceutical cocrystals changed in 2018,[8] when the guideline stopped classifying cocrystals as drug product intermediates. From a regulatory perspective, a cocrystal is now treated in the same way as a new polymorph of the same API and not as a different chemical entity. The new classification effectively simplifies the cocrystal regulatory landscape, as it is possible to use existing regulatory documents to establish potency, purity and stability of a cocrystal API. This aligns the FDA guidelines with those of the EMA, offering cocrystals exciting opportunities in the pharmaceutical industry.

There have been a significant number of pharmaceutical cocrystals approved as marketed products by the FDA, with some of the recent ones being Entresto® for the treatment of symptomatic heart failure, Odomzo® for skin cancer in 2015 as well as Steglatro® for diabetes in 2017.[9] Increased demand and commercialization of cocrystals inevitably leads to a need for robust screening and manufacturing cocrystal methods. As more testing is carried out experimentally, reliable and automated detection methods are required for cocrystal recognition and discrimination from large databases of starting materials. Cocrystal screening methods can be solid state and solution based. Solid state grinding has been successfully used to generate cocrystals, both by neat grinding (NG) or liquid assisted grinding (LAG). Ultrasonication using different molar ratios of caffeine and maleic acid has been used to obtain pure cocrystals of different stoichiometries.[10] In some cases, spontaneous cocrystals were obtained by contact formation, without grinding the components.[11] High throughput LAG has been achieved using a modified plenary mill, where up to 48 cocrystal systems can be ground in parallel.[12] Cocrystals were also obtained in 96 well plates subjected to solvent mediated sonication or vortexing.[13] Resonant acoustic mixing was implemented for cocrystal screening and scale-up, using beads to accelerate the kinetics of cocrystal formation.[14] All the above cocrystal screening methods use on/off-plate X-ray powder diffraction (XRPD) as an unambiguous way to differentiate between newly obtained phases and physical mixtures of starting materials.

Raman spectroscopy is also used in solid state analysis as it provides fast, non-destructive, non-contact measurements of physical properties and compositional changes. The analysis can be done on only a fraction of the amount of material required for XRPD and is also fully amenable to automation. Other advantages of Raman spectroscopy include high molecular specificity and minimal need for sample pre-treatment. A validated, high throughput cocrystal slurry approach was developed on 96 well plates, where resulting solids were analysed by Raman spectroscopy.[15] LAG followed by FTIR and Raman spectroscopy were recently used in conjunction with Differential Scanning Calorimetry (DSC) as a tool to discover

new cocrystals. The spectral and thermal characteristics were used to differentiate starting materials from new phases.[16] Another high-throughput ultrasound-assisted cocrystallisation screen of hydrochlorothiazide used mid-infrared spectroscopy and multivariate data analysis to assess cocrystal formation and purity. The study reported analysis problems due to fluorescence when Raman microspectroscopy was used.[17] For a theophylline-benzoic acid system, the rate of cocrystal conversion in slurry was also monitored by Raman spectroscopy, giving nucleation time and temperature information.[18] With the development of experimental high throughput screening methods generating increasingly vast amounts of data that need to be analysed, there is an ever-increasing need for fast and reliable automated data analysis. Such automation would save scientists considerable time and effort when manually/visually comparing and interpreting Raman data. Traditional Raman analysis for discovering new solid forms involves recording the spectra of starting materials and new phases. A visual comparison is then carried out in a search for Raman bands that are shifted with respect to the reagents, indicating the formation of a new solid form. These systems are classified as cocrystal leads, and cocrystal scale-up followed by detailed solid state characterization can be used to confirm cocrystal formation. In a high throughput screening (HTS) context, the visual analysis of the obtained spectra is a time consuming step due to the large number of experiments performed in well plates. The relatively low success rate for obtaining cocrystal leads means that during a screen, one will mostly obtain physical mixtures of the reagents. Visual interpretation of Raman results for a 96-well plate can take several hours and up to days for difficult systems that show only subtle differences in their Raman spectra. With such extended manual data analysis protocols operator fatigue can be an issue and cocrystal leads could be missed based on visual spectral comparison as the differences in Raman spectra are often more subtle than XRPD differences. To facilitate Raman spectroscopy in a HTS fashion, multivariate analysis is required to deconvolute the spectral lines.[19]

Since the advent of digital spectroscopy many multivariate techniques,[20] such as multivariable statistical techniques, have been proposed to identify important spectral compo-

nents, *e.g.* principal components[21] and independent components.[22] The underlying structure of Raman spectra (the peaks that correspond to molecular vibrational modes) facilitate the identification or classification of molecules or mixtures. However, multivariate techniques often need multiple measurements of the samples, due to the statistical inference nature of the methods, to be able to blindly find the components. An alternative to this non-biased approach, is to use our prior knowledge about spectral signals (i.e. where the peaks are), with the aim of reducing the number of necessary measurements, down to a single measurement. The use of a small number of peak positions to represent the overall spectrum of a molecule is known as sparse decomposition,[23] and has previously been demonstrated for the analysis of complex mixtures using Raman sepctroscopy.[24]

Our proposed cocrystal detection method builds on the concept of sparse decomposition and is based on the observation that the spectra of physical mixtures and cocrystals are different from each other (ie the spectra of cocrystals often have different peak positions compared to the spectra of the individual components).[25]

In this paper, we demonstrate that we can automatically decompose the spectra and compare them to the spectra of the individual components. If a spectrum is not well modelled by the combination of the spectra of the individual components (*i.e.* the fit of the spectrum to the model has a large residual energy) it generates a high novelty score and is likely to be a new physical form. In this work, the method is validated for cocrystal formation using known cocrystals of two APIs, (nalidixic acid and resveratrol),[26] but it can also be applied for screening new polymorphs, salts, solvates, hydrates and other molecular adducts.

Experimental Section

Generation of cocrystals and physical mixtures

Materials

Nalidixic acid (NLD), resveratrol (RES) and all cocrystal formers (CCFs) selected for experimental screening were purchased from Sigma-Aldrich and used without further purification. Analytical grade solvents were used for liquid-assisted grinding experiments. Three different methods of cocrystal formation were used: ball milling in stainless steel jars on a 20 mg scale, grinding in glass vials on a 5 mg scale and ultrasonication in 96 quartz well plates on a 1-2 mg scale.

Cocrystallisation Experiments

In a ball mill: A weighed amount of nalidixic acid or resveratrol (20-30 mg) along with the corresponding CCF was combined in a 5 mL stainless steel grinding jar containing a 7 mm diameter grinding ball. In liquid-assisted grinding (LAG) experiments, 20-30 μL of ethanol, methanol or n-heptane was also added. No solvent was added to the neat grinding (NG) experiments. The mixtures were ground on a Retsch MM 200 mixer mill for 20-45 min at 25 Hz. NLD was ground with tert-butylhydroquinone (1:1), hydroquinone (1:1), phloroglucinol (1:1), orcinol (1:1), resorcinol (1:1) and propyl gallate (1:1) to obtain cocrystals with stoichiometry indicated in brackets.[26] RES was ground with methenamine (1:1), phenazine (1:3.5), 4,4' -bipyridine (1:1.5), 4-dimethylaminopyridine (1:2) and piperazine (1:1) to obtain cocrystals with stoichiometry indicated in brackets. Additionally, RES was also ground with carbamazepine (1:1), theophylline (1:1) and trimesic acid (1:1) as control experiments. In a previous study, these CCFs were found to give physical mixtures of the starting materials following grinding experiments.[27] Resulting materials were characterized by XRPD and Raman spectroscopy.

In a pulverisette: Milling was carried out in Automaxion's 12-slot vial attachment com-

patible with the Fritsch Pulverisette for cocrystal screening.[12] The cocrystal components were weighed into 2 mL glass vials containing 3 mm stainless steel beads for grinding. 5 mg API and the corresponding stoichiometric amount of the CCF were used. Following catalytic solvent addition, the samples were subjected to LAG for 2 hours. Resulting materials were characterized by XRPD and Raman spectroscopy.

Ultrasonication: Stock solutions of the APIs and CCFs were made up at 0.025 M to 0.1 M in chloroform (for NLD) and MeOH (for RES). Stoichiometric amounts of API and CCF solution were pipetted into the wells of a 96 quartz well plate and solvent was evaporated under ambient conditions. The plate was then ultrasonicated to aid cocrystal formation. Ultrasonication was carried out in a Qsonica 700-Watt Sonicator system with a 431MPX microplate horn accessory at an amplitude of 80 percent. Three ultrasonication cycles of 20 minutes each were carried out, with a 15 minute pause in between them to avoid overheating. Resulting materials were characterized by XRPD and Raman spectroscopy.

Physical mixtures API and CCF

For the systems named "Physical mixture", the corresponding stating materials were gently mixed together (without grinding) to encourage a physical mixture and avoid cocrystal formation. Resulting materials were characterized by XRPD and Raman spectroscopy. Please see Tables 2 and 3 for the experimental conditions used for each of the tested systems.

Characterisation

All experiments were prepared on a large enough scale to allow XRPD analysis to confirm cocrystal formation or presence of physical mixtures. X-ray Powder Diffraction (XRPD) Measurements: XRPD data were collected on a Bruker D4 Endeavor diffractometer in reflectance mode. The powder samples were smeared onto zero-background silicon wafer sample holders. Each sample was exposed to Cu Kα_1 and Cu Kα_2 radiation with an average wavelength of 1.5418 Å, for 0.12 seconds per 0.02° 2θ increment (continuous scan mode) over

the range 2° to 40°. The operating voltage was 40kV and the operating current was 40mA.

Raman Measurements: Raman spectra were acquired using a Thermo DXR2 dispersive Raman microscope with a 785nm laser excitation and a 400 lines/mm grating. Laser power was set to the maximum (30 mW). Raman spectra were collected in the range 300-1900 cm^{-1}, using two acquisitions of between 10-20 seconds each, depending on the sample. A 10x magnification was used for measurements in quartz well plates. Samples generated in the ball mill and pulverisette were analysed on glass microscope slides using a 20x microscope objective. Alignment and calibration of the instrument was carried out prior to sample characterisation.

For the pure components, XRPD and Raman analysis was done on the powder API and CCF as received. This data was then used for the comparison to the LAG, NG, physical mixture and ultrasonication results. The XRPD was measured to compare it to previously reported cocrystals and ensure the cocrystal formed consistently using these methods. Raman was then measured to generate a library of the cocrystals obtained using different methods.

Data analysis

Assuming that all of our spectra are recorded over the same wavenumber range and at the same spectral resolution, we can represent them as column vectors where the API and CCF spectra respectively are $\mathbf{m}_i, 0 \leq i \leq N_a$ and $\widetilde{\mathbf{m}}_j, 0 \leq j \leq N_c$, where N_a and N_c are the numbers of different APIs and CCFs respectively. We can then put all spectra $\widehat{\mathbf{m}}_i$ and $\widetilde{\mathbf{m}}_j$ in the columns of a larger matrix $\mathbf{M}_{d\times(N_a+N_c)} = [\widehat{\mathbf{M}}_{d\times N_a}, \widetilde{\mathbf{M}}_{d\times N_c}]$, where $\widehat{\mathbf{M}}_{d\times N_a}$ and $\widetilde{\mathbf{M}}_{d\times N_c}$ are respectively API and CCF library matrices. The spectrum of a physical mixture can be modelled as a weighted sum of one $\mathbf{m}_i$ and one $\mathbf{m}_j$. The contribution of each component spectrum to the overall mixture can be characterised by a "positive" coefficient α_j, *i.e.* zero means no contribution and larger values, more contributions. The recorded spectrum of a mixture can then be represented by $\mathbf{y}_d$, where y_i is the measurement at ith wavenumber.

The generation of $\mathbf{y}$ can be modelled as:

$$\mathbf{y} = \mathbf{M}\boldsymbol{\alpha} + \mathbf{b} + \boldsymbol{\omega}, \tag{1}$$

where $\boldsymbol{\alpha} = [\alpha_j]_{j=1:N}$, $N = N_a + N_c$, $\mathbf{b}$ is the Fluorescence background signal and $\boldsymbol{\omega}$ includes the measurement noise. The background signal can be removed using the fact that Raman spectra $\mathbf{M}\alpha$ and background $\mathbf{b}$ are morphologically different, *i.e.* Raman spectra have sharp peaks whereas Fluorescence gives a broad background. By removing the background signal from measurements, the spectral mixture model can be simplified as,

$$\mathbf{y} = \mathbf{M}\boldsymbol{\alpha} + \mathbf{w}, \tag{2}$$

where the model-mismatch is small and it is absorbed in $\mathbf{w}$. Because the spectrum of a cocrystal is not well represented as a weighted sum of the API and CCF (because there is a new phase in the sample), we can use the poorness of fit between the spectrum and equation 2 to derive a *Novelty Score* based on the energy of residual signal. The energy of a signal, in the discrete spectral domain, is defined as the square-root of the sum of square of spectral values at different wave-numbers. The defined novelty score potentially enables us to classify physical mixtures from new phase cocrystals.

The sparse decomposition of $\boldsymbol{y}$, using the noisy linear model (2), generates a representation $\boldsymbol{\alpha}$ with only a few non-zero components. Sparse decomposition in this context does not mean that the spectra has only few non-zero elements, but it can be parsimoniously represented using the columns of $\mathbf{M}$. Such a task is combinatorial and computationally complex task, which cannot be exactly solved by conventional algorithms.[28] Various practical algorithms have been proposed to approximately solve this problem with convex relaxation, iterative updates, greedy selection, message passing and Bayesian approaches as the most well known algorithms.[29] The greedy selection methods are among the computationally cheapest methods for the sparsity diets of Raman spectral decomposition, *i.e.* very sparse.[24]

1: **input:** K, $\mathbf{M}_{d\times(N_a+N_c)}$ and $\mathbf{y}$
2: **initialisation:** $s = \emptyset$, $k = 0$, $\mathbf{x} = \mathbf{0}$ and $\mathbf{r}_0 = \mathbf{y}$
3: **while** $k < K$ & $\max(\mathbf{M}^T\mathbf{r}_k) > 0$ **do**
4: $(\zeta, \iota) \leftarrow \max(\mathbf{M}^T\mathbf{r}_k)$
5: $s \leftarrow FS(s, \iota)$
6: $\mathbf{x}_s \leftarrow \arg\min_{\boldsymbol{\theta}>0} \|\mathbf{y} - \mathbf{M}_s\boldsymbol{\theta}\|_2$
7: $\mathbf{r}_{k+1} \leftarrow \mathbf{y} - \mathbf{M}_s\mathbf{x}_s$
8: $k \leftarrow k + 1$
9: **end while**
10: **output:** $\mathbf{x}|_s \leftarrow \mathbf{x}_s$

Figure 1: Non-Negative Orthogonal Matching Pursuit: the selection step FS determines the algorithm to be either canonical or two-library version of the algorithm.

The sparse decomposition algorithm is also known as a non-negative orthogonal matching pursuit, which is presented in Figure 1.[30] In simple terms, we select a new column of M, i.e. a spectrum in this case, at each step and find the best set of positive weights to match the library spectra to the experimental spectrum. We only have one API and one CCF in each sample and it is therefore reasonable to select only one from each of the API and CCF libraries. This avoids inaccuracy due to the measurement noise and potentially high similarity between spectra within two libraries. Since there is a chance that we will choose the first API and/or CCF incorrectly, we can thus set K, *i.e.* the number of iterations to more than two, then have a backward cancellation step, if needed. In this case, the proposed two-library NNOMP (TNNOMP) needs a controlling step in the forward selection function $FS(s, \iota)$. The mapping table of FS is presented in Table 1. In simple terms lines 1 to 3 describe the process of building the library and initialising the algorithm. Line 4 finds the most correlated spectrum in the library and line 5 implements the *forward selection* function $FS(s, \iota) = s \cup \iota$.[30] described above. Line 6 find the best coefficients to fit to the experimental spectrum and line 7 updates the residual spectrum which is used to calculate a novelty score (see Yaghoobi *et al.*[30] for more information and the fast implementation of NNOMP).

The residual error $\epsilon = \|y - \mathbf{M}_s\boldsymbol{\alpha}\|$ is the basis for the novelty score, as it is low when the spectra are physical mixtures, and high when spectral components corresponding to new phases are present. By normalising the spectrum under investigation $\mathbf{y}$, and the library to

Table 1: Forward selection mapping table for Two Library NNOMP.

If	Then
$s = \emptyset$ or $\lvert s\rvert \geq 2$	$FS(s, \iota) = s \cup \iota$
$\lvert s\rvert = 1, s \in [1, N_a]$	$FS(s, \iota) = s \cup \widetilde{\iota}$, where $(\widetilde{\zeta}, \widetilde{\iota}) \leftarrow \max(\widetilde{\mathbf{M}}^T \mathbf{r}_k)$
$\lvert s\rvert = 1, s \in [N_a + 1, N_a + N_c]$	$FS(s, \iota) = s \cup \widehat{\iota}$, where $(\widehat{\zeta}, \widehat{\iota}) \leftarrow \max(\widehat{\mathbf{M}}^T \mathbf{r}_k)$

have unit column norms, we have $0 \leq \epsilon \leq 1$. By taking $100 \times \epsilon$ as the novelty score we get a number between zero, for a perfect physical mixture, and 100 for definite new phase (which could be a cocrystal or a new solid form of one of the starting materials).

The proposed screening algorithm TNNOMP allows us to input multiple spectra for each API and CCF, to compensate for spectral variabilities resulting from the measurement (*i.e.* solid *vs* solution). Multiple reasons cause the spectral variability in Raman spectroscopy including, different Fluorescent background, instrument related artefacts and laboratory measurement settings, which can potentially affect the accuracy of spectral decomposition and novelty detection methods, *i.e.* the variability can be interpreted as an indication of a new phase. Including multiple spectral measurements, whenever available, does not essentially lead to the selection of variants of the same API or CCF in the algorithm, as the algorithm selects one from each set, in the first two steps. As a result, we use multiple versions of each API and CCF in the spectral library to reduce the novelty score for spectral mixtures, while not significantly changing it for the new phase cocrystals.

The classifier here is a simple binary classifier with one learning parameter, *i.e.* the threshold ρ. The optimal value for ρ can, in principle, be found by learning using real physical mixtures and new phase cocrystals. However, in the absence of a well characterised training data set, we propose the use of synthetic physical mixtures and setting ρ based on a fixed false alarm rate. We practically observe that such a ρ can be close to optimal value based on limited real data observation.

Results and discussion

Spectral separation and automatic screening

Building a library of API and CCF spectra. Raman spectra of all of the APIs and CCFs were recorded using the same instrumental settings, allowing us to build a library against which to test spectra of the studied systems (both physical mixtures and cocrystals) using the TNNOMP method described in the experimental section. All spectra were baseline corrected with airPLS,[31] then normalised. We included multiple versions of API and CCFs in the library, when available, to mitigate the effect of measurement conditions, *e.g.* powder from ball milling or solid obtained following solvent evaporation. This was found to improve the robustness of the algorithm to slight spectral variabilities and made the classes more separable.

Testing the algorithm on well characterised exemplars of physical mixtures and cocrystals. Our first objective was to test whether the TNNOMP method could be applied to well characterised samples which are representative of either physical mixtures or new cocrystals.

The first of these cases is a physical mixture of resveratrol and theophylline. The nature of the physical mixture was confirmed using XRPD measurements (Figure S1 in the Supplementary Information). Figure 2.a shows the spectrum of the physical mixture and the modelled spectrum based on the superposition of the spectrum of the API (Figure 2.b) and the spectrum of the CCF (Figure 2.c). The contributions of the API and CCF to the modelled spectrum are 98 percent and 2 percent respectively. Because the two spectra in 2.a are very closely matched (as expected for a physical mixture) they generate a low novelty score (7.2558).

The strength of the TNNOMP method is clear when applying it to the spectrum of a system which is known to form a new cocrystal. In this case the API is nalidixic acid and the CCF is hydroquinone, which formed a cocrystal under ultrasonication. Again, in this case

the new cocrystal was confirmed using XRPD (Figure S2 in the Supplementary Information). While spectral features of the API and coformer (Figures 3.b and 3.c) are clearly represented in Figure 3.a, there are significant new peaks (such as those between 1250 and 1300 cm-1). The presence of such peaks increases the energy of the residual spectrum and leads to a novelty score of 61.110 which is significantly higher than those seen for the well characterised physical mixture.

Testing the algorithm on a range of experimental spectra Since our primary objective in carrying out this work is to develop an automated method of screening new forms which is faster than manual processing (but not necessarily more accurate), we used our TNNOMP method to generate novelty scores for a further 29 experiments (making 31 in total) and compared the accuracy of classification with that found when the spectra were manually analysed by a skilled Raman spectroscopist. The various experiments are summarised in Tables 2 and 3 and the difficulty of manually classifying the spectra as either physical mixtures or cocrystals ranked as either easy, medium or hard. For all of these experiments, the classification was confirmed using XRPD.

The automated TNNOMP processing of the spectra was performed in a minute compared to manual processing which took several hours. The output of the TNNOMP method was a range of novelty scores from 5.18 to 68.80.

Determining a suitable threshold level In order to automatically classify the spectra as either cocrystal or physical mixture we needed to impose a suitable threshold level ρ for the novelty score (higher than the threshold is a cocrystal, lower than the threshold is a physical mixture). To avoid missing difficult-to-detect cocrystals, we estimated that we could accept a false positive rate of 2%. We calculated ρ in the absence of a large training set of physical mixtures, by using synthetically generated mixtures (randomly combining an API and a coformer from the library, with different contribution weights, *i.e.* positive weights between 0.1 and 1). We randomly generated 1000 mixtures and sorted the novelty scores in a descending order, Figure 5. We then empirically found the novelty score threshold,

corresponding to the 20^{th} trial, *i.e.* $2\% \times 1000$, which is approximately $\rho = 25$. We chose this ρ to classify the spectra from the rest of the experiments.

The identity of the 19 investigated cocrystals and 12 physical mixtures are shown in Tables **??** and **??** in the Supplementary Information, respectively. The spectra for all of these cocrystals and physical mixtures can be found in Figures **??**-**??**. The novelty scores for these systems are shown in Figure 4. The average novelty score and a dotted horizontal line of 25 (indicating the threshold) are shown in each panel. Looking at Figure 4 the first observation worth remarking on is that there is no overlap between the two classes (*i.e.* the lowest scoring cocrystals have a higher novelty score than the highest scoring physical mixtures). Furthermore, as a consequence of the separation, by using the ρ of 25 we can achieve a perfect classification of cocrystals and physical mixtures.

While it is not surprising that the TNNOMP method can quickly classify physical mixtures or cocrystals that a skilled spectroscopist would characterise as "easy", the strength of TNNOMP can clearly be seen when examining the spectral profiles that the algorithm correctly identifies as either cocrystals or physical mixtures and which a trained spectroscopist classifies as "hard" or "medium". Looking at cocrystal 3 in Figure 4, which has a novelty score of 27.9 and shown in Figure 6 and is rated as "hard" the subtle shifts in peak positions around 600 cm^{-1} and 1200 cm^{-1}, are particularly difficult to detect by eye in the unprocessed spectrum (Figure S3 in the Supplementary Information). Comparing this with physical mixture 5 in Figure 4, which scored 23.7 and was rated "medium", the difference by eye is challenging to discern (Figures S8 and S4 in the Supplementary Information). The benefit of the automated process is therefore the quantitative basis for the novelty score and the ability to set a threshold that allows a clear classification.

Conclusion

In summary we have demonstrated the use of an automated algorithm for screening large numbers of potential pharmaceutical cocrystals. The value of this algorithm is in its accu-

Table 2: Nalidixic acid API experiments: coformer (second colum), experimental condition (third column), expected form (fourth column), and visual detection difficulty by an experienced operator (last column).

	Coformer	Experimental conditions	Expected result	Detection
1	Propyl gallate	Ultrasonication	Cocrystal	Easy
2	Propyl gallate	Physical mixture	Physical mixture	Easy
3	Propyl gallate	LAG	Cocrystal	Easy
4	Propyl gallate	NG	Cocrystal	Easy
5	t-butyl hydroquinone	Ultrasonication	Cocrystal	Easy
6	t-butyl hydroquinone	LAG	Cocrystal	Easy
7	t-butyl hydroquinone	NG	Cocrystal	Easy
8	Hydroquinone	Ultrasonication	Cocrystal	Easy
9	Hydroquinone	Physical mixture	Physical mixture	Easy
10	Hydroquinone	LAG	Cocrystal	Easy
11	Hydroquinone	NG	Cocrystal	Easy
12	Resorcinol	NG	Cocrystal	Easy
13	Orcinol	NG	Cocrystal	Hard
14	Phloroglucinol	Physical mixture	Physical mixture	Easy
15	Phloroglucinol	LAG	Cocrystal	Medium

Table 3: Resveratrol API experiments: coformer (second colum), experimental condition (third column), expected form (fourth column), and visual detection difficulty by an experienced operator (last column).

	Coformer	Experimental conditions	Expected result	Detection
1	$4,4'$-bipyridine	LAG	Cocystal	Easy
2	$4,4'$-bipyridine	Physical mixture	Physical mixture	Easy
3	$4,4'$-bipyridine	Ultrasonication	Cocrystal	Easy
4	Phenazine	LAG	Cocrystal	Hard
5	Phenazine	Physical mixture	Physical mixture	Easy
6	Phenazine	Ultrasonication	Cocrystal	Medium
7	Methenamine	Physical mixture	Physical mixture	Easy
8	Methenamine	Ultrasonication	Cocrystal	Easy
9	Piperazine	Ultrasonication	Cocrystal	Hard
10	4-Dimethylaminopyridine	Ultrasonication	Cocrystal	Easy
11	Theophylline	LAG	Physical mixture	Easy
12	Theophylline	Physical mixture	Physical mixture	Easy
13	Carbamazepine	LAG	Physical mixture	Medium
14	Carbamazepine	Physical mixture	Physical mixture	Medium
15	Trimesic acid	LAG	Physical mixture	Easy
16	Trimesic acid	Physical mixture	Physical mixture	Easy

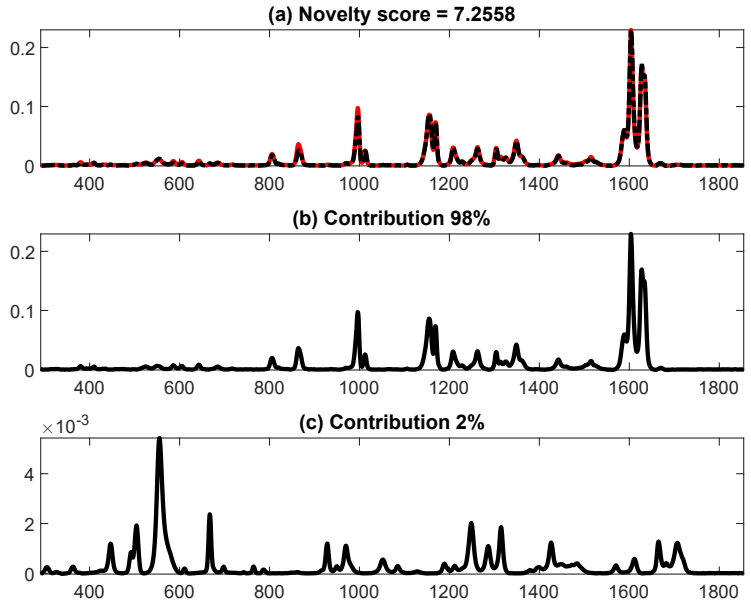


Figure 2: Resveratrol-Theophylline:LAG:Physical mixture. Acquired spectrum in black and reconstruction in dashed red (top), API (middle) and coformer (bottom).

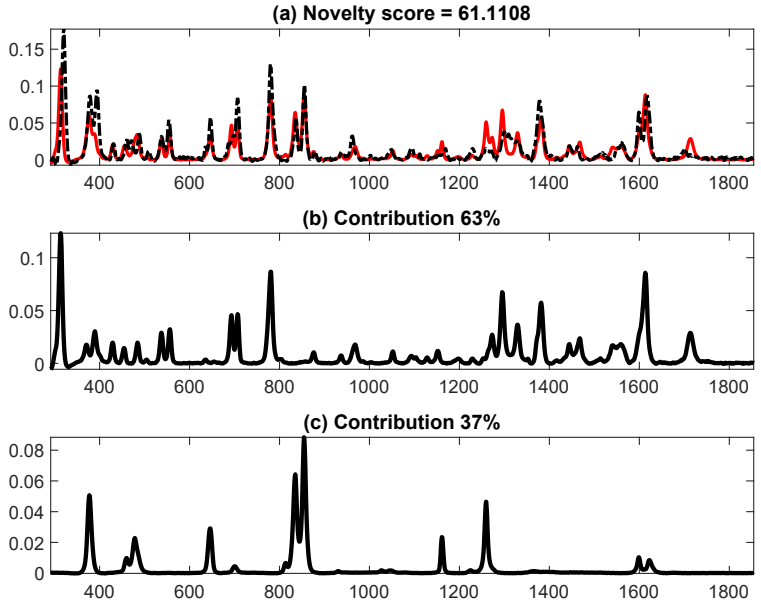


Figure 3: Nalidixic acid-Hydroquinone:NG:Cocrystal. Acquired spectrum in black and reconstruction in dashed red (top), API (middle) and coformer (bottom).

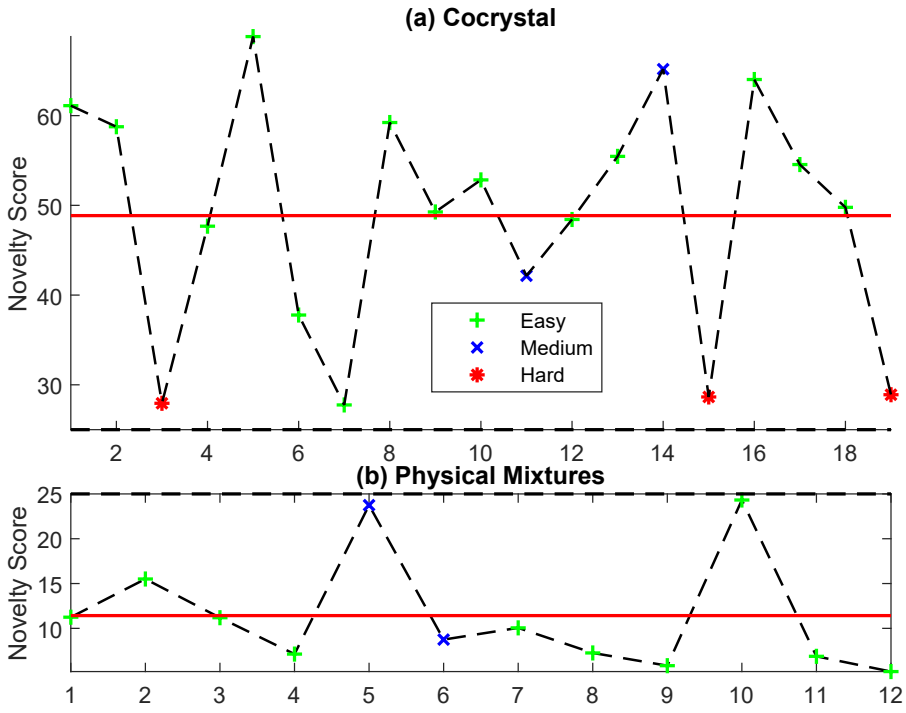


Figure 4: Novelty scores for cocrystals (top panel) and physical mixtures (bottom panel). The average has been indicated with a red line.

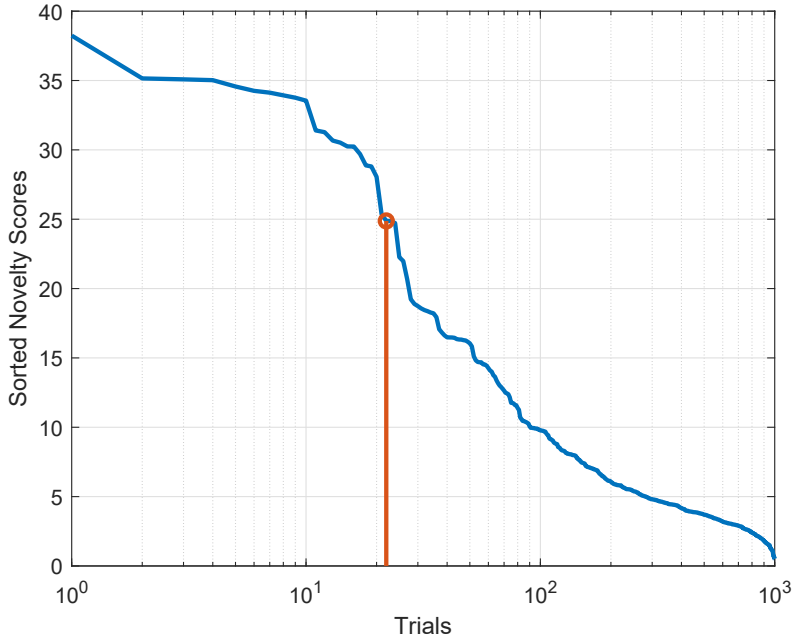


Figure 5: Sorted novelty scores for synthetically generated physical mixtures. The indicated threshold $\rho = 25$ corresponds to roughly 2% percent misdetection

.

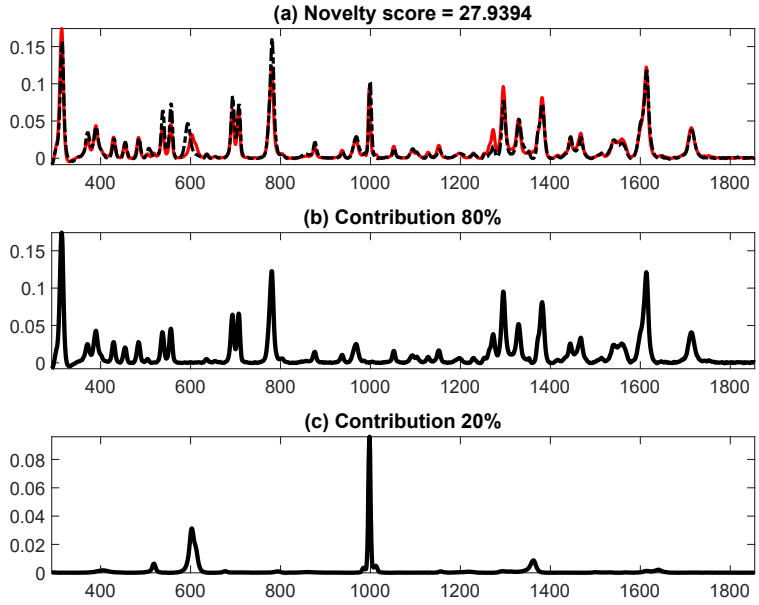


Figure 6: Nalidixic acid-Orcinol:NG:Cocrystal. Acquired spectrum in black and reconstruction in dashed red (top), API (middle) and coformer (bottom).

racy, which is as good as a trained spectroscopist, and in the time that it takes to screen and analyse multiple spectra (a time saving of several hours based on a relatively small data set). Moreover, for large data sets of 1000's of spectra normally seen during early solid state screening programs, this algorithm offers orders of magnitude time savings from potentially several days to a few minutes. Another advantage is that the algorithm would allow initial screening and data analysis to be carried out by non-Raman experts. The automated process not only carries out signal processing tasks such as background subtraction, but also compares the spectrum of an unknown (either cocrystal or physical mixture) to a library of its potential constituents and builds the best possible model of the spectrum based on the predicted constituents. By measuring the difference between the model and the experimental spectra, a novelty score can be calculated. We have demonstrated that the novelty score is a basis for classifying the experimental spectrum as belonging to either a cocrystal or physical mixture. We anticipate that the same methodology developed here for cocrystal can be applied to analysis of other molecular adducts such as solvates, salts and even polymorphic

screens, which is left for the future study.

Acknowledgement

MY acknowledges EPSRC IAA PIII077 support in enabling a visit to AstraZeneca, in February 2020.

Supplementary Information

Supplementary information of this paper includes: a) complete list of physical mixtures and cocrystals, b) XRPDs and Raman overlay of some samples, c) complimentary Set of Experimental Results.

References

(1) Aitipamula, S.; Banerjee, R.; Bansal, A. K.; Biradha, K.; Cheney, M. L.; Choudhury, A. R.; Desiraju, G. R.; Dikundwar, A. G.; Dubey, R.; Duggirala, N., et al. Polymorphs, salts, and cocrystals: what's in a name? Crystal growth & design **2012**, 12, 2147–2152.

(2) Arora, K.; Zaworotko, M. Polymorphism in pharmaceutical Solids; 2009.

(3) Karimi-Jafari, M.; Padrela, L.; Walker, G. M.; Croker, D. M. Creating cocrystals: a review of pharmaceutical cocrystal preparation routes and applications. Crystal Growth & Design **2018**, 18, 6370–6387.

(4) Schultheiss, N.; Newman, A. Pharmaceutical cocrystals and their physicochemical properties. Crystal growth and design **2009**, 9, 2950–2967.

(5) Shan, N.; Zaworotko, M. J. The role of cocrystals in pharmaceutical science. Drug discovery today **2008**, 13, 440–446.

(6) European Medicines Agency, Reflection paper on the use of cocrystals of active substances in medicinal products. **2015**.

(7) Childs, S. L.; Stahly, G. P.; Park, A. The salt- cocrystal continuum: the influence of crystal structure on ionization state. Molecular pharmaceutics **2007**, 4, 323–338.

(8) FDA, USA, Regulatory Classification of Pharmaceutical Co-Crystals Guidance for Industry. **2018**.

(9) Kavanagh, O. N.; Croker, D. M.; Walker, G. M.; Zaworotko, M. J. Pharmaceutical cocrystals: from serendipity to design to application. Drug Discovery Today **2019**, 24, 796–804.

(10) Aher, S.; Dhumal, R.; Mahadik, K.; Paradkar, A.; York, P. Ultrasound assisted cocrystallization from solution (USSC) containing a non-congruently soluble cocrystal component pair: Caffeine/maleic acid. European Journal of pharmaceutical sciences **2010**, 41, 597–602.

(11) Ervasti, T.; Ketolainen, J.; AAltonen, J. Spontaneous Formation of Theophylline–Nicotinamide Cocrystals. Scientia Pharmaceutica **2010**, 78, 622.

(12) Bysouth, S. R.; Bis, J. A.; Igo, D. Cocrystallization via planetary milling: enhancing throughput of solid-state screening methods. International journal of pharmaceutics **2011**, 411, 169–171.

(13) Luu, V.; Jona, J.; Stanton, M. K.; Peterson, M. L.; Morrison, H. G.; Nagapudi, K.; Tan, H. High-throughput 96-well solvent mediated sonic blending synthesis and on-plate solid/solution stability characterization of pharmaceutical cocrystals. International journal of pharmaceutics **2013**, 441, 356–364.

(14) Nagapudi, K.; Umanzor, E. Y.; Masui, C. High-throughput screening and scale-up

of cocrystals using resonant acoustic mixing. International Journal of Pharmaceutics **2017**, 521, 337–345.

(15) Kojima, T.; Tsutsumi, S.; Yamamoto, K.; Ikeda, Y.; Moriwaki, T. High-throughput cocrystal slurry screening by use of in situ Raman microscopy and multi-well plate. International journal of pharmaceutics **2010**, 399, 52–59.

(16) Garbacz, P.; Wesolowski, M. Benzodiazepines co-crystals screening using FTIR and Raman spectroscopy supported by differential scanning calorimetry. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **2020**, 118242.

(17) Rodrigues, M.; Lopes, J.; Guedes, A.; Sarraguça, J.; Sarraguça, M. Considerations on high-throughput cocrystals screening by ultrasound assisted cocrystallization and vibrational spectroscopy. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **2020**, 229, 117876.

(18) Huang, Y.; Zhou, L.; Yang, W.; Li, Y.; Yang, Y.; Zhang, Z.; Wang, C.; Zhang, X.; Yin, Q. Preparation of Theophylline-Benzoic Acid Cocrystal and On-Line Monitoring of Cocrystallization Process in Solution by Raman Spectroscopy. Crystals **2019**, 9, 329.

(19) Strachan, C. J.; Rades, T.; Gordon, K. C.; Rantanen, J. Raman spectroscopy for quantitative analysis of pharmaceutical solids. Journal of pharmacy and pharmacology **2007**, 59, 179–192.

(20) Anderson, T. W. An introduction to multivariate statistical analysis; 1962.

(21) Jolliffe, I. T.; Cadima, J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **2016**, 374, 20150202.

(22) Comon, P. Independent component analysis, a new concept? Signal processing **1994**, 36, 287–314.

(23) Chen, S. S.; Donoho, D. L.; Saunders, M. A. Atomic decomposition by basis pursuit. SIAM review **2001**, 43, 129–159.

(24) Yaghoobi, M.; Wu, D.; Clewes, R. J.; Davies, M. E. Fast sparse Raman spectral unmixing for chemical fingerprinting and quantification. Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XII. **2016**; p 99950E.

(25) Elbagerma, M. A.; Edwards, H. G.; Munshi, T.; Hargreaves, M. D.; Matousek, P.; Scowen, I. J. Characterization of new cocrystals by Raman spectroscopy, powder X-ray diffraction, differential scanning calorimetry, and transmission Raman spectroscopy. Crystal growth & design **2010**, 10, 2360–2371.

(26) Grecu, T.; Adams, H.; Hunter, C. A.; McCabe, J. F.; Portell, A.; Prohens, R. Virtual screening identifies new cocrystals of nalidixic acid. Crystal growth & design **2014**, 14, 1749–1755.

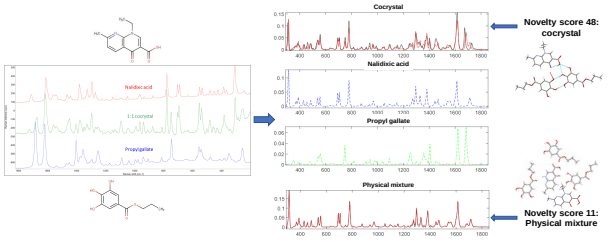
(27) Mehta, B. K.; Singh, S. S.; Chaturvedi, S.; Wahajuddin, M.; Thakur, T. S. Rational Coformer Selection and the Development of New Crystalline Multicomponent Forms of Resveratrol with Enhanced Water Solubility. Crystal Growth & Design **2018**, 18, 1581–1592.

(28) Natarajan, B. K. Sparse approximate solutions to linear systems. SIAM journal on computing **1995**, 24, 227–234.

(29) Elad, M. Sparse and redundant representations: from theory to applications in signal and image processing; Springer Science & Business Media, **2010**.

(30) Yaghoobi, M.; Wu, D.; Davies, M. E. Fast non-negative orthogonal matching pursuit. IEEE Signal Processing Letters **2015**, 22, 1229–1233.

(31) Zhang, Z.-M.; Chen, S.; Liang, Y.-Z. Baseline correction using adaptive iteratively reweighted penalized least squares. Analyst **2010**, 135, 1138–1146.


Nalidixic acid
1:1 cocrystal
Propylgallate
Cocrystal
Nalidixic acid
Propyl gallate
Physical mixture
Novelty score 48: cocrystal
Novelty score 11: Physical mixture