Chemometric-Assisted Cocrystallization: Supervised Pattern Recognition for Predicting the Formation of New Functional Cocrystals

Fabio Fornari [a], Fabio Montisci [a], Federica Bianchi [a,b,*], Marina Cocchi [c], Claudia Carraro [a], Francesca Cavaliere [d], Pietro Cozzini [d], Francesca Peccati [e], Paolo P. Mazzeo [a,f], Nicolò Riboni [a], Maria Careri [a,g], Alessia Bacchi [a,f]

[a] *University of Parma, Department of Chemistry, Life Sciences and Environmental Sustainability, Parco Area delle Scienze 17/A, 43124, Parma, Italy*

[b] *University of Parma, Interdepartmental Center for Packaging (CIPACK), Parco Area delle Scienze, 43124, Parma, Italy*

[c] *University of Modena and Reggio Emilia, Department of Chemical and Geological Sciences, Via Giuseppe Campi 103, 41125, Modena, Italy*

[d] *University of Parma, Department of Food and Drug, Parco Area delle Scienze 17/A, 43124, Parma, Italy*

[e] *Basque Research and Technology Alliance (BRTA), Center for Cooperative Research in Biosciences (CIC bioGUNE), Bizkaia Technology Park 801A, 48160, Derio, Spain*

[f] *University of Parma, Biopharmanet-TEC, Parco Area delle Scienze 27/A, 43124, Parma, Italy*

[g] *University of Parma, Interdepartmental Center on Safety, Technologies, and Agri-Food Innovation (SITEIA.PARMA), Parco Area delle Scienze, 43124, Parma, Italy*

* Author to whom correspondence should be addressed:

e-mail address: federica.bianchi@unipr.it; University of Parma, Department of Chemistry, Life Sciences and Environmental Sustainability, Parco Area delle Scienze 17/A, 43124, Parma, Italy

Abstract

Owing to the antimicrobial and insecticide properties, the use of natural compounds like essential oils and their active components has proven to be an effective alternative to synthetic chemicals in different fields ranging from drug delivery to agriculture and from nutrition to food preservation. Their limited application due to the high volatility and scarce water solubility can be expanded by using crystal engineering approaches to tune some properties of the active molecule by combining it with a suitable partner molecule (coformer). However, the selection of coformers and the experimental effort required for discovering cocrystals are the bottleneck of cocrystal engineering. This study explores the use of chemometrics to aid the discovery of cocrystals of active ingredients suitable for various applications. Partial Least Squares–Discriminant Analysis is used to discern cocrystals from binary mixtures based on the molecular features of the coformers. For the first time, by including failed cocrystallization data and considering a variety of chemically diverse compounds, the proposed method resulted in a successful prediction rate of ~~83~~85% for the test set in the model validation phase and of ~~62~~74% for the external test set.

Keywords

cocrystal, crystal engineering, chemoinformatics, chemometrics, partial least square discriminant analysis, Quantitative Structure–Property Relationship

List of abbreviations (sorted alphabetically)

ACC%: Accuracy

BM: Binary Mixture

CC: Cocrystal

CCDC: Cambridge Crystallographic Data Centre

CSD: Cambridge Structural Database

EO: Essential Oil

FDA: Food and Drug Administration

GRAS: Generally Recognized As Safe

LV: Latent Variable

MEP: Molecular Electrostatic Potential

~~NER%: Non-Error classification Rate~~

PC: Principal Component

PCA: Principal Component Analysis

PLS–DA: Partial Least Squares–Discriminant Analysis

PXRD: Powder X-Ray Diffraction

QSPR: Quantitative Structure–Property Relationship

SEN%: Sensitivity

VIP: Variable Importance in Projection.

1. Introduction

In the last few decades, the use of agrochemicals and food preservatives has grown exponentially as a direct consequence of the rapid increase of the world population [1,2]. Owing to their potential adverse effects on both human health and environment [2–4], alternative strategies based on the use of more sustainable chemicals have been proposed to support the food system. Being able to exert antimicrobial, insecticidal, and antioxidant properties [5,6], essential oils (EOs) and their active components have been used as green substitutes of synthetic chemicals to extend the shelf-life of foodstuff and in pests control [7–9]. These compounds are Generally Recognized As Safe (GRAS) by the Food and Drug Administration (FDA) [10], however, despite their appealing properties, their use is limited by their high volatility and poor stability [7,9,11,12].

In fact, physicochemical properties of materials play a key role in determining whether a chemical is suitable for a specific purpose, thus strongly affecting its field of application. Scientists have always desired to obtain materials with target properties, and crystal engineering is one of the most interesting approaches to synthesize a great variety of crystalline materials for applications in various fields, ranging from pharmaceuticals to agrochemicals, and from nutraceuticals to cosmetics [13–16]. The basic idea of crystal engineering is related to the possibility of controlling the crystal structure of molecules and, therefore, the properties of the resulting solids. Polymorphism, vitrification and cocrystallization are some of the available strategies to modify the intrinsic properties of molecules without the need of synthetic modifications [17–21].

Cocrystals are multicomponent crystalline solid materials in which the constituents (i.e., coformers) are bound in a well-defined stoichiometric ratio [22,23] *via* non-covalent interactions (e.g., hydrogen bonds, halogen bonds, π–π stacking) within the same crystal structure. Cocrystallization allows for the combination of the desired molecule of interest with properly selected partner molecules, paving the way to an array of potential materials with enhanced properties [19,24,25]. Within this frame of reference, cocrystals based on the active components of EOs have been proposed as active ingredients for food packaging, agrochemical and pharmaceutical applications [7,19,26,27].

Despite the great advantages offered by cocrystallization, the proper degree of complementarity between the two partner molecules required to obtain crystalline materials with the desired properties is not easy to assess [28–31]. In this context, the selection of coformers and the great effort required for both the systematic experimental screening and careful characterization of the products derived from the combination of all the possible coformer pairs represent the major bottleneck of cocrystal engineering. Computational techniques represent a powerful tool to reduce the experimental effort required for the discovery of new cocrystals, enabling to evaluate beforehand whether a cocrystal can be obtained starting from pre-selected coformers.

These *in silico* strategies can be based on the calculation of a variety of parameters useful to predict the formation of a cocrystal, such as lattice energy [32], solubility [33], hydrogen bond propensity along with the quantitation of molecular interaction energy [29,30], and molecular complementarity [34].

Despite the massive efforts spent to develop a method to predict cocrystal formation, at present none of the proposed strategies has proven to be both totally reliable and easy to apply.

Chemometrics could play a pivotal role in cocrystal discovery: up to now only few Machine Learning methods have been proposed in predicting cocrystal formation, enabling the screening of new cocrystals once a supervised model is properly trained and validated. In the study proposed by Devogelaer et al., information of successful cocrystallization experiments was directly taken from the Cambridge Structural Database (CSD) [35] and Artificial Neural Networks (ANN) were used to predict the formation of new cocrystals [36]. Similarly, Wang et al. relied on a consensus method based on multiple Random Forest algorithms, in which the successful cocrystallization dataset was integrated with randomly generated failed cocrystallization data [37]. These approaches are reported in the literature as network-based methods. Additional studies were based on the use of successful and unsuccessful cocrystallization datasets obtained from experimentation, literature, and/or the CSD for screening specific classes of coformers. Within this framework, Przybyłek et al. used Multivariate Adaptive Regression Splines to predict the formation of dicarboxylic and phenolic acid-based cocrystals [38,39], whereas Wicker et al. focused on variously substituted benzoic acids and benzamides using a Support Vector Machine algorithm [24]. Vriza et al. used an ensemble one-class classification method to aid the discovery of π–π cocrystals, thus giving a great contribution in enriching one of the most under-represented classes of cocrystals in the CSD [40]. Most recently, Mswahili et al. developed a cocrystal screening method based on ANN by using both successful and unsuccessful experimental cocrystallization data retrieved from the literature and a plethora of molecular descriptors calculated using Mordred [41,42].

In the frame of a research activity dealing with the synthesis of new functional cocrystals based on the active constituents of EOs and other GRAS molecules to broaden their applicability in the industrial field [7,19], we propose a chemometric approach to aid the discovery of new cocrystalline materials.

For the first time, a training set based on the results of failed (binary mixtures, BM) and successful (cocrystal, CC) cocrystallization experiments was used for the computation of a Quantitative Structure–Property Relationship-like (QSPR) model based on Partial Least Squares–Discriminant Analysis (PLS–DA), after preliminary exploratory analysis by Principal Component Analysis (PCA). The PLS-DA approach, with respect to network-based methods offers the advantages of having

only one parameter to optimize, i.e., the model dimensionality, and the direct interpretation of the importance of descriptors in classification, while highlighting their interplay (by inspection of weights and loadings plots).

The effectiveness of the study relies on the use of compounds belonging to different chemical classes and a reduced number of 1D, 2D, and 3D molecular descriptors of various nature (e.g., constitutional, geometric, physical, topological, and surface area-based descriptors) [24,38,39,43], enabling the high-throughput screening of novel cocrystalline materials and offering maximum flexibility and effectiveness at a minimum computational and experimental cost.

2. Experimental Procedures

2.1. Mechanochemical protocol and class assignation

All the molecules in the dataset were chosen among the list of GRAS molecules drawn up by the FDA [10]. Selected pair of molecules among the chosen ones were assigned either to the CC class or to the BM class. Pairs of molecules in the dataset for which a cocrystal structure was already described in literature were individuated in the Cambridge Structural Database (CSD) [35] with the Cambridge Crystallographic Data Centre (CCDC) software ConQuest [44] and visualized with Mercury [45]. They are reported in Section 3 of the Supplementary Material, together with their unique CSD refcode and the reference to the original publications.

Cocrystallization for all the pairs with no known structure in literature was instead attempted with the following mechanochemical protocol. All the reagents employed were commercially available and used as such in all the experiments. Equimolar amounts of each reagent were directly mixed in an agate mortar and subjected to manual grinding for 10-15 minutes, without using any solvent. The resulting powder samples were collected in closed vials. Assignation to CC or BM classes was performed by comparing the Powder X-ray Diffraction (PXRD) pattern of the ground sample with those of the pure reagents. Possible occurrence of polymorphic transitions for the reagents was excluded by comparing the experimental PXRD data after milling with the calculated pattern of all the known crystalline forms of the reagents. The occurrence of new peaks, unexplained by the presence of unreacted reagents, was taken as indication that cocrystallization had occurred and the sample was assigned to the CC class. In case no additional peaks appeared in the PXRD pattern the sample was instead classified as a BM.

2.2. Powder X-ray diffraction

Typically, PXRD data were collected on a Rigaku Smartlab XE diffractometer in θ-θ Bragg-Brentano geometry with Cu Kα radiation. The samples were placed on glass supports and exposed to radiation ($1.5° \leq 2\theta \leq 50°$) at a scan rate of 10°/min. The diffracted beam was collected on a 2D Hypix 3000 solid state detector. 5° radiant soller were used as a compromise for high flux and moderate peak asymmetry at low angles. Beam stopper and anti-scatterer air component were used to mitigate the profile at low angle. In some rare cases, the data were collected on a Thermo Fisher Scientific ARL X'TRA diffractometer in θ-θ Bragg-Brentano geometry with Cu Kα radiation ($3° \leq 2\theta \leq 30°$ at a scan rate of 5°/min, or $3° \leq 2\theta \leq 40°$ at a scan rate of 0.3°/min).

3. Computational Methods

3.1. Molecular descriptors calculation

For each molecule 31 molecular descriptors were calculated (Table S1). A theoretical background for the less known descriptors is given in Section 4 of the Supplementary Material.

The molecular weight, the number of atoms, the number of bonds, the number of hydrogen bond donor sites and the number of hydrogen bond acceptor sites were calculated with FLAP software (Fingerprint for Ligand and Protein) [46] at pH 7.0, using the 3D structures of all molecules in SDF format as input (downloaded from the PubChem database). The number of rotatable bonds, the number of rings, the hydrophobicity (accounted as the number of hydrophobic centers), the logP (logarithm of octanol/water partition coefficient), the molecular volume, the total molecular dipole moment (based on point charge distribution in the molecule), and its components along the axes (using the principal axes of the molecular graph) were then calculated for the same structures using Sybyl 8.1 [47] (www.tripos.com) and taking in consideration the protonation state of molecules. The same software was also used to estimate the strain energy of the molecule without performing any geometry optimization. This energy term relies on an electrostatic calculation from atomic charges using the internal Tripos force field [48]. For the estimation of molecular volume and dipole moment, a specific SPL script was employed. The calculated volume is enclosed in a water-accessible surface computed at a repulsive interaction energy of 0.20 kcal/mol with a water probe. A custom Python script was used to automatically calculate the Solvent Accessible Surface Area (SASA) in PyMol 2.0 [49], with the dot density parameter set to 4. The number of heteroatoms, the number of valence electrons, and the indexes ${}^0\chi$, ${}^0\chi^n$, ${}^0\chi^v$, α_{HK}, ${}^1\kappa_\alpha$, LabuteASA, SMR_VSA, PEOE_VSA, and TPSA were obtained running a Python 3.7 code with

the open-source cheminformatics toolkit RDKit Q4 2013 [50]. The average isotropic polarizability α_{iso}, the polarizability anisotropy $\Delta\alpha$, and the Molecular Electrostatic Potential (MEP) were calculated with Gaussian 16 [51] following the *in vacuo* Density-Functional Theory optimization of all the molecules, employing the hybrid functional B3LYP and the People double-z basis set 6-31+g(d,p).

Postprocessing of the MEP to extract critical points at a given electron density isosurface was performed with a custom Python 3.6.1 script on a three-dimensional map (cube format) with a sampling density of 6 points/Bohr along the three directions. The MEP was analyzed at an electron density isosurface of 0.002 a.u. with a tolerance of 0.001 a.u., meaning that only MEP values corresponding to regions of space with electron density in the 0.001–0.003 a.u. range were considered. A first set of critical points was identified comparing MEP values of each cube point with those of its 6 nearest neighbors. A point was considered a local minimum if the number of nearest neighbors with higher MEP was greater or equal to a given integer (4). Likewise, a point was considered a local maximum if the number of nearest neighbors with lower MEP was greater or equal to the same integer. This first step yielded a large number of candidate critical points encompassing a wide range of MEP values. Since our focus was on identifying the regions of the molecules likely to be involved in strong hydrogen bonds within the cocrystal, in a second step this first set of points was filtered based on the MEP values of the global minimum and maximum. This was done as follows: for each local minimum (maximum), the ratio between its MEP value and that of the global minimum (maximum) was computed, and the point was kept only if the ratio exceeded a given threshold (0.1). In this way, only points corresponding to shallow critical points were discarded. This step allowed to identify the MEP isosurface regions corresponding to hydrogen bond donors and acceptors. However, due to the rugged character of the MEP map, multiple critical points of the same type could appear in close proximity. To univocally map a given region of the isosurface to a MEP value, critical points close to each other (below a distance threshold of 1.0 Bohr) were merged iteratively, keeping only the lower MEP point for minima and higher MEP point for maxima. The algorithm then provided the final set of MEP critical points at the given electron density isosurface.

3.2. Data analysis

The entire data analysis was carried out in MATLAB R2019a environment (Mathworks, Natick, Massachusetts, USA) with the aid of the PLS_Toolbox 8.7.1 (Eigenvector Research Inc., Washington, USA) chemometric package.

~~was used to carry out preprocessing, Principal Component Analysis (PCA) and PLS-DA computation, and to split the original dataset into calibration and test set. The proper number of latent variables (LVs) to be retained was evaluated by running a homemade MATLAB routine.~~

3.2.1. Data preprocessing

Each sample was described by $m = 31$ variables (Table S1): the absolute value of the difference between ~~difference in absolute value of~~ the molecular descriptors of the two partner molecules was calculated, thus obtaining the predictor matrix $\boldsymbol{X}$ (181 × 31). The class membership was binary encoded (1: belonging to the class; 0: otherwise) in a dummy matrix $\boldsymbol{Y}$ (181 × 2) with each column representing one of the two modelled classes. The dataset was split in two subsets by using the ~~Kennard-Stone sampling~~Duplex algorithm [52]: ~~80~~70% of the data were used as calibration set, $\boldsymbol{X}_{\mathbf{cal}}$ (~~146~~ 127 × 31) and $\boldsymbol{Y}_{\mathbf{cal}}$ (~~146~~ 127 × 2), whereas the remaining ~~20~~30% were used as test set, $\boldsymbol{X}_{\mathbf{test}}$ (~~35~~ 54 × 31) and $\boldsymbol{Y}_{\mathbf{test}}$ (~~(35~~54 × 2). The calibration set and the test set are reported in Table S2 and Table S3, respectively.

Before carrying out both exploratory multivariate data analysis and the computation of the supervised model, the calibration matrix $\boldsymbol{X}_{\mathbf{cal}}$ was preprocessed column-wise by performing mean centering and scaling to unit variance. Mean centering was applied on the response matrix $\boldsymbol{Y}_{\mathbf{cal}}$ to ensure the stability of the model.

3.2.2. Exploratory multivariate data analysis

PCA [53–55] was carried out preliminarily on the calibration set $\boldsymbol{X}_{\mathbf{cal}}$ to assess the distribution of the samples and to check for potential data structures. Reduction of data dimensionality is carried out through the linear combination of the original variables in a set of orthogonal ones, i.e., Principal Components (PCs), which identify the direction of maximum variance. This is summarized in the decomposition equation:

$$\boldsymbol{X}_{\mathbf{cal}} = \boldsymbol{T}\boldsymbol{P}^{\mathbf{T}} + \boldsymbol{E}$$

where $\boldsymbol{T}$ and $\boldsymbol{P}$ represent, respectively, the coordinates of the samples projected in the reduced space, i.e., the scores, and the weights each original variable has on a given PC, i.e., the loadings. The deviations from the model are accounted in the error matrix $\boldsymbol{E}$.

3.2.3. Supervised pattern recognition

PLS–DA [56,57] was used to discriminate pairs of partner molecules whose combination forms CCs from the ones giving BMs. PLS–DA is based on PLS regression [58]. Briefly, this supervised technique decomposes the predictor matrix $\boldsymbol{X}_{\mathbf{cal}}$ and the dependent variables matrix $\boldsymbol{Y}_{\mathbf{cal}}$ in a PCA-like way and imposes inner linear relationships between the $\boldsymbol{X}_{\mathbf{cal}}$ and the $\boldsymbol{Y}_{\mathbf{cal}}$ scores as follows:

$$\boldsymbol{U} = \boldsymbol{bT}$$

where $\boldsymbol{T}$ and $\boldsymbol{U}$, are the $\boldsymbol{X}_{\mathbf{cal}}$ and the $\boldsymbol{Y}_{\mathbf{cal}}$ scores, respectively. This is accomplished by rotating the Latent Variable (LV) space of $\boldsymbol{X}_{\mathbf{cal}}$ through a weight matrix $\boldsymbol{W}$ in a way that maximizes the covariance between $\boldsymbol{T}$ and $\boldsymbol{U}$. The PLS regression model is summarized as:

$$\boldsymbol{Y}_{\mathbf{cal}} = \boldsymbol{X}_{\mathbf{cal}}\boldsymbol{B} + \boldsymbol{E}$$

where $\boldsymbol{E}$ is the error matrix and $\boldsymbol{B}$ is the pseudo-regression coefficient matrix expressed according to the following equation:

$$\boldsymbol{B} = \boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}\,\mathrm{diag}(\boldsymbol{b})\,\boldsymbol{Q}$$

where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are the $\boldsymbol{X}_{\mathbf{cal}}$ and the $\boldsymbol{Y}_{\mathbf{cal}}$ loadings, respectively.

In this case, the dependent variables in the $\boldsymbol{Y}_{\mathbf{cal}}$ matrix are defined as dummy variables, one for each modelled class, taking values of 1 if the sample belongs to the class and 0 otherwise. Current implementation of PLS–DA may differ on the basis of how the classification rule is defined. In this work, a pure discriminant rule (samples are assigned univocally to only one category) was applied, and thus a sample is assigned to the class for which the predicted response $\hat{\boldsymbol{y}}$ is the highest (i.e., $\hat{\boldsymbol{Y}}$ values are continuous and not dummy as they were codified).

The proper number of LVs was chosen according to the maximum ~~Non-Error-classification-Rate~~accuracy (~~NER~~ACC$_{\%}$; i.e., the percentage of samples correctly assigned to the respective class) in leave-more-out cross validation, adopting a Venetian blinds cancellation scheme with 10 splits (blind thickness: 1). This operation was carried out by running a custom MATLAB routine. The performance of the classification model was evaluated both on the calibration and the test sets in terms of ~~NER~~ACC$_{\%}$ as well as showing the confusion matrix. In addition, the sensitivity (SEN$_{\%}$, i.e., the percentage of samples within a class that were correctly assigned to their class) was calculated for both classes.

The importance of each predictor was estimated in terms of Variable Importance in Projection (VIP score) [56]. The VIP score of the jth variable in the $\boldsymbol{X}$ space is defined as the component-wise sum of its PLS weight w_j on the fth component multiplied by the fraction of variance of the $\boldsymbol{Y}$ explained by that component, according to the following equation:

$$VIP_j^2 = \frac{1}{SSY_{\text{tot}} F} \sum_{f=1}^{F} w_{jf}^2 SSY_f J$$

where J is the number of variables in the $\boldsymbol{X}$ space and F is the number of LVs that were retained. Since:

$$\sum_{j=1}^{J} VIP_j^2 = J$$

the proposed threshold for determining whether a variable could be considered important is set to 1..

Finally, the predictive capability of the model was evaluated on an external set of N = 58 binary combinations of partner molecules. An overview of the involved samples is reported in Table S4 along with their estimated $\hat{y}$ values.

The number of entries in the BM and CC classes for training, test, and external validation sets is reported in Table 1.

Table 1. Number of CC and BM samples in the calibration, test, and external validation set. The last column reports the total number N of samples *per* set.

	CC	**BM**	***N***
Calibration set	71	56	127
Test set	30	24	54
External validation set	31	27	58

4. Results and Discussion

The cocrystallization experiments were carried out mechanochemically by manual neat grinding of the two substances. This method was selected among many possible others due to its simplicity and promptness, allowing us to screen several molecular pairs in a standardized way [21,59]. The classification as BM or CC was assessed by PXRD patterns (available in Section 3 of the Supplementary Material). Cocrystals already present in the CSD were also included in our dataset.

4.1. Data analysis

Data handling prior to analysis can affect the way the model is trained, thus having consequences on its interpretation. Since samples are the result of the combination of two partner molecules, each one described by its own set of descriptors, the concatenation strategy, i.e., listing the descriptors of the first coformer followed by those of the second coformer, was

discarded due to the lack of commutation between the two sets of descriptors. In fact, the order in which the molecular descriptors are listed represents an *a priori* decision on which compound is acting as molecule of interest or partner molecule. In our case this would be sub-optimal since many of the molecules in our dataset could assume both roles. Therefore, in order to address the problem described above a commutative strategy capable of avoiding the production of indeterminate forms should be chosen. Considering that in the dataset many constitutional molecular descriptors were characterized by a few non-zero values, the calculation of both products and ratios between the descriptors was discarded since additional zero values or indeterminate forms could be generated. Also, the division is non-commutative.

In order to combine the two partner molecules without imposition on their role, the absolute value of the difference ~~difference in absolute value~~ between the molecular descriptors of the partner molecules was calculated and used to describe each binary combination [38,39], giving maximum flexibility to the model. The information achieved is still relevant and easily interpretable since it is related to the dissimilarity between the descriptors. In fact, ~~the differences in absolute value~~absolute value of the differences between descriptors for each case are the elements of the Manhattan distance [60], one of the possible indexes to account for the dissimilarity between cases in a multivariate way.

In the present study, the use of basic linear modelling methods was preferred with respect to non-linear modelling, such as ANN [61,62], to keep the calculations as simple as possible, and to ensure a certain degree of interpretability of the results. Furthermore, the number of samples is too limited to ensure proper tuning of the ANN hyperparameters.

For a fruitful discussion, samples and variables are reported in the text according to their identification number as follows: i) samples are written in plain text; ii) variables are underlined. The key is available in Tables S1–S3.

4.1.1. Exploratory multivariate data analysis

After data preprocessing, PCA was used in an exploratory way to assess the presence of potential data structures in the calibration set.

Four PCs were retained explaining 67% of the variance. As shown in the score plots (Figure 1), a mild ~~segregation~~ separation was present in the PC ~~3~~ 2 *vs.* PC ~~4~~ 3 score plot, with the groups separated by the bisecting line of the II and the ~~III~~ IV quadrant. In addition, most of the CC samples occupied the III quadrant of the PC 2 *vs.* PC 3 score plot and were, in general, less scattered than BM samples.

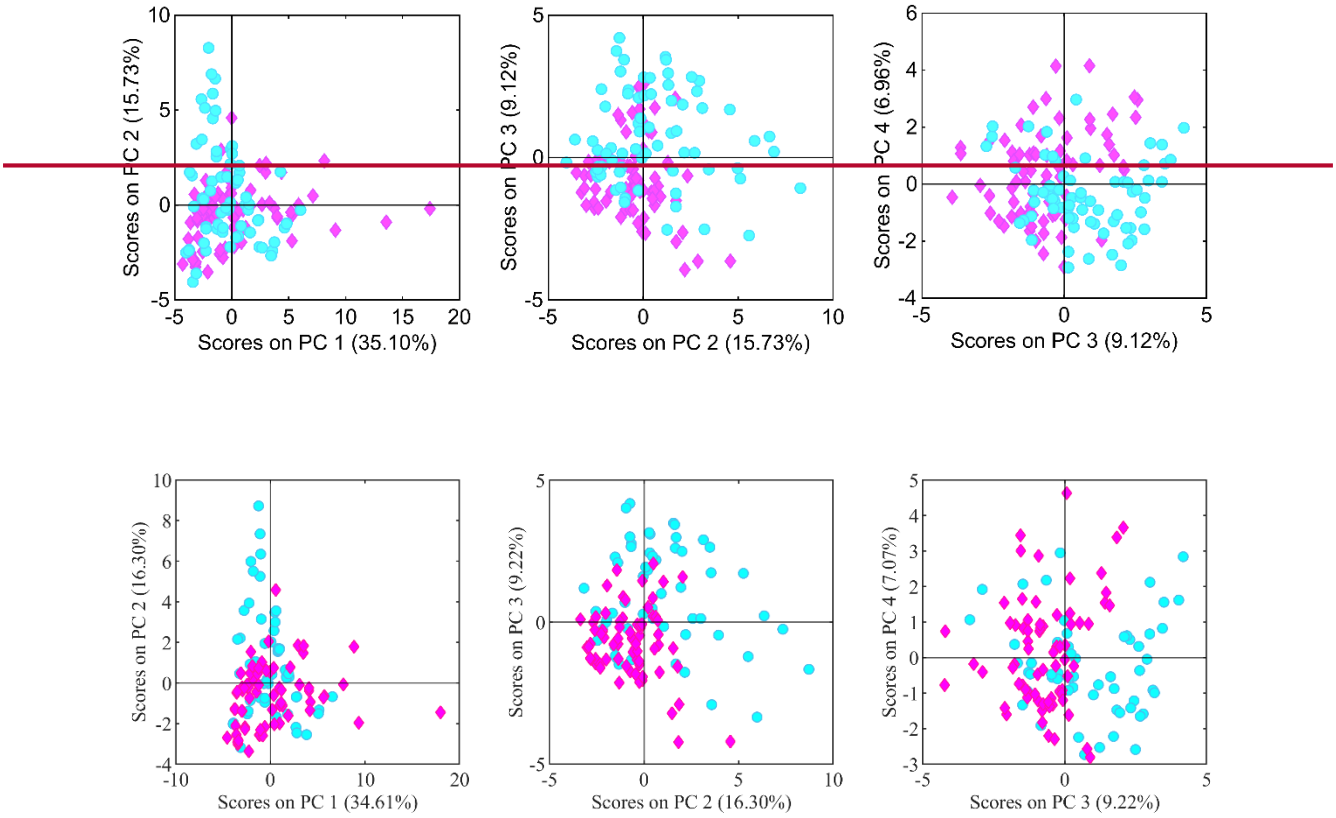


Figure 1. Scores of the samples on the first 4 PCs: PC 1 *vs.* PC 2 (left), PC 2 *vs.* PC 3 (center), PC 3 *vs.* PC 4 (right). The fraction of variance explained by a given component is reported as a percentage value in parenthesis on the corresponding axis. Samples are marked according to their class (~~empty~~ magenta diamonds: cocrystals; ~~black~~ cyan circles: binary mixtures).

As for the loading plots depicted in Figure ~~S1~~2, it can be observed that PC 1 explains the features related both to differences in molecular dimensions (e.g., 1, 2, 3, 9, 16, 29) and connectivity (23, 24, 25, 26). PC 2 considers the dissimilarities in the electronic properties (e.g., 6, 7, 19, 22, 30) of the two partner molecules as well as the difference in their number of heteroatoms (27). PC 3 accounts for more specific features, such as differences in molecular complexity (21), molecular refractivity (31), and energy (15). Lastly, PC 4 considers the dissimilarity in: i) component of the dipole along the *x* axis (11), ii) total dipole (14) and iii) minimum of the MEP surface (20). Regarding the samples belonging to the CC class, these compounds were characterized by partner molecules with a similar behavior in terms of molecular complexity (21) and octanol/water partition coefficient (10).

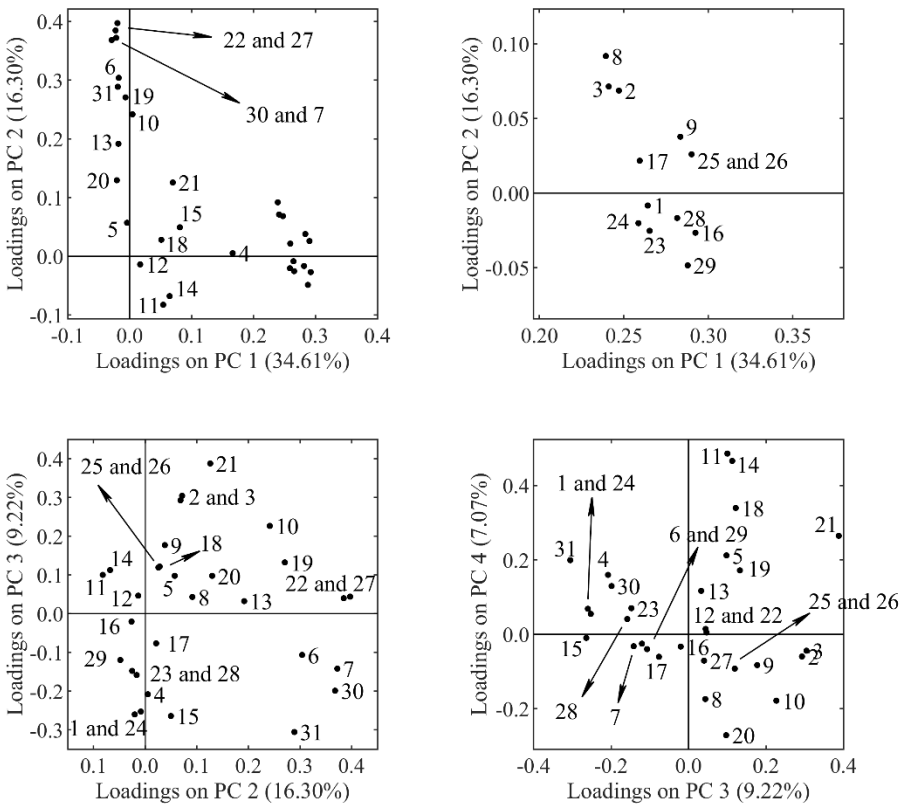
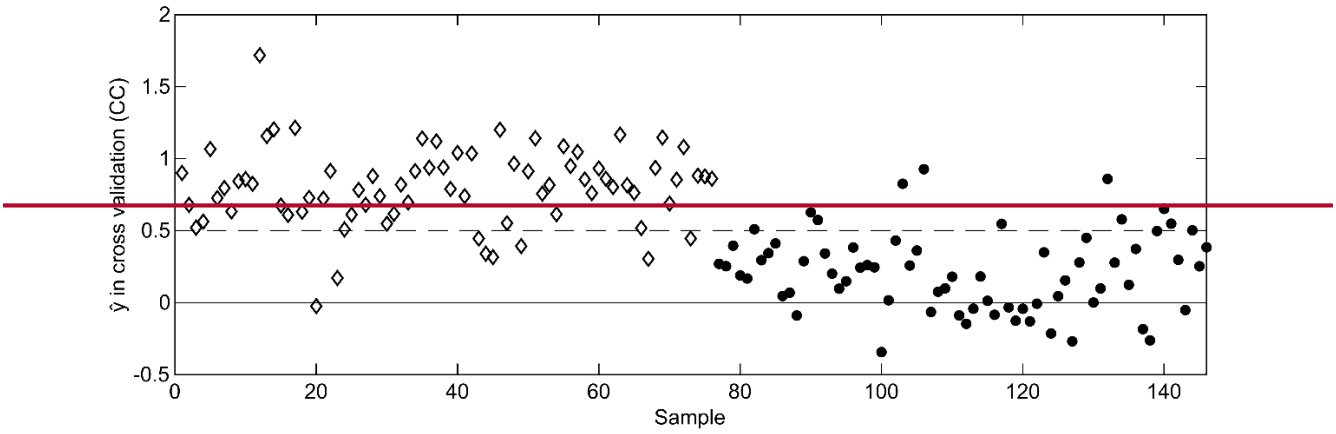


Figure 2. Loading plots related to the PCA decomposition. PC 1 *vs.* PC 2 (top-left), magnification of the PC 1 *vs.* PC 2 (top-right), PC 2 *vs.* PC 3 (bottom-left), PC 3 *vs.* PC 4 (bottom-right). The fraction of variance explained by a given component is reported as a percentage value in parenthesis on the corresponding axis.

4.1.2. Supervised pattern recognition

The relationship between the class membership and the variables was exploited by means of PLS–DA. ~~Four~~ Six LVs were retained according to maximum ~~NER~~ACC% in cross validation. The PLS–DA model captured the ~~63~~72% and ~~64~~69% of the variance of the $\boldsymbol{X}_{\mathbf{cal}}$ and $\boldsymbol{Y}_{\mathbf{cal}}$, respectively. The values of the predicted response $\hat{\boldsymbol{y}}$ in cross validation related to CC samples are plotted in Figure ~~2~~S1.



~~**Figure 2: *ŷ* in cross validation for the CC samples. The dashed horizontal line shows the hard classification threshold of 0.50, halfway between the codified 0 (BM) and 1 (CC). Samples are marked according to their class (empty diamonds: cocrystals; black circles: binary mixtures).**~~

A summary of the performance of the obtained model is reported in the confusion matrix (Table ~~1~~2), whereas a graphical representation of the estimated and predicted values ***ŷ*** for the CC class is reported in Figure S2. As reported in Table ~~1~~2, all the BM samples belonging to the test set were correctly classified except for ~~one~~ 5 samples, whereas only ~~5~~ 3 out of ~~25~~ 30 CC samples were wrongly assigned to the BM class, obtaining a ~~NER~~ACC$_{\%}$ of ~~83~~85%. Similarly, a high ~~NER~~ACC$_{\%}$ of 92% was obtained when the samples belonging to the calibration set were predicted by the model. The achieved results are extremely satisfactory, allowing for the *a priori* selection of the partner molecules required for the synthesis of novel cocrystals.

Table ~~1~~2. Confusion matrix of the calibration and the test sets for the PLS–DA model. The second-last line shows the SEN$_{\%}$ for the modelled classes and last line shows the ~~NER~~ACC$_{\%}$ for the calibration and the test sets.

	Calibration set		**Test set**	
	Predicted as		Predicted as	
	CC	BM	CC	BM
True CC	~~71~~66	5	~~20~~27	~~5~~3
True BM	~~7~~5	~~63~~51	~~1~~5	~~9~~19
SEN%	93%	91%	90%	79%

~~NER~~ACC%	92%	~~83~~85%

The distribution of the samples in the reduced space of the LVs can be observed by inspecting the score plot (Figure 3), whereas information regarding suspicious and/or influential samples can be retrieved by the squared residuals *Q vs.* Hotelling's T^2 and the residuals *vs.* leverage plots (Figure 4).

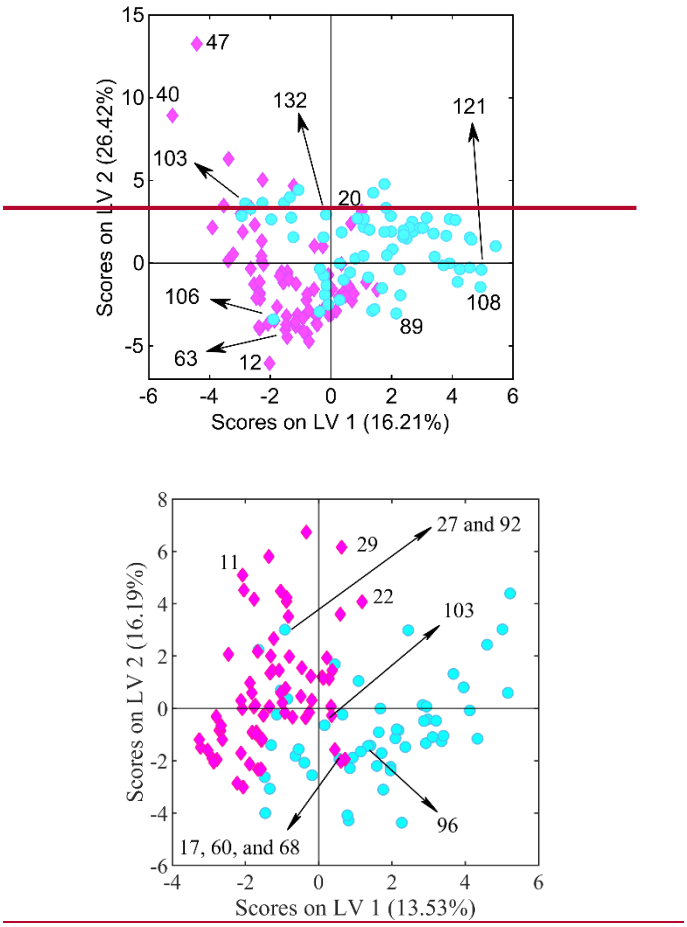


Figure 3. Distribution of the samples in the score space (LV 1 *vs.* LV 2). The fraction of variance explained by a given component is reported as a percentage value in parenthesis on the corresponding axis. Samples are marked according to their class (~~empty~~ magenta diamonds: cocrystals; ~~black~~ cyan circles: binary mixtures).

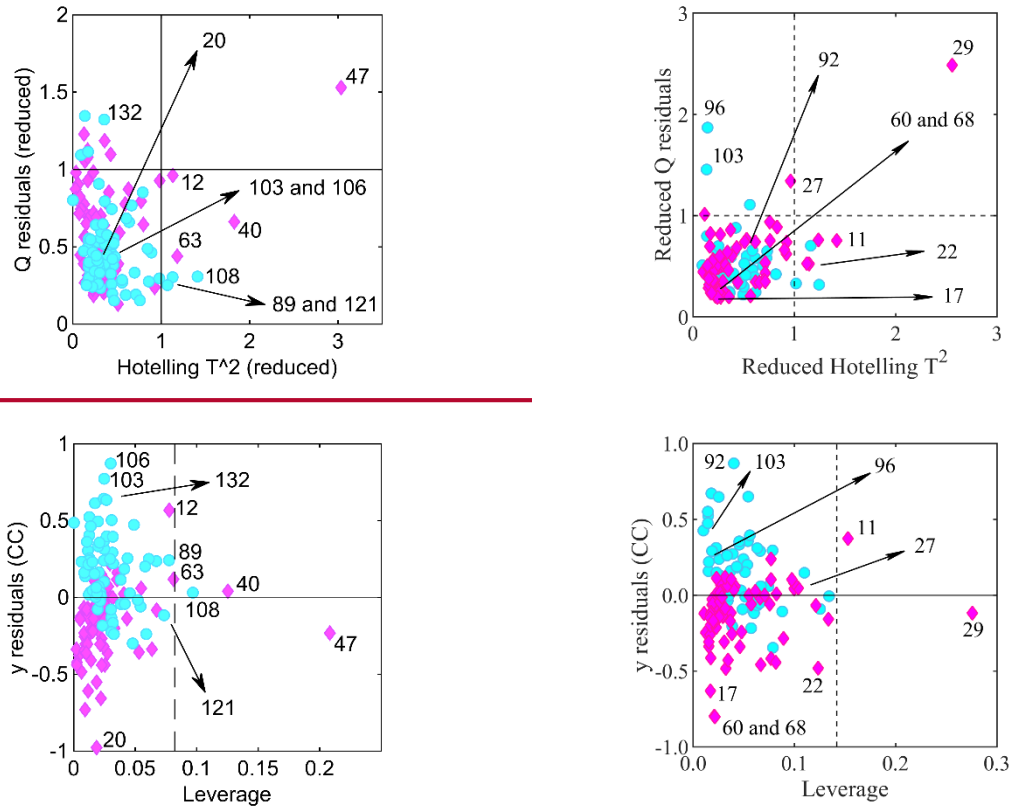


Figure 4. Reduced squared residuals *Q vs.* reduced Hotelling's T^2 plot, the dashed horizontal and vertical lines show the amplitude of the 95% confidence interval for both parameters (top). Residuals of CC samples *vs.* leverage plot, the dashed vertical line shows the leverage limit (bottom). Samples are marked according to their class (~~empty~~ magenta diamonds: cocrystals; ~~black~~ cyan circles: binary mixtures).

The maximum separation in the score space was provided by the first two LVs, with the CC samples ~~well grouped~~located mainly at negative scores ~~both~~ on LV 1 ~~and LV 2, especially in the III quadrant~~. By contrast, BM samples were more scattered and localized mostly on positive scores on LV 1.

~~A peculiar behavior was observed for samples (40) and (47): sample 40 showed Hotelling's T^2 value outside the 95% confidence interval, whereas sample 47 showed both high Hotelling's T^2 and high squared residuals Q together with high leverage in the $\mathbf{Y}_{cal}$ space. These CC samples were obtained by pairing fatty acids (lauric acid, 40, and palmitic acid, 47) with low molecular weight coformers, i.e., pyrazine and nicotinamide, respectively. The variables responsible for this behavior were related to the discrepancy in molecular dimensions between the two partner molecules.~~

~~Another BM sample, i.e., the limonene/ascorbic acid (108), was characterized by high leverage and high Hotelling's T^2 value, due to different behavior in terms of hydrogen bond propensity and octanol/water partition coefficient. A similar behavior was observed also by two additional non-influential BM samples based on ascorbic acid paired with cinnamaldehyde (89) and menthone (121).~~

~~Finally, also the tartaric acid/pyrazine (63) and the adipic acid/hexamethyleneamine (12) CC samples showed Hotelling's T^2 values outside the 95% confidence interval and the latter resulted in having the highest residuals among CC samples.~~

~~Nevertheless, all these anomalous samples were correctly assigned in cross validation, and the variables with high contribution (not shown) on their Hotelling's T^2 and high squared residuals Q were characterized by low PLS weights.~~

~~On the other hand, one not anomalous CC sample and two not anomalous BM samples, namely carveol/isonicotinamide (20), eugenol/pyrazine (103), geraniol/menthol (106), respectively, were characterized by very high residuals in absolute value, thus being misclassified. This behavior can be ascribed to the fact that their features were inversely related to their respective class. Also, the BM sample urea/hexamethylenetetramine (132) was misclassified: its features did not align with those of the other samples; in fact, this sample hold high Q squared residuals.~~ A peculiar behavior was observed for samples 27 and 29. Sample 27 showed squared residual Q outside the 95% confidence interval, whereas sample 29 showed both high Hotelling's T^2 and high squared residuals Q together with high leverage in the $\boldsymbol{Y}_{\mathbf{cal}}$ space. These CC samples were obtained by pairing fatty acids (lauric acid, 27, and palmitic acid, 29) with a low-molecular weight coformer, i.e., nicotinamide. The variables responsible for this behavior can be related to the discrepancy in molecular dimensions between the two partner molecules.

The adipic acid/hexamethylenetetramine (11) CC sample was characterized by high leverage and high Hotelling's T^2 value, due to different behavior in terms of ruotable bonds and number of rings present in the structure. A similar behaviour, in terms of difference in the number of rings, was observed also by two additional non-influential BM samples based on hexamethylenetetramine paired with limonene (96) and menthone (103). In addition, these samples held high squared residuals Q and, therefore, were characterized by features that did not align with the ones of the other samples.

Finally, the ferulic acid/pyrazine (22) CC sample showed Hotelling's T^2 values outside the 95% confidence interval due to the different behaviour of the partner molecules in terms of molecular weight, connectivity, surface area, and electronic properties, i.e., isotropic and anisotropic polarizability and number of valence electrons.

Suspicious samples 22 (ferulic acid/pyrazine, CC) and 103 (menthone/hexamethylenetetramine, BM) appeared also to have high residuals in absolute value and were wrongly assigned to their class in cross validation. Along with them, also three not anomalous CC samples and one not anomalous BM samples, namely cinnamaldehyde/4-hydroxybenzoic acid (17),

carvacrol/nicotinamide (60), thymol/tetramethylpyrazine (68) and eugenol/pyrazine (92), respectively, were misclassified. In fact, their features were inversely related to their respective class. Nevertheless, the exclusion of the samples discussed above from the calibration set would not have produced any difference in terms of rotation of the LV space due to their low leverage. The correlation between class membership, coded in the $\boldsymbol{Y}_{\mathbf{cal}}$, and the predictors contained in the $\boldsymbol{X}_{\mathbf{cal}}$ space can be observed in the PLS weights plot (Figure S3). ~~The variables involved in the discrimination are those whose weights follow the discriminant direction: BM samples reach positive values of LV 1 and LV 2 in the $\boldsymbol{Y}_{\mathbf{cal}}$ loading space (not shown), whereas it is the opposite for CC samples. Therefore, it can be inferred that significant differences in descriptors related to polarizability, exposed surface, and volume, such as atom and bond count (2, 3), molecular volume (9), octanol-water partition coefficient (10), heteroatom count (27), topological polar surface area (22) are likely to prevent the formation of a cocrystal.~~

Information regarding the contribution of each predictor involved in the discrimination of the modelled classes can be inferred by inspecting the pseudo-regression coefficients and the VIP score plots, reported in Figure 5. The latter parameter denotes the relative importance of each predictor of the $\boldsymbol{X}_{\mathbf{cal}}$ space in the PLS–DA model in explaining the class membership encoded in the $\boldsymbol{Y}_{\mathbf{cal}}$ and may guide variable selection. Generally, a variable can be considered important with a VIP score > 1; by contrast, a VIP score significantly lower than 1 indicates that a given variable is a good candidate for exclusion. According to the negative sign of the pseudo-regression coefficients related to CC class, ~~Therefore,~~ it can be stated~~inferred~~ that significant differences in descriptors related to polarizability and exposed surface, such as atom and bond count (2, 3), octanol-water partition coefficient (10), topological polar surface area (22) and heteroatom count (27) are likely to prevent the formation of a cocrystal. It should be noted that this consideration agrees with what has emerged earlier from unsupervised modelling, and it is largely in agreement with widely applied rules of thumb in crystal engineering.

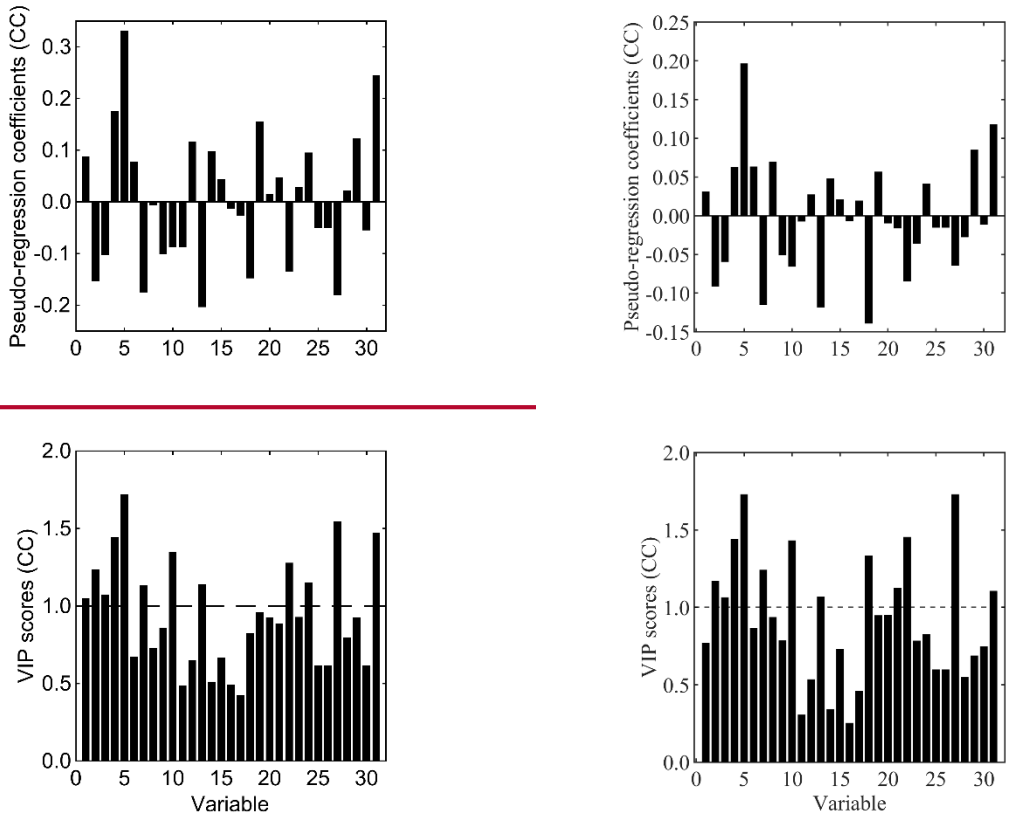


Figure 5. Pseudo-regression coefficients plot for the CC class (top). VIP scores related to each variable included in the $\boldsymbol{X}_{\mathbf{cal}}$ space for the CC class, the significance threshold of 1 is depicted as a dashed horizontal line (bottom). Both plots report the variable identification number on the *x* axis.

~~The variables that contributed most to the PLS weights were characterized by a VIP score > 1. A~~Finally, a reduced PLS–DA model based only on the important variables in agreement with the VIP scores was computed because of its ease of interpretation. A 4 LVs model was calculated according to the maximum ~~NER~~ACC% in cross validation obtaining a classification performance very similar to that achieved by including all descriptors. The weights plot of the reduced model is provided in Figure 6, in which the weights of the variables involved in the discrimination are located on the positive and negative sides of the first LV.

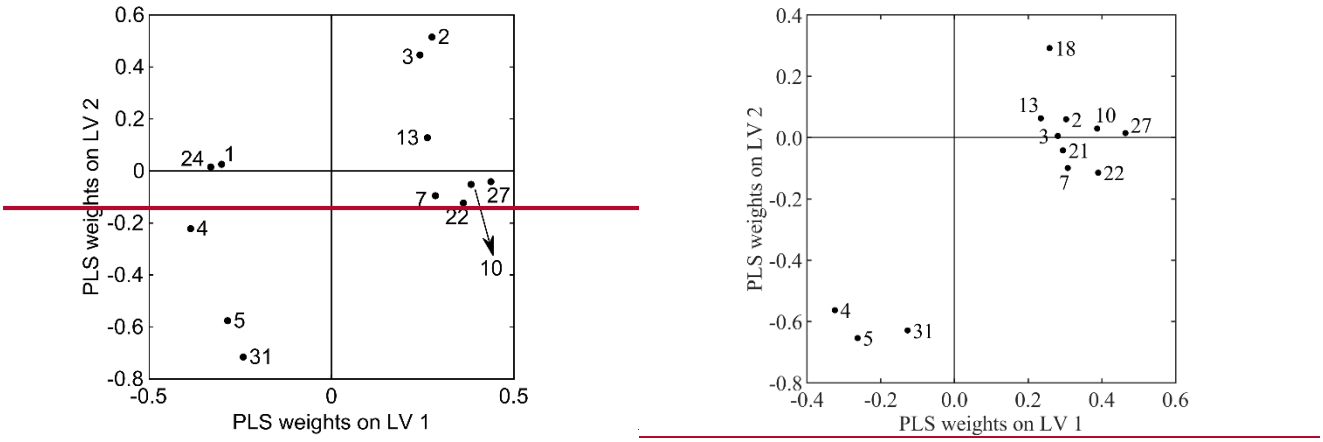


Figure 6. PLS weights of the variables for LV 1 *vs.* LV 2 of the reduced model.

~~Despite the interpretation of the correlation pattern among the variables being not so straightforward, some main considerations can be drawn. As a general comment, it can be stated that a good balance in the hydrogen bond propensity between the two partner molecules has to be achieved, being the difference in hydrogen bond acceptors (7) and number of heteroatoms (27) on positive PLS weights on LV 1. In addition, the coformers should have a similar behavior in terms of polarity, being the dissimilarities in polar surface area (22) and octanol/water partition coefficient (10) on positive weights on LV 1. It should be noted that this consideration agrees with what has emerged earlier from unsupervised modelling, and it is largely in agreement with widely applied rules of thumb in crystal engineering.~~

4.2. ~~Prediction of unknown samples~~Benchmarking on an external validation set

~~In order to evaluate the predictive capability of the model, an external set of $N = 58$ binary combinations of partner molecules was used. An overview of the involved samples is reported in Table S4 along with their estimated $\hat{y}$ values.~~

The~~is~~ external validation set consists of 27 pairs classified experimentally as BMs and 31 pairs classified as CCs, i.e., 11 with our mechanochemistry/PXRD protocol and 20 retrieved from the CSD. The latter ones are reported in Section 3 of the Supplementary Information together with their CSD refcode and the reference to the original publications.

A graphical representation of the predicted values $\hat{y}$ for the CC class is reported in Figure 7, as well as the squared residuals *Q vs.* Hotelling's T^2 plot. Although there were some samples that did not conform to the model space, not all of them have been systematically misclassified. The confusion matrix is reported in Table 3 to summarize the results. In total, about ~~62~~74% of the predictions were in agreement with the experimental results~~.~~.

Table 3. Confusion matrix of the external validation set for the PLS–DA model. The second-last line shows the $SEN_{\%}$ for the modelled classes and last line shows the $ACC_{\%}$.

	External validation set	
	Predicted as	
	CC	BM
Experimental CC	29	2
Experimental BM	13	14
$SEN_{\%}$	94%	52%
$ACC_{\%}$	74%	

Specifically, ~~26~~ 29 CC cases out of 31 were correctly classified. On the other hand, there were ~~17~~ 13 false positive results and ~~10~~ 14 cases in which the pairs were correctly identified as BMs. The model appears therefore to be quite conservative in discarding the possibility of cocrystallization; hence, fewer potential new materials could be overlooked. ~~Similar behavior was observed also in the test set, meaning that the missed discovery rate does not get worse when working on completely external data, thus demonstrating the stability of the model.~~ In addition, the fraction of misclassified CC samples (see Table 2 and Table 3) does not change significantly in the external validation set with respect to the test set.

As shown in Figure 7, Table 3 and Table S4, only 2 pairs of partner molecules predicted as BMs actually formed CCs, thus producing 2 false negative results. It can be stated that these~~Moreover, all the 5 false negative~~ results were borderline cases when the estimated error [63] on the prediction was taken into consideration (data not shown): in fact their $\hat{y}$ value was close to the classification threshold of 0.53. The same is true for ~~8~~ 6 of the false positive cases, while in the remaining ~~9~~ 7 cases the model confidently classified the pairs as CCs, in contrast to our experimental results. These findings could be ascribed to the use of different preparation methods other than mechanochemical grinding. This hypothesis is somehow supported by the fact that 19 out of the 20 CCs retrieved from the CSD, then prepared with a variety of methods, were correctly identified.

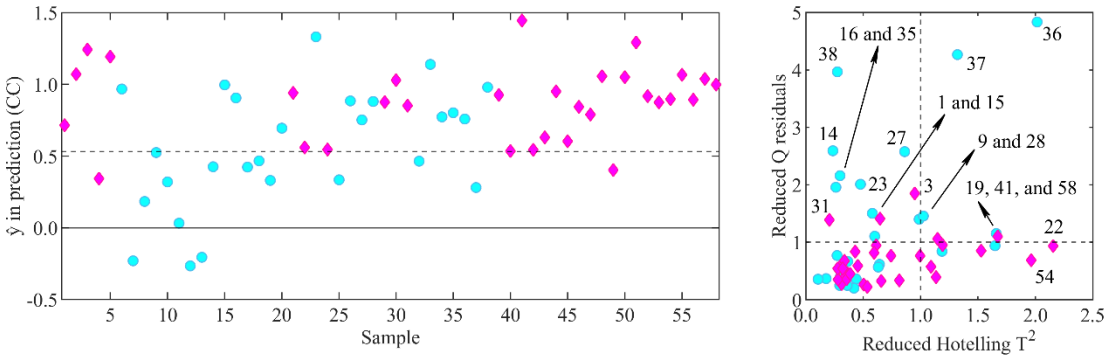


Figure 7. $\hat{y}$ (CC samples) for the external validation set, the dashed horizontal line shows the hard classification threshold of 0.53 (left). Reduced squared residuals *Q vs.* reduced Hotelling's T^2 plot, the dashed horizontal and vertical lines show the amplitude of the 95% confidence interval for both parameters (right). Samples are marked according to their class (magenta diamonds: cocrystals; cyan circles: binary mixtures).

5. Conclusion

This study highlighted the ability of a simple PLS–DA model to predict cocrystal formation without any *a priori* knowledge of the specific role of the involved partner molecules. Information deriving from both successful and unsuccessful cocrystallization experiments was used. The major advantage of the proposed methodology relies on the reduction of the experimental effort required for both the synthesis and characterization of new crystalline structures.

The model allows us to predict cocrystallization propensity with a ~~62~~74% of the predictions in agreement with the experimental results. Considering that the model was obtained on a training set spanning different molecular characteristics, it can be stated that it is suitable for a fairly general applicability.

Indeed, once in possess of the set of chemical descriptors for the molecules of interest, it is sufficient to calculate the absolute value of their difference and perform a linear combination using the pseudo-regression coefficients to obtain a prediction on cocrystal formation. The precalculated values for the set of descriptors comprising 2193 GRAS molecules are available in the Supplementary Material. By applying the proposed methodology, ~~seven~~ ten new cocrystals were discovered and an additional ~~four~~ compound~~s~~ was~~were~~ obtained by chance.

Another figure of merit of the proposed approach is the possibility of understanding through the inspection of PLS weights how the degree of similarity in terms of molecular features of the two partner molecules is correlated with the possibility of obtaining a cocrystal.

On a closing note, we would like to strongly encourage scientists to report failed attempts at cocrystallization along with the technique used, as access to this information could play a pivotal role in refining predictive models, making them less sensitive to selective reporting bias.

Acknowledgements

This work was funded by the Ministero delle Politiche Agricole, Alimentari, Forestali e del Turismo (MIPAAFT) by granting for the project PAC/Packaging Attivo Cristallino. This work has benefited from the framework of the COMP-HUB initiative, funded by the "Department of Excellence" program of the Italian Ministry for Education, University and Research (MIUR, 2018–2022).

Authorship Contribution Statement

Fabio Fornari: Conceptualization (equal), Formal Analysis (lead), Methodology (equal), Validation (equal), Visualization (equal), Writing – Original Draft Preparation (lead), Writing – Review & Editing (equal).

Fabio Montisci: Investigation (equal), Data Curation (lead), Validation (equal), Visualization (equal), Writing – Original Draft Preparation (supporting), Writing – Review & Editing (equal).

Federica Bianchi: Conceptualization (equal), Methodology (lead), Resources (equal), Supervision (equal), Writing – Review & Editing (equal).

Marina Cocchi: Conceptualization (equal), Methodology (equal), Resources (equal), Software (lead), Writing – Review & Editing (equal).

Claudia Carraro: Investigation (equal), Data Curation (equal).

Francesca Cavaliere: Software (equal), Writing – Review & Editing (supporting).

Pietro Cozzini: Software (equal), Writing – Review & Editing (supporting).

Francesca Peccati: Software (supporting), Writing – Review & Editing (supporting).

Paolo P. Mazzeo: Conceptualization (equal), Investigation (supporting), Data Curation (supporting), Validation (supporting), Writing – Review & Editing (supporting).

Nicolò Riboni: Investigation (supporting), Writing – Review & Editing (supporting).

Maria Careri: Supervision (equal), Funding Acquisition (equal), Resources (equal), Writing – Review & Editing (equal).

Alessia Bacchi: Conceptualization (lead), Funding Acquisition (equal), Resources (equal), Supervision (equal), Writing – Review & Editing (equal).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E.V.R. Campos, P.L.F. Proença, J.L. Oliveira, M. Bakshi, P.C. Abhilash, L.F. Fraceto, Use of botanical insecticides for sustainable agriculture: Future perspectives, Ecol. Indic. 105 (2019) 483–495. doi:10.1016/J.ECOLIND.2018.04.038.

[2] Y. Kourkoutas, M. Angane, S. Swift, K. Huang, C.A. Butts, S.Y. Quek, Essential Oils and Their Major Components: An Updated Review on Antimicrobial Activities, Mechanism of Action and Their Potential Application in the Food Industry, Foods 2022, Vol. 11, Page 464. 11 (2022) 464. doi:10.3390/FOODS11030464.

[3] N.S. Singh, R. Sharma, T. Parween, P.K. Patanjali, Pesticide Contamination and Human Health Risk Factor, Mod. Age Environ. Probl. Their Remediat. (2018) 49–68. doi:10.1007/978-3-319-64501-8_3.

[4] Climate Change, Intercropping, Pest Control and Beneficial Microorganisms, Springer Netherlands, 2009. doi:10.1007/978-90-481-2716-0.

[5] F. Bakkali, S. Averbeck, D. Averbeck, M. Idaomar, Biological effects of essential oils - A review, Food Chem. Toxicol. 46 (2008) 446–475. doi:10.1016/j.fct.2007.09.106.

[6] M. Alonso-Gato, G. Astray, J.C. Mejuto, J. Simal-Gandara, Essential Oils as Antimicrobials in Crop Protection, Antibiotics. 10 (2021) 34. doi:10.3390/antibiotics10010034.

[7] F. Bianchi, F. Fornari, N. Riboni, C. Spadini, C.S. Cabassi, M. Iannarelli, C. Carraro, P.P. Mazzeo, A. Bacchi, S. Orlandini, S. Furlanetto, M. Careri, Development of novel cocrystal-based active food packaging by a Quality by Design approach, Food Chem. 347 (2021) 129051. doi:10.1016/j.foodchem.2021.129051.

[8] A.T.H. Mossa, Green Pesticides: Essential oils as biopesticides in insect-pest management, J. Environ. Sci. Technol. 9 (2016) 354–378. doi:10.3923/jest.2016.354.378.

[9] S. Sharma, S. Barkauskaite, A.K. Jaiswal, S. Jaiswal, Essential oils as additives in active food packaging, Food Chem. (2020) 128403. doi:10.1016/j.foodchem.2020.128403.

[10] Food and Drug Administration, 54960 Federal Register / Vol . 81 , No . 159 / Wednesday , August 17 , 2016 / Rules and Regulations, 81 (2016) 54960–55055.

[11] J. Wieczyńska, I. Cavoski, Antimicrobial, antioxidant and sensory features of eugenol, carvacrol and trans-anethole in active packaging for organic ready-to-eat iceberg lettuce, Food Chem. 259 (2018) 251–260. doi:10.1016/j.foodchem.2018.03.137.

[12] L. Pavoni, D.R. Perinelli, G. Bonacucina, M. Cespi, G.F. Palmieri, An overview of micro-and nanoemulsions as

vehicles for essential oils: Formulation, preparation and stability, Nanomaterials. 10 (2020) 135. doi:10.3390/nano10010135.

[13] G.R. Desiraju, Crystal Engineering: A Holistic View, Angew. Chemie Int. Ed. 46 (2007) 8342–8356. doi:10.1002/anie.200700534.

[14] A. Bacchi, P.P. Mazzeo, Cocrystallization as a tool to stabilize liquid active ingredients, Crystallogr. Rev. (2021) 1–22. doi:10.1080/0889311X.2021.1978079.

[15] D. Balestri, P.P. Mazzeo, R. Perrone, F. Fornari, F. Bianchi, M. Careri, A. Bacchi, P. Pelagatti, Deciphering the Supramolecular Organization of Multiple Guests Inside a Microporous MOF to Understand their Release Profile, Angew. Chemie Int. Ed. 60 (2021) 10194–10202. doi:10.1002/ANIE.202017105.

[16] D. Balestri, P.P. Mazzeo, C. Carraro, N. Demitri, P. Pelagatti, A. Bacchi, Stepwise Evolution of Molecular Nanoaggregates Inside the Pores of a Highly Flexible Metal–Organic Framework, Angew. Chemie. 131 (2019) 17503–17511. doi:10.1002/ange.201907621.

[17] P.P. Mazzeo, S. Canossa, C. Carraro, P. Pelagatti, A. Bacchi, Systematic coformer contribution to cocrystal stabilization: energy and packing trends, CrystEngComm. 22 (2020) 7341–7349. doi:10.1039/D0CE00291G.

[18] P.P. Mazzeo, C. Carraro, A. Arns, P. Pelagatti, A. Bacchi, Diversity through Similarity: A World of Polymorphs, Solid Solutions, and Cocrystals in a Vial of 4,4′-Diazopyridine, Cryst. Growth Des. 20 (2020) 636–644. doi:10.1021/acs.cgd.9b01052.

[19] P.P. Mazzeo, C. Carraro, A. Monica, D. Capucci, P. Pelagatti, F. Bianchi, S. Agazzi, M. Careri, A. Raio, M. Carta, F. Menicucci, M. Belli, M. Michelozzi, A. Bacchi, Designing a Palette of Cocrystals Based on Essential Oil Constituents for Agricultural Applications, ACS Sustain. Chem. Eng. 7 (2019) 17929–17940. doi:10.1021/acssuschemeng.9b04576.

[20] D. Capucci, D. Balestri, P.P. Mazzeo, P. Pelagatti, K. Rubini, A. Bacchi, Liquid Nicotine Tamed in Solid Forms by Cocrystallization, Cryst. Growth Des. 17 (2017) 4958–4964. doi:10.1021/acs.cgd.7b00887.

[21] P.P. Mazzeo, M. Pioli, F. Montisci, A. Bacchi, P. Pelagatti, Mechanochemical Preparation of Dipyridyl-Naphthalenediimide Cocrystals: Relative Role of Halogen-Bond and π–π Interactions, Cryst. Growth Des. 21 (2021) 5687–5696. doi:10.1021/acs.cgd.1c00531.

[22] N.K. Duggirala, M.L. Perry, Ö. Almarsson, M.J. Zaworotko, Pharmaceutical cocrystals: Along the path to improved medicines, Chem. Commun. 52 (2016) 640–655. doi:10.1039/c5cc08216a.

595 [23] J.W. Steed, The role of co-crystals in pharmaceutical design., Trends Pharmacol. Sci. 34 (2013) 185–93.
596 doi:10.1016/j.tips.2012.12.003.
597 [24] J.G.P. Wicker, L.M. Crowley, O. Robshaw, E.J. Little, S.P. Stokes, R.I. Cooper, S.E. Lawrence, Will they co-
598 crystallize?, CrystEngComm. 19 (2017) 5336–5340. doi:10.1039/C7CE00587C.
599 [25] A. Bacchi, D. Capucci, M. Giannetto, M. Mattarozzi, P. Pelagatti, N. Rodriguez-Hornedo, K. Rubini, A. Sala,
600 Turning Liquid Propofol into Solid (without Freezing It): Thermodynamic Characterization of Pharmaceutical
601 Cocrystals Built with a Liquid Drug, Cryst. Growth Des. 16 (2016) 6547–6555. doi:10.1021/acs.cgd.6b01241.
602 [26] Y. Xiao, L. Zhou, H. Hao, Y. Bao, Q. Yin, C. Xie, Cocrystals of propylthiouracil and nutraceuticals toward
603 sustained-release: Design, structure analysis, and solid-state characterization, Cryst. Growth Des. 21 (2021) 1202–
604 1217. doi:10.1021/acs.cgd.0c01519.
605 [27] O. Shemchuk, S. D'Agostino, C. Fiore, V. Sambri, S. Zannoli, F. Grepioni, D. Braga, Natural Antimicrobials Meet
606 a Synthetic Antibiotic: Carvacrol/Thymol and Ciprofloxacin Cocrystals as a Promising Solid-State Route to
607 Activity Enhancement, Cryst. Growth Des. 20 (2020) 6796–6803. doi:10.1021/acs.cgd.0c00900.
608 [28] M.D. Perera, J. Desper, A.S. Sinha, C.B. Aakeröy, Impact and importance of electrostatic potential calculations for
609 predicting structural patterns of hydrogen and halogen bonding, CrystEngComm. 18 (2016) 8631–8636.
610 doi:10.1039/c6ce02089e.
611 [29] M.C. Etter, Encoding and Decoding Hydrogen-Bond Patterns of Organic Compounds, Acc. Chem. Res. 23 (1990)
612 120–126. doi:10.1021/ar00172a005.
613 [30] C.A. Hunter, Quantifying intermolecular interactions: Guidelines for the molecular recognition toolbox, Angew.
614 Chemie - Int. Ed. 43 (2004) 5310–5324. doi:10.1002/anie.200301739.
615 [31] M. Karimi-Jafari, L. Padrela, G.M. Walker, D.M. Croker, Creating cocrystals: A review of pharmaceutical cocrystal
616 preparation routes and applications, Cryst. Growth Des. 18 (2018) 6370–6387. doi:10.1021/acs.cgd.8b00933.
617 [32] N. Issa, P.G. Karamertzanis, G.W.A. Welch, S.L. Price, Can the formation of pharmaceutical cocrystals be
618 computationally predicted? I. Comparison of lattice energies, Cryst. Growth Des. 9 (2009) 442–453.
619 doi:10.1021/cg800685z.
620 [33] M.A. Mohammad, A. Alhalaweh, S.P. Velaga, Hansen solubility parameter as a tool to predict cocrystal formation,
621 Int. J. Pharm. 407 (2011) 63–71. doi:10.1016/j.ijpharm.2011.01.030.
622 [34] L. Fábián, Cambridge Structural Database Analysis of Molecular Complementarity in Cocrystals, Cryst. Growth

Des. 9 (2009) 1436–1443. doi:10.1021/cg800861m.

[35] C.R. Groom, I.J. Bruno, M.P. Lightfoot, S.C. Ward, The Cambridge structural database, Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater. 72 (2016) 171–179. doi:10.1107/S2052520616003954.

[36] J. Devogelaer, H. Meekes, P. Tinnemans, E. Vlieg, R. Gelder, Co- crystal Prediction by Artificial Neural Networks**, Angew. Chemie Int. Ed. 59 (2020) 21711–21718. doi:10.1002/anie.202009467.

[37] D. Wang, Z. Yang, B. Zhu, X. Mei, X. Luo, Machine-Learning-Guided Cocrystal Prediction Based on Large Data Base, Cryst. Growth Des. 20 (2020) 6610–6621. doi:10.1021/acs.cgd.0c00767.

[38] M. Przybyłek, T. Jeliński, J. Słabuszewska, D. Ziółkowska, K. Mroczyńska, P. Cysewski, Application of Multivariate Adaptive Regression Splines (MARSplines) Methodology for Screening of Dicarboxylic Acid Cocrystal Using 1D and 2D Molecular Descriptors, Cryst. Growth Des. 19 (2019) 3876–3887. doi:10.1021/acs.cgd.9b00318.

[39] M. Przybyłek, P. Cysewski, Distinguishing Cocrystals from Simple Eutectic Mixtures: Phenolic Acids as Potential Pharmaceutical Coformers, Cryst. Growth Des. 18 (2018) 3524–3534. doi:10.1021/acs.cgd.8b00335.

[40] A. Vriza, A.B. Canaj, R. Vismara, L.J. Kershaw Cook, T.D. Manning, M.W. Gaultois, P.A. Wood, V. Kurlin, N. Berry, M.S. Dyer, M.J. Rosseinsky, One class classification as a practical approach for accelerating π–π co-crystal discovery, Chem. Sci. 12 (2021) 1702–1719. doi:10.1039/d0sc04263c.

[41] M.E. Mswahili, M.J. Lee, G.L. Martin, J. Kim, P. Kim, G.J. Choi, Y.S. Jeong, Cocrystal prediction using machine learning models and descriptors, Appl. Sci. 11 (2021) 1–12. doi:10.3390/app11031323.

[42] H. Moriwaki, Y.S. Tian, N. Kawashita, T. Takagi, Mordred: A molecular descriptor calculator, J. Cheminform. 10 (2018) 4. doi:10.1186/s13321-018-0258-y.

[43] J.G.P. Wicker, R.I. Cooper, Will it crystallise? Predicting crystallinity of molecular materials, CrystEngComm. 17 (2015) 1927–1934. doi:10.1039/c4ce01912a.

[44] I.J. Bruno, J.C. Cole, P.R. Edgington, M. Kessler, C.F. Macrae, P. McCabe, J. Pearson, R. Taylor, New software for searching the Cambridge Structural Database and visualizing crystal structures, Acta Crystallogr. Sect. B Struct. Sci. 58 (2002) 389–397. doi:10.1107/S0108768102003324.

[45] C.F. Macrae, I.J. Bruno, J.A. Chisholm, P.R. Edgington, P. McCabe, E. Pidcock, L. Rodriguez-Monge, R. Taylor, J. van de Streek, P.A. Wood, Mercury CSD 2.0 – new features for the visualization and investigation of crystal structures, J. Appl. Crystallogr. 41 (2008) 466–470. doi:10.1107/S0021889807067908.

[46] M. Baroni, G. Cruciani, S. Sciabola, F. Perruccio, J.S. Mason, A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application, J. Chem. Inf. Model. 47 (2007) 279–294. doi:10.1021/ci600253e.

[47] Sybyl 8.1, Tripos International, St. Louis, USA, 2009.

[48] M. Clark, R.D. Cramer, N. Van Opdenbosch, Validation of the general purpose tripos 5.2 force field, J. Comput. Chem. 10 (1989) 982–1012. doi:10.1002/jcc.540100804.

[49] Schrödinger LLC, The PyMOL Molecular Graphics System, PyMOL Mol. Graph. Syst. Version 2.0. (2010).

[50] G. Landrum, RDKit: Open-Source Cheminformatics Software, Http://Www.Rdkit.Org/. (2021).

[51] G.A. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Petersson, Gaussian 16, Rev. A.03, Gaussian, Inc., Wallingford, CT. (2016).

[52] R.D. Snee, Validation of Regression Models: Methods and Examples, Technometrics. 19 (1977) 415–428. doi:10.1080/00401706.1977.10489581.

[53] M. Li Vigni, C. Durante, M. Cocchi, Exploratory Data Analysis, in: Data Handl. Sci. Technol., Elsevier Ltd, 2013: pp. 55–126. doi:10.1016/B978-0-444-59528-7.00003-X.

[54] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52. doi:10.1016/0169-7439(87)80084-9.

[55] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1039/c3ay41907j.

[56] M. Cocchi, A. Biancolillo, F. Marini, Chemometric Methods for Classification and Feature Selection, in: Compr. Anal. Chem., Elsevier B.V., 2018: pp. 265–299. doi:10.1016/bs.coac.2018.08.006.

[57] U.G. Indahl, H. Martens, T. Næs, From dummy regression to prior probabilities in PLS-DA, J. Chemom. 21 (2007) 529–536. doi:10.1002/cem.1061.

[58] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, in: Chemom. Intell. Lab. Syst., Elsevier, 2001: pp. 109–130. doi:10.1016/S0169-7439(01)00155-1.

[59] D. Braga, E. Dichiarante, F. Grepioni, G.I. Lampronti, L. Maini, P.P. Mazzeo, S. D'Agostino, Mechanical Preparation of Crystalline Materials. An Oxymoron?, in: Supramol. Chem., John Wiley & Sons, Ltd, Chichester, UK, 2012. doi:10.1002/9780470661345.smc115.

[60] J. Fürnkranz, P.K. Chan, S. Craw, C. Sammut, W. Uther, A. Ratnaparkhi, X. Jin, J. Han, Y. Yang, K. Morik, M. Dorigo, M. Birattari, T. Stützle, P. Brazdil, R. Vilalta, C. Giraud-Carrier, C. Soares, J. Rissanen, R.A. Baxter, I.

Bruha, R.A. Baxter, G.I. Webb, L. Torgo, A. Banerjee, H. Shan, S. Ray, P. Tadepalli, Y. Shoham, R. Powers, Y. Shoham, R. Powers, G.I. Webb, S. Ray, S. Scott, H. Blockeel, L. De Raedt, Manhattan Distance, Encycl. Mach. Learn. (2011) 639–639. doi:10.1007/978-0-387-30164-8_506.

[61] L.E. Juarez-Orozco, O. Martinez-Manzanera, S. V. Nesterov, S. Kajander, J. Knuuti, The machine learning horizon in cardiac hybrid imaging, Eur. J. Hybrid Imaging. 2 (2018) 1–15. doi:10.1186/S41824-018-0033-3/FIGURES/5.

[62] A. Singh, N. Thakur, A. Sharma, A review of supervised machine learning algorithms, in: 3rd Int. Conf. Comput. Sustain. Glob. Dev., Institute of Electrical and Electronics Engineers, New Delhi, India, 2016: pp. 1310–1315. https://ieeexplore.ieee.org/abstract/document/7724478.

[63] N.M. Faber, R. Bro, Standard error of prediction for multiway PLS: 1. Background and a simulation study, Chemom. Intell. Lab. Syst. 61 (2002) 133–149. doi:10.1016/S0169-7439(01)00204-0.

Highlights

A QSPR model for the discovery of cocrystals made by EOs and other GRAS molecules

Training set based on failed and successful cocrystallization experiments

Correct classification rate of 85% on the test set

Broad applicability and reduced experimental effort

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Author Statement

Fabio Fornari: Conceptualization (equal), Formal Analysis (lead), Methodology (equal), Validation (equal), Visualization (equal), Writing – Original Draft Preparation (lead), Writing – Review & Editing (equal).

Fabio Montisci: Investigation (equal), Data Curation (lead), Validation (equal), Visualization (equal), Writing – Original Draft Preparation (supporting), Writing – Review & Editing (equal).

Federica Bianchi: Conceptualization (equal), Methodology (lead), Resources (equal), Supervision (equal), Writing – Review & Editing (equal).

Marina Cocchi: Conceptualization (equal), Methodology (equal), Resources (equal), Software (lead), Writing – Review & Editing (equal).

Claudia Carraro: Investigation (equal), Data Curation (equal).

Francesca Cavaliere: Software (equal), Writing – Review & Editing (supporting).

Pietro Cozzini: Software (equal), Writing – Review & Editing (supporting).

Francesca Peccati: Software (supporting), Writing – Review & Editing (supporting).

Paolo P. Mazzeo: Conceptualization (equal), Investigation (supporting), Data Curation (supporting), Validation (supporting), Writing – Review & Editing (supporting).

Nicolò Riboni: Investigation (supporting), Writing – Review & Editing (supporting).

Maria Careri: Supervision (equal), Funding Acquisition (equal), Resources (equal), Writing – Review & Editing (equal).

Alessia Bacchi: Conceptualization (lead), Funding Acquisition (equal), Resources (equal), Supervision (equal), Writing – Review & Editing (equal).

Click here to access/download

Supplementary Material

GRAS_molecules_dataset.xlsx

Chemometric-Assisted Cocrystallization: Supervised Pattern Recognition for Predicting the Formation of New Functional Cocrystals

Fabio Fornari [a], Fabio Montisci [a], Federica Bianchi [a,b,*], Marina Cocchi [c], Claudia Carraro [a], Francesca Cavaliere [d], Pietro Cozzini [d], Francesca Peccati [e], Paolo P. Mazzeo [a,f], Nicolò Riboni [a], Maria Careri [a,g], Alessia Bacchi [a,f]

[a] *University of Parma, Department of Chemistry, Life Sciences and Environmental Sustainability, Parco Area delle Scienze 17/A, 43124, Parma, Italy*

[b] *University of Parma, Interdepartmental Center for Packaging (CIPACK), Parco Area delle Scienze, 43124, Parma, Italy*

[c] *University of Modena and Reggio Emilia, Department of Chemical and Geological Sciences, Via Giuseppe Campi 103, 41125, Modena, Italy*

[d] *University of Parma, Department of Food and Drug, Parco Area delle Scienze 17/A, 43124, Parma, Italy*

[e] *Basque Research and Technology Alliance (BRTA), Center for Cooperative Research in Biosciences (CIC bioGUNE), Bizkaia Technology Park 801A, 48160, Derio, Spain*

[f] *University of Parma, Biopharmanet-TEC, Parco Area delle Scienze 27/A, 43124, Parma, Italy*

[g] *University of Parma, Interdepartmental Center on Safety, Technologies, and Agri-Food Innovation (SITEIA.PARMA), Parco Area delle Scienze, 43124, Parma, Italy*

* Author to whom correspondence should be addressed:

e-mail address: federica.bianchi@unipr.it; University of Parma, Department of Chemistry, Life Sciences and Environmental Sustainability, Parco Area delle Scienze 17/A, 43124, Parma, Italy

Abstract

Owing to the antimicrobial and insecticide properties, the use of natural compounds like essential oils and their active components has proven to be an effective alternative to synthetic chemicals in different fields ranging from drug delivery to agriculture and from nutrition to food preservation. Their limited application due to the high volatility and scarce water solubility can be expanded by using crystal engineering approaches to tune some properties of the active molecule by combining it with a suitable partner molecule (coformer). However, the selection of coformers and the experimental effort required for discovering cocrystals are the bottleneck of cocrystal engineering. This study explores the use of chemometrics to aid the discovery of cocrystals of active ingredients suitable for various applications. Partial Least Squares–Discriminant Analysis is used to discern cocrystals from binary mixtures based on the molecular features of the coformers. For the first time, by including failed cocrystallization data and considering a variety of chemically diverse compounds, the proposed method resulted in a successful prediction rate of 85% for the test set in the model validation phase and of 74% for the external test set.

Keywords

cocrystal, crystal engineering, chemoinformatics, chemometrics, partial least square discriminant analysis, Quantitative Structure–Property Relationship

List of abbreviations (sorted alphabetically)

$ACC_{\%}$; Accuracy

BM: Binary Mixture

CC: Cocrystal

CCDC: Cambridge Crystallographic Data Centre

CSD: Cambridge Structural Database

EO: Essential Oil

FDA: Food and Drug Administration

GRAS: Generally Recognized As Safe

LV: Latent Variable

MEP: Molecular Electrostatic Potential

PC: Principal Component

PCA: Principal Component Analysis

PLS–DA: Partial Least Squares–Discriminant Analysis

PXRD: Powder X-Ray Diffraction

QSPR: Quantitative Structure–Property Relationship

$SEN_{\%}$: Sensitivity

VIP: Variable Importance in Projection.

1. Introduction

In the last few decades, the use of agrochemicals and food preservatives has grown exponentially as a direct consequence of the rapid increase of the world population [1,2]. Owing to their potential adverse effects on both human health and environment [2–4], alternative strategies based on the use of more sustainable chemicals have been proposed to support the food system. Being able to exert antimicrobial, insecticidal, and antioxidant properties [5,6], essential oils (EOs) and their active components have been used as green substitutes of synthetic chemicals to extend the shelf-life of foodstuff and in pests control [7–9]. These compounds are Generally Recognized As Safe (GRAS) by the Food and Drug Administration (FDA) [10], however, despite their appealing properties, their use is limited by their high volatility and poor stability [7,9,11,12].

In fact, physicochemical properties of materials play a key role in determining whether a chemical is suitable for a specific purpose, thus strongly affecting its field of application. Scientists have always desired to obtain materials with target properties, and crystal engineering is one of the most interesting approaches to synthesize a great variety of crystalline materials for applications in various fields, ranging from pharmaceuticals to agrochemicals, and from nutraceuticals to cosmetics [13–16]. The basic idea of crystal engineering is related to the possibility of controlling the crystal structure of molecules and, therefore, the properties of the resulting solids. Polymorphism, vitrification and cocrystallization are some of the available strategies to modify the intrinsic properties of molecules without the need of synthetic modifications [17–21].

Cocrystals are multicomponent crystalline solid materials in which the constituents (i.e., coformers) are bound in a well-defined stoichiometric ratio [22,23] *via* non-covalent interactions (e.g., hydrogen bonds, halogen bonds, π–π stacking) within the same crystal structure. Cocrystallization allows for the combination of the desired molecule of interest with properly selected partner molecules, paving the way to an array of potential materials with enhanced properties [19,24,25]. Within this frame of reference, cocrystals based on the active components of EOs have been proposed as active ingredients for food packaging, agrochemical and pharmaceutical applications [7,19,26,27].

Despite the great advantages offered by cocrystallization, the proper degree of complementarity between the two partner molecules required to obtain crystalline materials with the desired properties is not easy to assess [28–31]. In this context, the selection of coformers and the great effort required for both the systematic experimental screening and careful characterization of the products derived from the combination of all the possible coformer pairs represent the major bottleneck of cocrystal engineering. Computational techniques represent a powerful tool to reduce the experimental effort required for the discovery of new cocrystals, enabling to evaluate beforehand whether a cocrystal can be obtained starting from pre-selected coformers. These *in silico* strategies can be based on the calculation of a variety of parameters useful to predict the formation of a

cocrystal, such as lattice energy [32], solubility [33], hydrogen bond propensity along with the quantitation of molecular interaction energy [29,30], and molecular complementarity [34].

Despite the massive efforts spent to develop a method to predict cocrystal formation, at present none of the proposed strategies has proven to be both totally reliable and easy to apply.

Chemometrics could play a pivotal role in cocrystal discovery: up to now only few Machine Learning methods have been proposed in predicting cocrystal formation, enabling the screening of new cocrystals once a supervised model is properly trained and validated. In the study proposed by Devogelaer et al., information of successful cocrystallization experiments was directly taken from the Cambridge Structural Database (CSD) [35] and Artificial Neural Networks (ANN) were used to predict the formation of new cocrystals [36]. Similarly, Wang et al. relied on a consensus method based on multiple Random Forest algorithms, in which the successful cocrystallization dataset was integrated with randomly generated failed cocrystallization data [37]. These approaches are reported in the literature as network-based methods. Additional studies were based on the use of successful and unsuccessful cocrystallization datasets obtained from experimentation, literature, and/or the CSD for screening specific classes of coformers. Within this framework, Przybyłek et al. used Multivariate Adaptive Regression Splines to predict the formation of dicarboxylic and phenolic acid-based cocrystals [38,39], whereas Wicker et al. focused on variously substituted benzoic acids and benzamides using a Support Vector Machine algorithm [24]. Vriza et al. used an ensemble one-class classification method to aid the discovery of π–π cocrystals, thus giving a great contribution in enriching one of the most under-represented classes of cocrystals in the CSD [40]. Most recently, Mswahili et al. developed a cocrystal screening method based on ANN by using both successful and unsuccessful experimental cocrystallization data retrieved from the literature and a plethora of molecular descriptors calculated using Mordred [41,42].

In the frame of a research activity dealing with the synthesis of new functional cocrystals based on the active constituents of EOs and other GRAS molecules to broaden their applicability in the industrial field [7,19], we propose a chemometric approach to aid the discovery of new cocrystalline materials.

For the first time, a training set based on the results of failed (binary mixtures, BM) and successful (cocrystal, CC) cocrystallization experiments was used for the computation of a Quantitative Structure–Property Relationship-like (QSPR) model based on Partial Least Squares–Discriminant Analysis (PLS–DA), after preliminary exploratory analysis by Principal Component Analysis (PCA). The PLS-DA approach, with respect to network-based methods offers the advantages of having only one parameter to optimize, i.e., the model dimensionality, and the direct interpretation of the importance of descriptors in classification, while highlighting their interplay (by inspection of weights and loadings plots).

The effectiveness of the study relies on the use of compounds belonging to different chemical classes and a reduced number of 1D, 2D, and 3D molecular descriptors of various nature (e.g., constitutional, geometric, physical, topological, and surface area-based descriptors) [24,38,39,43], enabling the high-throughput screening of novel cocrystalline materials and offering maximum flexibility and effectiveness at a minimum computational and experimental cost.

2. Experimental Procedures

2.1. Mechanochemical protocol and class assignation

All the molecules in the dataset were chosen among the list of GRAS molecules drawn up by the FDA [10]. Selected pair of molecules among the chosen ones were assigned either to the CC class or to the BM class. Pairs of molecules in the dataset for which a cocrystal structure was already described in literature were individuated in the Cambridge Structural Database (CSD) [35] with the Cambridge Crystallographic Data Centre (CCDC) software ConQuest [44] and visualized with Mercury [45]. They are reported in Section 3 of the Supplementary Material, together with their unique CSD refcode and the reference to the original publications.

Cocrystallization for all the pairs with no known structure in literature was instead attempted with the following mechanochemical protocol. All the reagents employed were commercially available and used as such in all the experiments. Equimolar amounts of each reagent were directly mixed in an agate mortar and subjected to manual grinding for 10-15 minutes, without using any solvent. The resulting powder samples were collected in closed vials. Assignation to CC or BM classes was performed by comparing the Powder X-ray Diffraction (PXRD) pattern of the ground sample with those of the pure reagents. Possible occurrence of polymorphic transitions for the reagents was excluded by comparing the experimental PXRD data after milling with the calculated pattern of all the known crystalline forms of the reagents. The occurrence of new peaks, unexplained by the presence of unreacted reagents, was taken as indication that cocrystallization had occurred and the sample was assigned to the CC class. In case no additional peaks appeared in the PXRD pattern the sample was instead classified as a BM.

2.2. Powder X-ray diffraction

Typically, PXRD data were collected on a Rigaku Smartlab XE diffractometer in θ-θ Bragg-Brentano geometry with Cu Kα radiation. The samples were placed on glass supports and exposed to radiation ($1.5° \leq 2\theta \leq 50°$) at a scan rate of 10°/min. The diffracted beam was collected on a 2D Hypix 3000 solid state detector. 5° radiant soller were used as a compromise for high flux and moderate peak asymmetry at low angles. Beam stopper and anti-scatterer air component were used to mitigate the profile at low angle. In some rare cases, the data were collected on a Thermo Fisher Scientific ARL X'TRA diffractometer in θ-θ Bragg-Brentano geometry with Cu Kα radiation ($3° \leq 2\theta \leq 30°$ at a scan rate of 5°/min, or $3° \leq 2\theta \leq 40°$ at a scan rate of 0.3°/min).

3. Computational Methods

3.1. Molecular descriptors calculation

For each molecule 31 molecular descriptors were calculated (Table S1). A theoretical background for the less known descriptors is given in Section 4 of the Supplementary Material.

The molecular weight, the number of atoms, the number of bonds, the number of hydrogen bond donor sites and the number of hydrogen bond acceptor sites were calculated with FLAP software (Fingerprint for Ligand and Protein) [46] at pH 7.0, using the 3D structures of all molecules in SDF format as input (downloaded from the PubChem database). The number of rotatable bonds, the number of rings, the hydrophobicity (accounted as the number of hydrophobic centers), the logP (logarithm of octanol/water partition coefficient), the molecular volume, the total molecular dipole moment (based on point charge distribution in the molecule), and its components along the axes (using the principal axes of the molecular graph) were then calculated for the same structures using Sybyl 8.1 [47] (www.tripos.com) and taking in consideration the protonation state of molecules. The same software was also used to estimate the strain energy of the molecule without performing any geometry optimization. This energy term relies on an electrostatic calculation from atomic charges using the internal Tripos force field [48]. For the estimation of molecular volume and dipole moment, a specific SPL script was employed. The calculated volume is enclosed in a water-accessible surface computed at a repulsive interaction energy of 0.20 kcal/mol with a water probe. A custom Python script was used to automatically calculate the Solvent Accessible Surface Area (SASA) in PyMol 2.0 [49], with the dot density parameter set to 4. The number of heteroatoms, the number of valence electrons, and the indexes ${}^0\chi$, ${}^0\chi^n$, ${}^0\chi^v$, α_{HK}, ${}^1\kappa_\alpha$, LabuteASA, SMR_VSA, PEOE_VSA, and TPSA were obtained running a Python 3.7 code with the open-source cheminformatics toolkit RDKit Q4 2013 [50]. The average isotropic polarizability α_{iso}, the polarizability

anisotropy $\Delta\alpha$, and the Molecular Electrostatic Potential (MEP) were calculated with Gaussian 16 [51] following the *in vacuo* Density-Functional Theory optimization of all the molecules, employing the hybrid functional B3LYP and the People double-z basis set 6-31+g(d,p).

Postprocessing of the MEP to extract critical points at a given electron density isosurface was performed with a custom Python 3.6.1 script on a three-dimensional map (cube format) with a sampling density of 6 points/Bohr along the three directions. The MEP was analyzed at an electron density isosurface of 0.002 a.u. with a tolerance of 0.001 a.u., meaning that only MEP values corresponding to regions of space with electron density in the 0.001–0.003 a.u. range were considered. A first set of critical points was identified comparing MEP values of each cube point with those of its 6 nearest neighbors. A point was considered a local minimum if the number of nearest neighbors with higher MEP was greater or equal to a given integer (4). Likewise, a point was considered a local maximum if the number of nearest neighbors with lower MEP was greater or equal to the same integer. This first step yielded a large number of candidate critical points encompassing a wide range of MEP values. Since our focus was on identifying the regions of the molecules likely to be involved in strong hydrogen bonds within the cocrystal, in a second step this first set of points was filtered based on the MEP values of the global minimum and maximum. This was done as follows: for each local minimum (maximum), the ratio between its MEP value and that of the global minimum (maximum) was computed, and the point was kept only if the ratio exceeded a given threshold (0.1). In this way, only points corresponding to shallow critical points were discarded. This step allowed to identify the MEP isosurface regions corresponding to hydrogen bond donors and acceptors. However, due to the rugged character of the MEP map, multiple critical points of the same type could appear in close proximity. To univocally map a given region of the isosurface to a MEP value, critical points close to each other (below a distance threshold of 1.0 Bohr) were merged iteratively, keeping only the lower MEP point for minima and higher MEP point for maxima. The algorithm then provided the final set of MEP critical points at the given electron density isosurface.

3.2. Data analysis

The entire data analysis was carried out in MATLAB R2019a environment (Mathworks, Natick, Massachusetts, USA) with the aid of the PLS_Toolbox 8.7.1 (Eigenvector Research Inc., Washington, USA) chemometric package.

3.2.1. Data preprocessing

Each sample was described by m = 31 variables (Table S1): the absolute value of the difference between the molecular descriptors of the two partner molecules was calculated, thus obtaining the predictor matrix $\boldsymbol{X}$ (181 × 31). The class membership was binary encoded (1: belonging to the class; 0: otherwise) in a dummy matrix $\boldsymbol{Y}$ (181 × 2) with each column representing one of the two modelled classes. The dataset was split in two subsets by using the Duplex algorithm [52]: 70% of the data were used as calibration set, $\boldsymbol{X}_{\mathbf{cal}}$ (127 × 31) and $\boldsymbol{Y}_{\mathbf{cal}}$ (127 × 2), whereas the remaining 30% were used as test set, $\boldsymbol{X}_{\mathbf{test}}$ (54 × 31) and $\boldsymbol{Y}_{\mathbf{test}}$ (54 × 2). The calibration set and the test set are reported in Table S2 and Table S3, respectively.

Before carrying out both exploratory multivariate data analysis and the computation of the supervised model, the calibration matrix $\boldsymbol{X}_{\mathbf{cal}}$ was preprocessed column-wise by performing mean centering and scaling to unit variance. Mean centering was applied on the response matrix $\boldsymbol{Y}_{\mathbf{cal}}$ to ensure the stability of the model.

3.2.2. Exploratory multivariate data analysis

PCA [53–55] was carried out preliminarily on the calibration set $\boldsymbol{X}_{\mathbf{cal}}$ to assess the distribution of the samples and to check for potential data structures. Reduction of data dimensionality is carried out through the linear combination of the original variables in a set of orthogonal ones, i.e., Principal Components (PCs), which identify the direction of maximum variance. This is summarized in the decomposition equation:

$$\boldsymbol{X}_{\mathbf{cal}} = \boldsymbol{T}\boldsymbol{P}^{\mathbf{T}} + \boldsymbol{E}$$

where $\boldsymbol{T}$ and $\boldsymbol{P}$ represent, respectively, the coordinates of the samples projected in the reduced space, i.e., the scores, and the weights each original variable has on a given PC, i.e., the loadings. The deviations from the model are accounted in the error matrix $\boldsymbol{E}$.

3.2.3. Supervised pattern recognition

PLS–DA [56,57] was used to discriminate pairs of partner molecules whose combination forms CCs from the ones giving BMs. PLS–DA is based on PLS regression [58]. Briefly, this supervised technique decomposes the predictor matrix $\boldsymbol{X}_{\mathbf{cal}}$ and the dependent variables matrix $\boldsymbol{Y}_{\mathbf{cal}}$ in a PCA-like way and imposes inner linear relationships between the $\boldsymbol{X}_{\mathbf{cal}}$ and the $\boldsymbol{Y}_{\mathbf{cal}}$ scores as follows:

$$\boldsymbol{U} = \boldsymbol{b}\boldsymbol{T}$$

where $\boldsymbol{T}$ and $\boldsymbol{U}$, are the $\boldsymbol{X}_{\mathbf{cal}}$ and the $\boldsymbol{Y}_{\mathbf{cal}}$ scores, respectively. This is accomplished by rotating the Latent Variable (LV) space of $\boldsymbol{X}_{\mathbf{cal}}$ through a weight matrix $\boldsymbol{W}$ in a way that maximizes the covariance between $\boldsymbol{T}$ and $\boldsymbol{U}$. The PLS regression model is summarized as:

$$\boldsymbol{Y}_{\mathbf{cal}} = \boldsymbol{X}_{\mathbf{cal}}\boldsymbol{B} + \boldsymbol{E}$$

where $\boldsymbol{E}$ is the error matrix and $\boldsymbol{B}$ is the pseudo-regression coefficient matrix expressed according to the following equation:

$$\boldsymbol{B} = \boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}\,\mathrm{diag}(\boldsymbol{b})\,\boldsymbol{Q}$$

where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are the $\boldsymbol{X}_{\mathbf{cal}}$ and the $\boldsymbol{Y}_{\mathbf{cal}}$ loadings, respectively.

In this case, the dependent variables in the $\boldsymbol{Y}_{\mathbf{cal}}$ matrix are defined as dummy variables, one for each modelled class, taking values of 1 if the sample belongs to the class and 0 otherwise. Current implementation of PLS–DA may differ on the basis of how the classification rule is defined. In this work, a pure discriminant rule (samples are assigned univocally to only one category) was applied, and thus a sample is assigned to the class for which the predicted response $\hat{\boldsymbol{y}}$ is the highest (i.e., $\hat{\boldsymbol{Y}}$ values are continuous and not dummy as they were codified).

The proper number of LVs was chosen according to the maximum accuracy ($ACC_{\%}$; i.e., the percentage of samples correctly assigned to the respective class) in leave-more-out cross validation, adopting a Venetian blinds cancellation scheme with 10 splits (blind thickness: 1). This operation was carried out by running a custom MATLAB routine. The performance of the classification model was evaluated both on the calibration and the test sets in terms of $ACC_{\%}$ as well as showing the confusion matrix. In addition, the sensitivity ($SEN_{\%}$, i.e., the percentage of samples within a class that were correctly assigned to their class) was calculated for both classes.

The importance of each predictor was estimated in terms of Variable Importance in Projection (VIP score) [56]. The VIP score of the j^{th} variable in the $\boldsymbol{X}$ space is defined as the component-wise sum of its PLS weight w_j on the f^{th} component multiplied by the fraction of variance of the $\boldsymbol{Y}$ explained by that component, according to the following equation:

$$VIP_j^2 = \frac{1}{SSY_{\mathrm{tot}}F}\sum_{f=1}^{F} w_{jf}^2 SSY_f J$$

where J is the number of variables in the $\boldsymbol{X}$ space and F is the number of LVs that were retained. Since:

$$\sum_{j=1}^{J} VIP_j^2 = J$$

the proposed threshold for determining whether a variable could be considered important is set to 1..

Finally, the predictive capability of the model was evaluated on an external set of $N = 58$ binary combinations of partner molecules. An overview of the involved samples is reported in Table S4 along with their estimated $\hat{y}$ values.

The number of entries in the BM and CC classes for training, test, and external validation sets is reported in Table 1.

Table 1. Number of CC and BM samples in the calibration, test, and external validation set. The last column reports the total number *N* of samples *per* set.

	CC	**BM**	***N***
Calibration set	71	56	127
Test set	30	24	54
External validation set	31	27	58

4. Results and Discussion

The cocrystallization experiments were carried out mechanochemically by manual neat grinding of the two substances. This method was selected among many possible others due to its simplicity and promptness, allowing us to screen several molecular pairs in a standardized way [21,59]. The classification as BM or CC was assessed by PXRD patterns (available in Section 3 of the Supplementary Material). Cocrystals already present in the CSD were also included in our dataset.

4.1. Data analysis

Data handling prior to analysis can affect the way the model is trained, thus having consequences on its interpretation.

Since samples are the result of the combination of two partner molecules, each one described by its own set of descriptors, the concatenation strategy, i.e., listing the descriptors of the first coformer followed by those of the second coformer, was discarded due to the lack of commutation between the two sets of descriptors. In fact, the order in which the molecular descriptors are listed represents an *a priori* decision on which compound is acting as molecule of interest or partner molecule. In our case this would be sub-optimal since many of the molecules in our dataset could assume both roles. Therefore, in order to address the problem described above a commutative strategy capable of avoiding the production of indeterminate forms should be chosen. Considering that in the dataset many constitutional molecular descriptors were characterized by a few non-zero values, the calculation of both products and ratios between the descriptors was discarded since additional zero values or indeterminate forms could be generated. Also, the division is non-commutative.

In order to combine the two partner molecules without imposition on their role, the absolute value of the difference between the molecular descriptors of the partner molecules was calculated and used to describe each binary combination [38,39], giving maximum flexibility to the model. The information achieved is still relevant and easily interpretable since it is related to the dissimilarity between the descriptors. In fact, absolute value of the differences between descriptors for each case are the elements of the Manhattan distance [60], one of the possible indexes to account for the dissimilarity between cases in a multivariate way.

In the present study, the use of basic linear modelling methods was preferred with respect to non-linear modelling, such as ANN [61,62], to keep the calculations as simple as possible, and to ensure a certain degree of interpretability of the results. Furthermore, the number of samples is too limited to ensure proper tuning of the ANN hyperparameters.

For a fruitful discussion, samples and variables are reported in the text according to their identification number as follows: i) samples are written in plain text; ii) variables are underlined. The key is available in Tables S1–S3.

4.1.1. Exploratory multivariate data analysis

After data preprocessing, PCA was used in an exploratory way to assess the presence of potential data structures in the calibration set.

Four PCs were retained explaining 67% of the variance. As shown in the score plots (Figure 1), a mild separation was present in the PC 2 *vs.* PC 3 score plot, with the groups separated by the bisecting line of the II and the IV quadrant. In addition, most of the CC samples occupied the III quadrant of the PC 2 *vs.* PC 3 score plot and were, in general, less scattered than BM samples.

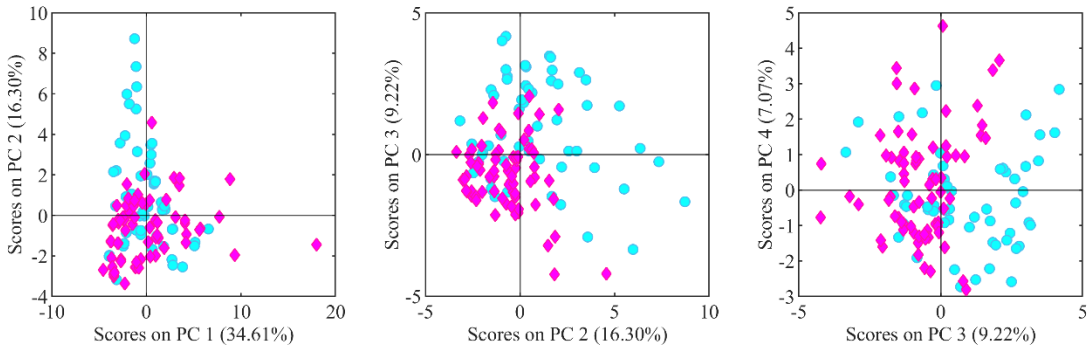


Figure 1. Scores of the samples on the first 4 PCs: PC 1 *vs.* PC 2 (left), PC 2 *vs.* PC 3 (center), PC 3 *vs.* PC 4 (right). The fraction of variance explained by a given component is reported as a percentage value in parenthesis on the corresponding axis. Samples are marked according to their class (magenta diamonds: cocrystals; cyan circles: binary mixtures).

As for the loading plots depicted in Figure 2, it can be observed that PC 1 explains the features related both to differences in molecular dimensions (e.g., 1, 2, 3, 9, 16, 29) and connectivity (23, 24, 25, 26). PC 2 considers the dissimilarities in the electronic properties (e.g., 6, 7, 19, 22, 30) of the two partner molecules as well as the difference in their number of heteroatoms (27). PC 3 accounts for more specific features, such as differences in molecular complexity (21), molecular refractivity (31), and energy (15). Lastly, PC 4 considers the dissimilarity in: i) component of the dipole along the *x* axis (11), ii) total dipole (14) and iii) minimum of the MEP surface (20). Regarding the samples belonging to the CC class, these compounds were characterized by partner molecules with a similar behavior in terms of molecular complexity (21) and octanol/water partition coefficient (10).

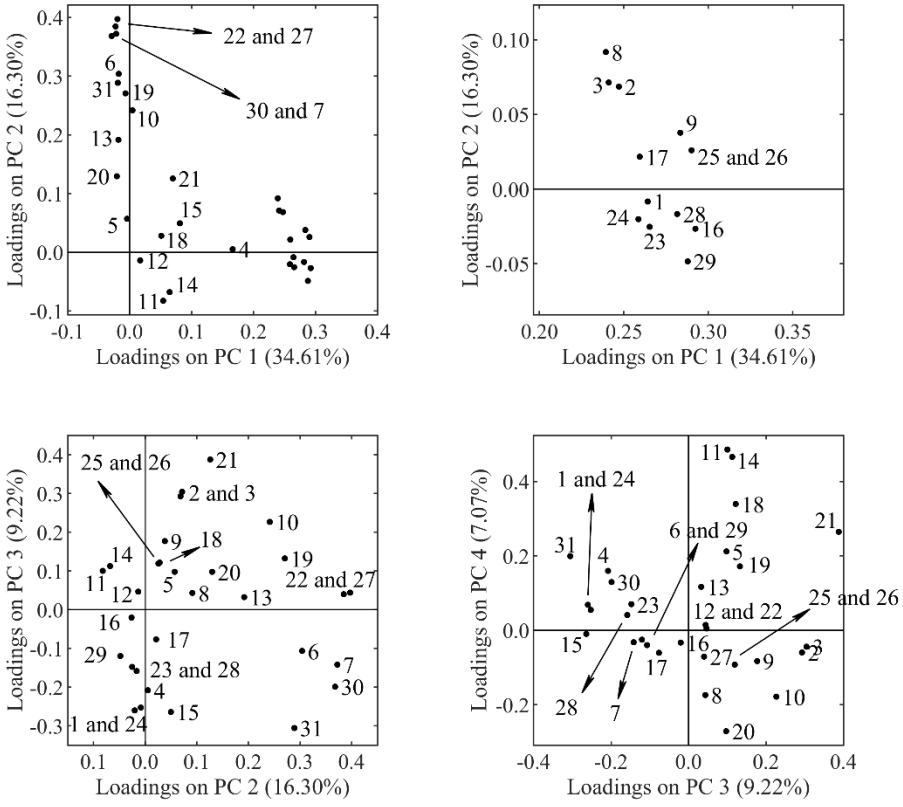


Figure 2. Loading plots related to the PCA decomposition. PC 1 *vs.* PC 2 (top-left), magnification of the PC 1 *vs.* PC 2 (top-right), PC 2 *vs.* PC 3 (bottom-left), PC 3 *vs.* PC 4 (bottom-right). The fraction of variance explained by a given component is reported as a percentage value in parenthesis on the corresponding axis.

4.1.2. Supervised pattern recognition

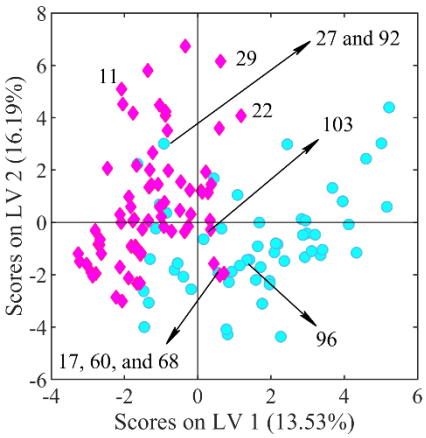
The relationship between the class membership and the variables was exploited by means of PLS–DA. Six LVs were retained according to maximum $ACC_{\%}$ in cross validation. The PLS–DA model captured the 72% and 69% of the variance of the $\boldsymbol{X}_{\mathbf{cal}}$ and $\boldsymbol{Y}_{\mathbf{cal}}$, respectively. The values of the predicted response $\hat{\boldsymbol{y}}$ in cross validation related to CC samples are plotted in Figure S1.

A summary of the performance of the obtained model is reported in the confusion matrix (Table 2), whereas a graphical representation of the estimated and predicted values $\hat{\boldsymbol{y}}$ for the CC class is reported in Figure S2. As reported in Table 2, all the BM samples belonging to the test set were correctly classified except for 5 samples, whereas only 3 out of 30 CC samples were wrongly assigned to the BM class, obtaining a $ACC_{\%}$ of 85%. Similarly, a high $ACC_{\%}$ of 92% was obtained when the samples belonging to the calibration set were predicted by the model. The achieved results are extremely satisfactory, allowing for the *a priori* selection of the partner molecules required for the synthesis of novel cocrystals.

Table 2. Confusion matrix of the calibration and the test sets for the PLS–DA model. The second-last line shows the $SEN_{\%}$ for the modelled classes and last line shows the $ACC_{\%}$ for the calibration and the test sets.

	Calibration set		**Test set**	
	Predicted as		Predicted as	
	CC	BM	CC	BM
True CC	66	5	27	3
True BM	5	51	5	19
SEN%	93%	91%	90%	79%
$ACC_{\%}$	92%		85%	

346 The distribution of the samples in the reduced space of the LVs can be observed by inspecting the score plot (Figure 3),
347 whereas information regarding suspicious and/or influential samples can be retrieved by the squared residuals *Q vs.*
348 Hotelling's T^2 and the residuals *vs.* leverage plots (Figure 4).
349



350

351 **Figure 3.** Distribution of the samples in the score space (LV 1 *vs.* LV 2). The fraction of variance explained by a given
352 component is reported as a percentage value in parenthesis on the corresponding axis. Samples are marked according to their
353 class (magenta diamonds: cocrystals; cyan circles: binary mixtures).
354

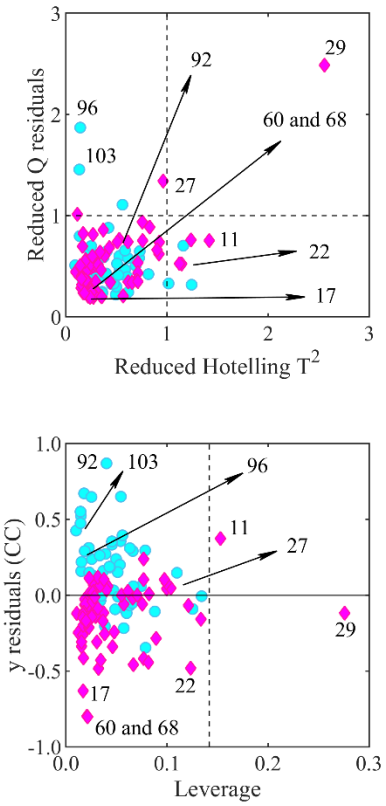


Figure 4. Reduced squared residuals *Q vs.* reduced Hotelling's T^2 plot, the dashed horizontal and vertical lines show the amplitude of the 95% confidence interval for both parameters (top). Residuals of CC samples *vs.* leverage plot, the dashed vertical line shows the leverage limit (bottom). Samples are marked according to their class (magenta diamonds: cocrystals; cyan circles: binary mixtures).

The maximum separation in the score space was provided by the first two LVs, with the CC samples located mainly at negative scores on LV 1. By contrast, BM samples were more scattered and localized mostly on positive scores on LV 1.

A peculiar behavior was observed for samples 27 and 29. Sample 27 showed squared residual *Q* outside the 95% confidence interval, whereas sample 29 showed both high Hotelling's T^2 and high squared residuals *Q* together with high leverage in the $\boldsymbol{Y}_{\mathbf{cal}}$ space. These CC samples were obtained by pairing fatty acids (lauric acid, 27, and palmitic acid, 29) with a low-molecular weight coformer, i.e., nicotinamide. The variables responsible for this behavior can be related to the discrepancy in molecular dimensions between the two partner molecules.

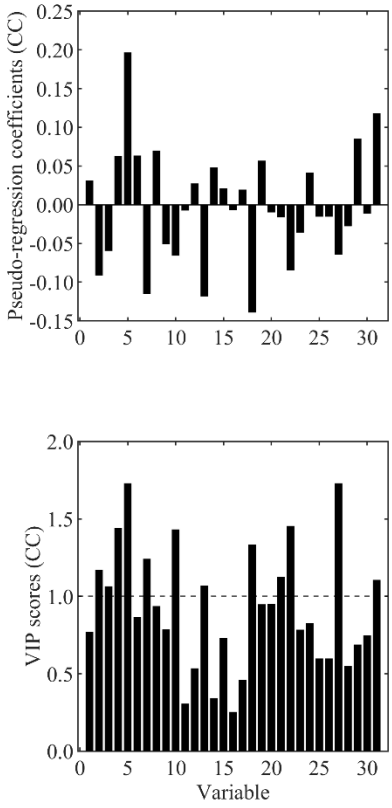
The adipic acid/hexamethylenetetramine (11) CC sample was characterized by high leverage and high Hotelling's T^2 value, due to different behavior in terms of ruotable bonds and number of rings present in the structure. A similar behaviour, in terms of difference in the number of rings, was observed also by two additional non-influential BM samples based on hexamethylenetetramine paired with limonene (96) and menthone (103). In addition, these samples held high squared residuals *Q* and, therefore, were characterized by features that did not align with the ones of the other samples.

Finally, the ferulic acid/pyrazine (22) CC sample showed Hotelling's T^2 values outside the 95% confidence interval due to the different behaviour of the partner molecules in terms of molecular weight, connectivity, surface area, and electronic properties, i.e., isotropic and anisotropic polarizability and number of valence electrons.

Suspicious samples 22 (ferulic acid/pyrazine, CC) and 103 (menthone/hexamethylenetetramine, BM) appeared also to have high residuals in absolute value and were wrongly assigned to their class in cross validation. Along with them, also three not anomalous CC samples and one not anomalous BM samples, namely cinnamaldehyde/4-hydroxybenzoic acid (17), carvacrol/nicotinamide (60), thymol/tetramethylpyrazine (68) and eugenol/pyrazine (92), respectively, were misclassified. In fact, their features were inversely related to their respective class. Nevertheless, the exclusion of the samples discussed above from the calibration set would not have produced any difference in terms of rotation of the LV space due to their low leverage.

The correlation between class membership, coded in the $\boldsymbol{Y}_{\mathbf{cal}}$, and the predictors contained in the $\boldsymbol{X}_{\mathbf{cal}}$ space can be observed in the PLS weights plot (Figure S3).

Information regarding the contribution of each predictor involved in the discrimination of the modelled classes can be inferred by inspecting the pseudo-regression coefficients and the VIP score plots, reported in Figure 5. The latter parameter denotes the relative importance of each predictor of the $\boldsymbol{X}_{\mathbf{cal}}$ space in the PLS–DA model in explaining the class membership encoded in the $\boldsymbol{Y}_{\mathbf{cal}}$ and may guide variable selection. Generally, a variable can be considered important with a VIP score > 1; by contrast, a VIP score significantly lower than 1 indicates that a given variable is a good candidate for exclusion. According to the negative sign of the pseudo-regression coefficients related to CC class, it can be stated that significant differences in descriptors related to polarizability and exposed surface, such as atom and bond count (2, 3), octanol-water partition coefficient (10), topological polar surface area (22) and heteroatom count (27) are likely to prevent the formation of a cocrystal. It should be noted that this consideration agrees with what has emerged earlier from unsupervised modelling, and it is largely in agreement with widely applied rules of thumb in crystal engineering.



394

395 **Figure 5.** Pseudo-regression coefficients plot for the CC class (top). VIP scores related to each variable included in the $\boldsymbol{X}_{\mathbf{cal}}$
396 space for the CC class, the significance threshold of 1 is depicted as a dashed horizontal line (bottom). Both plots report the
397 variable identification number on the *x* axis.

398

399 Finally, a reduced PLS–DA model based only on the important variables in agreement with the VIP scores was computed
400 because of its ease of interpretation. A 4 LVs model was calculated according to the maximum $ACC_{\%}$ in cross validation
401 obtaining a classification performance very similar to that achieved by including all descriptors. The weights plot of the
402 reduced model is provided in Figure 6, in which the weights of the variables involved in the discrimination are located on the
403 positive and negative sides of the first LV.

404

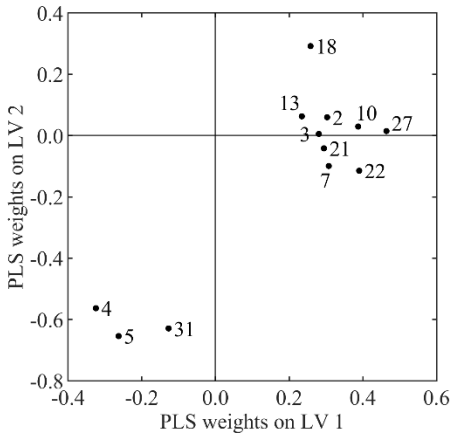


Figure 6. PLS weights of the variables for LV 1 *vs.* LV 2 of the reduced model.

4.2. Benchmarking on an external validation set

The external validation set consists of 27 pairs classified experimentally as BMs and 31 pairs classified as CCs, i.e., 11 with our mechanochemistry/PXRD protocol and 20 retrieved from the CSD. The latter ones are reported in Section 3 of the Supplementary Information together with their CSD refcode and the reference to the original publications.

A graphical representation of the predicted values $\hat{\boldsymbol{y}}$ for the CC class is reported in Figure 7, as well as the squared residuals *Q vs.* Hotelling's T^2 plot. Although there were some samples that did not conform to the model space, not all of them have been systematically misclassified. The confusion matrix is reported in Table 3 to summarize the results. In total, about 74% of the predictions were in agreement with the experimental results**.**

Table 3. Confusion matrix of the external validation set for the PLS–DA model. The second-last line shows the $SEN_{\%}$ for the modelled classes and last line shows the $ACC_{\%}$.

	External validation set	
	Predicted as	
	CC	BM
Experimental CC	29	2
Experimental BM	13	14
$SEN_{\%}$	94%	52%

$ACC_{\%}$	74%

Specifically, 29 CC cases out of 31 were correctly classified. On the other hand, there were 13 false positive results and 14 cases in which the pairs were correctly identified as BMs. The model appears therefore to be quite conservative in discarding the possibility of cocrystallization; hence, fewer potential new materials could be overlooked. In addition, the fraction of misclassified CC samples (see Table 2 and Table 3) does not change significantly in the external validation set with respect to the test set.

As shown in Figure 7, Table 3 and Table S4, only 2 pairs of partner molecules predicted as BMs actually formed CCs, thus producing 2 false negative results. It can be stated that these results were borderline cases when the estimated error [63] on the prediction was taken into consideration (data not shown): in fact their $\hat{y}$ value was close to the classification threshold of 0.53. The same is true for 6 of the false positive cases, while in the remaining 7 cases the model confidently classified the pairs as CCs, in contrast to our experimental results. These findings could be ascribed to the use of different preparation methods other than mechanochemical grinding. This hypothesis is somehow supported by the fact that 19 out of the 20 CCs retrieved from the CSD, then prepared with a variety of methods, were correctly identified.

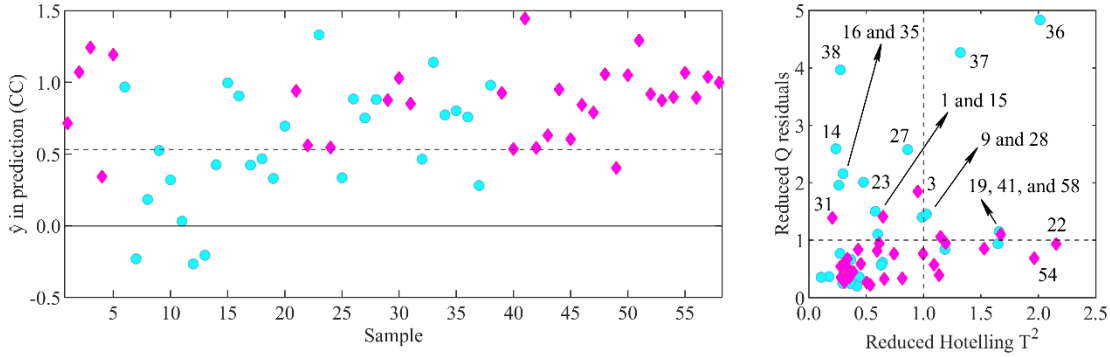


Figure 7. $\hat{y}$ (CC samples) for the external validation set, the dashed horizontal line shows the hard classification threshold of 0.53 (left). Reduced squared residuals *Q vs.* reduced Hotelling's T^2 plot, the dashed horizontal and vertical lines show the amplitude of the 95% confidence interval for both parameters (right). Samples are marked according to their class (magenta diamonds: cocrystals; cyan circles: binary mixtures).

5. Conclusion

This study highlighted the ability of a simple PLS–DA model to predict cocrystal formation without any *a priori* knowledge of the specific role of the involved partner molecules. Information deriving from both successful and unsuccessful cocrystallization experiments was used. The major advantage of the proposed methodology relies on the reduction of the experimental effort required for both the synthesis and characterization of new crystalline structures.

The model allows us to predict cocrystallization propensity with a 74% of the predictions in agreement with the experimental results. Considering that the model was obtained on a training set spanning different molecular characteristics, it can be stated that it is suitable for a fairly general applicability.

Indeed, once in possess of the set of chemical descriptors for the molecules of interest, it is sufficient to calculate the absolute value of their difference and perform a linear combination using the pseudo-regression coefficients to obtain a prediction on cocrystal formation. The precalculated values for the set of descriptors comprising 2193 GRAS molecules are available in the Supplementary Material. By applying the proposed methodology, ten new cocrystals were discovered and an additional compound was obtained by chance.

Another figure of merit of the proposed approach is the possibility of understanding through the inspection of PLS weights how the degree of similarity in terms of molecular features of the two partner molecules is correlated with the possibility of obtaining a cocrystal.

On a closing note, we would like to strongly encourage scientists to report failed attempts at cocrystallization along with the technique used, as access to this information could play a pivotal role in refining predictive models, making them less sensitive to selective reporting bias.

Acknowledgements

This work was funded by the Ministero delle Politiche Agricole, Alimentari, Forestali e del Turismo (MIPAAFT) by granting for the project PAC/Packaging Attivo Cristallino. This work has benefited from the framework of the COMP-HUB initiative, funded by the "Department of Excellence" program of the Italian Ministry for Education, University and Research (MIUR, 2018–2022).

Authorship Contribution Statement

Fabio Fornari: Conceptualization (equal), Formal Analysis (lead), Methodology (equal), Validation (equal), Visualization (equal), Writing – Original Draft Preparation (lead), Writing – Review & Editing (equal).

Fabio Montisci: Investigation (equal), Data Curation (lead), Validation (equal), Visualization (equal), Writing – Original Draft Preparation (supporting), Writing – Review & Editing (equal).

Federica Bianchi: Conceptualization (equal), Methodology (lead), Resources (equal), Supervision (equal), Writing – Review & Editing (equal).

Marina Cocchi: Conceptualization (equal), Methodology (equal), Resources (equal), Software (lead), Writing – Review & Editing (equal).

Claudia Carraro: Investigation (equal), Data Curation (equal).

Francesca Cavaliere: Software (equal), Writing – Review & Editing (supporting).

Pietro Cozzini: Software (equal), Writing – Review & Editing (supporting).

Francesca Peccati: Software (supporting), Writing – Review & Editing (supporting).

Paolo P. Mazzeo: Conceptualization (equal), Investigation (supporting), Data Curation (supporting), Validation (supporting), Writing – Review & Editing (supporting).

Nicolò Riboni: Investigation (supporting), Writing – Review & Editing (supporting).

Maria Careri: Supervision (equal), Funding Acquisition (equal), Resources (equal), Writing – Review & Editing (equal).

Alessia Bacchi: Conceptualization (lead), Funding Acquisition (equal), Resources (equal), Supervision (equal), Writing – Review & Editing (equal).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E.V.R. Campos, P.L.F. Proença, J.L. Oliveira, M. Bakshi, P.C. Abhilash, L.F. Fraceto, Use of botanical insecticides for sustainable agriculture: Future perspectives, Ecol. Indic. 105 (2019) 483–495. doi:10.1016/J.ECOLIND.2018.04.038.

[2] Y. Kourkoutas, M. Angane, S. Swift, K. Huang, C.A. Butts, S.Y. Quek, Essential Oils and Their Major Components: An Updated Review on Antimicrobial Activities, Mechanism of Action and Their Potential Application in the Food Industry, Foods 2022, Vol. 11, Page 464. 11 (2022) 464. doi:10.3390/FOODS11030464.

[3] N.S. Singh, R. Sharma, T. Parween, P.K. Patanjali, Pesticide Contamination and Human Health Risk Factor, Mod. Age Environ. Probl. Their Remediat. (2018) 49–68. doi:10.1007/978-3-319-64501-8_3.

[4] Climate Change, Intercropping, Pest Control and Beneficial Microorganisms, Springer Netherlands, 2009. doi:10.1007/978-90-481-2716-0.

[5] F. Bakkali, S. Averbeck, D. Averbeck, M. Idaomar, Biological effects of essential oils - A review, Food Chem. Toxicol. 46 (2008) 446–475. doi:10.1016/j.fct.2007.09.106.

[6] M. Alonso-Gato, G. Astray, J.C. Mejuto, J. Simal-Gandara, Essential Oils as Antimicrobials in Crop Protection, Antibiotics. 10 (2021) 34. doi:10.3390/antibiotics10010034.

[7] F. Bianchi, F. Fornari, N. Riboni, C. Spadini, C.S. Cabassi, M. Iannarelli, C. Carraro, P.P. Mazzeo, A. Bacchi, S. Orlandini, S. Furlanetto, M. Careri, Development of novel cocrystal-based active food packaging by a Quality by Design approach, Food Chem. 347 (2021) 129051. doi:10.1016/j.foodchem.2021.129051.

[8] A.T.H. Mossa, Green Pesticides: Essential oils as biopesticides in insect-pest management, J. Environ. Sci. Technol. 9 (2016) 354–378. doi:10.3923/jest.2016.354.378.

[9] S. Sharma, S. Barkauskaite, A.K. Jaiswal, S. Jaiswal, Essential oils as additives in active food packaging, Food Chem. (2020) 128403. doi:10.1016/j.foodchem.2020.128403.

[10] Food and Drug Administration, 54960 Federal Register / Vol . 81 , No . 159 / Wednesday , August 17 , 2016 / Rules and Regulations, 81 (2016) 54960–55055.

[11] J. Wieczyńska, I. Cavoski, Antimicrobial, antioxidant and sensory features of eugenol, carvacrol and trans-anethole in active packaging for organic ready-to-eat iceberg lettuce, Food Chem. 259 (2018) 251–260. doi:10.1016/j.foodchem.2018.03.137.

[12] L. Pavoni, D.R. Perinelli, G. Bonacucina, M. Cespi, G.F. Palmieri, An overview of micro-and nanoemulsions as

vehicles for essential oils: Formulation, preparation and stability, Nanomaterials. 10 (2020) 135. doi:10.3390/nano10010135.

[13] G.R. Desiraju, Crystal Engineering: A Holistic View, Angew. Chemie Int. Ed. 46 (2007) 8342–8356. doi:10.1002/anie.200700534.

[14] A. Bacchi, P.P. Mazzeo, Cocrystallization as a tool to stabilize liquid active ingredients, Crystallogr. Rev. (2021) 1–22. doi:10.1080/0889311X.2021.1978079.

[15] D. Balestri, P.P. Mazzeo, R. Perrone, F. Fornari, F. Bianchi, M. Careri, A. Bacchi, P. Pelagatti, Deciphering the Supramolecular Organization of Multiple Guests Inside a Microporous MOF to Understand their Release Profile, Angew. Chemie Int. Ed. 60 (2021) 10194–10202. doi:10.1002/ANIE.202017105.

[16] D. Balestri, P.P. Mazzeo, C. Carraro, N. Demitri, P. Pelagatti, A. Bacchi, Stepwise Evolution of Molecular Nanoaggregates Inside the Pores of a Highly Flexible Metal–Organic Framework, Angew. Chemie. 131 (2019) 17503–17511. doi:10.1002/ange.201907621.

[17] P.P. Mazzeo, S. Canossa, C. Carraro, P. Pelagatti, A. Bacchi, Systematic coformer contribution to cocrystal stabilization: energy and packing trends, CrystEngComm. 22 (2020) 7341–7349. doi:10.1039/D0CE00291G.

[18] P.P. Mazzeo, C. Carraro, A. Arns, P. Pelagatti, A. Bacchi, Diversity through Similarity: A World of Polymorphs, Solid Solutions, and Cocrystals in a Vial of 4,4′-Diazopyridine, Cryst. Growth Des. 20 (2020) 636–644. doi:10.1021/acs.cgd.9b01052.

[19] P.P. Mazzeo, C. Carraro, A. Monica, D. Capucci, P. Pelagatti, F. Bianchi, S. Agazzi, M. Careri, A. Raio, M. Carta, F. Menicucci, M. Belli, M. Michelozzi, A. Bacchi, Designing a Palette of Cocrystals Based on Essential Oil Constituents for Agricultural Applications, ACS Sustain. Chem. Eng. 7 (2019) 17929–17940. doi:10.1021/acssuschemeng.9b04576.

[20] D. Capucci, D. Balestri, P.P. Mazzeo, P. Pelagatti, K. Rubini, A. Bacchi, Liquid Nicotine Tamed in Solid Forms by Cocrystallization, Cryst. Growth Des. 17 (2017) 4958–4964. doi:10.1021/acs.cgd.7b00887.

[21] P.P. Mazzeo, M. Pioli, F. Montisci, A. Bacchi, P. Pelagatti, Mechanochemical Preparation of Dipyridyl-Naphthalenediimide Cocrystals: Relative Role of Halogen-Bond and π–π Interactions, Cryst. Growth Des. 21 (2021) 5687–5696. doi:10.1021/acs.cgd.1c00531.

[22] N.K. Duggirala, M.L. Perry, Ö. Almarsson, M.J. Zaworotko, Pharmaceutical cocrystals: Along the path to improved medicines, Chem. Commun. 52 (2016) 640–655. doi:10.1039/c5cc08216a.

[23] J.W. Steed, The role of co-crystals in pharmaceutical design., Trends Pharmacol. Sci. 34 (2013) 185–93. doi:10.1016/j.tips.2012.12.003.

[24] J.G.P. Wicker, L.M. Crowley, O. Robshaw, E.J. Little, S.P. Stokes, R.I. Cooper, S.E. Lawrence, Will they co-crystallize?, CrystEngComm. 19 (2017) 5336–5340. doi:10.1039/C7CE00587C.

[25] A. Bacchi, D. Capucci, M. Giannetto, M. Mattarozzi, P. Pelagatti, N. Rodriguez-Hornedo, K. Rubini, A. Sala, Turning Liquid Propofol into Solid (without Freezing It): Thermodynamic Characterization of Pharmaceutical Cocrystals Built with a Liquid Drug, Cryst. Growth Des. 16 (2016) 6547–6555. doi:10.1021/acs.cgd.6b01241.

[26] Y. Xiao, L. Zhou, H. Hao, Y. Bao, Q. Yin, C. Xie, Cocrystals of propylthiouracil and nutraceuticals toward sustained-release: Design, structure analysis, and solid-state characterization, Cryst. Growth Des. 21 (2021) 1202–1217. doi:10.1021/acs.cgd.0c01519.

[27] O. Shemchuk, S. D'Agostino, C. Fiore, V. Sambri, S. Zannoli, F. Grepioni, D. Braga, Natural Antimicrobials Meet a Synthetic Antibiotic: Carvacrol/Thymol and Ciprofloxacin Cocrystals as a Promising Solid-State Route to Activity Enhancement, Cryst. Growth Des. 20 (2020) 6796–6803. doi:10.1021/acs.cgd.0c00900.

[28] M.D. Perera, J. Desper, A.S. Sinha, C.B. Aakeröy, Impact and importance of electrostatic potential calculations for predicting structural patterns of hydrogen and halogen bonding, CrystEngComm. 18 (2016) 8631–8636. doi:10.1039/c6ce02089e.

[29] M.C. Etter, Encoding and Decoding Hydrogen-Bond Patterns of Organic Compounds, Acc. Chem. Res. 23 (1990) 120–126. doi:10.1021/ar00172a005.

[30] C.A. Hunter, Quantifying intermolecular interactions: Guidelines for the molecular recognition toolbox, Angew. Chemie - Int. Ed. 43 (2004) 5310–5324. doi:10.1002/anie.200301739.

[31] M. Karimi-Jafari, L. Padrela, G.M. Walker, D.M. Croker, Creating cocrystals: A review of pharmaceutical cocrystal preparation routes and applications, Cryst. Growth Des. 18 (2018) 6370–6387. doi:10.1021/acs.cgd.8b00933.

[32] N. Issa, P.G. Karamertzanis, G.W.A. Welch, S.L. Price, Can the formation of pharmaceutical cocrystals be computationally predicted? I. Comparison of lattice energies, Cryst. Growth Des. 9 (2009) 442–453. doi:10.1021/cg800685z.

[33] M.A. Mohammad, A. Alhalaweh, S.P. Velaga, Hansen solubility parameter as a tool to predict cocrystal formation, Int. J. Pharm. 407 (2011) 63–71. doi:10.1016/j.ijpharm.2011.01.030.

[34] L. Fábián, Cambridge Structural Database Analysis of Molecular Complementarity in Cocrystals, Cryst. Growth

Des. 9 (2009) 1436–1443. doi:10.1021/cg800861m.

[35] C.R. Groom, I.J. Bruno, M.P. Lightfoot, S.C. Ward, The Cambridge structural database, Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater. 72 (2016) 171–179. doi:10.1107/S2052520616003954.

[36] J. Devogelaer, H. Meekes, P. Tinnemans, E. Vlieg, R. Gelder, Co- crystal Prediction by Artificial Neural Networks**, Angew. Chemie Int. Ed. 59 (2020) 21711–21718. doi:10.1002/anie.202009467.

[37] D. Wang, Z. Yang, B. Zhu, X. Mei, X. Luo, Machine-Learning-Guided Cocrystal Prediction Based on Large Data Base, Cryst. Growth Des. 20 (2020) 6610–6621. doi:10.1021/acs.cgd.0c00767.

[38] M. Przybyłek, T. Jeliński, J. Słabuszewska, D. Ziółkowska, K. Mroczyńska, P. Cysewski, Application of Multivariate Adaptive Regression Splines (MARSplines) Methodology for Screening of Dicarboxylic Acid Cocrystal Using 1D and 2D Molecular Descriptors, Cryst. Growth Des. 19 (2019) 3876–3887. doi:10.1021/acs.cgd.9b00318.

[39] M. Przybyłek, P. Cysewski, Distinguishing Cocrystals from Simple Eutectic Mixtures: Phenolic Acids as Potential Pharmaceutical Coformers, Cryst. Growth Des. 18 (2018) 3524–3534. doi:10.1021/acs.cgd.8b00335.

[40] A. Vriza, A.B. Canaj, R. Vismara, L.J. Kershaw Cook, T.D. Manning, M.W. Gaultois, P.A. Wood, V. Kurlin, N. Berry, M.S. Dyer, M.J. Rosseinsky, One class classification as a practical approach for accelerating π–π co-crystal discovery, Chem. Sci. 12 (2021) 1702–1719. doi:10.1039/d0sc04263c.

[41] M.E. Mswahili, M.J. Lee, G.L. Martin, J. Kim, P. Kim, G.J. Choi, Y.S. Jeong, Cocrystal prediction using machine learning models and descriptors, Appl. Sci. 11 (2021) 1–12. doi:10.3390/app11031323.

[42] H. Moriwaki, Y.S. Tian, N. Kawashita, T. Takagi, Mordred: A molecular descriptor calculator, J. Cheminform. 10 (2018) 4. doi:10.1186/s13321-018-0258-y.

[43] J.G.P. Wicker, R.I. Cooper, Will it crystallise? Predicting crystallinity of molecular materials, CrystEngComm. 17 (2015) 1927–1934. doi:10.1039/c4ce01912a.

[44] I.J. Bruno, J.C. Cole, P.R. Edgington, M. Kessler, C.F. Macrae, P. McCabe, J. Pearson, R. Taylor, New software for searching the Cambridge Structural Database and visualizing crystal structures, Acta Crystallogr. Sect. B Struct. Sci. 58 (2002) 389–397. doi:10.1107/S0108768102003324.

[45] C.F. Macrae, I.J. Bruno, J.A. Chisholm, P.R. Edgington, P. McCabe, E. Pidcock, L. Rodriguez-Monge, R. Taylor, J. van de Streek, P.A. Wood, Mercury CSD 2.0 – new features for the visualization and investigation of crystal structures, J. Appl. Crystallogr. 41 (2008) 466–470. doi:10.1107/S0021889807067908.

[46] M. Baroni, G. Cruciani, S. Sciabola, F. Perruccio, J.S. Mason, A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application, J. Chem. Inf. Model. 47 (2007) 279–294. doi:10.1021/ci600253e.

[47] Sybyl 8.1, Tripos International, St. Louis, USA, 2009.

[48] M. Clark, R.D. Cramer, N. Van Opdenbosch, Validation of the general purpose tripos 5.2 force field, J. Comput. Chem. 10 (1989) 982–1012. doi:10.1002/jcc.540100804.

[49] Schrödinger LLC, The PyMOL Molecular Graphics System, PyMOL Mol. Graph. Syst. Version 2.0. (2010).

[50] G. Landrum, RDKit: Open-Source Cheminformatics Software, Http://Www.Rdkit.Org/. (2021).

[51] G.A. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Petersson, Gaussian 16, Rev. A.03, Gaussian, Inc., Wallingford, CT. (2016).

[52] R.D. Snee, Validation of Regression Models: Methods and Examples, Technometrics. 19 (1977) 415–428. doi:10.1080/00401706.1977.10489581.

[53] M. Li Vigni, C. Durante, M. Cocchi, Exploratory Data Analysis, in: Data Handl. Sci. Technol., Elsevier Ltd, 2013: pp. 55–126. doi:10.1016/B978-0-444-59528-7.00003-X.

[54] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52. doi:10.1016/0169-7439(87)80084-9.

[55] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1039/c3ay41907j.

[56] M. Cocchi, A. Biancolillo, F. Marini, Chemometric Methods for Classification and Feature Selection, in: Compr. Anal. Chem., Elsevier B.V., 2018: pp. 265–299. doi:10.1016/bs.coac.2018.08.006.

[57] U.G. Indahl, H. Martens, T. Næs, From dummy regression to prior probabilities in PLS-DA, J. Chemom. 21 (2007) 529–536. doi:10.1002/cem.1061.

[58] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, in: Chemom. Intell. Lab. Syst., Elsevier, 2001: pp. 109–130. doi:10.1016/S0169-7439(01)00155-1.

[59] D. Braga, E. Dichiarante, F. Grepioni, G.I. Lampronti, L. Maini, P.P. Mazzeo, S. D'Agostino, Mechanical Preparation of Crystalline Materials. An Oxymoron?, in: Supramol. Chem., John Wiley & Sons, Ltd, Chichester, UK, 2012. doi:10.1002/9780470661345.smc115.

[60] J. Fürnkranz, P.K. Chan, S. Craw, C. Sammut, W. Uther, A. Ratnaparkhi, X. Jin, J. Han, Y. Yang, K. Morik, M. Dorigo, M. Birattari, T. Stützle, P. Brazdil, R. Vilalta, C. Giraud-Carrier, C. Soares, J. Rissanen, R.A. Baxter, I.

Bruha, R.A. Baxter, G.I. Webb, L. Torgo, A. Banerjee, H. Shan, S. Ray, P. Tadepalli, Y. Shoham, R. Powers, Y. Shoham, R. Powers, G.I. Webb, S. Ray, S. Scott, H. Blockeel, L. De Raedt, Manhattan Distance, Encycl. Mach. Learn. (2011) 639–639. doi:10.1007/978-0-387-30164-8_506.

[61] L.E. Juarez-Orozco, O. Martinez-Manzanera, S. V. Nesterov, S. Kajander, J. Knuuti, The machine learning horizon in cardiac hybrid imaging, Eur. J. Hybrid Imaging. 2 (2018) 1–15. doi:10.1186/S41824-018-0033-3/FIGURES/5.

[62] A. Singh, N. Thakur, A. Sharma, A review of supervised machine learning algorithms, in: 3rd Int. Conf. Comput. Sustain. Glob. Dev., Institute of Electrical and Electronics Engineers, New Delhi, India, 2016: pp. 1310–1315. https://ieeexplore.ieee.org/abstract/document/7724478.

[63] N.M. Faber, R. Bro, Standard error of prediction for multiway PLS: 1. Background and a simulation study, Chemom. Intell. Lab. Syst. 61 (2002) 133–149. doi:10.1016/S0169-7439(01)00204-0.