CRYSTAL GROWTH & DESIGN

pubs.acs.org/crystal

Article

Speeding Up the Cocrystallization Process: Machine Learning-Combined Methods for the Prediction of Multicomponent Systems

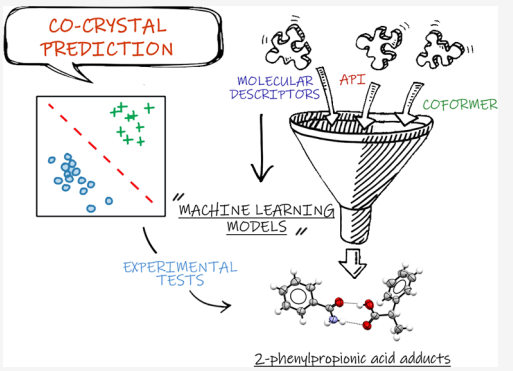
Rebecca Birolo, Federica Bravetti, Eugenio Alladio, Emanuele Priola, Gianluca Bianchini, Rubina Novelli, Andrea Aramini, Roberto Gobetto,* and Michele R. Chierotti*

Cite This: *Cryst. Growth Des.* 2023, 23, 7898−7911

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Pharmaceutical cocrystals are crystalline materials composed of at least two molecules, *i.e.*, an active pharmaceutical ingredient (API) and a coformer, assembled by noncovalent forces. Cocrystallization is successfully applied to improve the physicochemical properties of APIs, such as solubility, dissolution profile, pharmacokinetics, and stability. However, choosing the ideal coformer is a challenging task in terms of time, efforts, and laboratory resources. Several computational tools and machine learning (ML) models have been proposed to mitigate this problem. However, the challenge of achieving a robust and generalizable predictive method is still open. In this study, we propose a new approach to quickly predict the formation of cocrystals, employing partial least squares-discriminant analysis, random forest, and neural networks. The models were based on the data sets of 13 structurally different APIs with both positive and negative cocrystallization outcomes. At the same time, the features were specially selected from a variety of molecular descriptors to explain the phenomenon of the cocrystallization. All of the proposed ML models showed a cross-validation accuracy higher than 83%. Furthermore, this approach was successfully applied to drive the cocrystallization experimental tests of 2-phenylpropionic acid, showcasing the high potential of the ML models in practice.



INTRODUCTION

Pharmaceutical cocrystals are multicomponent systems in which at least one component is an active pharmaceutical ingredient (API) and the other, *i.e.*, the coformer, is a pharmaceutically acceptable ingredient.[1] The cocrystallization of a drug with a coformer is a well-established and effective approach to improve the physicochemical properties of APIs,[2] such as solubility, dissolution profile, pharmacokinetics, and stability.[3,4] Cocrystals are of great interest in both the academic and industrial world, and the hot topic of recent years is to overcome the limiting step in the preparation of these new multicomponent forms: that is, given an API, to successfully identify the coformers that are most likely to form a supramolecular adduct, *i.e.*, a cocrystal or a molecular salt. Actually, the most applied method for coformer selection is the trial-and-error approach, which is time- and reagent-consuming. Indeed, despite the large number of molecules that are usable for the cocrystallization, searchable among GRAS (generally recognized as safe) or EAFUS (everything added to food in the United States) molecules or nontoxic chemicals (*e.g.*, nutraceuticals, flavonoids, vitamins, etc.), only a limited number of these actually will form the adduct with the API of interest.[5] However, implementing robust predictive tools would minimize the waste of reagents, time, and costs, turning the cocrystal design workflow green and sustainable. For this purpose, a series of *in silico* methods were developed to predict the outcome of the cocrystallizations. The currently available predictive methods can be subdivided into three main classes: *ab initio*, property-based, and machine learning (ML). *Ab initio* methods, such as CSP and molecular dynamics, directly model hypothetical solid structures, taking into account the properties of the crystal lattice.[6] These methods are computationally demanding due to the use of high-level quantum-chemical calculations and thus are not often employed. The second group relies only on the physical properties, such as the miscibility, the hydrogen bond tendency, and geometrics descriptors of the interacting molecules and ignores the characteristics of the crystal structure. Hydrogen bond energy (HBE),[7] hydrogen bond propensity,[8] Hansen solubility parameters (HSP),[9] COSMO,[10] and molecular complementarity (MC)[11] are among the first proposed. Their advantage is the fast computation, which makes them suitable for treating a vast number of API-coformer pairs; on the other hand, the

Received: June 7, 2023
Revised: September 27, 2023
Published: October 19, 2023

level of approximation of these methods is remarkably high, and their prediction performance often depends on the chemical class to which the API belongs. For these methods, it was estimated a variable accuracy in the range of 30−80% depending on the API.[12] To overcome the poor accuracy of the property-based methods, a combination of different tools was also proposed, showing an improvement in the coformer selection of specific systems.[8,13−15] Recently, data-driven ML approaches have become increasingly popular due to the rapidity of calculation and promising predictive accuracy.[16−22] Several algorithms were evaluated, such as support vector machine (SVM), random forest (RF), neural networks (NN), and partial least squares-discriminant analysis (PLS-DA), and also, a wide variety of molecular representations were considered, including molecular descriptors,[23] fingerprint vectors,[24] and molecular graphs.[25] To mention a few studies, Fornari et al. proposed using QSAR descriptors and the PLS-DA model to discriminate between the formation of cocrystals and physical mixtures.[26] Wang et al. tested several algorithms on a set of 14480 data to estimate which was the most promising for coformer selection.[17] Molecular fingerprints were used to represent molecules of the data set, and the compared algorithms were logistic regression, RF, AdaBoost, gradient boosting, multinomial Naïve Bayes, and deep NN. The best performer was the RF, with an ROC-AUC value of 0.844. Similarly, Gelder and co-workers proposed the application of artificial NN to guide the selection of coformers, represented with molecular graphs, in experimental tests, with a model accuracy of 80%.[25] Zheng et al. observed that one of the problems of artificial NN could be the applicability in predicting coformers for new chemical entities that do not have any binary cocrystal reported in the literature. They developed a new tool named SMINBR,[27] combining network and cheminformatics methods to achieve a high generalization capability of the model. Although many different approaches have been proposed, only a few of them are easily reproducible and applicable due to the lack of data sharing. Among the most promising and usable methods certainly include that of Jiang et al.[28] (named CCGNet) and that of Zheng et al.[27] (named SMINBR). Regardless, it is clear that the ML algorithms are the most promising ones for coformer selection; however, their applicability over different classes of APIs remains an open challenge. Specifically, the desirable goal is to improve the generalization ability of ML methods, which involves optimizing three variables: (i) the data set, (ii) the feature representation, and (iii) the model algorithm. Regarding the data set, ML algorithms require thousands of data items possibly balanced between positive and negative cocrystallization tests to avoid biased results toward the largest group. However, while positive cocrystallization data are widely reported in the literature and on the Cambridge Structural Database (CSD), the negative ones, although fundamental for prediction, are generally rarely reported and/or may be unreliable. A negative result might be due to either the impossibility of forming the cocrystal or an insufficient number of experimental trials and/or the choice of inappropriate synthesis techniques. To cope with the limiting number of NO data issues, negative cocrystallization cases could be computationally generated using link-prediction methods,[25] but the weak point of this approach is the possibility of introducing a large number of false negatives into the training set. Moreover, the training set data should represent the samples to be predicted; otherwise, the ML model will have low predictive ability. Instead, as observed from the literature,[17,18,22] the features can be chosen from a plethora of descriptors and also the employed algorithms are several; however, the algorithm should be chosen coherently with the set of features and data.

The purpose of this work is to develop a new predictive strategy and compare it with several methods already reported in the literature, emphasizing the advantages of the proposed approach and the points that make the prediction of new crystal forms challenging. The workflow follows four steps:

1. Data set creation: negative and positive cocrystallization cases were collected both from literature and from in-house experiments.

2. Evaluation of property-based tools and design of a new predictive method: the predictive performances of HSP, HBE, and MC tools were deeply evaluated by testing them on the data sets of 13 structurally different APIs. Setting as a further goal the development of a predictive procedure with greater accuracy and generalizability, 22 descriptors were selected from the three used predictive methods (HSP, HBE, and MC) as features for new ML models. The variety of molecular descriptors, representing the miscibility, the possibility of establishing hydrogen bonds, and the size and shape of the molecules, were specially chosen to represent the phenomenon of the cocrystallization. Moreover, PLS-DA, RF, and NN were employed as ML algorithms for this study, and their predictive performance was evaluated in cross-validation on the same cocrystallization cases collected for the 13 APIs of the data set.

3. Validation of the method on a real case study: the new approach was successfully applied to the 2-phenylpropionic acid (PPA) case study to aid the discovery of new cocrystals. In this case, we used a small and specific training set with cocrystallization data of molecules structurally similar to PPA.

4. Comparison of predictive methods: the results of the prediction for PPA adducts obtained by our method were compared with several approaches, including property-based (HSP, HBE, and MC) and machine learning (CCGNet, SMINBR, and models using QSAR descriptors) tools to evaluate the performance of our approach in relation to the most promising methods in literature.

EXPERIMENTAL SECTION

PPA Cocrystal Synthesis. Thirteen coformers, belonging to different chemical classes, were selected, and an experimental screening was conducted to investigate the formation of PPA cocrystals. The selected coformers are acetamide, benzamide (BA), 4-chlorobenzamide (4ClBA), L-histidine, L-proline (PRO), 4-nitroaniline, 2-aminopyridine (2APY), 4-aminopyridine (4APY), 3-hydroxypyridine, isonicotinamide (INA), nicotinamide (NA), nicotinic acid, and 1,2-bis(4-pyridyl)ethane (BPY). PPA and BPY were purchased from Tokyo Chemical Industry (TCI, Milan, Italy) with a declared purity of >99%. Histidine, acetamide, nicotinic acid, PRO, 4APY, 2APY, NA, and INA were purchased from Sigma-Aldrich, while 4-nitroaniline, 4ClBA, 3-hydroxypyridine, and BA were from Jassen Chemical, with a declared purity of all products over 98%. All of the starting materials were used as such in all the supramolecular syntheses.

An experimental screening strategy covering grinding, liquid-assisted grinding (LAG), and solvent evaporation was used employing the following protocol. Slurry experiments were performed but mostly produced sticky substances. For grinding, the reagents were mixed in a specific stoichiometric ratio (see the Supporting Information, Table S15) in an agate mortar and subjected to manual grinding for 15 min. For LAG, the reagents were mixed in the mortar, and the grinding was assisted by adding five drops of acetone repeated three times. In-solution cocrystallization techniques were also employed: PPA and

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

relative coformers were dissolved in 5 mL of ethanol, and the solution was allowed to evaporate at RT. In more detail, the PPA adducts were obtained as follows:

-PPA-PRO, PPA-4ClBA, and PPA-BA: PPA and the relative coformer were taken in a 1:1 stoichiometric ratio, using 100 mg of PPA (0.66 mmol) and the respective amount of the coformer (0.66 mmol). The mixed powder was ground in a mortar for 15 min by adding five drops of acetone. Addition of the solvent was repeated three times.

-PPA-4APY and PPA-BPY: PPA and the relative coformer were taken in a 2:1 stoichiometric ratio, using 100 mg of PPA (0.66 mmol) and the respective amount of the coformer (0.33 mmol). The mixed suspension was ground in a mortar for 15 min by adding five drops of acetone. Addition of the solvent was repeated three times.

-PPA-NA and PPA-2APY: PPA and the relative coformer were taken in a 1:1 stoichiometric ratio, using 100 mg of PPA (0.66 mmol) and the respective amount of the coformer (0.66 mmol). The starting materials were mixed and milled manually in a mortar for 15 min, obtaining a crystalline powder.

-PPA-INA: PPA (100 mg, 0.66 mmol) and INA (40.6 mg, 0.33 mmol) were mixed and milled manually in a mortar for 15 min, obtaining a crystalline powder.

For PPA-BA, PPA-4ClBA, PPA-2APY, and PPA-BPY, it was possible to obtain crystals suitable for single-crystal X-ray diffraction (SCXRD), solubilizing 10 mg of PPA and the respective milligram of the coformer (following the same stoichiometric ratio as in mechanochemical synthesis) in 10 mL of ethanol; complete dissolution was facilitated by heating to 50 °C and adding a spatula tip of the adduct previously synthesized by the mechanochemical technique to act as a seed for the crystal growth.

For all synthetic techniques, the obtained powder samples were analyzed by FTIR-ATR spectroscopy. If the possible formation of the adduct was observed, then a deep solid-state analysis was carried out by solid-state NMR (SSNMR) and powder XRD (PXRD).

All of the synthesis techniques were tested twice for each API-coformer system to ensure the result and the reproducibility of the data. Each system was then assigned to one of the two classes, *i.e.*, YES or NO, depending on the outcome of the cocrystallization reactions.

IR Spectroscopy. Fourier transform infrared (FTIR) spectra were recorded on an Equinox 55 (Bruker) spectrometer with an ATR reflectance accessory. Spectra were collected in the 400−4000 cm^{-1} range with a resolution of 2 cm^{-1} and 16 scans. The FTIR-ATR spectra are reported in the Supporting Information (Figures S1−S8).

Powder X-ray Diffraction. X-ray powder patterns were recorded on an Xpert Pro (45 kV, 40,000 μA) diffractometer in the Bragg−Brentano geometry, using Cu−Kα radiation (λ = 1.5418 Å) in the 2θ range between 5 and 50° (continuous scan mode, step size 0.0167°, counting time 40 s). The PXRD patterns are reported in the Supporting Information (Figures S9−S17).

Solid-State NMR. Solid-state NMR spectra were acquired with a Bruker Avance II 400 Ultra Shield instrument, operating at 400.23, 100.63, and 40.56 MHz, respectively, for ^{1}H, ^{13}C, and ^{15}N nuclei. The powdered samples were packed into cylindrical zirconia rotors with a 4 mm o.d. and a 80 μL volume. A certain amount of the sample was collected from the batch and used without further preparations to fill the rotor. ^{13}C CPMAS spectra were acquired at a spinning speed of 12 kHz, using a ramp cross-polarization pulse sequence with a 90° ^{1}H pulse of 3.60 μs, a contact time of 3 ms, optimized recycle delays between 3 and 6 s, and a number of scans in the range of 60−400, depending on the sample. ^{15}N CPMAS spectra were acquired at a spinning speed of 9 kHz using a ramp cross-polarization pulse sequence with a 90° ^{1}H pulse of 3.60 μs, a contact time of 4 ms, optimized recycle delays between 3 and 6 s, and a number of scans in the range of 20000−50000, depending on the sample. A two-pulse phase modulation (TPPM) decoupling scheme was used for all spectra, with a radiofrequency field of 69.4 kHz. For ^{13}C T_1 ^{1}H analyses, 13 increments were acquired for 400 scans with different τ values ranging from 0.02 to 60 s. The ^{13}C and ^{15}N chemical shift scales were calibrated through the signals of γ-glycine (^{13}C methylenic peak at 43.7 ppm and ^{15}N peak at 33.4 ppm with reference to NH_3).

Single-Crystal X-ray Diffraction. For PPA-BA, PPA-4ClBA, PPA-2APY, and PPA-BPY, it was possible to obtain single crystals to perform SCXRD analysis. A single crystal of compounds PPA-BA, PPA-4ClBA, PPA-BPY, and PPA-2APY was mounted on a glass fiber, and diffraction data were collected at room temperature on an Xcalibur, AtlasS2, Gemini Ultra diffractometer using graphite monochromated Mo−Kα radiation (λ = 0.71073 Å) for PPA-BA, PPA-4ClBA, PPA-BPY, and Cu−Kα radiation, (λ = 1.5406 Å) for PPA-2APY. Data reduction and proper absorption correction were performed with the CrysAlisPro (Rigaku OD, 2021) software package.[29] All the structures were solved by direct methods with the SHELXS-2008 program[30] and refined by full-matrix least squares procedures using the SHELXTL program. The hydrogen atoms were placed at the calculated positions and constrained to ride to the atoms to which they were attached. Drawings were performed with the program Mercury.[31] The details of crystallographic data and refinements are given in Tables S16−S27 in the Supporting Information. The X-ray data were deposited in the CCDC/FIZ Karlsruhe service. The deposition numbers were 2264630, 2264632, 2264633, and 2264634.

COMPUTATIONAL METHODS

The HSP tool, implemented in the HSPiP software,[32] was developed from an intrinsic definition of a cocrystal, which is a homogeneous molecular mixture consisting of an API and a coformer.[33] It evaluates the miscibility of two substances and correlates it with the adduct formation. To use miscibility as a predictor, it is necessary, given an API and a list of coformers, to calculate the solubility parameter[34] of all the substances, which is divided into three contributions representing the energy density from dispersion bonds, δ_D, the dipolar intermolecular force, δ_P, and the energy from hydrogen bonds between the molecules, δ_H.

In agreement with the principle of this method, the total solubility parameter (δ) is defined as (eq 1)

$$\delta = \sqrt{{\delta_D}^2 + {\delta_P}^2 + {\delta_H}^2} \quad (1)$$

Two molecules with similar δ values will theoretically be miscible.

The three Hansen's parameters δ_D, δ_P, and δ_H can also be treated as the coordinates of a point in a three-dimensional space, defined as Hansen's space.[9] To assess the miscibility of two molecules, three different formulas, combining Hansen's parameters in different ways, have been proposed in the literature:

Absolute difference of the total solubility parameter (eq 2)

$$\Delta\delta = |\delta_{API} - \delta_{COFORMER}| \quad (2)$$

Euclidean distance of two points in Hansen's space (eq 3)

$$\Delta\delta t = \sqrt{(\Delta\delta_D)^2 + (\Delta\delta_P)^2 + (\Delta\delta_H)^2} \quad (3)$$

Euclidean distance of two points in Hansen's space (eq 4), which takes into account a correction factor that emphasizes the contribution of dispersion forces

$$Ra = \sqrt{4\cdot(\Delta\delta_D)^2 + (\Delta\delta_P)^2 + (\Delta\delta_H)^2} \quad (4)$$

Mathematically, constant 4 represents the solubility data as a sphere, which is a convenient way to display the HSP characteristic.

Several papers report using these parameters to predict the cocrystallization, all with different cutoff values depending on the studied system (see Table S31 in the Supporting Information). To evaluate the performance of this method, we chose to employ the *Ra* parameter to discriminate between the prediction of the YES instead of the NO. The cutoff values were optimized for each data set.

HBE is a computational approach based on calculating the molecular electrostatic potential surface (MEPS), which treats intermolecular interactions as contact points between specific polar sites on the molecular surface.[7] The MEPS of a molecule is calculated in the gas phase and allowed one to derive a set of interaction points

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

Table 1. Calculated Parameters for the Application of the Predictive Methods[a]

METHOD	PARAMETER	DESCRIPTION
HSP	δ_P	polar interactions
	δ_D	dispersion forces
	δ_H	hydrogen bonding
	H_a	hydrogen bond acceptor
	H_d	hydrogen bond donor
	MVol	molar volume
	Δδ	solubility difference
	Δδt	Euclidean distance of two points in Hansen space
	Ra	distance of two points in Hansen space
HBE	$E_{coformer}$	coformer-coformer hydrogen bond interaction energy
	E_{adduct}	API-coformer hydrogen bond interaction energy
	ΔE	difference of all possible interactions through hydrogen bonding
CM	M/L	shape descriptor
	S-axis	dimension descriptor
	S/L	shape descriptor
	FNO	number of N and O atoms/number of heavy atoms, polarity descriptor
	d.m.	dipole moment
	ΔM/L	API-coformer form descriptor difference
	ΔS-axis	difference API-coformer dimension descriptor
	ΔS/L	API-coformer form descriptor difference
	ΔFNO	difference number of N and O atoms API-coformer
	Δ*d.m.*	API-coformer dipole moment difference

[a]Additional descriptors implemented only for ML models are highlighted in yellow.

describing the hydrogen bond donor (α) or acceptor (β) groups. The α and β parameters, also called Hunter parameters, are obtained by converting the local maxima and minima of the MEPS (eqs 5 and 6).

$$\alpha = 0.0000162\mathrm{MEP}_{max}^2 + 0.00962\mathrm{MEP}_{max} \quad (5)$$

$$\beta = 0.000146\mathrm{MEP}_{min}^2 + 0.00930\mathrm{MEP}_{min} \quad (6)$$

The site with the highest value of α_i interacts with the site with the highest value of β_i, the next α_i, in order of decreasing value, interacts with the next β_i, and so on. The total interaction energy is estimated from the sum of all contacts, according to eq 7.

$$E = \sum_{ij} \alpha_i \beta_j \quad (7)$$

where α_i and β_j are defined above. Excess positive or negative sites, which thus remain unpaired, are considered noncontributors to the crystal energy. A multicomponent crystalline solid can be treated the same way as a pure solid. In the case of a two-component cocrystal, the α and β parameters of both molecules are combined in a single list in descending order, and the best hydrogen bond donors are sequentially paired with the best hydrogen bond acceptors to obtain the coupling energy of the cocrystal interaction sites.[7] The probability of cocrystal formation is evaluated by comparing the energy difference between the cocrystal and the two pure crystal forms (eq 8).

$$\Delta E = E_{cc} - nE_1 - mE_2 \quad (8)$$

Because of how the interaction site coupling system is built, cocrystal formation occurs when the interaction energy is enhanced compared to the pure individual forms. Thus, the cocrystal is the favorite form for values of $\Delta E > 0$, while, conversely, for $\Delta E < 0$ values, API−API and coformer−coformer interactions are preferred, no new synthons are formed, and there is simply a physical mixing of the two components.[8]

MC is a routine developed after a chemometric study[11] that identified which molecular descriptors positively influence the formation of supramolecular adducts. According to the study, two molecules with a high molecular weight do not tend to cocrystallize with each other; instead, interactions between small molecules or between a large molecule and a small one are favored. In addition, molecules with similar polarity will be more likely to cocrystallize. The tool is available on Mercury and allows the calculation of the molecular complementarity between APIs and coformers by simply entering their molecular structure. The descriptors considered for the calculation are polar descriptors (FNO = (number of N atoms + number of O atoms)/number of heavy atoms and dipole moment), one size descriptor (S axis), and shape descriptors (S/L ratio and M/L ratio). These descriptors refer to the box model of crystalline packing, according to which the entire volume of molecules is ideally enclosed in a rectangular box. The three axes (S, M, and L) indicate the size of the molecule, while the ratio of the axes describes its shape. The results are given in terms of percentages, and theoretically, the closer to 100%, the greater the probability of forming the cocrystal.

Molecular Descriptors Calculation. All the descriptors cited above, used in this work and calculated to build our data set together with the additional ones implemented only for ML models, are summarized for convenience in Table 1. The method and tools used to calculate the descriptors are described in detail in the Supporting Information.

QSAR descriptors were also evaluated as different features to compare our new strategy with one of the most widely used descriptors reported in previous articles.[18,21] We calculated these descriptors for all molecules in the same data set that we have already used as a case study for the PPA cocrystal prediction, following these steps:

1. For each API and coformer, molecular descriptors were calculated by Mordred, excluding 3D descriptors.
2. The same descriptors, calculated for the API and coformer, were paired and summed.
3. Descriptors equal to zero for all molecules were eliminated, resulting in 925 molecular descriptors, which are included in the following classes: 2D, ABCIndex, AcidBase, AdjacencyMatrix, Aromatic, AtomCount, Autocorrelation, BCUT, BalabanJ, BaryszMatrix, BertzCT, BondCount, CarbonTypes, Chi, Constitutional, DetourMatrix, DistanceMatrix, Estate, EccentricConnectivityIndex, ExtendedTopochemicalAtom, FragmentComplexity, Framework, HydrogenBond, InformationContent, KappaShapeIndex, Lipinski, McGowanVolume, MoeTypea, MolecularDistanceEdge, MolecularId, PathCount, Polarizability, RingCount, RotatableBond, *S* Log *P*, TopoPSA, TopologicalCharge, TopologicalIndex, VdwVolumeABC, VertexAdjacencyInformation, WalkCount, Weight, WienerIndex, and ZagrebIndex.

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

All the descriptors are considered as features for the model training, using the same preprocessing and algorithms employed for the models trained with HSP, HBE, and MC descriptors.

Machine Learning Models. The current study used various discriminant analysis models, machine learning approaches, and deep learning models, such as PLS-DA,[35] RF,[36] and NN.[37] These models were selected to represent a benchmark of the ML models available to solve this task. All performance metrics of the models in terms of sensitivity, specificity, and accuracy were evaluated by repeated double cross-validation[38] using the R chemometrics[39] and caret[40] packages, with a number of 4-fold and 10 repetitions. Moreover, all of the collected data were autoscaled before building the ML models. Before building the ML models, multicollinearity was evaluated in terms of Pearson correlation. A correlation matrix is reported in Figure S37 in the Supporting Information. As it can be observed, no severe multicollinearity was observed as the only strongly correlated parameters are those related to δ_H, H_d, H_a, $\Delta\delta$, $\Delta\delta t$, and *Ra*, as expected. Despite several *p*-values turned lower than 0.05 in terms of correlation, the overall data set seems suitable for machine learning evaluations, and no single descriptor holds a disproportionately strong representation. Moreover, the employed machine learning approach such as, for instance, PLS-DA and RF, can compute the robust model (if properly validated) with multicollinearity. Therefore, the hypothetical intrinsic issue of multicollinearity does not give rise to issues concerning the robustness and interpretability of our developed models.

The tested models are concisely described as follows.

PLS-DA, as a model of discriminant analysis, aims to compute specific boundaries in a multidimensional space that allows separation of the different objects within their corresponding classes. In particular, these models always provide a result related to the classification of the objects. Moreover, PLS-DA can also be called a probabilistic (parametric) method since its classification algorithm is based on estimating the parameters describing the probability density functions (*i.e.*, arithmetic mean, variance, and covariance) of the studied features and their relative distributions. In particular, PLS-DA computes new components, called latent variables (LV), which are computed by simultaneously evaluating the *X* and *Y* matrices.[41] From a geometric point of view, the latent variables represent a slightly rotated version of the principal components of PCA modeling, whereas PCA maximizes the variance of the *X* matrix, and the PLS approach iteratively maximizes the covariance between *X* and *Y*. To this end, the components computed for the *Y* response(s) are rotated to maximize the covariance concerning the components computed for *X*. The PLS approach iteratively maximizes the covariance between the two matrices.[42] The iterative process ends when no helpful information can be extracted from the *X* and *Y* matrices. The discriminant model is computed by classifying the objects by the regression (PLS) of *X* with respect to a matrix *Y* containing binary responses. In particular, *Y* consists of *N* columns (two columns in our case) corresponding to the number of categories to be evaluated (in our case, YES vs NO). Each column contains the class membership information on the corresponding *n* observations (objects/individuals). Since the response is binary, if a subject belongs to a particular *n*th category, then the *n*th column will show a response equal to 1. Otherwise, the response is coded as 0. In the present case, the correct number of latent variables for each PLS-DA was determined by optimizing the root mean square error in cross-validation (RMSECV) parameter, *i.e.*, the lower the RMSECV value, the higher the discriminant power of the developed model. In the case of the PPA model, this analysis is reported in the Supporting Information (Figure S38). The developed model was computed using the Rpls[43] package.

RF is an ML algorithm that belongs to the ensemble learning family. Ensemble learning is a technique that uses multiple models to achieve a better predictive performance than could be achieved with a single model. In particular, RF is a type of ensemble learning algorithm in which a large number of decision trees are created at training time, and the class representing the mode of the classes predicted by each tree is the output.[36] Specifically, RF has several advantages over other ML algorithms, including the following: (i) it can be used for both regression and classification tasks; (ii) it is not biased toward any particular feature, which makes it resistant to overfitting; (iii) it can handle large data sets with high-dimensional data; and (iv) it is relatively fast to train and makes predictions. Despite these advantages, RF also presented some drawbacks, especially regarding our predictive task, including the limited number of instances/objects available for our casework, the difficulty of interpreting the results of a model due to its complex structure, and the fact that it can be computationally expensive when adequately trained and tuned.[44] RF is traditionally used for classification, regression, and other tasks that work with data consisting of a set of input features and corresponding labels. The algorithm builds a model consisting of a collection of decision trees. In particular, the individual decision trees are created by using a random subset of the input features (feature randomness) and samples (bootstrap aggregation). The final predictions are made by evaluating most of the predictions of all of the individual decision trees. When RF is used for classification, the algorithm first randomly selects a subset of the training data. It then uses this subset to create a decision tree, and the process is repeated for each tree in the forest until finally the predictions of all trees are combined into the final prediction. The most important hyperparameters in the RF are the number of trees in the forest (ntree) and the number of variables randomly selected as candidates at each split (mtry). Increasing the number of trees usually not only increases the accuracy of the model but also makes the training process longer. If the number of trees is set too high, then the model will overfit the training data. Generally, RF requires large data sets, especially in the case of high-dimensional features, to produce robust cross-validated models. Due to the limited number of samples available for the current classification task, a repeated double cross-validation approach was employed to avoid overfitting and obtain reliable RF models. The developed model was computed using the R randomForest[45] package. In the present study, the models of RF were trained using a grid-search approach that includes the estimation of the parameter ntree (from 100 to 1000) and the parameter mtry (from 1 to 15). The feature importance for the PPA model is reported in the Supporting Information (Figure S39).

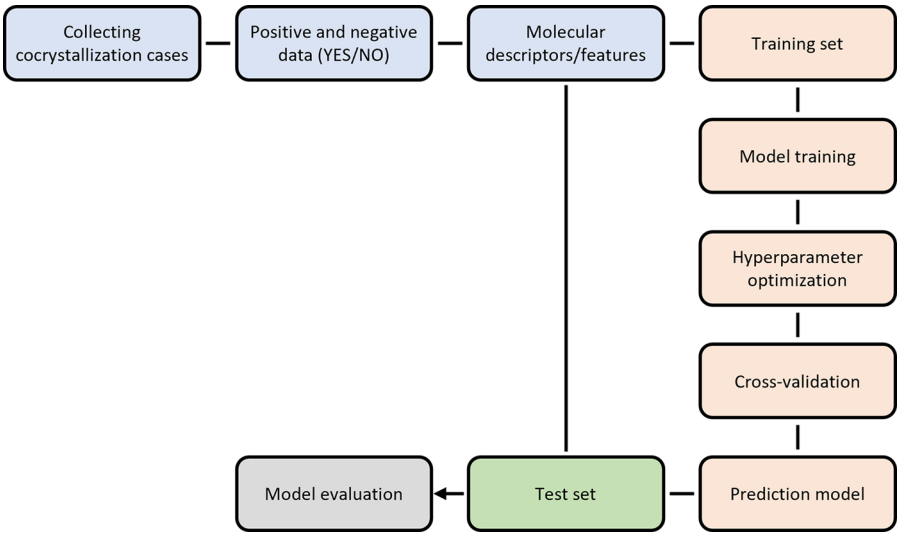
NNs are used to model complex patterns in data.[46] NNs are arranged in layers consisting of interconnected nodes that compute an activation function that evaluates the network's output. There are many different types of neural networks, but they all share the same basic structure. Namely, a neural network consists of (i) an input layer that receives the input data, (ii) an output layer that produces the desired output, and (iii) one or more hidden layers that process the data and pass it between the input and output layers. Specifically, NNs learn by adjusting the weights of the connections between their neurons until the network produces the desired output for a given input. In our case, a repeated double cross-validation approach was employed to avoid overfitting and obtain reliable NN models despite the limited number of samples available. In the present study, single/multilayer feedforward network models were trained using a research grid approach that involves estimating the size parameter (*i.e.*, the number of hidden layers between the input and output layers, from 1 to 10) and the decay parameter (*i.e.*, the regularization parameter to avoid overfitting, from 0.1 to 0.5), and a linear activation function was used employing the R package nnet.[47]

First, PLS-DA, RF, and NN were selected as ML algorithms for evaluating their predictive performance on 13 data sets of cocrystallization cases collected for structurally different APIs. Then, these three algorithms were employed to design the needed ML models for predicting the PPA cocrystals, using as the training set the cocrystallization data of nonsteroidal anti-inflammatory drugs (NSAIDs), *i.e.*, training NSAIDs. The algorithms were evaluated separately and statistically combined for this test set.

Sensitivity, specificity, and accuracy were calculated as performance measures for all tested ML models to compare the obtained results in terms of predictions. In this study, sensitivity represents the ability to correctly identify the pairs that form the cocrystal. On the other hand, specificity represents the ability to correctly predict the pairs for which the cocrystal formation does not occur. Finally, accuracy is a measure

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

Scheme 1. Flowchart for Building a Machine Learning Model for Cocrystal Screening



of the ability of the models to correctly identify the pairs that provide the formation of the cocrystal among the samples that are characterized as positive for cocrystal formation. Scheme 1 shows the flowchart to represent the various steps followed for building and validating the models applied in this study.

SMINBR: To apply this machine learning tool, we just followed the procedure proposed in the referenced article[27] *via* the web site shared at the link: http://lmmd.ecust.edu.cn/SMINBR/. As input, the SMILES code of PPA was entered, and the output provided was a list of the most promising coformer for the cocrystallization screening. The weaknesses of this approach include the absence of a list detailing the coformers used by the tool and the inability to select coformers according to one's preferences. Therefore, to evaluate the accuracy of this tool and to compare it with our approach, the coformers that SMINBR selected as the 50 most promising were labeled as YES, while "not available" (NA) labels were assigned to all the others.

CCGNet: all the data and source codes, required for the use of this model, were downloaded at the link: https://github.com/Saoge123/CCGNet. The code was tested by reproducing the data reported in the reference article.[28] Then, it was used, without changing the training set data, the features, and any part of the code, for predicting the PPA cocrystallization cases. For this purpose, the structure files of PPA and all of the coformers were generated and used as input. After running the code, the output was an excel file that we reported in the Supporting Information. CCGNet associates to each adduct a score indicating the probability of forming the cocrystal: API-coformer pairs with positive scores were classified as YES; conversely, those with negative scores were labeled as NO.

RESULTS AND DISCUSSION

Data Set Creation. The first step to evaluate the predictive tools is collecting experimental data from cocrystallization trials. In this study, 13 APIs, including riluzole, diclofenac, indomethacin, nalidixic acid, piracetam, carbamazepine, acetazolamide, furosemide, caffeine, pyrazinamide, paracetamol, sulpiride, and piroxicam (Scheme 2) and more than 300 coformers, were selected to test the predictive methods. Each API has its data set composed of the experimental outcomes of cocrystallization tests with both positive (YES) and negative (NO) results (see Tables S1−S13, in the Supporting Information). As already stated (see in the Introduction), we are aware that negative data might be the weak point of the data set. The 13 APIs were chosen according to the following criteria:

-A large number of data reported in the literature, at least more than 15 adducts, to have a statistically significant population.

-Chemical structures as different as possible from each other with different functional groups (acid, basic, aromatic, and halogen) to assess how all possible supramolecular interactions (hydrogen bonds, halogen bonds, van der Waals forces, and $\pi-\pi$ stacking interactions) affect the cocrystallization and the prediction outcomes.

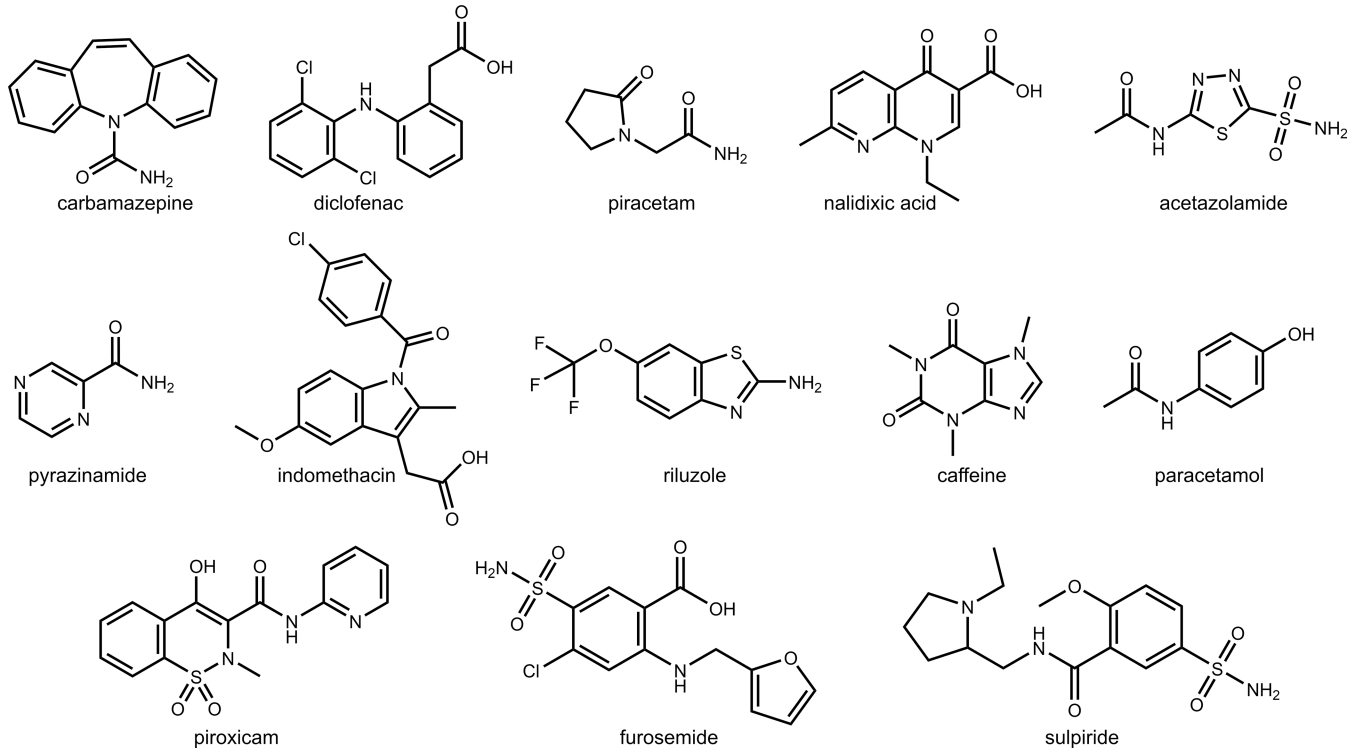
-A comparative number of YES and NO experimental outcomes. This criterion would be desirable, but finding sufficient data in literature was not always possible.

Evaluation of Property-Based Tools and Design of a New Method. For the first time, the predictive methods of HSP, HBE, and MC were tested on 13 structurally different APIs. Numerous papers report the predictive results of these methods on individual data sets, while the focus of this work was to verify the performance of these approaches on several classes of APIs. The results were classified according to true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Specifically, TP and TN are obtained when the results of the predictive methods show agreement with the experimental results; otherwise, the data are classified as FP or FN. Then, to determine the precision and validity of the predictive methods, sensitivity, specificity, and accuracy were calculated.

The HSP method shows 80% accuracy for the caffeine data set, while the same parameter drops to 40% for the piracetam one (see Supporting Information, Table S32). This data is in line with previous reports in the literature since Wu et al. reported that Hansen's approach was unpromising for cocrystal prediction, with an overall success rate of 49%.[13] The limitation of this method is that miscibility is not a sufficient condition for the cocrystallization to occur. Numerous examples of the failure of the cocrystallization by potentially miscible systems are reported. However, opposite trends were found for the HBE and MC methods: the HBE method showed particular promise for detecting negative outcomes (*i.e.*, high specificity), while MC generally showed high sensitivity but very low specificity. In general, none of the three methods exhibit an average accuracy higher than 60%, which is rather close to a random selection of coformers. In

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

Scheme 2. Chemical Structures of the 13 APIs Selected for the Data Set



other words, these prediction tools are extremely approximated and, although promising on individual classes of compounds, not easily generalizable. Each predictive method is based on the evaluation of specific properties of the interacting molecules, such as miscibility (HSP), availability of hydrogen bond donor and acceptor groups (HBE), or geometric (MC) aspects: each of these may be predominant in the cocrystallization of some systems and negligible in others. In fact, the cocrystallization depends on multiple factors and it is difficult to predict which one will lead to the formation of the product. Thus, a procedure that combines the variables achievable from these three predictive methods through ML algorithms was tested. The PLS-DA, RF, and NN algorithms were selected for the new predictive model since they may also be used on systems with collinear data and correlated variables. Indeed, the analysis of the data sets by ML models may be conducted using all variables provided by the three predictive methods, which are often mutually dependent. This procedure was applied to the same data sets, and sensitivity, specificity, and accuracy were evaluated in cross-validation. The comparison of the predictive performance, calculated as the average of the results obtained for each of the 13 API subsets, of standard methods and ML models is provided in Table 2. As reported in the Machine Learning Models section, all the ML models have been tuned and optimized using proper grid-search approaches.

The ability to correctly identify coformers leading to adduct formation (*i.e.*, TP) increases significantly using PLS-DA, RF, and NN methods, achieving an average accuracy of about 85%. Indeed, thanks to the chemometric treatment of the data, it was possible to obtain predictive methods with comparable sensitivity and specificity values and not the discordant ones, as in the case of HBE and MC used alone. Thus, the proposed predictive procedure is considerably superior to the employment of individual predictive methods.

Table 2. Comparison of the Average over the 13 Data Sets of Sensitivity, Specificity, and Accuracy for the HSP, HBE, and MC Methods and ML Models[a]

predictive tool	sensitivity	specificity	accuracy
HSP	55%	58%	57%
HBE	44%	67%	59%
MC	72%	33%	49%
PLS-DA	80%	87%	83%
RF	100%	100%	100%
NN	80%	90%	86%

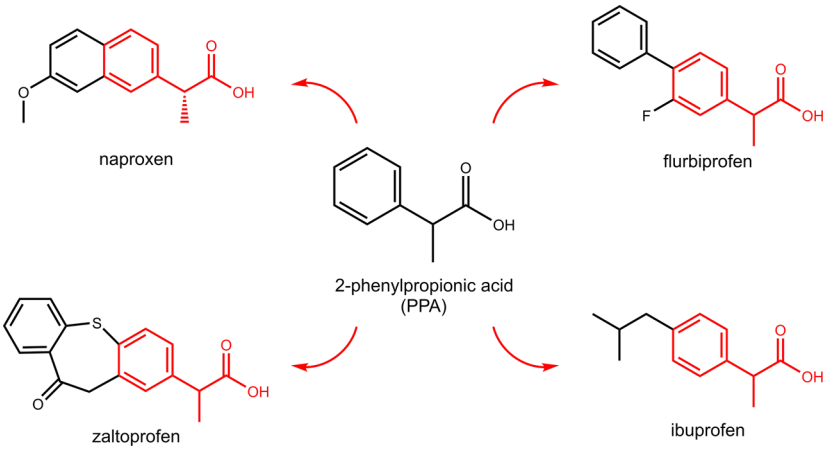
[a]The values reported for the ML models correspond to the average values obtained from the models built using repeated double cross-validation strategies.

Validation of the Method on a Real Case Study. Based on these promising cross-validation data, the new approach based on the combination of descriptors derivable from the three predictive methods (HSP, HBE, and MC) by ML models was applied to a real case study: the new PPA set, (a) to evaluate the possibility of using RF and NN for the prediction of new adducts on a real case study and (b) to verify that the excellent results during the model optimization phase were not due to overfitting given the limited number of samples used for the model. In particular, RF seems optimal as it provides results of 100% for sensitivity, specificity, and accuracy. However, this might suggest that the model is overfitting to the training data and may not generalize well to new unseen data. Indeed, overfitting may occur when a model is too complex and fits the training data too closely, including the noise in the data, and it is built on a relatively small number of training samples.

Training and Test Set Data Collection. It is worth noting that, to the best of our knowledge, no cocrystallization tests are available in the literature for PPA. So, the data of the training

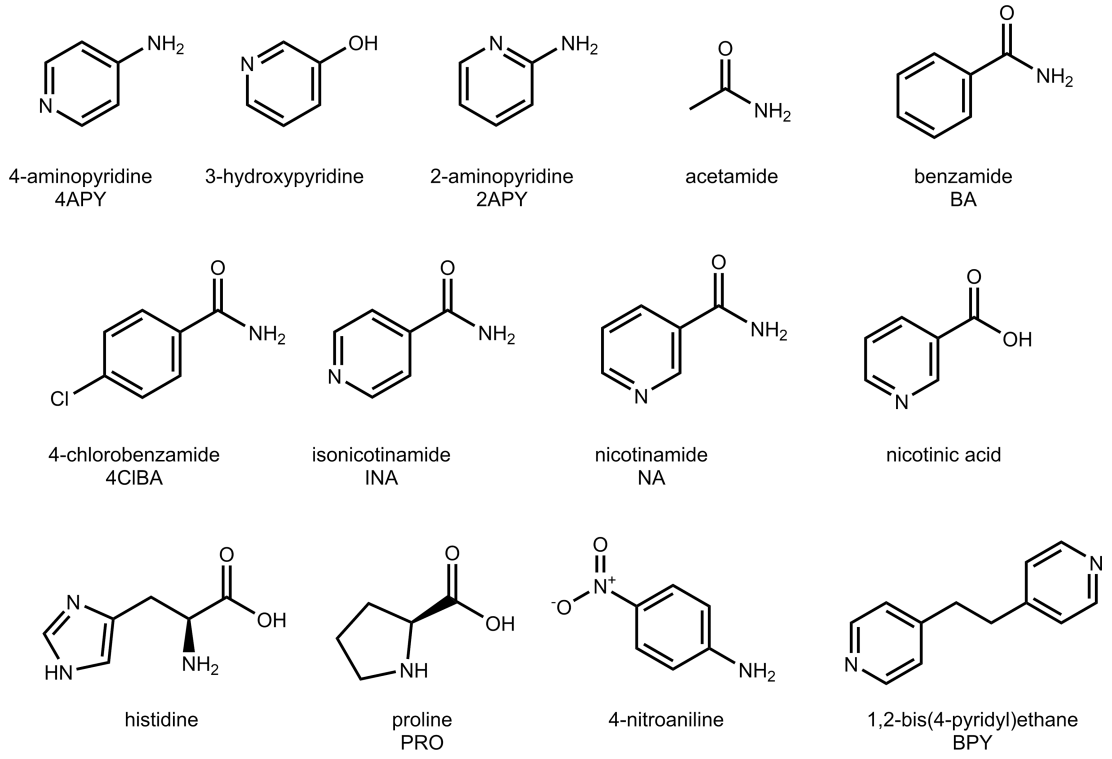
https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

Scheme 3. Chemical Structures of Ibuprofen, Flurbiprofen, Zaltoprofen, Naproxen (Training Set APIs), and PPA (Test Set)[a]



[a]In red, the structural similarity is highlighted

Scheme 4. Chemical Structures of the 13 Coformers Selected for the Cocrystallization Screening of PPA



set (training NSAIDs: 64 cases; 30 NO and 34 YES, see Table S14 in the Supporting Information) were collected from already reported cocrystallization tests of NSAIDs. In particular, flurbiprofen, ibuprofen, naproxen, and zaltoprofen were selected as these molecules are structurally similar to PPA (Scheme 3) and interact with the coformer *via* the same functional group, *i.e.*, the carboxylic acid. In this sense, the test also evaluates the robustness of the approach in predicting adducts of a given molecule even if the training set does not contain examples of that molecule.

The test set is composed of data obtained from the in-house PPA cocrystal screening (see the Experimental Section) with 13 coformers reported in Scheme 4 and consists of 5 NO, *i.e.*, physical mixtures, and 8 YES, *i.e.*, salts or cocrystals. The adducts PPA-4APY, PPA-2APY, PPA-BPY, PPA-BA, PPA-4ClBA, PPA-NA, PPA-INA, and PPA-PRO were obtained by mechanochemical syntheses and fully characterized in the solid state.

SSNMR is particularly informative for confirming the adduct formation, identifying the number of independent molecules in the unit cell, and verifying the protonation state of the adducts, *i.e.*, discriminating between salts and cocrystals. The ^{13}C CPMAS spectra of the eight adducts are reported in Figure 1.

All spectra point out the high degree of crystallinity of the new crystal forms, evidenced by the average full width at half-maximum value for the signals in the range of 90−130 Hz. The stochiometric API:coformer ratio is 1:1 for all the adducts, apart from PPA-BPY, PPA-4APY, and PPA-INA characterized by a 2:1 ratio. For PPA-4APY and PPA-INA, the API:coformer ratio is simply observable by the splitting of the PPA signals in

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

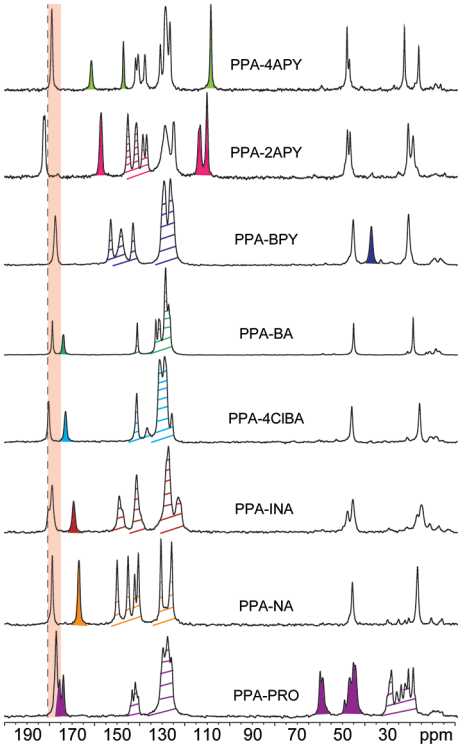


Figure 1. ^{13}C (100 MHz) CPMAS spectra of all eight new adducts of PPA, acquired with a spinning speed of 12 kHz at room temperature. The dashed line highlights the position of the COOH signal in pure PPA (solution spectrum), while the pink box highlights the COOH signal in the adducts. Filled colored peaks represent signals ascribable to the coformer. Striped peaks result from the overlap of coformer signals with those of PPA.

the aliphatic region. As for the number of the independent molecules in the unit cell (Z'), PPA-BA, PPA-4ClBA, and PPA-NA display one independent molecule for each component ($Z' = 1$), while PPA-PRO and PPA-2APY are characterized by two independent molecules of PPA and two of the coformer entities in the unit cell ($Z' = 2$), as observed from the characteristic splitting of all ^{13}C signals. The Z' value is also highlighted by the ^{15}N CPMAS spectra (Figure S26 in the Supporting Information) and, for PPA-2APY, supported by the SCXRD data (Table S16 in the Supporting Information). Confirmation that the PPA-PRO sample has a single phase and not a physical mixture was obtained by ^{13}C T_1 ^{1}H measurements. For PPA-BPY, due to the symmetry of the coformer, there is one molecule of acid and half molecule of BPY per asymmetric unit. Moreover, from the chemical shift value of the PPA carboxylic carbon, it is possible to determine the protonation state of the adducts. The chemical shift value of the carboxyl group of pure PPA (180.9 ppm) is referred to the spectrum recorded in solution in $CDCl_3$ since the PPA is liquid at RT. For PPA systems, the shift of the COOH peak in the adduct toward lower frequencies probably indicates a cocrystal formation (*i.e.*, neutral COOH group), while the shift to higher frequencies indicates a salt formation (*i.e.*, COO^- group).[48] For all adducts, except PPA-2APY, the carboxylic carbon resonates at lower frequencies than in the pure PPA, suggesting the formation of a neutral supramolecular synthon, *i.e.*, cocrystal. This is confirmed by the absence of protonable functional groups in the coformers for PPA-4ClBA, PPA-BA, and PPA-PRO and by the SCXRD data for PPA-BPY. To further investigate the nature of PPA-4APY, the ^{15}N CPMAS spectrum was also acquired (Figure S27 in the Supporting Information) as the chemical shift of pyridine nitrogen is highly informative for discriminating between salt and cocrystal. Based on the observed chemical shift values and the comparison with the free and protonated 4APY (Table S17 in the Supporting Information), it is possible to state that the adduct formation involves a proton transfer between PPA and 4APY. The protonic transfer between the acid group of PPA and the basic group of the coformer also occurs for PPA-2APY, as confirmed by the crystal structure (Figure 2d). For these new systems, the 3D molecular arrangement is driven by the supramolecular carboxylic acid−pyridine (PPA-BPY), carboxylic−aminopyridine (PPA-2APY), or carboxylic acid−amide (PPA-4ClBA and PPA-BA) heterosynthons, depending on the coformer (Figure 2). It is worth noting that the distinction between salts and cocrystals is crucial for the structural characterization of new adducts; however, from a predictive point of view, the proposed methods are unable to determine the position of the H atom along the hydrogen bond axes. Indeed, these predictive methods aim to identify the most promising coformers for the formation of the adduct regardless of the protonation state. For a reliable prediction of the synthesis outcome in terms of salt or cocrystal formation, we refer to the pK_a rule.[49,50] Similarly, also, the stoichiometry is not considered at all in this prediction approach since no descriptors in the training set include this information, nor the ML models can take it into account. The complete structure description of the newly achieved adducts is reported in the Supporting Information.

Model Evaluation. The comparison of the experimental results with those predicted by PLS-DA, RF, and NN algorithms is shown in Table 3, while the confusion matrix for each model is reported in Table 4, and the ROC curves in Figure S40 are shown in the Supporting Information. All the ML models used for predicting the novel compounds were trained using a repeated double cross-validation strategy.

The PLS-DA, RF, and NN models can correctly predict 8, 10, and 11 cases out of the 13 coformers, respectively. This result is extremely promising since the test set data are external to the training set and there are no PPA examples in the training set. For this data set, the best predictive performances are achievable with NN, where about 85% of the predictions agree with the experimental results. This result might be related to the fact that NN can predict samples, even if their related APIs are not present in the training data. However, the relatively limited number of instances requires further evaluations to confirm the reliability of the computed model. On the other hand, PLS-DA, with an accuracy of 62%, proved to be the least promising. However, if the results of the three algorithms are statistically combined, then only three coformers (*i.e.*, 4-nitroaniline, 4-aminopyridine, and 2-aminopyridine) are misclassified, resulting in a successful prediction rate of 77% for the external test set. Regarding the models here presented, stoichiometry appears to have no particular influence as well as proton transfer; while it is worth noting that the three coformers erroneously predicted (4APY, 2APY, and 4-nitroaniline) are characterized by an aniline moiety. Probably this is due to the fact that in the training set, all the examples containing the aniline moiety (*e.g.*, IBU and 4-aminosalicylic acid, IBU and 2-aminopyridine, ZAL and 4-aminobenzoic acid, and ZAL and 2-aminopyridine) are physical mixtures, so the algorithms label coformers 2APY,

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

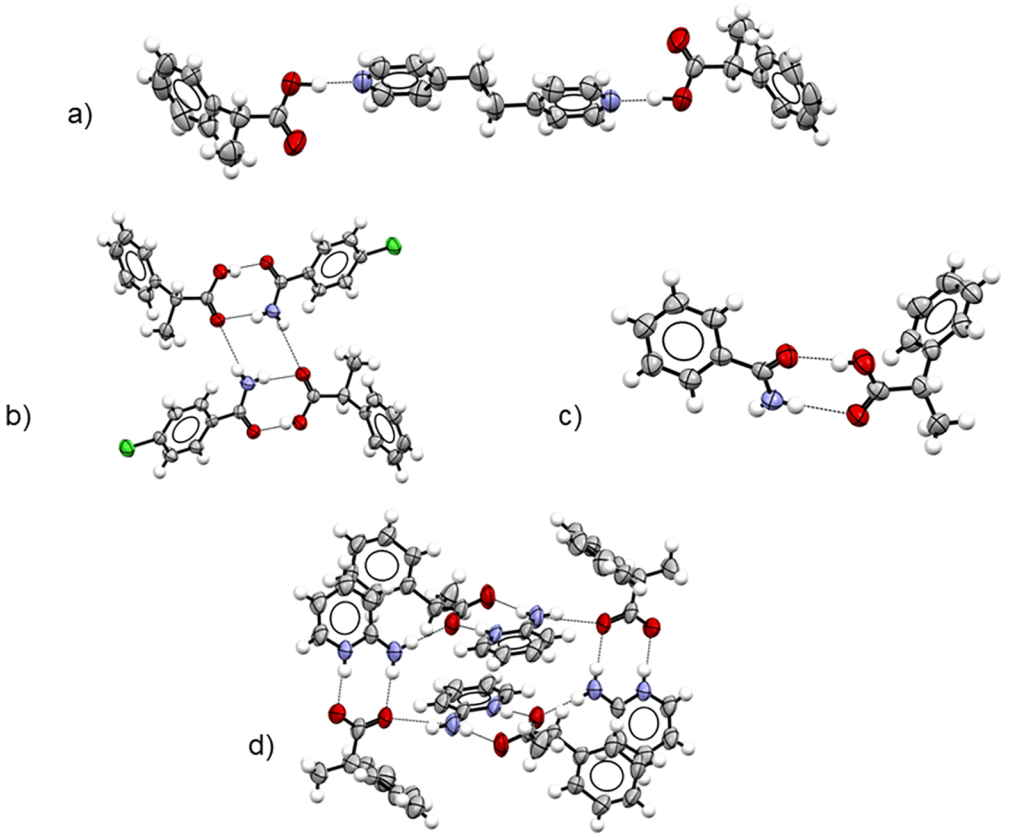


Figure 2. Structural motifs in the molecular PPA adducts (50% probability ellipsoids). (a) PPA-BPY, (b) PPA-4ClBA, (c) PPA-BA, and (d) PPA-2APY. Color code: white, hydrogen; gray, carbon; blue, nitrogen; green, chlorine; red, oxygen.

Table 3. Comparison of the Experimental and Predictive Results of PPA Adducts[a]

API	coformer	properties	EXP	PLS-DA	RF	NN
PPA	histidine	-	NO	NO	NO	NO
PPA	acetamide	-	NO	NO	NO	NO
PPA	4-nitroaniline	-	NO	YES	YES	NO
PPA	nicotinic acid	-	NO	YES	NO	NO
PPA	proline	1:1, cc	YES	NO	YES	YES
PPA	4-aminopyridine	2:1, s	YES	NO	NO	NO
PPA	nicotinamide	1:1, cc	YES	YES	YES	YES
PPA	2-aminopyridine	1:1, s	YES	YES	NO	NO
PPA	4-chlorobenzamide	1:1, cc	YES	YES	YES	YES
PPA	3-hydroxypyridine	-	NO	YES	NO	NO
PPA	isonicotinamide	2:1, cc	YES	YES	YES	YES
PPA	benzamide	1:1, cc	YES	YES	YES	YES
PPA	1,2-bis(4-pyridyl)ethane	2:1, cc	YES	YES	YES	YES

[a]Misclassifications are highlighted in red (cc = cocrystal; s = salt).

Table 4. Confusion Matrices and Performance Measures of the Models Designed for the PPA Cocrystal Prediction

PLS-DA

predicted \ Experimental outcome	YES	NO
YES	6	3
NO	2	2

sensitivity = 75%
specificity = 40%
accuracy = 62%

RF

predicted \ Experimental outcome	YES	NO
YES	6	1
NO	2	4

sensitivity = 75%
specificity = 80%
accuracy = 77%

NN

predicted \ Experimental outcome	YES	NO
YES	6	2
NO	2	5

sensitivity = 75%
specificity = 100%
accuracy = 85%

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

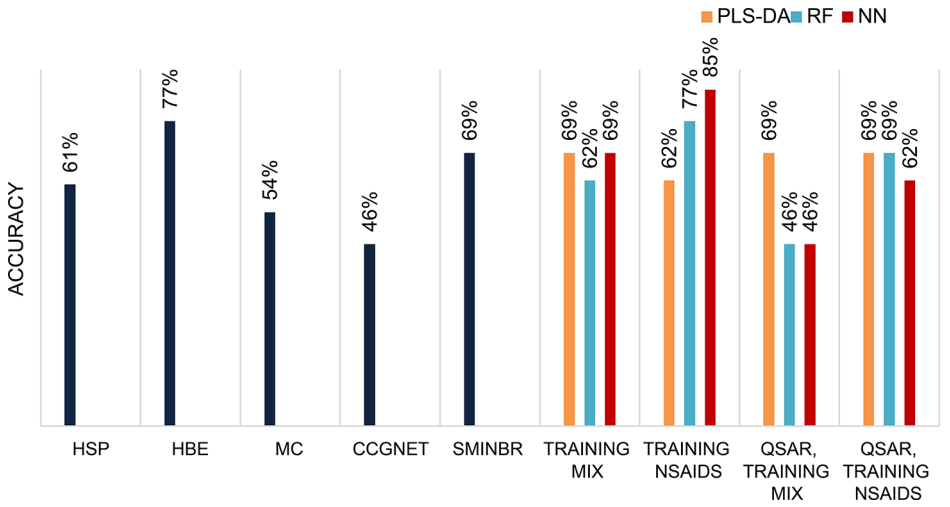


Figure 3. Benchmark analysis for the PPA case study. Training mix = generic training set data for the models; training NSAIDs = specific training set with cocrystallization data collected by searching for molecules structurally similar to PPA (ibuprofen, flurbiprofen, naproxen, and zaltoprofen); QSAR = QSAR molecular descriptors used as features for model training.

4APY, and 4-nitroaniline as unpromising for the cocrystallization.

As stated above, one of the risks of using ML models is overfitting, and the use of an external test set allows validating the model's performance. When the training data set is small, there is a risk that the model will not capture the full range of variability in the data, and the ML model cannot be extended to future predictions. To overcome this problem, it will be necessary to collect more data, either by obtaining new samples, by generating synthetic cases, or by using techniques such as data augmentation or transfer learning to help the model generalize better to new data. In summary, while using an external test set can help avoid overfitting, when the training data set is small, additional measures may be necessary to ensure that the model is capable of generalizing to new data.

Comparison of Predictive Methods. A benchmark analysis comparing different predictive methods for the prediction of PPA adducts was also performed. We selected HSP, HBE, and MC for the property-based tool class and CCGNet and SMINBR for the ML ones, both cited in the literature for their robustness and generalization. Moreover, we also tested QSAR descriptors as features for training the PLS-DA, RF, and NN models as they are also widely used molecular descriptors in the literature for this topic. We also evaluate the performance of the models using a generic training set, containing all the cocrystallization data (training mix) we collected, *i.e.*, 13 APIs and 543 cocrystallization cases (for details, see Tables S1−S13 and the relative section in Computational Methods) instead of the specific training NSAIDs. The accuracy of each method, for the PPA case study, is reported in Figure 3, while all of the confusion matrices are shown in the Supporting Information (Tables S14).

The analysis of the accuracy values highlights three important aspects: (i) the model trained using the training set of NSAIDs, descriptors derived from property-based methods, and the NN algorithm shows the best predictive performance followed by the model trained on the same data with the RF algorithm. The HBE method is comparable, but we have previously verified its poor robustness (see Table 1 and Table S32, in the Supporting Information); (ii) the QSAR descriptors perform worse than the specific cocrystallization descriptors derived from the HSP, HBE, and MC methods. In particular, models trained with QSAR descriptors show very low specificity (see the confusion matrices reported in the Supporting Information, Tables S33−S36), *i.e.*, the NO cases are misclassified. This implies that the model is not useful for decreasing the number of experimental tests; (iii) regardless of the descriptors used, whether QSAR or property-based, the use of specific data (*e.g.*, training NSAIDs), although involving fewer cases, shows better predictive accuracy than using larger training sets with structurally different molecules than the target API (training mix). It is important to note that the calculated accuracy is limited by the small number of data in the test set, which nevertheless is the result of the possible trials of the experimental screening.

Finally, comparing our strategy with SMINBR and CCGNet methods, it is possible to highlight that, although the performance of SMINBR is quite good, a high number of not available data is present, and these could be misclassified; moreover, the strong limitation concerning the impossibility of an *a priori* selection of the coformers remains. This is a significant limitation, especially for pharmaceutical companies, where coformers are selected based on their properties and the possible coadministration with the target API. Different is the case of CCGNet where most of the predicted results disagree with those obtained experimentally. This comparison points out that, so far, no generalized and robust methods able to reliably predict different types of molecules still exist. In this sense, our method, based on the fact that the training set contains examples of the target molecule or structurally similar molecules, represents a valid and promising alternative.

CONCLUSIONS

A deep analysis on 13 structurally different API data sets was performed to evaluate the predictive performance of property-based methods, such as HSP, HBE, and MC, highlighting that these tools are approximate, and their accuracy, varying in the range of 40−80%, strongly depends on the chemical class of the API under study. Although not reliable, the variables included by each describe accurately some of the properties required for the cocrystallization (*i.e.*, miscibility, hydrogen

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

bond donor and acceptor sites, and shape and size of interacting molecules). Thus, a new approach was designed through the combination of descriptors obtained from HSP, HBE, and MC using ML algorithms. This leads to a substantial improvement in the prediction outcomes (the average accuracies obtained from the 13 API models built using repeated double cross-validation strategies are PLS-DA = 83%, RF = 100%, and NN = 86%).

This new predictive approach was tested on an external test set of new PPA adducts, showing very promising results: it correctly predicts 77% of the experimental data. Finally, a benchmark analysis comparing several ML models and predictive tools, *i.e.*, CCGNet and SMINBR, reported in literature was performed on the PPA case study. The influence of the training set and the molecular descriptors was also verified. Specifically, the best predictive performance is obtained when a specific training set, including cocrystallization data of molecules structurally similar to the one to be predicted, together with HSP, HBE, and MC molecular descriptors are used. This study highlights that a specific, even if small, training set also allows a reliable prediction of molecules for which no data are available.

Cocrystal prediction still remains a challenging task because (i) the generalization of the methods is difficult to achieve since the training set should be specific for the target molecule to be predicted; (ii) the training set needs equal numbers of positive and negative outcomes, but while positive outcomes are easy to collect, the negative ones always remain the weak point; for this reason, we encourage the publication of the failed results of the cocrystallization too; (iii) the features have to be representative of the cocrystallization phenomena and well discriminant between the two classes. These are probably the reasons why, so far, there are no generalized and robust methods, even if the training set is built with thousands of examples and thousands of descriptors.

■ ASSOCIATED CONTENT

Supporting Information

The following data and files are available free of charge. The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.cgd.3c00696.

Data sets used for the training and the validation of the ML models (Tables S1−S15); detailed description for calculating the molecular descriptors used in ML models; FTIR-ATR spectra of PPA adducts (Figures S1−S8); powder X-ray diffraction patterns of PPA adducts (Figures S9−S17); ^{13}C and ^{15}N CPMAS SSNMR spectra of PPA adducts (Figures S18−S27); SCXRD data and structures of PPA adducts with a brief description of the main supramolecular interaction (Figures S28−S36 and Tables S18−S28); comparison of sensitivity, specificity, and accuracy of the HSP, HBE, and MC methods for each API in the data set (Table S32); correlation matrix in terms of Pearson correlation (Figure S37); RMSECV analysis to determine the number of components to use in the PLS-DA model for PPA cocrystal prediction (Figure S38); feature importance in the design of the RF model for the PPA cocrystal prediction (Figure S39); ROC curves for the PPA case study (Figure S40); confusion matrices for PPA cocrystal prediction (Tables S33−S36) (PDF)

Excel file with the CCGNet outputs of PPA cocrystal prediction (XLSX)

Excel file with the SMINBR outputs of PPA cocrystal prediction (XLSX)

Accession Codes

CCDC 2264630 and 2264632−2264634 contain the supplementary crystallographic data for this paper. These data can be obtained free of charge via www.ccdc.cam.ac.uk/data_request/cif, or by emailing data_request@ccdc.cam.ac.uk, or by contacting The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK; fax: +44 1223 336033.

■ AUTHOR INFORMATION

Corresponding Authors

Roberto Gobetto − *Department of Chemistry and NIS Centre, University of Torino, 10125 Torino, Italy;* orcid.org/0000-0002-2431-8051; Email: roberto.gobetto@unito.it

Michele R. Chierotti − *Department of Chemistry and NIS Centre, University of Torino, 10125 Torino, Italy;* orcid.org/0000-0002-8734-6009; Email: michele.chierotti@unito.it

Authors

Rebecca Birolo − *Department of Chemistry and NIS Centre, University of Torino, 10125 Torino, Italy*

Federica Bravetti − *Department of Chemistry and NIS Centre, University of Torino, 10125 Torino, Italy*

Eugenio Alladio − *Department of Chemistry and NIS Centre, University of Torino, 10125 Torino, Italy*

Emanuele Priola − *Department of Chemistry and NIS Centre, University of Torino, 10125 Torino, Italy;* orcid.org/0000-0002-0270-738X

Gianluca Bianchini − *Research and Early Development, Dompé Farmaceutici S.p.A, 67100 L'Aquila, Italy*

Rubina Novelli − *Research and Early Development, Dompé Farmaceutici S.p.A., 20122 Milano, Italy*

Andrea Aramini − *Research and Early Development, Dompé Farmaceutici S.p.A, 67100 L'Aquila, Italy;* orcid.org/0000-0001-8390-6614

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.cgd.3c00696

Author Contributions

Conceptualization: M.R.C., R.B., A.A., and R.G.; data curation: R.B., E.A., F.B., E.P., and G.B.; methodology: M.R.C., R.B., A.A., R.G., and E.A.; project administration, resources, and supervision: M.R.C., R.G., and A.A.; software: R.B., E.A., and F.B.; visualization: R.B.; formal analysis, validation, and writing—original draft and review and editing: all authors. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This research was funded by the Ministry of Education, Universities and Research PRIN 2020: project number 2020Y2CZJ2—Nature Inspired Crystal Engineering (NICE).

Notes

The authors declare the following competing financial interest(s): Andrea Aramini, Gianluca Bianchini, and Rubina Novelli are employees of Dompé Farmaceutici s.p.a., Italy. The company has interest in the development of co-crystals of

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

pharmaceutical compounds. The other authors declare no conflicts of interest.

■ ACKNOWLEDGMENTS

Serena Zampieri is sincerely acknowledged for synthesizing PPA-BPY, PPA-BA, PPA-4ClBA, PPA-NA, PPA-INA, and PPA-PRO, which were part of her master thesis; Marta Mencarelli is sincerely acknowledged for synthesizing PPA-4APY and PPA-2APY. The Ministry of Universities and Research is acknowledged (PRIN 2020: project number 2020Y2CZJ2).

■ ABBREVIATIONS

PLS-DA; partial least squares-discriminant analysis; RF; random forest; NN; neural networks; HSP; Hansen solubility parameters; HBE; hydrogen bond energy; MC; molecular complementarity; API; active pharmaceutical ingredient; PPA; 2-phenylpropionic acid; BA; benzamide; 4ClBA; 4-chlorobenzamide; PRO; proline; 2APY; 2-aminopyridine; 4APY; 4-aminopyridine; INA; isonicotinamide; NA; nicotinamide; BPY; 1,2-bis(4-pyridyl)ethane; SSNMR; solid-state NMR; SCXRD; single-crystal X-ray diffraction; PXRD; powder X-ray diffraction

■ REFERENCES

(1) Aitipamula, S.; Banerjee, R.; Bansal, A. K.; Biradha, K.; Cheney, M. L.; Choudhury, A. R.; Desiraju, G. R.; Dikundwar, A. G.; Dubey, R.; Duggirala, N.; Ghogale, P. P.; Ghosh, S.; Goswami, P. K.; Goud, N. R.; Jetti, R. R. K. R.; Karpinski, P.; Kaushik, P.; Kumar, D.; Kumar, V.; Moulton, B.; Mukherjee, A.; Mukherjee, G.; Myerson, A. S.; Puri, V.; Ramanan, A.; Rajamannar, T.; Reddy, C. M.; Rodriguez-Hornedo, N.; Rogers, R. D.; Row, T. N. G.; Sanphui, P.; Shan, N.; Shete, G.; Singh, A.; Sun, C. C.; Swift, J. A.; Thaimattam, R.; Thakur, T. S.; Kumar Thaper, R.; Thomas, S. P.; Tothadi, S.; Vangala, V. R.; Variankaval, N.; Vishweshwar, P.; Weyna, D. R.; Zaworotko, M. J. Polymorphs, Salts, and cocrystals: What's in a Name? *Cryst. Growth Des* **2012**, *12* (5), 2147−2152.

(2) Aramini, A.; Bianchini, G.; Lillini, S.; Tomassetti, M.; Pacchiarotti, N.; Canestrari, D.; Cocchiaro, P.; Novelli, R.; Dragani, M. C.; Palmerio, F.; Mattioli, S.; Bordignon, S.; d'Angelo, M.; Castelli, V.; d'Egidio, F.; Maione, S.; Luongo, L.; Boccella, S.; Cimini, A.; Brandolini, L.; Chierotti, M. R.; Allegretti, M. K. Lysine and Gabapentin Co-Crystal Magnifies Synergistic Efficacy and Tolerability of the Constituent Drugs: Pre-Clinical Evidences towards an Innovative Therapeutic Approach for Neuroinflammatory Pain. *Biomed. Pharmacother.* **2023**, *163*, No. 114845.

(3) Almarsson, Ö.; Peterson, M. L.; Zaworotko, M. The A to Z of Pharmaceutical cocrystals: A Decade of Fast-Moving New Science and Patents. *Pharm. Pat. Anal.* **2012**, *1* (3), 313−327.

(4) Aramini, A.; Bianchini, G.; Lillini, S.; Bordignon, S.; Tomassetti, M.; Novelli, R.; Mattioli, S.; Lvova, L.; Paolesse, R.; Chierotti, M. R.; Allegretti, M. Unexpected Salt/cocrystal Polymorphism of the Ketoprofen−Lysine System: Discovery of a New Ketoprofen−l-Lysine Salt Polymorph with Different Physicochemical and Pharmacokinetic Properties. *Pharmaceuticals* **2021**, *14* (6), 555.

(5) Cappuccino, C.; Cusack, D.; Flanagan, J.; Harrison, C.; Holohan, C.; Lestari, M.; Walsh, G.; Lusi, M. How Many cocrystals Are We Missing? Assessing Two Crystal Engineering Approaches to Pharmaceutical cocrystal Screening. *Cryst. Growth Des.* **2022**, *22* (2), 1390−1397.

(6) Sugden, I. J.; Braun, D. E.; Bowskill, D. H.; Adjiman, C. S.; Pantelides, C. C. Efficient Screening of coformers for Active Pharmaceutical Ingredient Cocrystallization. *Cryst. Growth Des.* **2022**, *22* (7), 4513−4527.

(7) Musumeci, D.; Hunter, C. A.; Prohens, R.; Scuderi, S.; McCabe, J. F. Virtual cocrystal Screening. *Chem. Sci.* **2011**, *2* (5), 883.

(8) Sarkar, N.; Gonnella, N. C.; Krawiec, M.; Xin, D.; Aakeröy, C. B. Evaluating the Predictive Abilities of Protocols Based on Hydrogen-Bond Propensity, Molecular Complementarity, and Hydrogen-Bond Energy for cocrystal Screening. *Cryst. Growth Des.* **2020**, *20* (11), 7320−7327.

(9) Salem, A.; Nagy, S.; Pál, S.; Széchenyi, A. Reliability of the Hansen Solubility Parameters as Co-Crystal Formation Prediction Tool. *Int. J. Pharm.* **2019**, *558*, 319−327.

(10) Abramov, Y. A.; Loschen, C.; Klamt, A. Rational coformer or Solvent Selection for Pharmaceutical Cocrystallization or Desolvation. *J. Pharm. Sci.* **2012**, *101* (10), 3687−3697.

(11) Fábián, L. Cambridge Structural Database Analysis of Molecular Complementarity in cocrystals. *Cryst. Growth Des.* **2009**, *9* (3), 1436−1443.

(12) Cappuccino, C.; Cusack, D.; Flanagan, J.; Harrison, C.; Holohan, C.; Lestari, M.; Walsh, G.; Lusi, M. How Many cocrystals Are We Missing? Assessing Two Crystal Engineering Approaches to Pharmaceutical cocrystal Screening. *Cryst. Growth Des.* **2022**, 1390 DOI: 10.1021/acs.cgd.1c01342.

(13) Wu, D.; Zhang, B.; Yao, Q.; Hou, B.; Zhou, L.; Xie, C.; Gong, J.; Hao, H.; Chen, W. Evaluation on cocrystal Screening Methods and Synthesis of Multicomponent Crystals: A Case Study. *Cryst. Growth Des.* **2021**, *21* (8), 4531−4546.

(14) Khalaji, M.; Potrzebowski, M. J.; Dudek, M. K. Virtual cocrystal Screening Methods as Tools to Understand the Formation of Pharmaceutical cocrystals—A Case Study of Linezolid, a Wide-Range Antibacterial Drug. *Cryst. Growth Des.* **2021**, *21* (4), 2301−2314.

(15) Nunes Costa, R.; Choquesillo-Lazarte, D.; Cuffini, S. L.; Pidcock, E.; Infantes, L. Optimization and Comparison of Statistical Tools for the Prediction of Multicomponent Forms of a Molecule: The Antiretroviral Nevirapine as a Case Study. *CrystEngComm* **2020**, *22* (43), 7460−7474.

(16) Yang, D.; Wang, L.; Yuan, P.; An, Q.; Su, B.; Yu, M.; Chen, T.; Hu, K.; Zhang, L.; Lu, Y.; Du, G. Cocrystal virtual screening based on the XGBoost machine learning model. *Chin Chem Lett* **2013**, *34* (8), 107964.

(17) Wang, D.; Yang, Z.; Zhu, B.; Mei, X.; Luo, X. Machine-Learning-Guided cocrystal Prediction Based on Large Data Base. *Cryst. Growth Des.* **2020**, *20* (10), 6610−6621.

(18) Mswahili, M. E.; Lee, M.-J.; Martin, G. L.; Kim, J.; Kim, P.; Choi, G. J.; Jeong, Y.-S. cocrystal Prediction Using Machine Learning Models and Descriptors. *Appl. Sci.* **2021**, *11*, 1323 DOI: 10.3390/app11031323.

(19) Kim, P.; Lee, I.-S.; Kim, J.-Y.; Mswahili, M. E.; Jeong, Y.-S.; Yoon, W.-J.; Yun, H.-S.; Lee, M.-J.; Choi, G. J. A Study to Discover Novel Pharmaceutical cocrystals of Pelubiprofen with a Machine Learning Approach Compared. *CrystEngComm* **2022**, *24* (21), 3938−3952.

(20) Guo, J.; Sun, M.; Zhao, X.; Shi, C.; Su, H.; Guo, Y.; Pu, X. General Graph Neural Network-Based Model To Accurately Predict cocrystal Density and Insight from Data Quality and Feature Representation. *J. Chem. Inf. Model.* **2023**, *63* (4), 1143−1156.

(21) Syed, T. A.; Ansari, K. B.; Banerjee, A.; Wood, D. A.; Khan, M. S.; Al Mesfer, M. K. Machine-learning Predictions of Caffeine Co-crystal Formation Accompanying Experimental and Molecular Validations. *J. Food Process Eng.* **2023**, *46* (2), No. e14230, DOI: 10.1111/jfpe.14230.

(22) Hao, Y.; Hung, Y. C.; Shimoyama, Y. Investigating Spatial Charge Descriptors for Prediction of cocrystal Formation Using Machine Learning Algorithms. *Cryst. Growth Des.* **2022**, *22* (11), 6608−6615.

(23) Cysewski, P.; Przybyłek, M. Selection of Effective cocrystals Former for Dissolution Rate Improvement of Active Pharmaceutical Ingredients Based on Lipoaffinity Index. *Eur. J. Pharm. Sci.* **2017**, *107*, 87−96.

(24) Xiao, F.; Cheng, Y.; Wang, J.-R.; Wang, D.; Zhang, Y.; Chen, K.; Mei, X.; Luo, X. cocrystal Prediction of Bexarotene by Graph Convolution Network and Bioavailability Improvement. *Pharmaceutics* **2022**, *14* (10), 2198.

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

(25) Devogelaer, J.; Meekes, H.; Tinnemans, P.; Vlieg, E.; Gelder, R. Co-crystal Prediction by Artificial Neural Networks**. *Angew. Chem.* **2020**, *132* (48), 21895−21902.
(26) Fornari, F.; Montisci, F.; Bianchi, F.; Cocchi, M.; Carraro, C.; Cavaliere, F.; Cozzini, P.; Peccati, F.; Mazzeo, P. P.; Riboni, N.; Careri, M.; Bacchi, A. chemometric-Assisted Cocrystallization: Supervised Pattern Recognition for Predicting the Formation of New Functional cocrystals. *Chemom. Intell. Lab. Syst.* **2022**, *226*, No. 104580.
(27) Zheng, L.; Zhu, B.; Wu, Z.; Liang, F.; Hong, M.; Liu, G.; Li, W.; Ren, G.; Tang, Y. SMINBR: An Integrated Network and Chemo-informatics Tool Specialized for Prediction of Two-Component Crystal Formation. *J. Chem. Inf. Model.* **2021**, *61* (9), 4290−4302.
(28) Jiang, Y.; Yang, Z.; Guo, J.; Li, H.; Liu, Y.; Guo, Y.; Li, M.; Pu, X. Coupling Complementary Strategy to Flexible Graph Neural Network for Quick Discovery of coformer in Diverse Co-Crystal Materials. *Nat. Commun.* **2021**, *12* (1), 5950.
(29) *Rigaku Oxford Diffraction, CrysAlisPro*; Oxford Diffraction Ltd.: Abingdon, Oxfordshire, England, 2015.
(30) Sheldrick, G. M. A Short History of *SHELX*. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2008**, *64* (1), 112−122.
(31) Macrae, C. F.; Sovago, I.; Cottrell, S. J.; Galek, P. T. A.; McCabe, P.; Pidcock, E.; Platings, M.; Shields, G. P.; Stevens, J. S.; Towler, M.; Wood, P. A. *Mercury 4.0*: From Visualization to Analysis, Design and Prediction. *J. Appl. Crystallogr.* **2020**, *53* (1), 226−235.
(32) Díaz de los Ríos, M.; Hernández Ramos, E. Determination of the Hansen Solubility Parameters and the Hansen Sphere Radius with the Aid of the Solver Add-in of Microsoft Excel. *SN Appl. Sci.* **2020**, *2* (4), 676.
(33) Aakeröy, C. B.; Salmon, D. J. Building Co-Crystals with Molecular Sense and Supramolecular Sensibility. *CrystEngComm* **2005**, *7* (72), 439.
(34) Mohammad, M. A.; Alhalaweh, A.; Velaga, S. P. Hansen Solubility Parameter as a Tool to Predict cocrystal Formation. *Int. J. Pharm.* **2011**, *407* (1−2), 63−71.
(35) Ballabio, D.; Consonni, V. Classification Tools in Chemistry. Part 1: Linear Models. PLS-DA. *Anal. Methods* **2013**, *5*, 3790−3798.
(36) Fratello, M.; Tagliaferri, R. Decision Trees and Random Forests. In *Encyclopedia of Bioinformatics and Computational Biology*; Elsevier, 2019; pp 374−383.
(37) Nastro, F.; Sorrentino, M.; Trifirò, A. A Machine Learning Approach Based on Neural Networks for Energy Diagnosis of Telecommunication Sites. *Energy* **2022**, *245*, No. 123266.
(38) Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated Double Cross Validation. *J. Chemom.* **2009**, *23* (4), 160−171.
(39) Filzmoser, P.; Varmuza, K. *Chemometrics: Multivariate Statistical Analysis in Chemometrics*; 2016.
(40) Kuhn, M. *Caret: Classification and Regression Training*; 2020.
(41) Ballabio, D.; Consonni, V. Classification Tools in Chemistry. Part 1: Linear Models. PLS-DA. *Anal. Methods* **2013**, *5* (16), 3790.
(42) Brereton, R. G. *Applied Chemometrics for Scientists*; 1st ed.; John Wiley & Sons, Ltd., Ed.; John Wiley & Sons, Ltd.: Chichester, UK, 2007.
(43) Mevik, B.-H.; Wehrens, R.; Liland, K. H. *Pls: Partial Least Squares and Principal Component Regression*; 2020.
(44) Scott, I. M.; Lin, W.; Liakata, M.; Wood, J. E.; Vermeer, C. P.; Allaway, D.; Ward, J. L.; Draper, J.; Beale, M. H.; Corol, D. I.; Baker, J. M.; King, R. D. Merits of Random Forests Emerge in Evaluation of chemometric Classifiers by External Validation. *Anal. Chim. Acta* **2013**, *801*, 22−33.
(45) Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2* (3), 18−22.
(46) Leardi, R. Chemometrics: From Classical to Genetic Algorithms. *Grasas Aceites* **2002**, *53* (1), 115−127.
(47) Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S*; 4th ed.; Springer: New York, 2002.
(48) Cerreia Vioglio, P.; Chierotti, M. R.; Gobetto, R. Pharmaceutical Aspects of Salt and cocrystal Forms of APIs and Characterization Challenges. *Adv. Drug Delivery Rev.* **2017**, *117*, 86−110.
(49) Bhogala, B. R.; Basavoju, S.; Nangia, A. Tape and Layer Structures in cocrystals of Some Di- and Tricarboxylic Acids with 4,4′-Bipyridines and isonicotinamide. From Binary to Ternary cocrystals. *CrystEngComm* **2005**, *7* (90), 551.
(50) Cruz-Cabeza, A. J. Acid−Base Crystalline Complexes and the PKa Rule. *CrystEngComm* **2012**, *14* (20), 6362.

https://doi.org/10.1021/acs.cgd.3c00696
Cryst. Growth Des. 2023, 23, 7898−7911

Speeding Up the Cocrystallization Process: Machine Learning-Combined Methods for the Prediction of Multicomponent Systems

Rebecca Birolo[a], Federica Bravetti[a], Eugenio Alladio[a], Emanuele Priola[a], Gianluca Bianchini[b], Rubina Novelli[c], Andrea Aramini[b], Roberto Gobetto[a], Michele R. Chierotti[a]**

[a]Department of Chemistry and NIS centre, University of Torino, Via P. Giuria 7, 10125 Torino, Italy

[b] Research and Early Development, Dompé Farmaceutici S.p.A, Via Campo di Pilel, 67100 L'Aquila, Italy

[c] Research and Early Development, Dompé Farmaceutici S.p.A., Via S. Lucia 6, 20122 Milano, Italy

Table S1. Riluzole dataset: 35 cases of which 13 NO and 22 YES. Codes in the reference column refer to the code of the adduct structure reported on the Cambridge Structural Database (CSD).

coformer	Exp	reference
adipic acid	YES	NAQNEB
ascorbic acid	NO	[1]
azelaic acid	YES	NAQPIH
benzoic acid	NO	[1]
caffeic acid	NO	[1]
chrysin	NO	[1]
cinnamic acid	YES	YEPKAI
citric acid	NO	[1]
2,4-dihydroxybenzoic acid	YES	[1]
4-dimethylaminopyridine	YES	NAQPUT
ferulic acid	YES	YEPHOT
fumaric acid	YES	YEPJOV
gallic acid	NO	[1]
glutaric acid	YES	NAQNAX
maleic acid	YES	[1]
malic acid	NO	[1]
malonic acid	YES	NAQMUQ
3,4,5-trimethoxybenzoic acid	YES	ZIYFOF

coformer	Exp	reference
nicotinic acid	YES	NAQPON
p-coumaric acid	NO	[1]
proline	YES	YEPJEL
quercitin	NO	[1]
salicylamide	NO	[1]
salicylic acid	NO	[1]
sebacic acid	YES	NAQPED
sylibinin	NO	[1]
sorbic acid	YES	NAQNOL
suberic acid	YES	NAQPAZ
succinic acid	YES	NAQMOK
syringic acid	YES	[1]
tartaric acid	YES	[1]
vanillic acid	YES	YEPJUB
vanillin	NO	[1]
pimelic acid	YES	NAQNUR
nicotinamide	YES	ZIYFUL

Table S2. Diclofenac dataset: 19 cases of which 5 NO and 14 YES. Codes in the reference column refer to the code of the adduct structure reported on CSD.

coformer	Exp	reference
2-aminopyridine	YES	[2]
2-amino-5-chloropyridine	YES	[2]
4,4'-bipyridine	YES	[2]
pyrazole	NO	[2]
3-aminopyridine	YES	[2]
2-amino-3,5-dibromopyridine	YES	[2]
2-amino-4-chloro-6-methylpyrimidine	YES	[2]
3-hydroxypyridine	YES	[2]
2-aminopyrimidine	YES	[2]
2-chloropyrimidine	NO	[2]

coformer	Exp	reference
2-amino-4,6-dimethylpyrimidine	YES	[2]
4-bromopyrazole	NO	[2]
4-chloro-2,6-diaminopyrimidine	YES	[2]
isonicotinamide	YES	[2]
3,5-dimethyl-4-chloropyrazole	NO	[2]
2-amino-4-hydroxy-6-methylpyrimidine	YES	[2]
3,5-dimethylpyrazole	NO	[2]
theophylline	YES	[3]
L-proline	YES	RETNEM

Table S3. Indomethacin dataset: 61 cases of which 47 NO and 14 YES. The codes in the reference column refer to the code of the adduct structure reported on CSD.

coformer	Exp	reference
4,4'-bypiridine	YES	[4]
benzoic acid	NO	[5]
saccharin	YES	UFERED
1-hydroxy-2-naphtoic acid	NO	[5]
L-tryptophan	NO	[5]
cinnamic acid	YES	[4]
4-aminobenzoic acid	NO	[4]
vanillic acid	NO	[4]
benzamide	YES	[5]
4-hydroxybenzoic acid	NO	[5]
hippuric acid	NO	[5]
nicotinamide	YES	SESKUY
isonicotinamide	NO	[4]
ethyl maltol	NO	[4]
lidocaine	YES	DEWNOJ
4-aminobenzamide	NO	[4]
gentisic acid	NO	[5]
mandelic acid	YES	[5]
4-hydroxybenzamide	NO	[4]
neotame	NO	[4]
sorbic acid	NO	[5]
cyclamic acid	NO	[4]
L-lysine	NO	[5]
adipic acid	NO	[5]
stearic acid	NO	[5]
L-leucine	NO	[5]
naphthalenesulfonic acid	NO	[5]
2-methoxy-5-nitroaniline	YES	JAMYUV
1,2-ethanedisulfonic acid	NO	[4]
2,5-dihydroxybenzoic acid	NO	[5]
2-hydroxy-4-methylpyridine	YES	[6]

coformer	Exp	reference
maltose	NO	[5]
lactose	NO	[5]
glycine	NO	[5]
succinic acid	NO	[5]
lactic acid	NO	[5]
arabinose	NO	[4]
tromethamine	YES	[5]
L-aspartic acid	NO	[5]
L-ascorbic acid	NO	[5]
malonic acid	NO	[5]
N-methyl-D-glucamine	YES	[5]
mannose	NO	[4]
glucose	NO	[4]
citric acid	NO	[5]
malic acid	NO	[5]
glycolamide	NO	[5]
glycolic acid	NO	[5]
tartaric acid	NO	[5]
oxalic acid	NO	[5]
D-mannitol	NO	[5]
fumaric acid	NO	[5]
caffeine	YES	JELJES
carbamazepine	YES	LEZKEI
salicylic acid	NO	[5]
lactamide	YES	[5]
urea	NO	[5]
p-toluenesulfonic acid	NO	[5]
L-arginine	NO	[5]
glutaric acid	NO	[5]
maleic acid	NO	[5]

Table S4. Nalidixic acid dataset: 35 cases of which 22 NO and 13 YES. Codes in the reference column refer to the code of the adduct structure reported on CSD.

coformer	Exp	reference
isonicotinamide	NO	[4]
nicotinamide	NO	[4]
cytosine	NO	[4]
thymine	NO	[4]
adenine	NO	[4]
nicotinic acid	NO	[4]
salicylic acid	NO	[4]
L-histidine	NO	[4]
phloroglucinol	YES	[4]
benzoic acid	NO	[4]
orcinol	YES	[4]
ferulic acid	NO	[4]
L-proline	NO	[4]
catechol	YES	[4]
pyrogallol	YES	[4]
resorcinol	YES	[4]
3,4-dihydroxybenzoic acid	YES	[7]
2,4-dihydroxybenzoic acid	YES	[7]

coformer	Exp	reference
acetamide	NO	[4]
L-threonine	NO	[4]
L-lysine	NO	[4]
hydroquinone	YES	[4]
t-butylhydroquinone	YES	DIYYOB
n-propyl gallate	YES	DIYYUH
biphenyl-2-ol	YES	DYZES
indole	YES	DIYZAO
skatole	YES	[7]
L-glutamine	NO	[4]
etidronic acid	NO	[7]
tartaric acid	NO	[7]
citric acid	NO	[7]
fumaric acid	NO	[7]
malic acid	NO	[7]
ascorbic acid	NO	[7]
urea	NO	[4]

Table S5. Piracetam dataset: 31 cases of which 17 NO and 14 YES. Codes in the reference column refer to the code of the adduct structure reported on CSD.

coformer	Exp	reference
3,4-dihydroxybenzoic acid	YES	[8]
3,5-dihydroxybenzoic acid	YES	[8]
hydroquinone	YES	[9]
gentisic acid	YES	DAVPAS
4-hydroxybenzoic acid	YES	DAVPEW
fumaric acid	NO	[4]
2,3-dihydroxybenzoic acid	YES	[8]
urea	NO	[4]
2,4-dihydroxybenzoic acid	YES	[8]
succinic acid	NO	[4]
mandelic acid	YES	RUCFIF
tartaric acid	YES	FIXROV
glucuronic acid	NO	[4]
saccharin	NO	[4]
4-acetamidobenzoic acid	NO	[4]
2,6-dihydroxybenzoic acid	NO	[4]

coformer	Exp	reference
maleic acid	NO	[4]
citric acid	YES	RUCFAX
glutaric acid	YES	[10]
glycine	NO	[4]
camphoric acid	NO	[4]
mannitol	NO	[4]
piperazine	NO	[4]
hippuric acid	NO	[4]
sulfaproxyline	NO	[4]
proline	NO	[4]
imidazole	NO	[4]
gallic acid	YES	AKISEU
pyridine-2,6-diamine	YES	DEDMAD
myricetin	YES	FIXROV
citric acid	YES	RUCFAX
2-amino-5-methylbenzoic acid	NO	[4]

Table S6. Carbamazepine dataset: 52 cases of which 18 NO and 34 YES. Codes in the reference column refer to the code of the adduct structure reported on CSD.

coformer	Exp	reference
5-nitroisophthalic acid	YES	UNIBEY
trimesic acid	YES	UNIBAU
saccharin	YES	EYEJAW
2,6-pyridinecarboxylic acid	YES	XAQRIR
aspirin	YES	TAZRAO
butyric acid	YES	UNEZUI
nicotinamide	YES	UNEZES
sulfuric acid	NO	[12]
sulfamic acid	NO	[12]
etidronic acid	NO	[12]
oxalic acid	YES	MOXWUS
fumaric acid	YES	WEYFEN
gentisic acid	YES	[13]
tartaric acid	YES	MOXWIG
ketoglutaric acid	NO	[13]
t-butylhydroquinone	NO	[12]
malonic acid	YES	MOXVUR
isocitric acid	NO	[12]
4-nitropyridine-N-oxide	YES	JIQKUS
5-chlorosalycylic acid	NO	[12]
4-hydroxybenzoic acid	YES	MOXVIF
1-hydroxy-2-naphtholic acid	YES	MOXWEC
salicylic acid	YES	MOXWAY
benzoic acid	YES	MOXVAX
nitromethane	YES	KIWBOI
glycolic acid	NO	[13]

coformer	Exp	reference
N-methylpyrrolidone	YES	KIWBIC
DL-mandelic acid	YES	[11]
indomethacin	YES	LEZKEI
adipic acid	YES	MOXVEB
isonicotinamide	YES	LOFKIB
succinic acid	YES	XOBCIB
thiourea	YES	UWAZID
caffeine	NO	[11]
flurbiprofen	NO	[11]
ketoprofen	YES	RAFGEO
lactulose	NO	[11]
paracetamol	NO	[11]
ibuprofen	NO	[13]
simvastatin	NO	[11]
theophylline	NO	[11]
camphoric acid	YES	MOXXAZ
p-aminosalicylic acid	YES	FAYXOV
benzene-1,4-diol	YES	ABOQUF
malic acid	YES	[12]
2-aminopyrimidine	YES	JIQLAZ
glutaric acid	YES	MOXVOL
4-hydroxybenzamide	YES	SOGSEP
pterostilbene	YES	YABHIU
argine	NO	[13]
lysine	NO	[13]
lactamide	NO	[13]

Table S7. Acetazolamide dataset: 37 cases of which 27 NO and 10 YES. Codes in the reference column refer to the code of the adduct structure reported on CSD.

coformer	Exp	reference
4-aminobenzamide	NO	[4]
3,5-dihydroxybenzoic acid	NO	[4]
3-aminobenzamide	NO	[4]
caffeine	NO	[4]
3,4-dihydroxybenzoic acid	NO	[4]
2-aminobenzamide	YES	DATFAH
2-hydroxybenzamide	YES	DATFEL
4-aminobenzoic acid	NO	[4]
2,3-dihydroxybenzoic acid	YES	DATDUZ
4-hydroxybenzoic acid	YES	RUYGIC
3-hydroxybenzamide	NO	[4]
4-hydroxybenzamide	NO	[4]
pyridine-2-carboxamide	YES	DATFIP
salicylic acid	NO	[4]
3-hydroxybenzoic acid	NO	[4]
nicotinamide	YES	MADTAP
gentisic acid	NO	[4]
2,4-dihydroxybenzoic acid	NO	[4]
2-aminobenzoic acid	NO	[4]

coformer	Exp	reference
3-aminobenzoic acid	NO	[4]
isonicotinic acid	NO	[4]
succinic acid	NO	[4]
nicotinic acid	NO	[4]
oxalic acid	NO	[4]
adipic acid	NO	[4]
isonicotinamide	NO	[4]
camphoric acid	NO	[4]
2-hydroxynicotinic acid	NO	[4]
malonic acid	NO	[4]
saccharin	NO	[4]
citric acid	NO	[4]
glutaric acid	NO	[4]
benzoic acid	NO	[4]
valerolactam	YES	MADGIK
2-pyridone	YES	MADSAO
6-methyl-2-pyridone	YES	MADSUI
theophylline	YES	YEVMUK

Table S8. Furosemide dataset: 37 cases of which 20 NO and 17 YES. Codes in the reference column refer to the code of the adduct structure reported on CSD.

coformer	Exp	reference
gallic acid	NO	[14]
cytosine	YES	XAVTIZ
urea	YES	XUDYED
asparagine	NO	[14]
aspirin	NO	[14]
caffeine	YES	XAVTEV
salycilic acid	NO	[14]
4-aminobenzoic acid	YES	[14]
fumaric acid	NO	[14]
aspartic acid	NO	[14]
tartaric acid	NO	[14]
acetamide	YES	[14]
saccharin	NO	[14]
2,3,5,6-tetramethylpyrazine	YES	BUQTAL
succinic acid	NO	[14]
glutamic acid	NO	[14]
citric acid	NO	[14]
nicotinamide	YES	YASGOQ
azepan-2-one	YES	NOLBEY

coformer	Exp	reference
benzamide	NO	[14]
glutamine	NO	[14]
nicotinic acid	NO	[14]
adipic acid	NO	[14]
4-hydroxybenzamide	NO	[14]
leucine	NO	[14]
phenylalanine	NO	[14]
benzoic acid	NO	[14]
ascorbic acid	NO	[14]
isonicotinamide	YES	[14]
4-4'-bipyridine	YES	BOKHAM
anthranilamide	YES	ESAVIF
piperazine	ES	ESAVOL
adenine	YES	[14]
pentoxifylline	YES	FEFYAS
2,2'-bipyridine	YES	HUQWAT
4-aminopyridine	YES	HUQWEX
triamterene	YES	HIQXEN

Table S9. Caffeine dataset: 62 cases of which 10 NO and 52 YES. Codes in the reference column refer to the code of the adduct structure reported on CSD.

coformer	Exp	reference
maleic acid	YES	GANYEA
glutaric acid	YES	EXUQUJ
malonic acid	YES	GANYAW
citric acid	YES	KIGKER
oxalic acid	YES	GANXUP
L-lactic acid	NO	[13]
octadecylamine	NO	[12]
etidronic acid	NO	[12]
gentisic acid	YES	MOZDIP
3-hydroxy-2-naphthoic acid	YES	KIGKOB
5-Fluorocytosine	YES	SIMBOI
epalrestat	YES	MAVQIM
genistein	YES	BOLBAH
indomethacin	YES	JELJES
lesinurad	YES	WOHWUO
1-hydroxy-2-naphthoic acid	YES	KIGKIV
2,3,4,5-tetrafluorobenzoic acid	YES	AFEMIJ
2,3-difluorobenzoic acid	YES	AFEQOT
2-chloro-5-nitroaniline	YES	LATGOE
2-fluorobenzoic acid	YES	AFERAG
3-(phenylthio)propanoic acid	YES	GOMDET
3,5-pyrazoledicarboxylic acid	YES	UNISUG
3-hydroxybenzoic acid	YES	MOZCOU
cinnamic acid	YES	AGIGEE
4-amino-salicylic acid	YES	ZEQCEG
4-chloro-3-nitroaniline	YES	LATGEU
4-chloro-3-nitrobenzoic acid	YES	DIPHUH
4-fluoro-3-nitroaniline	YES	LATGIY
4-fluoro-3-nitrobenzoic acid	YES	ARIFUE
2-fluoro-5-nitrobenzoic acid	YES	LATHIZ
6-hydroxy-2-naphthoic acid	YES	KIGKUH

coformer	Exp	reference
salicylic acid	YES	XOBCAT
anthranilic acid	YES	ZOBCOK
5-chlorosalicylic acid	YES	[12]
p-coumaric acid	YES	IJEZUT
3-nitrobenzoic acid	YES	[12]
pterostilbene	YES	YABHAM
paracetamol	YES	[15]
myricetin	YES	DOZGUX
saccharin	YES	[12]
indole	NO	[12]
adipic acid	YES	CESKAN
fumaric acid	NO	[12]
benzoic acid	YES	AFEREK
dapsone	YES	VOHKOU
D-tartaric acid	YES	TUGJOW
furosemide	YES	XAVTEV
isophthalic acid	YES	PUPGUD
L-malic acid	YES	HOLVOV
methyl gallate	YES	DIJVOH
niclosamide	YES	HEBDUP
sulfacetamide	YES	SACCAF
temozolomide	YES	KIJTEE
theophylline	YES	NEHJER
trimesic acid	YES	HUHQUZ
zonisamide	YES	TEZGIR
prochlorperazine	NO	[13]
tolfenamic acid	NO	[13]
acetylsalicylic acid	NO	[13]
ibuprofen	NO	[13]
piracetam	NO	[13]
4-nitroaniline	YES	LATGUK

Table S10. Pyrazinamide dataset: 50 cases of which 23 NO and 27 YES. Codes in the reference column refer to the code of the adduct structure reported on CSD.

coformer	Exp	reference
3,5-dihydroxybenzoic acid	YES	[16]
pyrazine-2-carboxylic acid	NO	[4]
gentisic acid	YES	[4]
4-hydroxybenzoic acid	YES	NUVFIV01
oxalic acid	YES	[17]
3-hydroxybenzoic acid	YES	[4]
1-hydroxy-2-naphthoic acid	YES	[4]
fumaric acid	YES	LATTIL
indole-2-carboxylic acid	YES	[4]
salicylic acid	YES	[4]
ketoglutaric acid	NO	[4]
malonic acid	YES	[18]
4-aminobenzoic acid	YES	VUTNAB
succinic acid	YES	LATTOR
indole-3-carboxylic acid	NO	[4]
undecylenic acid	NO	[4]
2-aminobenzoic acid	YES	[4]
pyruvic acid	NO	[4]
tartaric acid	NO	[4]
camphoric acid	NO	[4]
4-aminosalicylic acid	YES	[4]
sinapic acid	YES	THISAH
2,4-dihydroxybenzoic acid	YES	NEFFEM
2,6-dihydroxybenzoic acid	YES	NEFGEN
2,3-dihydroxybenzoic acid	YES	NEFGIR

coformer	Exp	reference
phenylalanine	NO	[4]
gallic acid	YES	[4]
azelaic acid	NO	[4]
vanillic acid	YES	[4]
hydrocinnamic acid	NO	[4]
saccharin	NO	[4]
nicotinic acid	NO	[4]
4-nitrobenzamide	YES	[4]
glutaric acid	YES	[18]
hippuric acid	NO	[4]
3-aminobenzoic acid	YES	[4]
benzoic acid	NO	[4]
glycolic acid	NO	[4]
pyroglutamic acid	NO	[4]
isonicotinamide	NO	[4]
nicotinamide	NO	[4]
cinnamic acid	NO	[4]
histidine	NO	[4]
benzamide	NO	[4]
pyrogallol	YES	HEDRAL
temozolomide	YES	KIJSED
adipic acid	YES	[19]
isonicotinic acid	NO	[4]
tyrosine	NO	[4]
theophylline	YES	RACFIN

Table S11. Paracetamol dataset: 38 cases of which 21 NO and 17 YES. Codes in the reference column refer to the code of the adduct structure reported on CSD.

coformer	Exp	reference
adipic acid	NO	[4]
cyclam	YES	[4]
morpholine	YES	AHEPUY
caffeine	YES	[4]
1,4-diaminocyclohexane	YES	WIGCEW
pyrazine	NO	[20]
imidazole	NO	[4]
succinic acid	NO	[4]
maleic acid	NO	[4]
nicotinamide	NO	[4]
isonicotinamide	NO	[4]
piperazine	YES	MUPPUI
N,N-dimethyl piperazine	YES	MUPPIW
benzoic acid	NO	[4]
malonic acid	NO	[4]
ascorbic acid	NO	[4]
2,4-pyridinedicarboxylic acid	YES	SUTVAF
2,5-dihydroxybenzoic acid	NO	[20]
4,4'-trimethylenedipyridine	NO	[4]

coformer	Exp	reference
naphthalene	YES	LUJSIT
4,4'-bipyridine	YES	MUPQAP
saccharin	NO	[4]
oxalic acid	YES	LUJTAM
phenazine	YES	LUJSIOZ
anthracene	NO	[4]
melamine	NO	[4]
5-nitroisophthalic acid	YES	[21]
4,4'-ethane-1,2-diyldipyridine	YES	WIGBUL
citric acid	YES	AMUBAM
theophylline	YES	KIGLUI
fumaric acid	NO	[20]
1-naphthol	NO	[4]
resorcinol	NO	[4]
malic acid	NO	[4]
3-isochromanone	NO	[4]
DABCO	YES	[4]
N-methylmorpholine	YES	MUPPOC
theobromine	NO	[4]

Table S12. Sulpiride dataset: 33 cases of which 21 NO and 12 YES. Codes in the reference column refer to the code of the adduct structure reported on CSD. "Unpublished" stands for cocrystallization tests performed by us but not reported in literature.

coformer	Exp	reference
adipic acid	YES	KENWOT
caffeic acid	YES	[22]
maleic acid	YES	KENVOS
malic acid	YES	[22]
quercitin	NO	unpublished
4-aminobenzoic acid	YES	[22]
hippuric acid	NO	unpublished
lactose	NO	unpublished
mannitol	NO	unpublished
caffeine	NO	unpublished
cytosine	NO	unpublished
thymine	NO	unpublished
trimesic_Acid	NO	unpublished
piracetam	NO	unpublished
ketoglutaric acid	NO	unpublished
succinic acid	YES	[22]
tyrosine	NO	unpublished

coformer	Exp	reference
GABA	NO	unpublished
acetylcysteine	NO	unpublished
fumaric acid	YES	KENWIN
malonic acid	YES	KENVUY
nicotinic acid	YES	[22]
proline	NO	unpublished
ascorbic acid	NO	unpublished
theophylline	NO	unpublished
lysine	NO	unpublished
n-propyl gallate	NO	unpublished
indomethacin	YES	KENWAF
ibuprofen	YES	[22]
acetazolamide	YES	KENWEJ
L-phenylalanine	NO	unpublished
glutamic acid	NO	unpublished
melatonin	NO	unpublished

Table S13. Piroxicam dataset: 53 cases of which 12 NO and 41 YES. Codes in the reference column refer to the code of the adduct structure reported on CSD.

coformer	Exp	reference
caprylic acid	YES	DIKCUW
camphoric acid	YES	[13]
sebacic acid	YES	[13]
benzoic acid	YES	DIKDOR
cinnamic acid	NO	[13]
azelaic acid	YES	[13]
suberic acid	YES	[13]
methylparaben	YES	[13]
pimelic acid	YES	[13]
isophthalic acid	YES	[13]
phenylsuccinic acid	YES	[13]
1-hydroxy-2-naphthoic acid	YES	DIKCOQ
vanillin	NO	[13]
catechol	NO	[13]
terephthalic acid	NO	[13]
anthranilic acid	NO	[13]
adipic acid	YES	[13]
salicylic acid	YES	CEKNEN
desaminotyrosinne	NO	[13]
ferulic acid	NO	[13]
p-aminobenzoic acid	NO	[13]
3-hydroxybenzoic acid	YES	[13]
4-hydroxybenzoic acid	YES	DIKDEH
trime YESc acid	YES	[13]
hippuric acid	YES	[13]
resorcinol	YES	[13]
furosemide	YES	[13]

coformer	Exp	reference
glutaric acid	YES	[13]
L-pyroglutamic	NO	[13]
nicotinamide	YES	[13]
hydrocaffeic acid	YES	[13]
genti YESc acid	YES	TUFNUF
succinic acid	YES	DIKCIK
ketoglutaric acid	YES	[13]
maleic acid	YES	[13]
fumaric acid	YES	DIKDIL
malonic acid	YES	DIKDAD
citric acid	NO	[13]
malic acid	YES	[13]
glycolic acid	NO	[13]
oxalic acid	YES	[13]
tartaric acid	NO	[13]
2-fluorobenzoic acid	YES	CEKLAH
2-methylbenzoic acid	YES	CEKLEL
3-chlorobenzoic acid	YES	CEKLOV
3-nitrobenzoic acid	YES	CEKMAI
febuxostat	YES	RUNSUR
ethanolamine	YES	SECDAF
bromanilic acid	YES	SOHVOC
benzotriazole	YES	SOHXAQ
triazole	YES	SOHWUJ
benzimidazole	YES	SOHWOD
2-methylimidazole	YES	SOHWIX

Table S14. Training set data. Codes in the reference column refer to the code of the adduct structure reported on CSD. FLU = flurbiprofen; IBU = ibuprofen; NAP = naproxen; ZAL = zaltoprofen.

API	coformer	Exp	reference
FLU	4,4'-bipyridine	YES	HUPPEN
FLU	4,4'-ethylenbipyridine	YES	HUPPIR
FLU	benzamide	YES	JOSPIT
FLU	salicylamide	YES	JOSPEP
FLU	picolinamide	YES	JOSPOZ
FLU	nicotinamide	YES	QOBGAR
IBU	citric acid	YES	[23]
IBU	oxalic acid	YES	[23]
IBU	nicotinamide	YES	SODDIZ, SOGLAC
IBU	isonicotinamide	YES	KIPPAD
IBU	piperazine	YES	[24]
IBU	pyridine-2-carboxamide	YES	SAJCOY
IBU	picolinamide	YES	[24]
IBU	L-proline	YES	[24]
IBU	2-aminopyrimidine	YES	TAWSOB
IBU	4,4'-bipyridine	YES	HUPPAJ
IBU	4,4'-ethylenbipyridine	YES	OWIGEH
IBU	caffeine	NO	[24]
IBU	urea	NO	[24]
IBU	nicotinic acid	NO	[24]
IBU	benzoic acid	NO	[24]
IBU	salicylamide	NO	[24]
IBU	glutaric acid	NO	[24]
IBU	saccharin	NO	[24]
IBU	L-alanine	NO	[24]
IBU	D-tyrosine	NO	[24]
IBU	D-tryptophan	NO	[24]
IBU	4-aminosalicylic acid	NO	[24]
IBU	2-aminopyridine	NO	[24]
IBU	resorcinol	NO	[24]
NAP	4,4'-ethane-1,2-diyldipyridine	YES	AFOPET
NAP	pyridine-2-carboxamide	YES	SAJCOY

API	coformer	Exp	reference
NAP	4,4'-trimethylenedipyridine	YES	AFOPOD
NAP	4,4'-azopyridine	YES	AFOPIX
NAP	4,4'-ethylenedipyridine	YES	COZYAS
NAP	piperidine	YES	TOBMEE
NAP	1,2,4,5-tetracyanobenzene	YES	YOCZUL
NAP	L-proline	YES	QILZET
NAP	L-alanine	YES	RODSEK
NAP	D-tryptophan	YES	RODSOU
NAP	D-tyrosine	YES	RODSUA
NAP	nicotinamide	YES	HEGGAD
NAP	isonicotinamide	YES	PAMQAX
NAP	4,4'-bipyridine	YES	TOBLUT
NAP	N-octyl-D-glucamine	YES	MICFEJ
ZAL	nicotinamide	YES	[24]
ZAL	isonicotinamide	YES	[24]
ZAL	2-aminopyridine	NO	[24]
ZAL	pyrazinamide	NO	[24]
ZAL	resorcinol	NO	[24]
ZAL	caffeine	NO	[24]
ZAL	saccharin	NO	[24]
ZAL	nicotinic acid	NO	[24]
ZAL	benzoic acid	NO	[24]
ZAL	malonic acid	NO	[24]
ZAL	glutaric acid	NO	[24]
ZAL	benzamide	NO	[24]
ZAL	4-aminosalicylic acid	NO	[24]
ZAL	picolinamide	NO	[24]
ZAL	4-aminobenzoic acid	NO	[24]
ZAL	L-proline	NO	[24]
ZAL	D-alanine	NO	[24]
ZAL	D-tyrosine	NO	[24]
ZAL	D-tryptophan	NO	[24]

Table S15. Test set data and synthetic information for the preparation of the adducts.

NAME	API	CPFORMER	EXP	Stoichiometry	Synthesis
-	PPA	histidine	NO	1:1	-
-	PPA	acetamide	NO	1:1	-
-	PPA	4-nitroaniline	NO	1:1	-
-	PPA	nicotinic acid	NO	2:1/1:1	-
PPA-PRO	PPA	proline (PRO)	YES	1:1	LAG
PPA-4APY	PPA	4-aminopyridine (4APY)	YES	2:1	LAG
PPA-NA	PPA	nicotinamide (NA)	YES	1:1	grinding
PPA-2APY	PPA	2-aminopyridine (2APY)	YES	1:1	grinding
PPA-4ClBA	PPA	4-chlorobenzamide (4ClBA)	YES	1:1	LAG
-	PPA	3-hydroxypyridine	NO	2:1/1:1	-
PPA-INA	PPA	isonicotinamide (INA)	YES	2:1	grinding
PPA-BA	PPA	benzamide (BA)	YES	1:1	LAG
PPA-BPY	PPA	1,2-bis(4-pyridyl)ethane (BPY)	YES	2:1	LAG

PROCEDURE FOR CALCULATING MOLECULAR DESCRIPTORS

HSP. The HSPiP software was employed to derive the Hansen solubility descriptors. The SMILES code of each molecule was entered into the programme, which automatically provides the three solubility parameters δ_P, δ_D, δ_H and the descriptors H_A, H_D and MVol. The solubility parameters that are obtainable from the software have two different degrees of approximation: experimental data (accurate) or predicted by the programme (approximation between 5% and 20%, generally the values of the δ_D parameter being the most different from the experimental ones). The experimental data from the software database were used whenever available. From the solubility descriptors of API and coformer, it is then possible to derive the descriptors Δδ, Δδt and Ra according to the equations 1-4 reported in the main text. The calculation of these parameters was performed by a self-made script in Python.

HBE. Each molecule in the dataset was drawn on GaussView to create the Gaussian-readable input file. The geometry of the gas-phase molecule was then optimised at the DFT level. The basis set used in most of the works reported in the literature for the HBE method is the B3LYP 6-311++G**, which is a modest basis set, generally used for rapid geometry optimisation. However, the def2-SVP (Split Valence)[31] basis set, which is better in terms of accuracy than the classical 6-31G sets, was used as an alternative in this work. Molecular electrostatic potential surfaces (MEPS) were calculated with an external software programme, Multiwfn,[32] and then derived the α and β parameters (Equations 5 and 6 in the main text). The values of the α and β parameters were finally combined according to the theory of the HBE predictive method for all the adducts in the dataset in order to obtain their interaction energy. The E_{API}, E_{cof}, E_{adduct} and ΔE parameters were calculated using a script in Python.

CM. The software used to derive the CM descriptors was Mercury 2020 3.0. The routine requires as input a *"mol2"* file describing the structure of APIs and coformers. Since the geometry for each molecule was optimized for the HBE method, the obtained structure was also used for this tool by simply converting the output file generated by Gaussian into a *"mol2"* file. Thus, the Molecular Complementarity Screening Wizard function was applied. The parameters set for the analysis are the standard parameters set by default configuration of the software.

TEST SET EXPERIMENTAL DATA

Each sample obtained during the experimental tests (Table S15) was characterized using different spectroscopic and diffraction techniques. All the experimental data are reported in this section. FTIR-ATR and PXRD were used for a rapid qualitative analysis on the adduct formation, comparing the spectrum or diffraction pattern of the adduct with those of the pure starting materials. In particular, FTIR-ATR spectroscopy is able to discriminate samples between adducts or simple physical mixtures in a highly precise way while PXRD is useful for checking the purity of the sample. Instead, SSNMR and SCXRD were employed to achieve a deep characterization of the new crystal forms and evaluate the protonic state of the adduct (*i.e.* salt or cocrystal).

FTIR-ATR

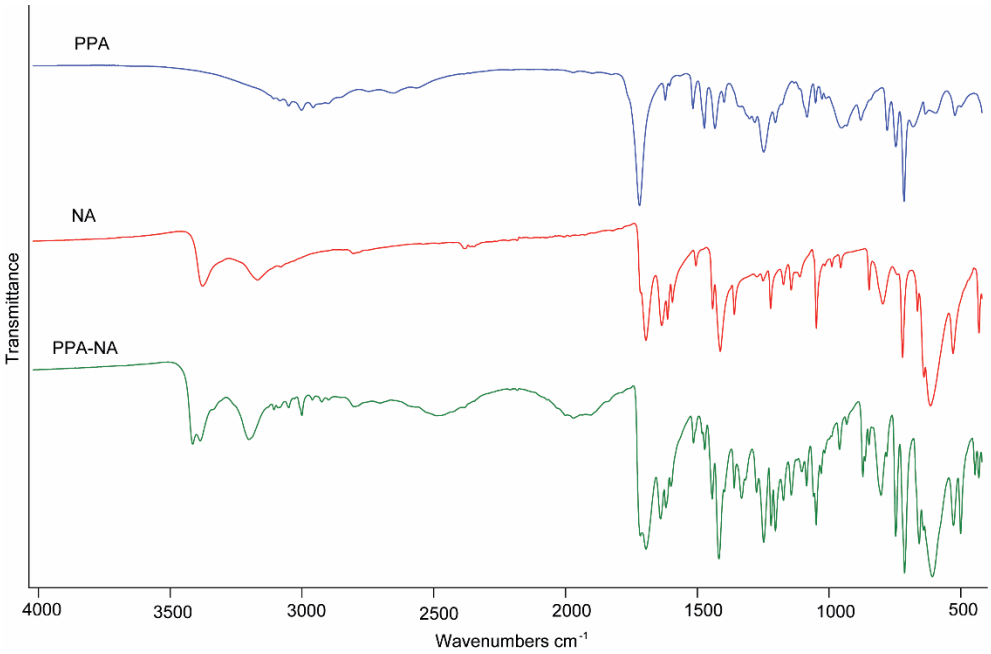


Figure S1. FTIR-ATR spectra of PPA-NA (green) compared with the pure starting materials PPA (blue) and NA (red).

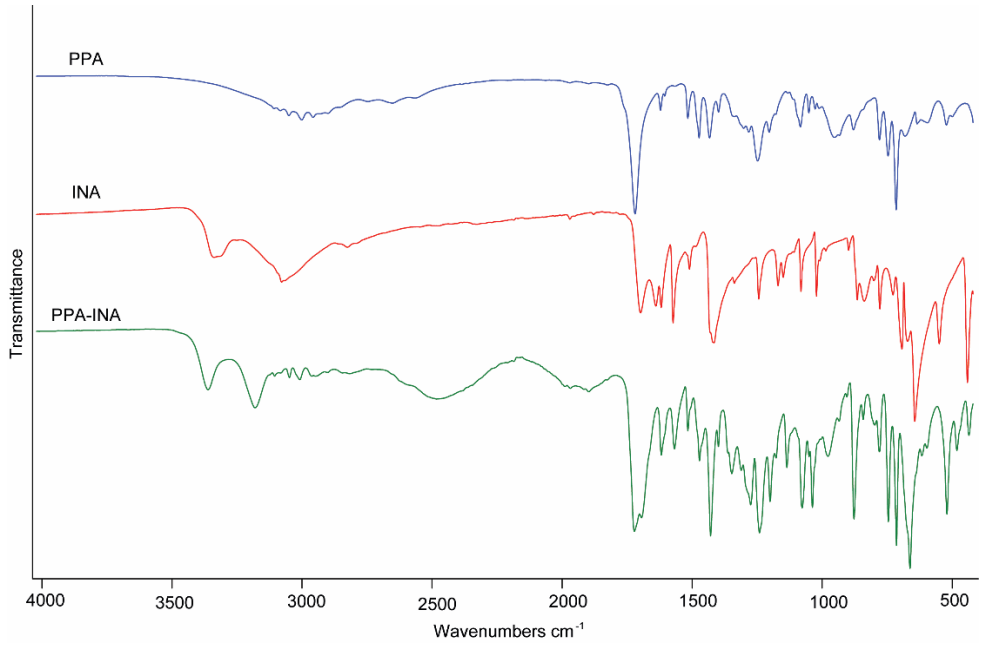


Figure S2. FTIR-ATR spectra of PPA-INA (green) compared with the pure starting materials PPA (blue) and INA (red).

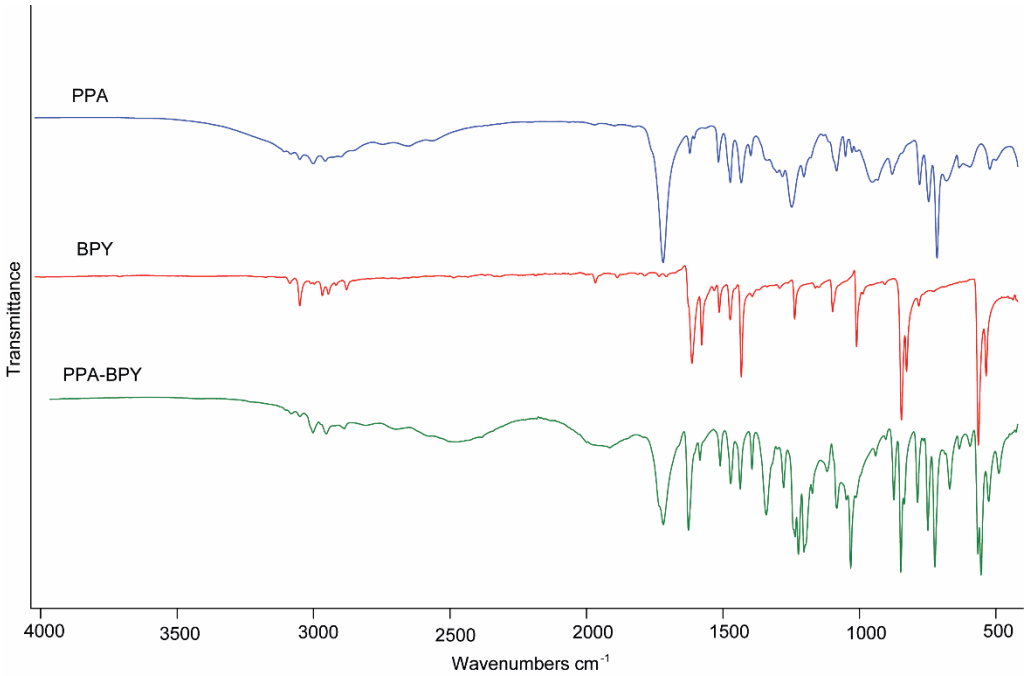


Figure S3. FTIR-ATR spectra of PPA-BPY (green) compared with the pure starting materials PPA (blue) and BPY (red).

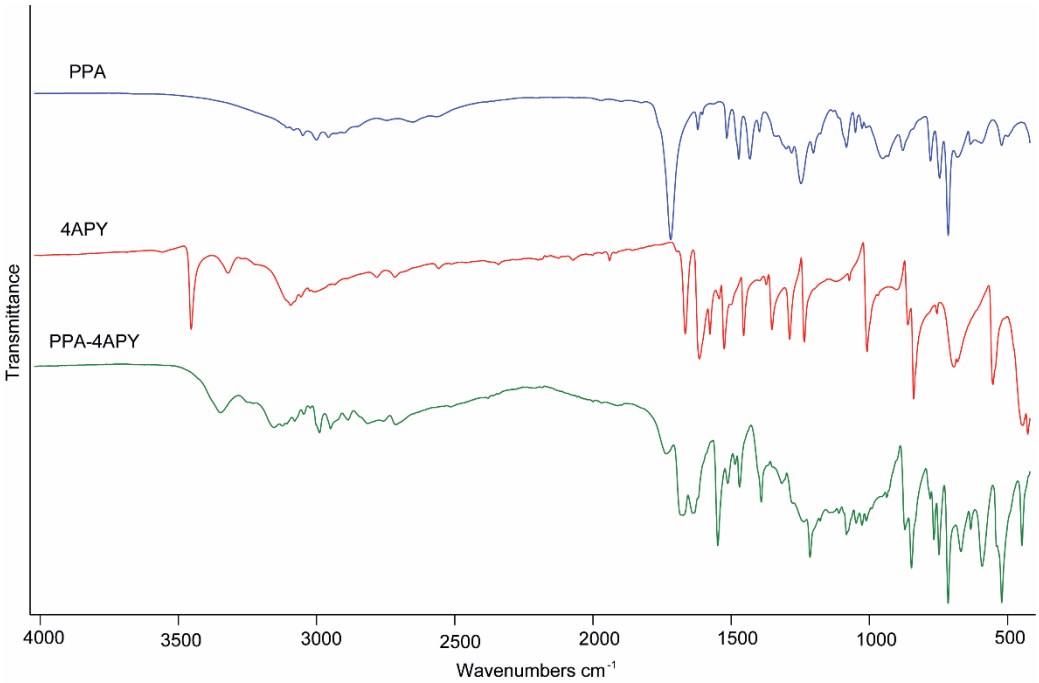


Figure S4. FTIR-ATR spectra of PPA-4APY (green) compared with the pure starting materials PPA (blue) and 4APY (red).

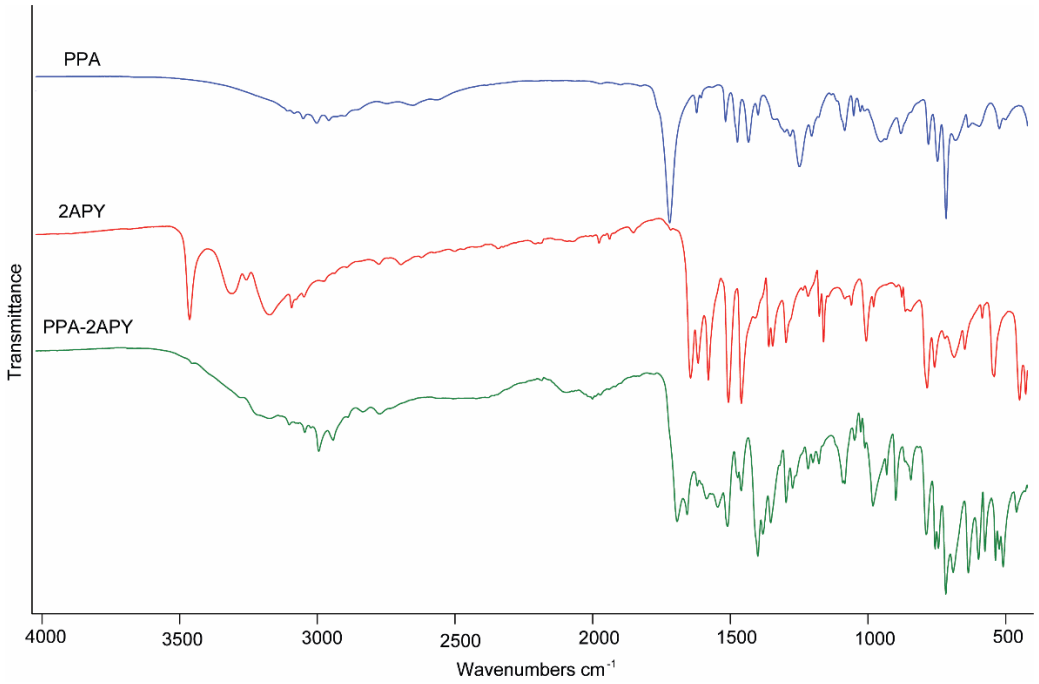


Figure S5. FTIR-ATR spectra of PPA-4APY (green) compared with the pure starting materials PPA (blue) and 4APY (red).

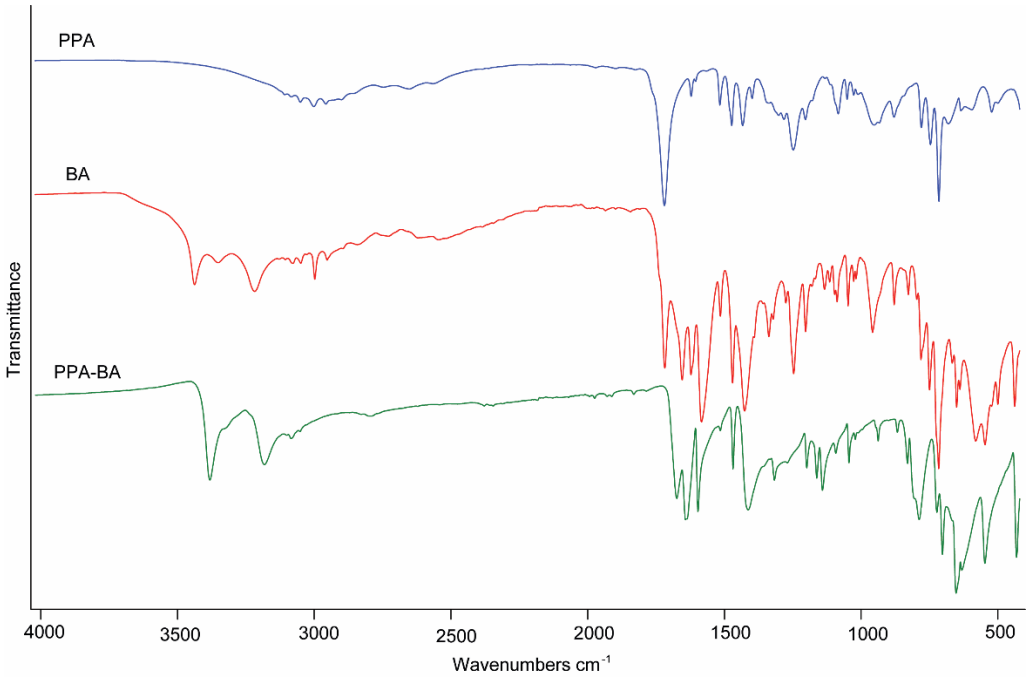


Figure S6. FTIR-ATR spectra of PPA-BA (green) compared with the pure starting materials PPA (blue) and BA (red).

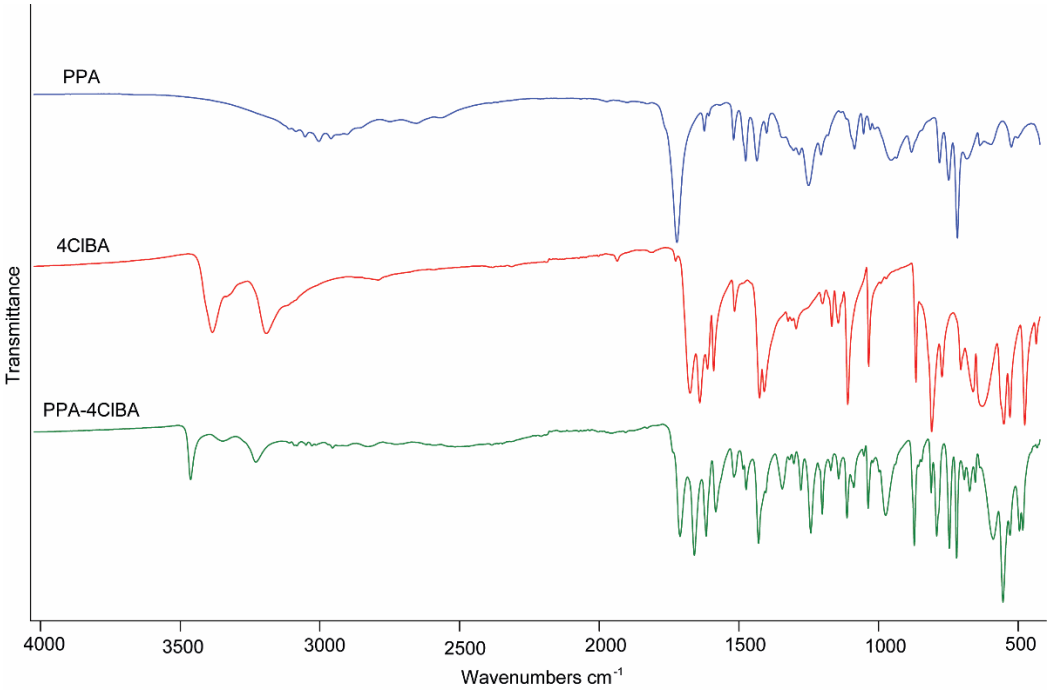


Figure S7. FTIR-ATR spectra of PPA-4ClBA (green) compared with the pure starting materials PPA (blue) and 4ClBA (red).

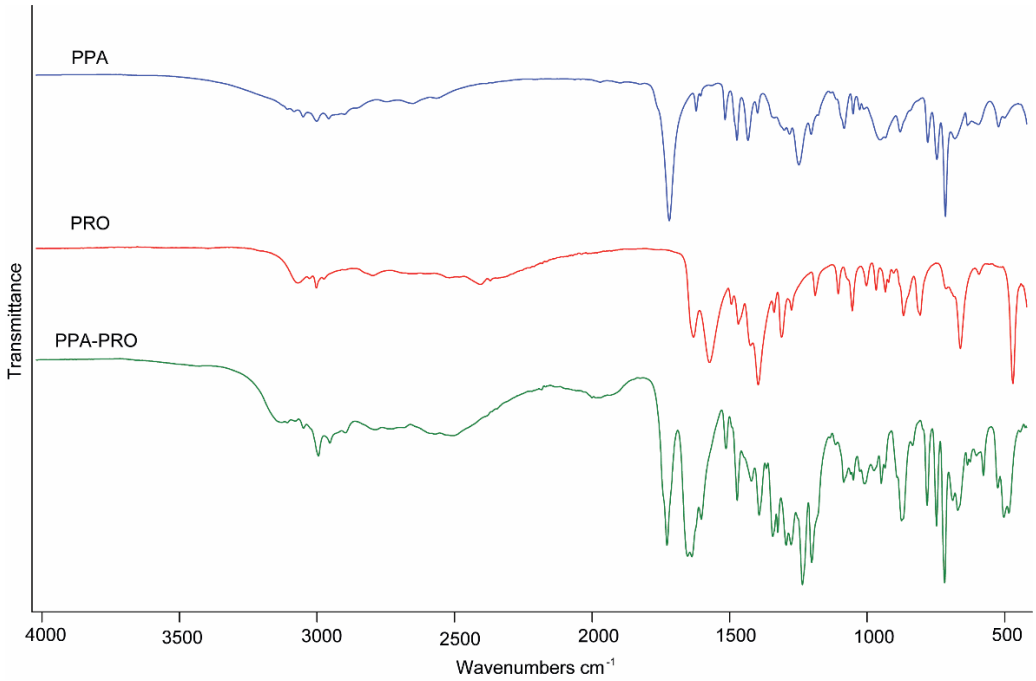


Figure S8. FTIR-ATR spectra of PPA-PRO (green) compared with the pure starting materials PPA (blue) and PRO (red).

PXRD

Powder X-ray diffraction patterns of new PPA adducts are reported in this section. For the samples whose crystal structure was obtained by SCXRD, a comparison of diffractograms from experimental powders with those calculated by Mercury from the relative structures are shown. This allowed to confirm that the single crystals selected for the analyses were representative of the bulk. For the other samples, the comparison was made with diffractograms of the starting materials. Since PPA is liquid at RT, its diffraction pattern was calculated with Mercury from the structure deposited on CSD with the name GOGPEY. In addition, for PPA-INA, the diffraction pattern was compared with that calculated from the cocrystal structure of the same system reported previously in the literature, with the RONDAA code in CSD. Since the two diffractograms are different, it is possible to conclude that a new PPA-INA polymorph was obtained; in fact, our new crystal form has a stoichiometric API:coformer ratio of 2:1, whereas the ratio of the cocrystal reported in the literature is 1:1.

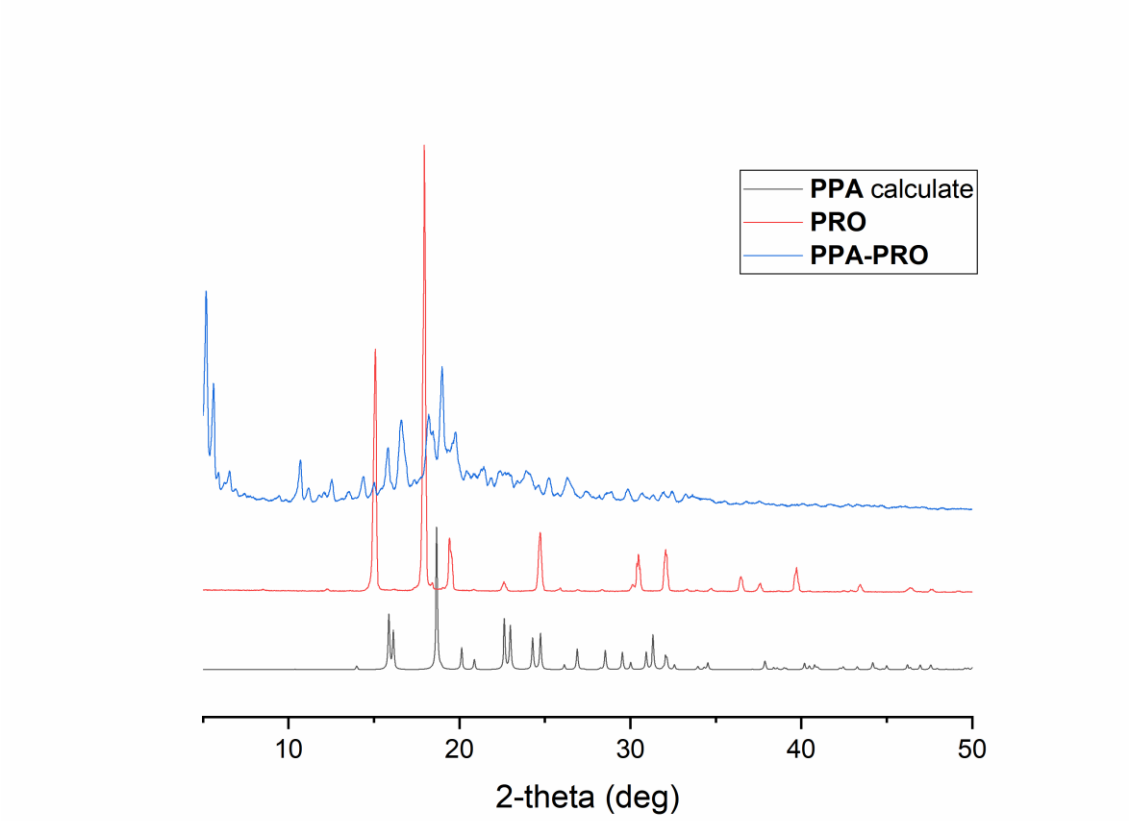


Figure S9. Powder X-ray diffraction patterns of PPA-PRO (blue) and PRO (red) and PPA (black).

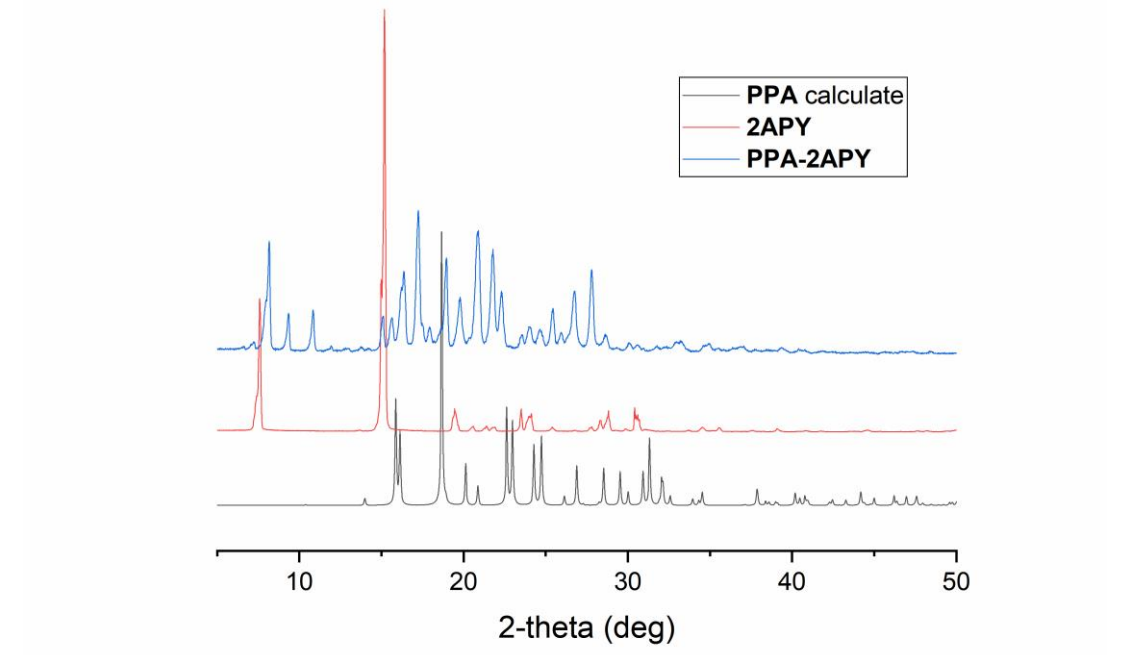


Figure S10. Powder X-ray diffraction patterns of PPA-2APY (blue) and 2APY (red) and PPA (black).

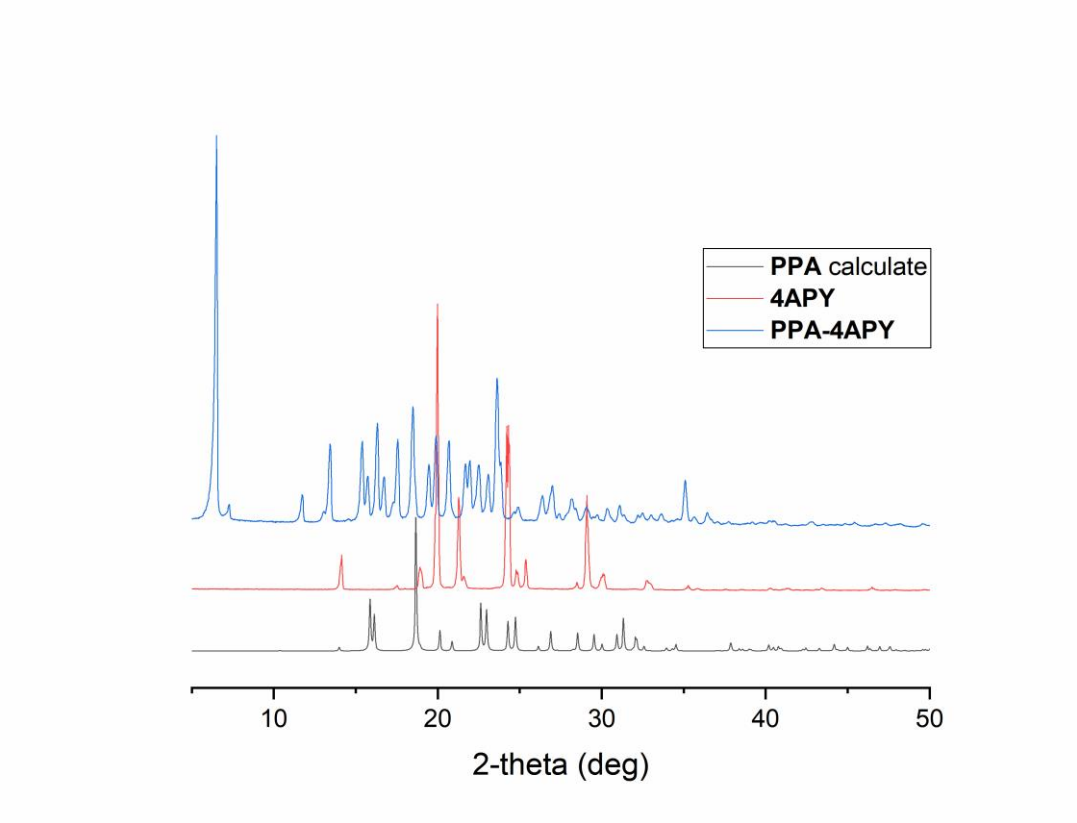


Figure S11. Powder X-ray diffraction patterns of PPA-4APY (blue) and 4APY (red) and PPA (black).

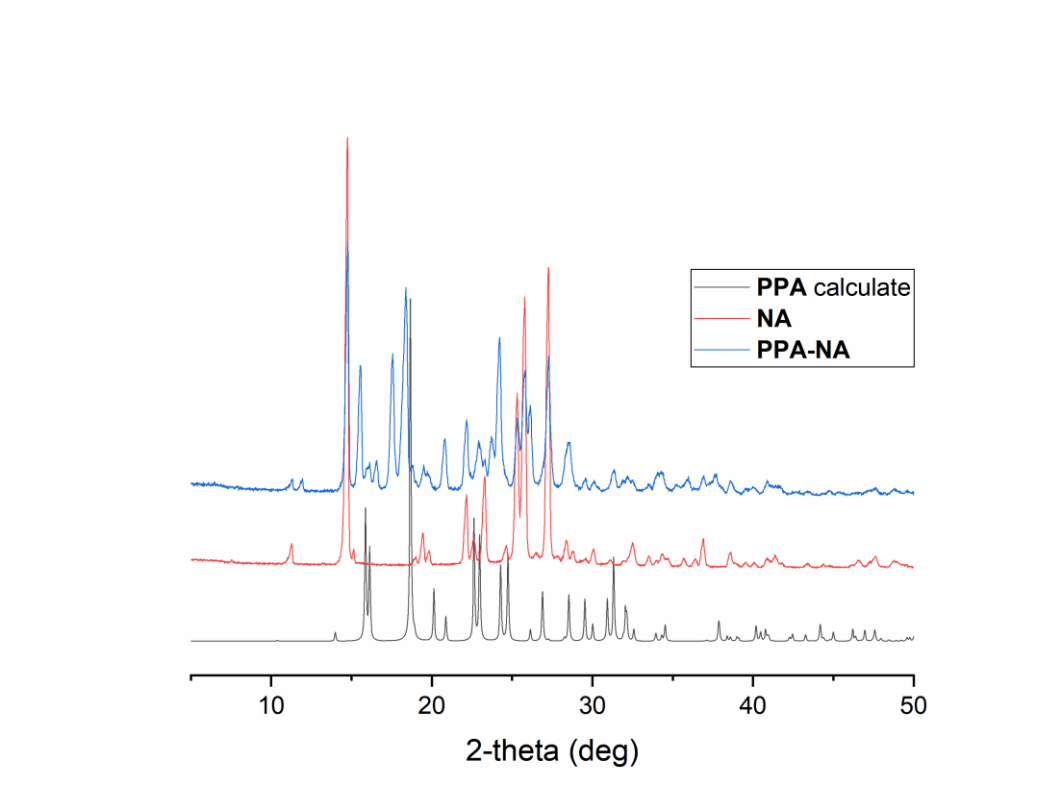


Figure S12. Powder X-ray diffraction patterns of PPA-NA (blue) and NA (red) and PPA (black).

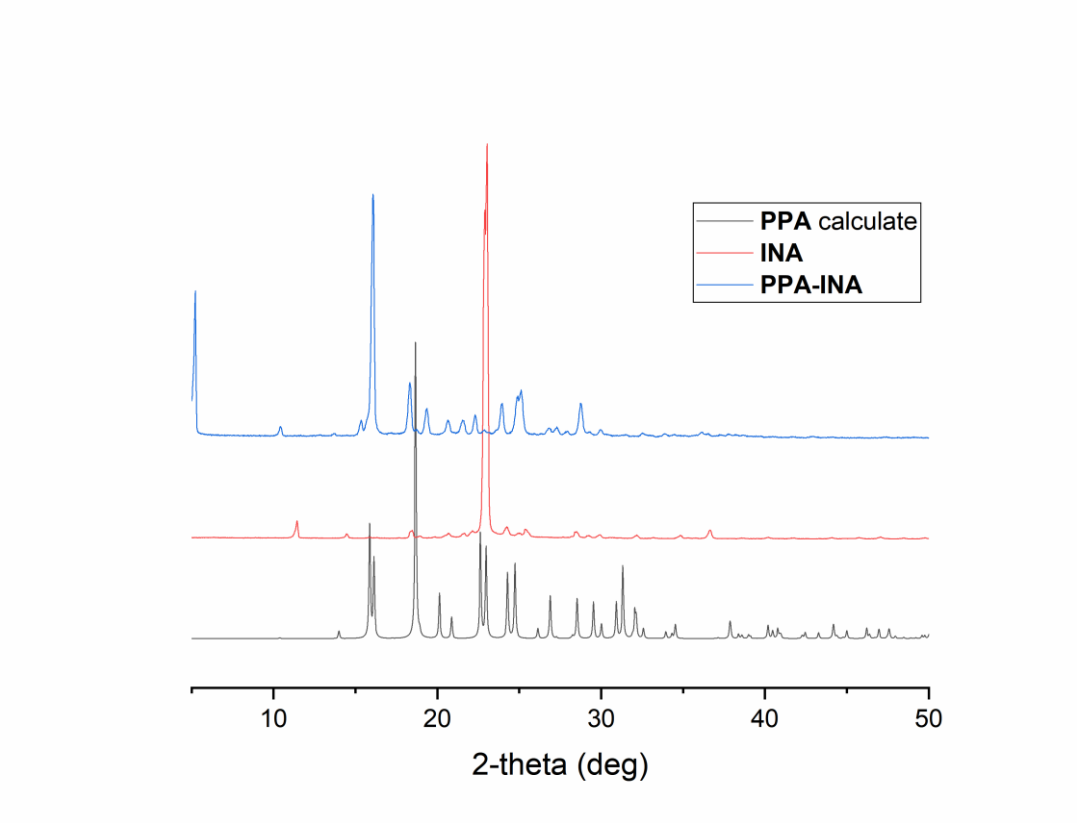


Figure S13. Powder X-ray diffraction patterns of PPA-INA (blue) and INA (red) and PPA (black).

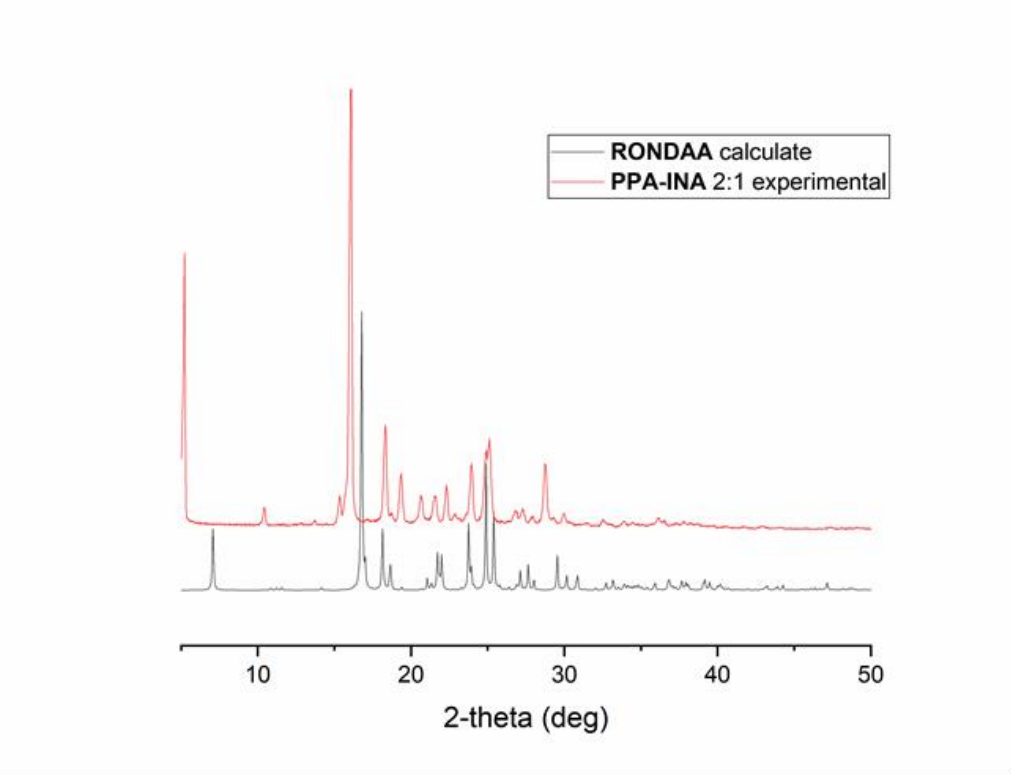


Figure S14. Superimposition of the experimental X-ray powder diffractogram collected on the bulk powder of PPA-INA and the simulated powder pattern calculated from the structure RONDAA on the CSD.

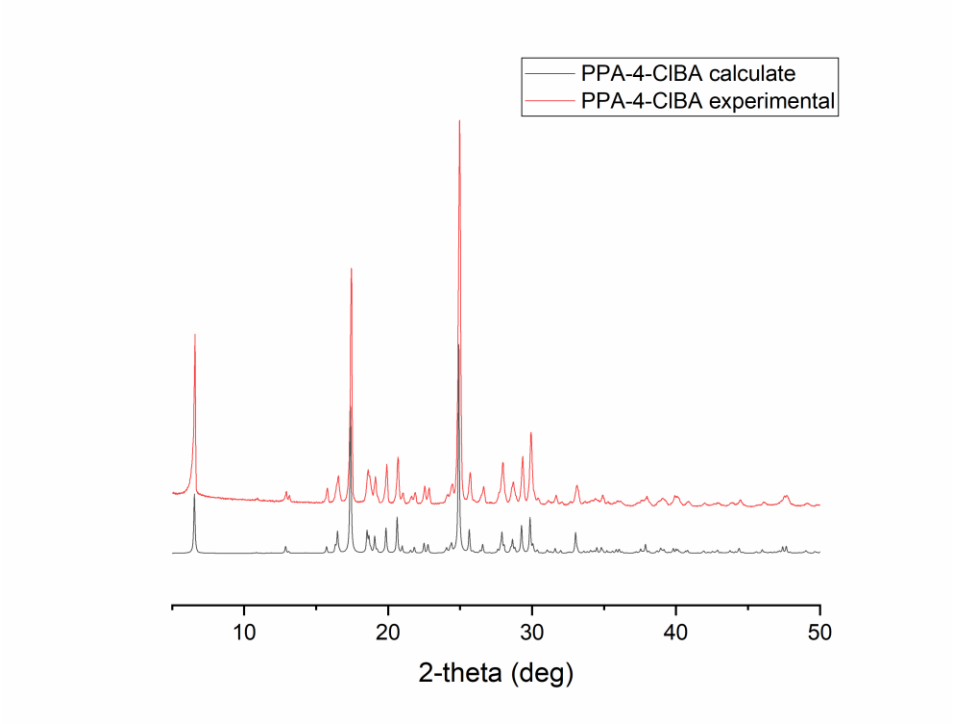


Figure S15. Superimposition of the experimental X-ray powder diffractogram collected on the bulk powder of PPA-4ClBA and the simulated powder pattern calculated from the structure solved via SCXRD.

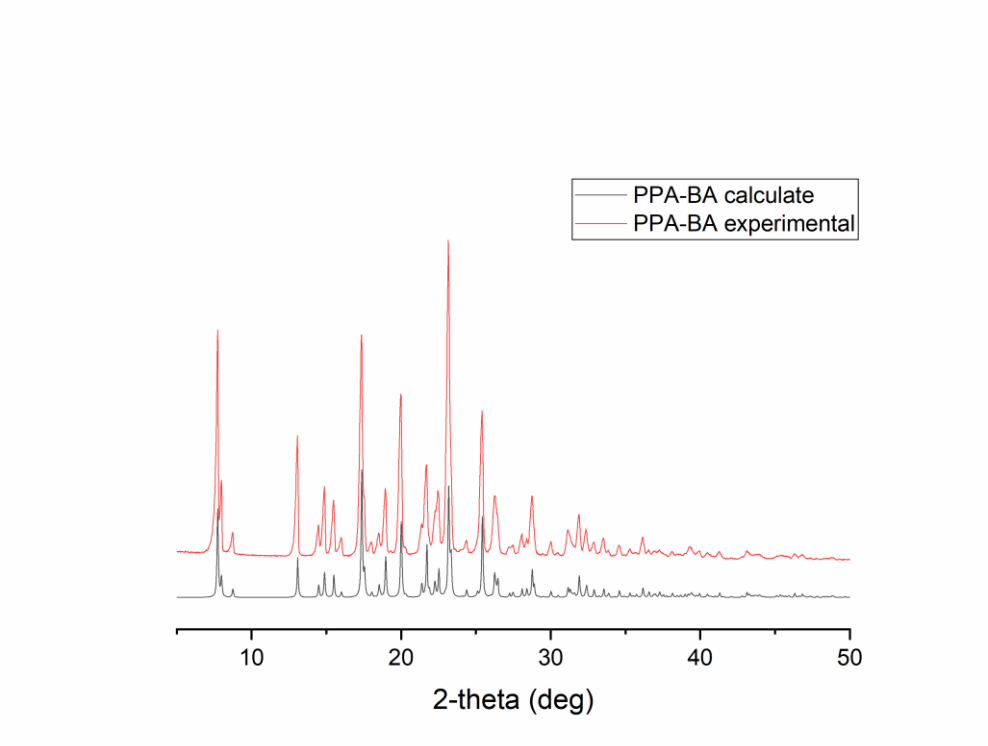


Figure S16. Superimposition of the experimental X-ray powder diffractogram collected on the bulk powder of PPA-BA and the simulated powder pattern calculated from the structure solved via SCXRD.

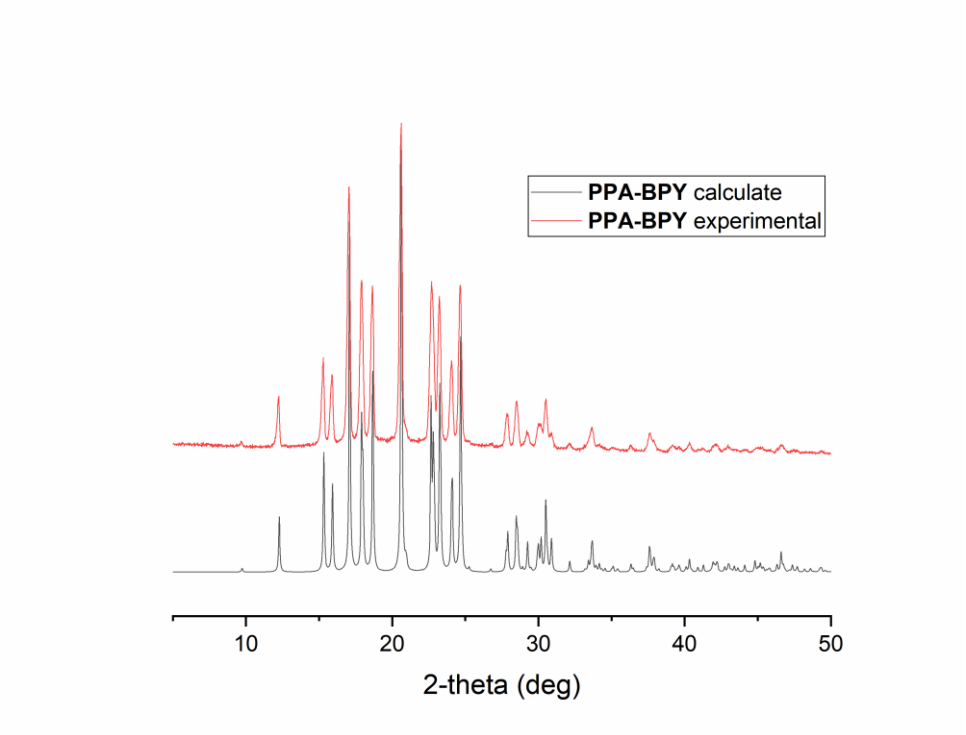
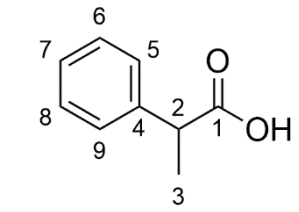


Figure S17. Superimposition of the experimental X-ray powder diffractogram collected on the bulk powder of PPA-BPY and the simulated powder pattern calculated from the structure solved via SCXRD.

SSNMR

Table S16. ^{13}C chemical shifts predicted by the software ChemDraw compared with experimental ones (in $CDCl_3$) reported in literature for PPA. Atom numbering refers to the Scheme.

	predicted	$CDCl_3$
Atom	**ppm**	
C3	13.4	18.07
C2	42.6	45.4
C7	127.5	127.3
C6	129.1	127.5
C8	129.1	127.5
C5	129.7	128.6
C9	129.7	128.6
C4	135.4	139.7
C1	181.2	180.9



2-phenylpropionic acid
(PPA)

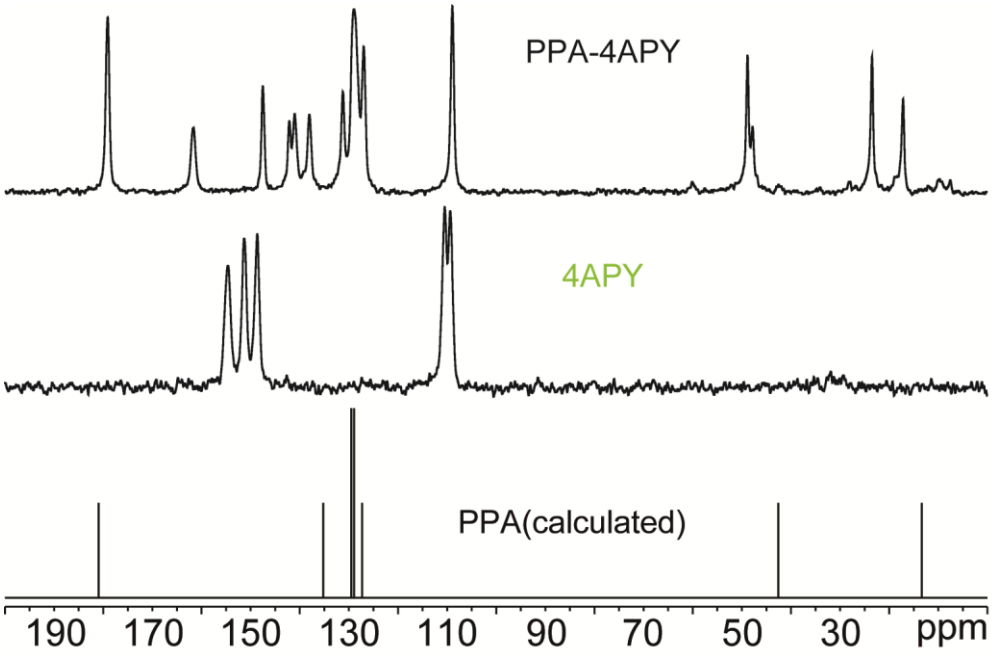


Figure S18. ^{13}C (100.63 MHz) CPMAS spectra of PPA-4APY and the pure 4APY, acquired with a spinning speed of 12 kHz at room temperature, compared with the ^{13}C NMR predicted chemical shift of PPA.

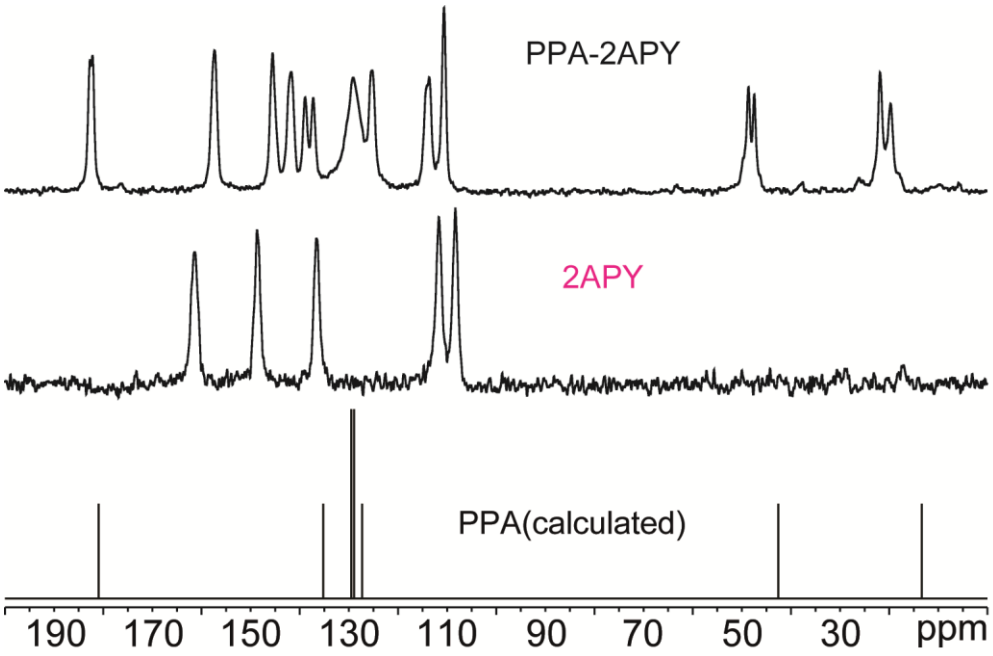


Figure S19. ^{13}C (100.63 MHz) CPMAS spectra of PPA-2APY and the pure 2APY, acquired with a spinning speed of 12 kHz at room temperature, compared with the ^{13}C NMR predicted chemical shift of PPA.

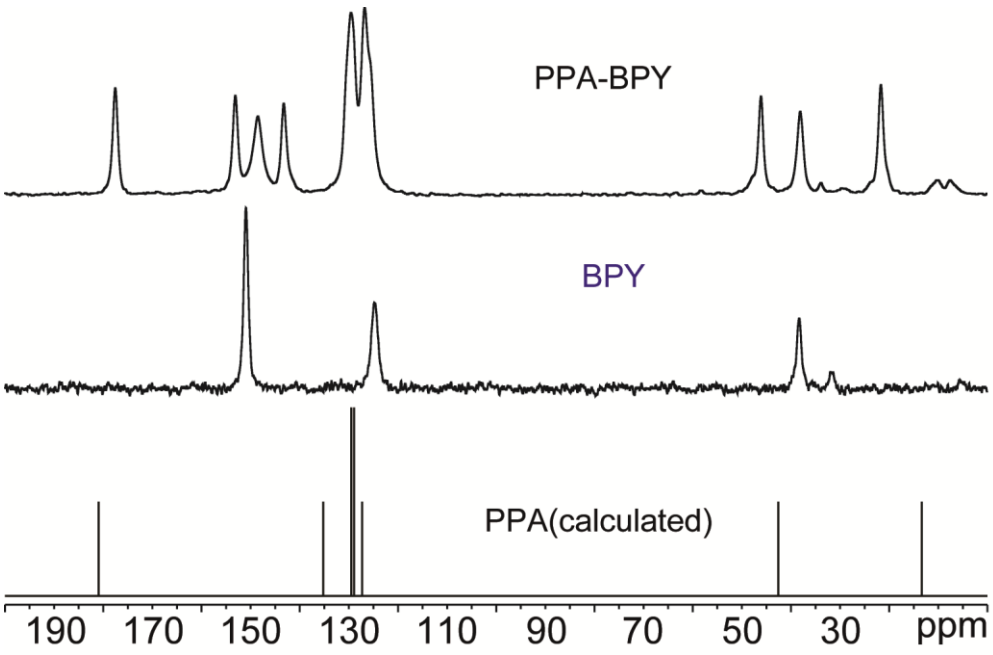


Figure S20. ^{13}C (100.63 MHz) CPMAS spectra of PPA-BPY and the pure BPY, acquired with a spinning speed of 12 kHz at room temperature, compared with the ^{13}C NMR predicted chemical shift of PPA.

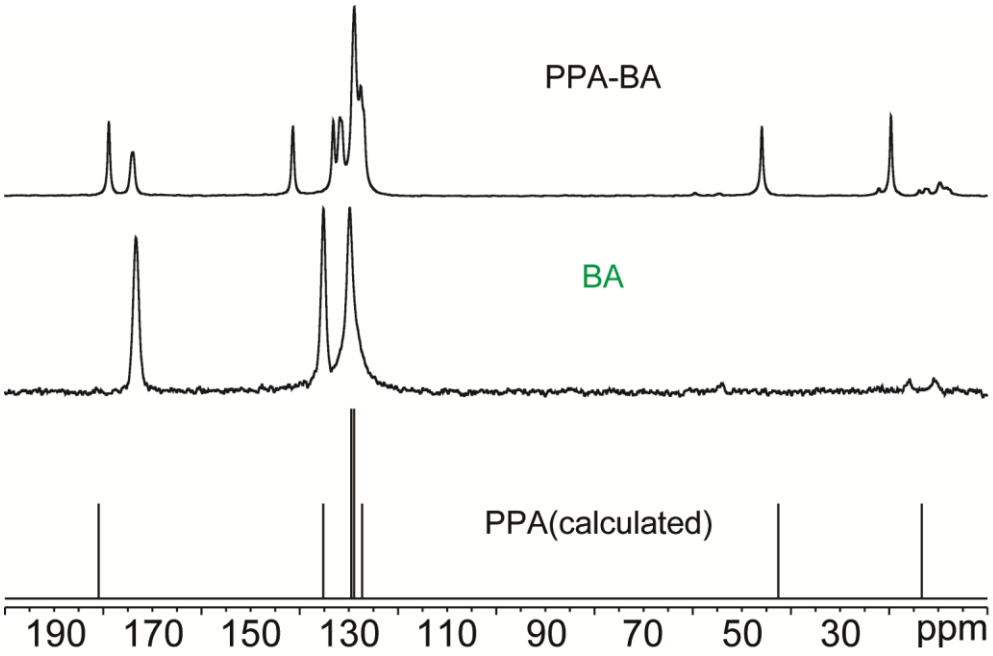


Figure S21. ^{13}C (100.63 MHz) CPMAS spectra of PPA-BA and the pure BA, acquired with a spinning speed of 12 kHz at room temperature, compared with the ^{13}C NMR predicted chemical shift of PPA.

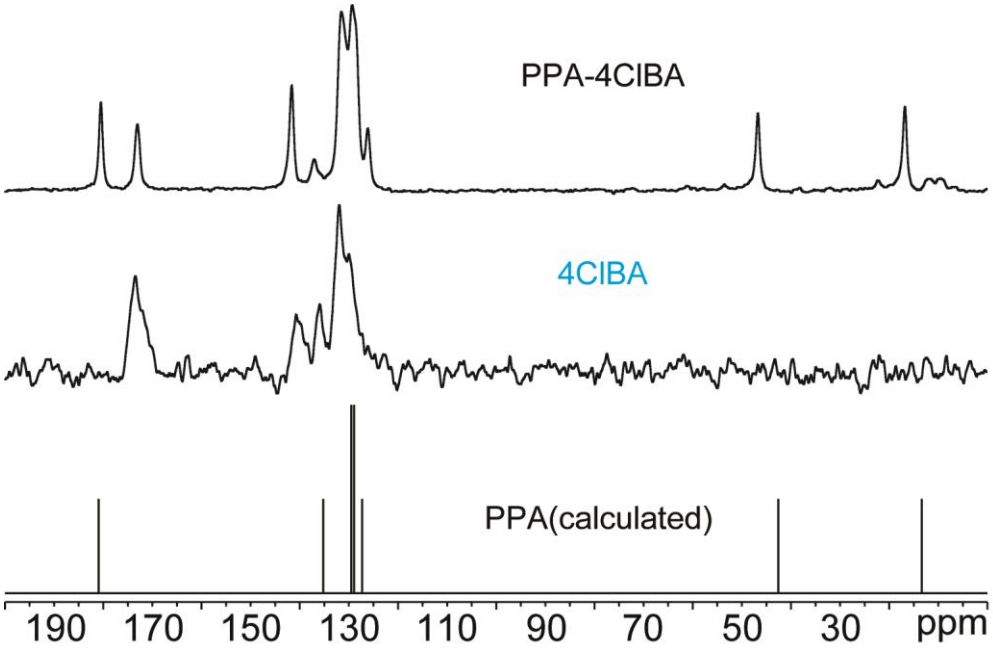


Figure S22. ^{13}C (100.63 MHz) CPMAS spectra of PPA-4ClBA and the pure 4ClBA, acquired with a spinning speed of 12 kHz at room temperature, compared with the ^{13}C NMR predicted chemical shift of PPA.

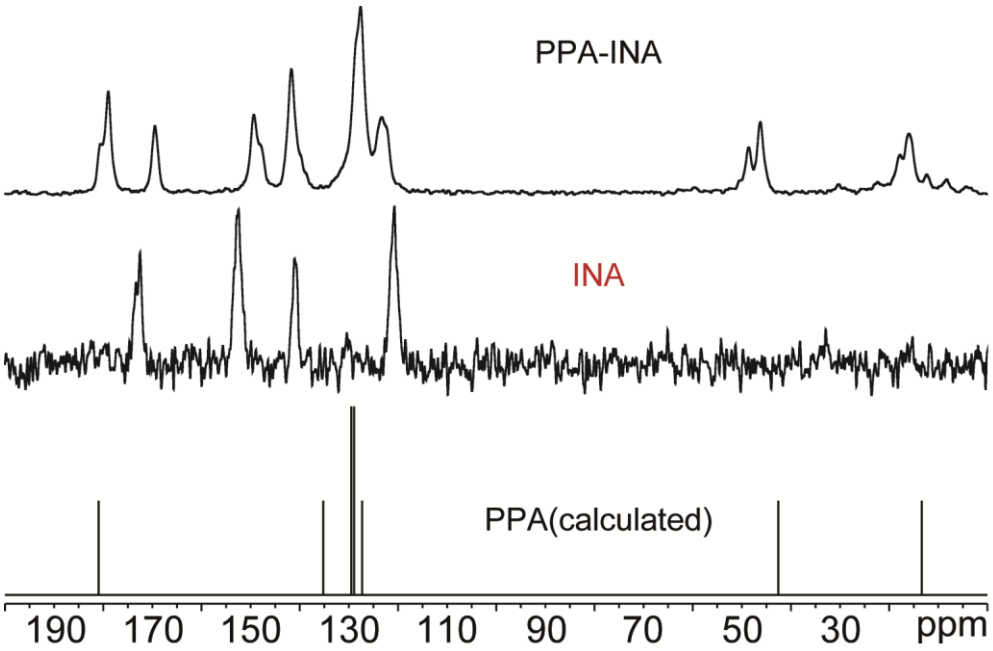


Figure S23. ^{13}C (100.63 MHz) CPMAS spectra of PPA-INA and the pure INA, acquired with a spinning speed of 12 kHz at room temperature, compared with the ^{13}C NMR predicted chemical shift of PPA.

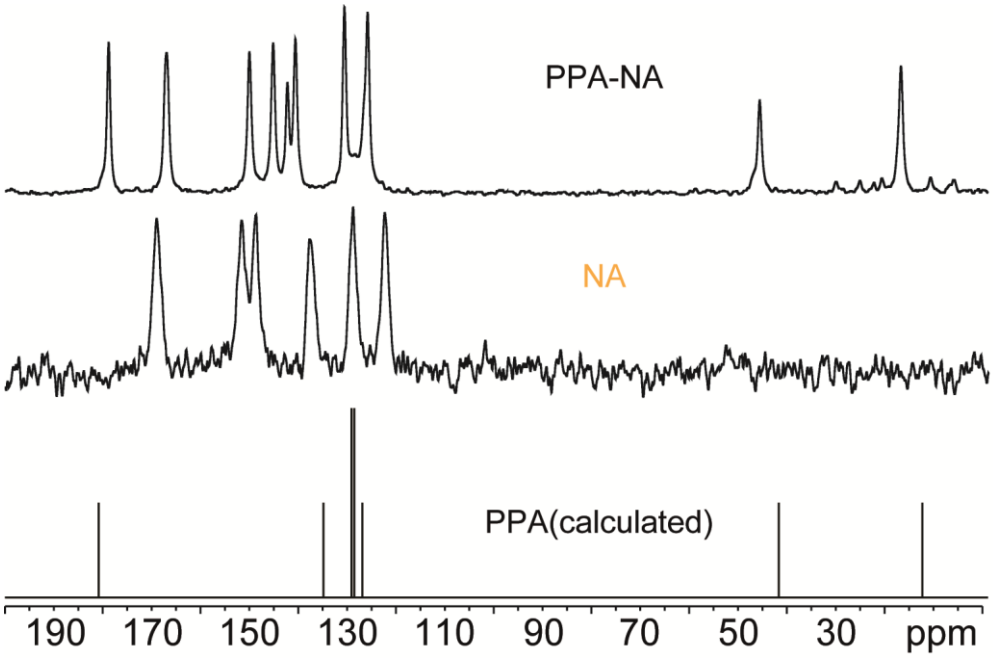


Figure S24. ^{13}C (100.63 MHz) CPMAS spectra of PPA-NA and the pure NA, acquired with a spinning speed of 12 kHz at room temperature, compared with the ^{13}C NMR predicted chemical shift of PPA.

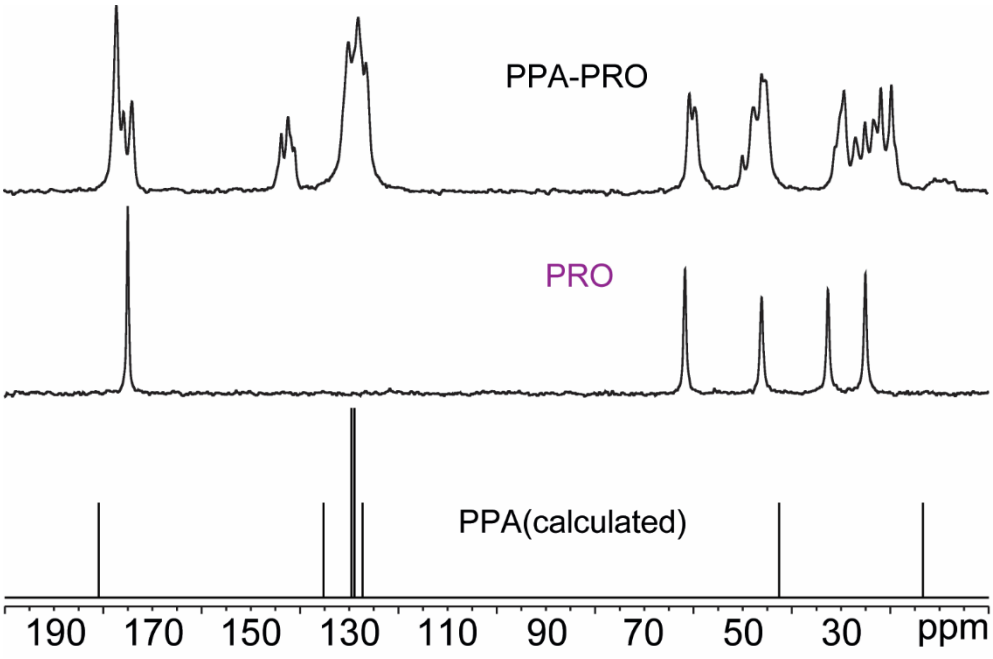


Figure S25. ^{13}C (100.63 MHz) CPMAS spectra of PPA-PRO and the pure PRO, acquired with a spinning speed of 12 kHz at room temperature, compared with the ^{13}C NMR predicted chemical shift of PPA.

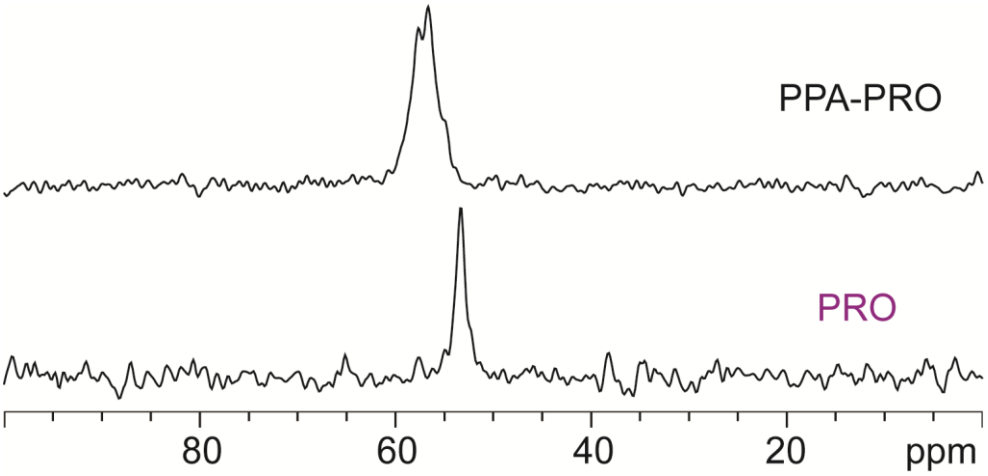


Figure S26. ^{15}N (40.56 MHz) CPMAS spectra of PPA-PRO and the pure PRO, acquired with a spinning speed of 9 kHz at room temperature.

Table S17. ^{15}N chemical shifts of free and protonated 4APY (from literature) compared with the chemical shift values of PPA-4APY registered by CPMAS SSNMR.

	ppm	
	N_{ar}	**NH_2**
4APY (free)[a]	265.0	60.0
4APY (protonated)[a]	159.3	90.9
PPA-4APY	164.5	98.3

a. ^{15}N chemical shifts (ppm) of free and protonated 4APY, in CDCl3/*d*6-DMSO (70:30 v/v) (P. Beltrame *et al*., *Spectrochimica Acta Part A* 58 **2002**, 2693–2697).

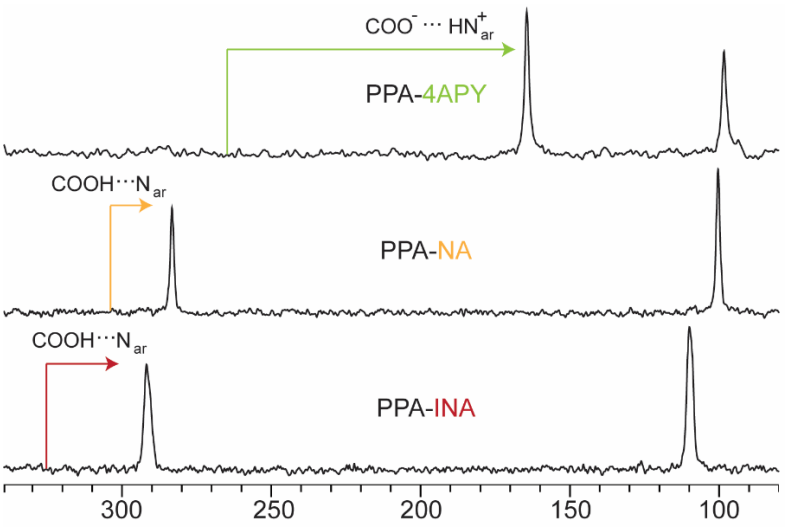


Figure S27. ^{15}N (40.56 MHz) CPMAS spectra of PPA-4PY, PPA-NA and PPA-INA, acquired with a spinning speed of 9 kHz at room temperature. The colored lines highlight the position of the N_{ar} signal in pure coformers, *i.e,* 265.0 ppm for 4APY (P. Beltrame *et al*., *Spectrochimica Acta Part A* **2002**, 58, 2693–2697), 302.6 ppm for NA (A.S. Tatton *et al.*, *Mol. Pharmaceutics* **2013**, 10, 999−1007), and 325.1 for INA (J. Li *et al.*, *Eur J Pharm Sci*, **2016**, 85, 47-52).

SCXRD

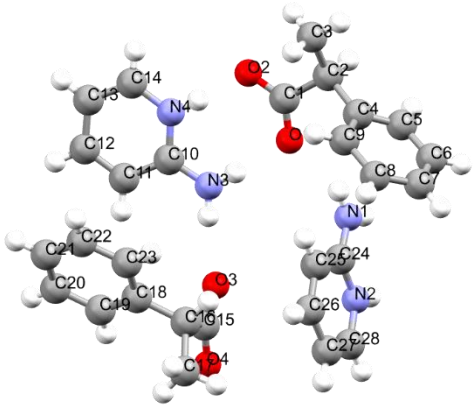


Figure S28. Asymmetric unit of PPA-2APY.

Table S18. Crystal data and structure refinement for PPA-2APY.

PPA-2APY	
Empirical formula	$C_{28}H_{32}N_4O_4$
Formula weight	488.37
Temperature/K	298.00
Crystal system	triclinic
Space group	P-1
a/Å	11.0851(6)
b/Å	11.2420(5)
c/Å	12.0891(4)
α/°	67.358(4)
β/°	79.991(4)
γ/°	77.384(4)
Volume/Å^3	1350.11(12)
Z	2
ρ_{calc}g/cm^3	1.201
μ/mm^{-1}	0.659
F(000)	520.0
Crystal size/mm^3	0.1 × 0.08 × 0.07
Radiation	Cu Kα (λ = 1.54184)
2Θ range for data collection/°	7.964 to 133.91
Index ranges	-13 ≤ h ≤ 13, -13 ≤ k ≤ 13, -11 ≤ l ≤ 14
Reflections collected	12624
Independent reflections	4695 [R_{int} = 0.0239, R_{sigma} = 0.0181]
Data/restraints/parameters	4695/456/346
Goodness-of-fit on F^2	1.032
Final R indexes [I>=2σ (I)]	R_1 = 0.0491, wR_2 = 0.1366
Final R indexes [all data]	R_1 = 0.0652, wR_2 = 0.1504
Largest diff. peak/hole / e Å^{-3}	0.23/-0.19

Table S19. Bond Lengths for PPA-2APY.

PPA-2APY						
Atom	**Atom**	**Length/Å**		**Atom**	**Atom**	**Length/Å**
O3	C15	1.266(2)		C18	C19	1.379(3)
N4	C10	1.343(2)		C18	C23	1.374(3)
N4	C14	1.347(2)		C12	C13	1.394(3)
O1	C1	1.231(2)		C4	C2	1.529(3)
O2	C1	1.243(2)		C4	C5	1.342(3)
N2	C24	1.343(2)		C4	C9	1.373(4)
N2	C28	1.349(3)		C4	C2A	1.584(10)
N3	C10	1.322(2)		C2	C3	1.529(4)
C10	C11	1.404(3)		C26	C25	1.351(3)
C24	N1	1.318(3)		C26	C27	1.389(4)
C24	C25	1.405(3)		C19	C20	1.379(3)
O4	C15	1.232(2)		C23	C22	1.377(3)
C1	C2	1.536(4)		C21	C22	1.367(4)
C1	C2A	1.657(11)		C21	C20	1.354(4)
C15	C16	1.520(3)		C5	C6	1.336(4)
C28	C27	1.338(3)		C7	C6	1.317(5)
C14	C13	1.343(3)		C7	C8	1.385(5)
C11	C12	1.355(3)		C9	C8	1.425(4)
C16	C18	1.512(3)		C2A	C3A	1.523(12)
C16	C17	1.521(3)				

Table S20. Bond Angles for PPA-2APY.

PPA-2APY								
Atom	**Atom**	**Atom**	**Angle/˚**		**Atom**	**Atom**	**Atom**	**Angle/˚**
C10	N4	C14	122.46(18)		C5	C4	C2	117.5(3)
C24	N2	C28	122.74(19)		C5	C4	C9	116.8(2)
N4	C10	C11	117.23(17)		C5	C4	C2A	151.6(5)
N3	C10	N4	118.72(18)		C9	C4	C2	125.6(3)
N3	C10	C11	124.04(17)		C9	C4	C2A	91.5(5)
N2	C24	C25	117.4(2)		C14	C13	C12	118.1(2)
N1	C24	N2	118.37(19)		C4	C2	C1	110.7(2)
N1	C24	C25	124.20(19)		C3	C2	C1	111.1(2)
O1	C1	O2	124.4(2)		C3	C2	C4	110.1(3)
O1	C1	C2	117.20(18)		C25	C26	C27	121.1(2)
O1	C1	C2A	119.3(3)		C20	C19	C18	121.2(2)
O2	C1	C2	118.16(18)		C26	C25	C24	119.5(2)
O2	C1	C2A	109.8(3)		C18	C23	C22	121.2(2)
O3	C15	C16	115.86(17)		C20	C21	C22	119.3(2)
O4	C15	O3	123.4(2)		C28	C27	C26	118.4(2)
O4	C15	C16	120.72(19)		C6	C5	C4	123.6(3)
C27	C28	N2	120.8(2)		C21	C22	C23	120.4(2)
C13	C14	N4	121.53(19)		C21	C20	C19	120.5(2)
C12	C11	C10	120.36(19)		C6	C7	C8	119.9(3)
C15	C16	C17	113.1(2)		C4	C9	C8	120.6(3)
C18	C16	C15	110.74(17)		C7	C6	C5	121.4(3)
C18	C16	C17	111.03(19)		C7	C8	C9	117.6(3)
C19	C18	C16	121.91(19)		C4	C2A	C1	102.1(6)
C23	C18	C16	120.7(2)		C3A	C2A	C1	100.8(8)
C23	C18	C19	117.4(2)		C3A	C2A	C4	103.1(8)
C11	C12	C13	120.4(2)					

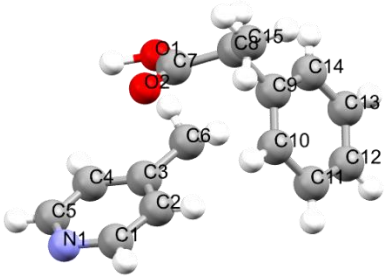


Figure S29. Asymmetric unit of PPA-BPY.

Table S21. Crystal data and structure refinement for PPA-BPY.

PPA-BPY	
Identification code	2F-1,2(4-pirazina)
Empirical formula	$C_{15}H_{16}NO_2$
Formula weight	242.29
Temperature/K	298.00
Crystal system	monoclinic
Space group	$P2_1/n$
a/Å	14.8175(12)
b/Å	6.3163(5)
c/Å	14.9017(14)
α/°	90
β/°	104.653(9)
γ/°	90
Volume/Å^3	1349.3(2)
Z	4
ρ_{calc}g/cm^3	1.193
μ/mm^{-1}	0.079
F(000)	516.0
Crystal size/mm^3	0.12 × 0.1 × 0.08
Radiation	Mo Kα (λ = 0.71073)
2Θ range for data collection/°	6.928 to 57.848
Index ranges	-19 ≤ h ≤ 18, -7 ≤ k ≤ 8, -20 ≤ l ≤ 17
Reflections collected	11752
Independent reflections	3249 [R_{int} = 0.0424, R_{sigma} = 0.0488]
Data/restraints/parameters	3249/0/168
Goodness-of-fit on F^2	1.031
Final R indexes [I>=2σ (I)]	R_1 = 0.0683, wR_2 = 0.1576
Final R indexes [all data]	R_1 = 0.1472, wR_2 = 0.2005
Largest diff. peak/hole / e Å^{-3}	0.32/-0.17

Table S22. Bond Lengths for PPA-BPY.

PPA-BPY						
Atom	**Atom**	**Length/Å**		**Atom**	**Atom**	**Length/Å**
N1	C5	1.305(4)		C9	C10	1.370(3)
N1	C1	1.312(4)		O1	C7	1.288(3)
C3	C4	1.377(4)		O2	C7	1.195(3)
C3	C2	1.376(4)		C8	C7	1.527(4)
C3	C6	1.507(4)		C8	C15	1.508(4)
C4	C5	1.370(4)		C14	C13	1.356(4)
C2	C1	1.385(4)		C10	C11	1.387(4)
C6	C6[1]	1.498(5)		C13	C12	1.342(5)
C9	C8	1.520(3)		C12	C11	1.384(5)
C9	C14	1.379(3)				

Table S23. Bond Angles for PPA-BPY.

PPA-BPY							
Atom	**Atom**	**Atom**	**Angle/˚**		**Atom**	**Atom**	**Angle/˚**
C5	N1	C1	117.2(2)		C9	C8	C7	109.2(2)
C4	C3	C6	121.2(3)		C15	C8	C9	112.3(2)
C2	C3	C4	116.3(2)		C15	C8	C7	111.8(2)
C2	C3	C6	122.6(3)		O1	C7	C8	112.4(2)
C5	C4	C3	120.0(3)		O2	C7	O1	123.2(3)
N1	C5	C4	123.7(3)		O2	C7	C8	124.3(3)
C3	C2	C1	119.4(3)		C13	C14	C9	121.1(3)
N1	C1	C2	123.5(3)		C9	C10	C11	120.4(3)
C6[1]	C6	C3	112.8(3)		C12	C13	C14	120.9(3)
C14	C9	C8	119.6(2)		C13	C12	C11	119.8(3)
C10	C9	C8	122.0(2)		C12	C11	C10	119.3(3)
C10	C9	C14	118.4(2)					

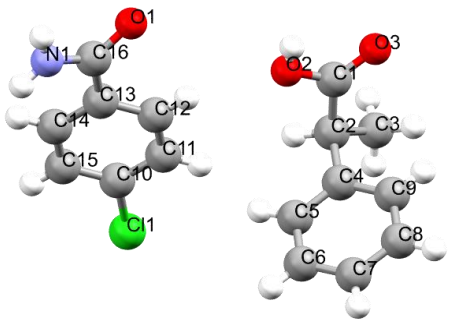


Figure S30. Asymmetric unit of PPA-4ClBA.

Table S24. Crystal data and structure refinement for PPA-4ClBA.

PPA-4ClBA	
Empirical formula	$C_{16}H_{16}ClNO_3$
Formula weight	305.75
Temperature/K	298.00
Crystal system	monoclinic
Space group	$P2_1/c$
a/Å	13.6917(12)
b/Å	10.1487(10)
c/Å	10.9882(9)
α/°	90
β/°	100.310(9)
γ/°	90
Volume/Å^3	1502.2(2)
Z	4
ρ_{calc}g/cm^3	1.352
μ/mm^{-1}	0.263
F(000)	640.0
Crystal size/mm^3	0.12 × 0.1 × 0.09
Radiation	Mo Kα (λ = 0.71073)
2Θ range for data collection/°	6.6 to 59.46
Index ranges	-18 ≤ h ≤ 18, -13 ≤ k ≤ 13, -15 ≤ l ≤ 13
Reflections collected	11768
Independent reflections	3659 [R_{int} = 0.0660, R_{sigma} = 0.0646]
Data/restraints/parameters	3659/0/192
Goodness-of-fit on F^2	1.028
Final R indexes [I>=2σ (I)]	R_1 = 0.0709, wR_2 = 0.1984
Final R indexes [all data]	R_1 = 0.1263, wR_2 = 0.2864
Largest diff. peak/hole / e Å^{-3}	0.49/-0.46

Table S25. Bond Lengths for PPA-4ClBA.

PPA-4ClBA						
Atom	**Atom**	**Length/Å**		**Atom**	**Atom**	**Length/Å**
Cl1	C10	1.745(3)		C1	C2	1.514(4)
O1	C16	1.238(4)		C4	C2	1.520(4)
C13	C16	1.494(4)		C4	C9	1.384(4)
C13	C14	1.387(4)		C4	C5	1.370(4)
C13	C12	1.390(4)		C2	C3	1.531(5)
C10	C11	1.365(4)		C11	C12	1.375(4)
C10	C15	1.370(4)		C8	C9	1.389(5)
O2	C1	1.306(4)		C8	C7	1.366(6)
O3	C1	1.212(4)		C5	C6	1.379(5)
C16	N1	1.304(4)		C6	C7	1.360(6)
C14	C15	1.379(4)				

Table S26. Bond Angles for PPA-BPY.

PPA-4ClBA								
Atom	**Atom**	**Atom**	**Angle/˚**		**Atom**	**Atom**	**Atom**	**Angle/˚**
C14	C13	C16	122.2(3)		C5	C4	C2	120.7(3)
C14	C13	C12	119.4(3)		C5	C4	C9	118.1(3)
C12	C13	C16	118.4(3)		C1	C2	C4	108.0(2)
C11	C10	Cl1	119.9(2)		C1	C2	C3	111.6(3)
C11	C10	C15	121.6(3)		C4	C2	C3	112.9(2)
C15	C10	Cl1	118.5(2)		C10	C11	C12	120.0(3)
O1	C16	C13	120.0(3)		C11	C12	C13	119.5(3)
O1	C16	N1	122.1(3)		C10	C15	C14	118.7(3)
N1	C16	C13	118.0(3)		C7	C8	C9	119.5(3)
C15	C14	C13	120.6(3)		C4	C9	C8	120.7(3)
O2	C1	C2	112.7(3)		C4	C5	C6	121.5(3)
O3	C1	O2	123.3(3)		C7	C6	C5	119.6(4)
O3	C1	C2	124.0(3)		C6	C7	C8	120.6(3)
C9	C4	C2	121.2(3)					

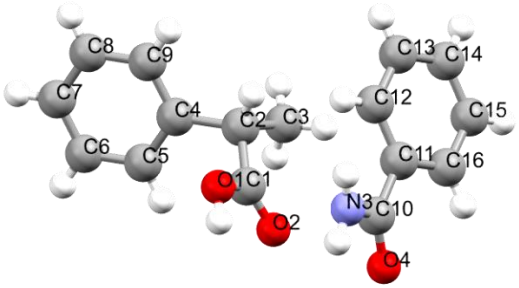


Figure S31. Asymmetric unit of PPA-BA.

Table S27. Crystal data and structure refinement for PPA-BA.

PPA-BA	
Empirical formula	$C_{16}H_{17}NO_3$
Formula weight	271.30
Temperature/K	298.00
Crystal system	monoclinic
Space group	$P2_1/n$
a/Å	13.5322(13)
b/Å	5.2730(5)
c/Å	20.2125(17)
α/°	90
β/°	91.939(7)
γ/°	90
Volume/Å^3	1441.4(2)
Z	4
ρ_{calc}g/cm^3	1.250
μ/mm^{-1}	0.086
F(000)	576.0
Crystal size/mm^3	0.12 × 0.1 × 0.08
Radiation	Mo Kα (λ = 0.71073)
2Θ range for data collection/°	6.85 to 58.79
Index ranges	-16 ≤ h ≤ 16, -7 ≤ k ≤ 6, -25 ≤ l ≤ 20
Reflections collected	10129
Independent reflections	3458 [R_{int} = 0.0446, R_{sigma} = 0.0504]
Data/restraints/parameters	3458/0/189
Goodness-of-fit on F^2	1.080
Final R indexes [I>=2σ (I)]	R_1 = 0.0539, wR_2 = 0.1300
Final R indexes [all data]	R_1 = 0.1126, wR_2 = 0.1787
Largest diff. peak/hole / e Å^{-3}	0.17/-0.18

Table S28. Bond Lengths for PPA-BA.

PPA-BA						
Atom	**Atom**	**Length/Å**		**Atom**	**Atom**	**Length/Å**
O4	C10	1.246(2)		C2	C1	1.505(3)
O1	C1	1.317(3)		C2	C3	1.517(3)
C10	N3	1.326(3)		C12	C13	1.390(3)
C10	C11	1.485(3)		C5	C6	1.377(3)
O2	C1	1.206(2)		C6	C7	1.363(4)
C11	C12	1.379(3)		C9	C8	1.386(3)
C11	C16	1.386(3)		C16	C15	1.383(3)
C4	C2	1.519(3)		C14	C15	1.364(3)
C4	C5	1.386(3)		C14	C13	1.370(3)
C4	C9	1.378(3)		C8	C7	1.369(4)

Table S29. Bond Angles for PPA-BPY.

PPA-BA								
Atom	**Atom**	**Atom**	**Angle/˚**		**Atom**	**Atom**	**Atom**	**Angle/˚**
O4	C10	N3	121.4(2)		O1	C1	C2	112.73(18)
O4	C10	C11	120.58(18)		O2	C1	O1	122.8(2)
N3	C10	C11	118.1(2)		O2	C1	C2	124.4(2)
C12	C11	C10	122.50(18)		C6	C5	C4	120.4(2)
C12	C11	C16	119.2(2)		C7	C6	C5	120.7(2)
C16	C11	C10	118.30(19)		C4	C9	C8	120.7(2)
C5	C4	C2	120.33(17)		C15	C16	C11	119.8(2)
C9	C4	C2	121.21(18)		C15	C14	C13	119.5(2)
C9	C4	C5	118.45(18)		C7	C8	C9	119.9(2)
C1	C2	C4	110.12(15)		C6	C7	C8	119.9(2)
C1	C2	C3	110.82(18)		C14	C15	C16	121.0(2)
C3	C2	C4	112.40(18)		C14	C13	C12	120.5(2)
C11	C12	C13	120.0(2)					

BRIEF COMMENT ON THE CRYSTAL STRUCTURES:

The PPA-2APY salt presents a triclinic P-1 space group with Z'=2 (two molecules of PPA and two molecules of 2AP in the asymmetric unit). All the APs are protonated on the pyridine nitrogen and the carboxylic group of PPAs are deprotonated, as can be seen by the C-O distances (C-O distances = 1.266 and 1.232 Å). In the crystal structure a multitude of charge assisted N^{+}-H···O^{-} hydrogen bonds (Table S30) form octamers with chair conformation in which PPA and 2APY are interchanged (see Figure S33). These oligomers are connected by weak C(ar)-H···O contacts, while no π··· π interactions are present in this system.

Table S30. Strong hydrogen bond in the PPA-2APY salt.

Interaction	distance
N4···O2	2.641 Å
N3···O1	2.780 Å
N3···O3	2.845 Å
N2···O3	2.601 Å
N1···O4	2.837 Å
N1···O1	2.843 Å

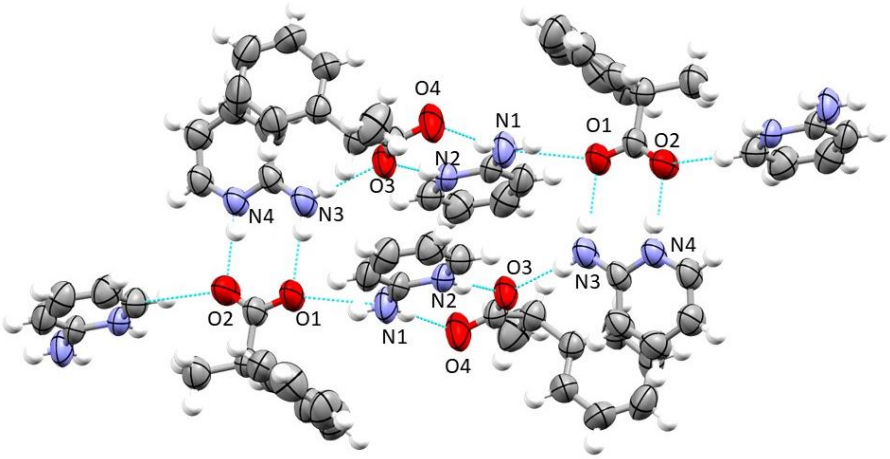


Figure S32. Hydrogen bond interactions in the PPA-2APY structure.

The PPA-BPY adduct presents a monoclinic $P2_1/n$ space group with a molecule of PPA and half BPY molecule in the asymmetric unit. In the crystal packing, the symmetry-completed BPY molecule connects two PPA with O-H···N hydrogen bonds (d(O1 ···N1) = 2.683(7) Å) defining a trimer along the a axis (Figure S34).

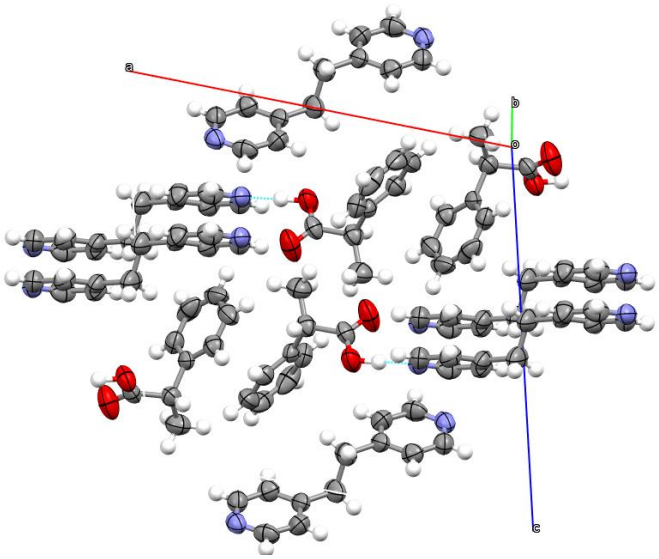


Figure S33. Unit cell of the PPA-BPY adduct.

The PPA-4ClBA adduct presents a monoclinic $P2_1/c$ space group with one neutral molecule of PPA and one neutral molecule of 4ClBA in the asymmetric unit. The strong N-H···O and O-H···O interactions between the amide and carboxylic groups of the two interacting molecules (d(N1 ···O3) = 2.972(7) Å, d(N1 ···O3) = 3.072(7), Å d(O2 ···O1) = 2.605(7) Å) make possible the formation of a centrosymmetric tetramer in the solid state (Figure S35).

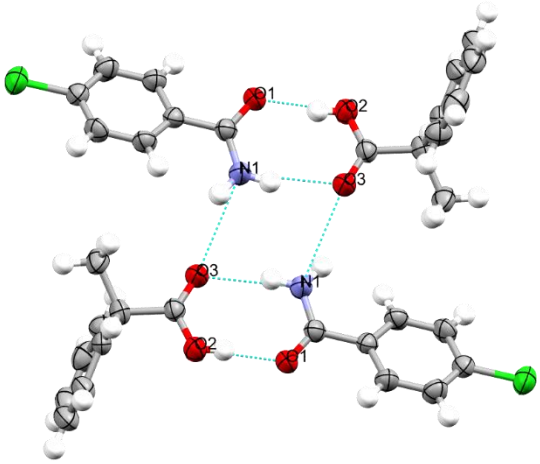


Figure S34. Hydrogen bond interactions in the PPA-2APY structure.

The molecules, although not presenting the correct geometry for π··· π interaction, show a herringbone disposition in the crystal structure, probably due to the directing angle of PPA (Figure S36).

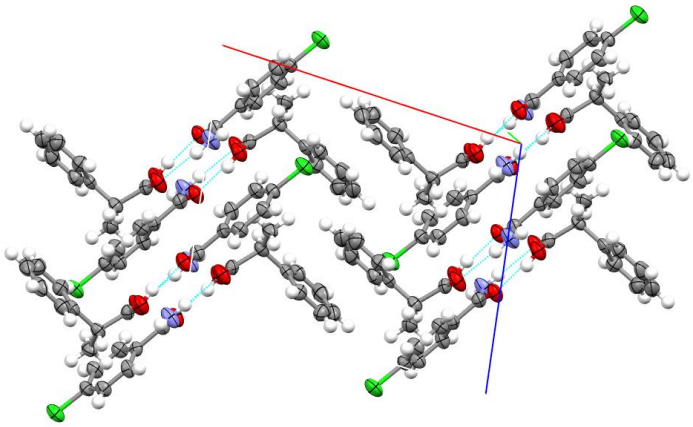


Figure S35. Unit cell of the PPA-4ClBA adduct.

The PPA-BA adduct presents a monoclinic $P2_1/n$ space group with one neutral molecule of PPA and one neutral molecule of BA in the asymmetric unit. In this case, only hydrogen-bonded dimeric aggregates between the amide and carboxylic groups are formed (d(N3 ···O2) = 2.952(5) Å and d(O4 ···O2) = 2.623(5) Å, Figure S37. These dimers interact to each other through C(ar)-H···π and π···π contacts between the aromatic moieties (Figure S37).

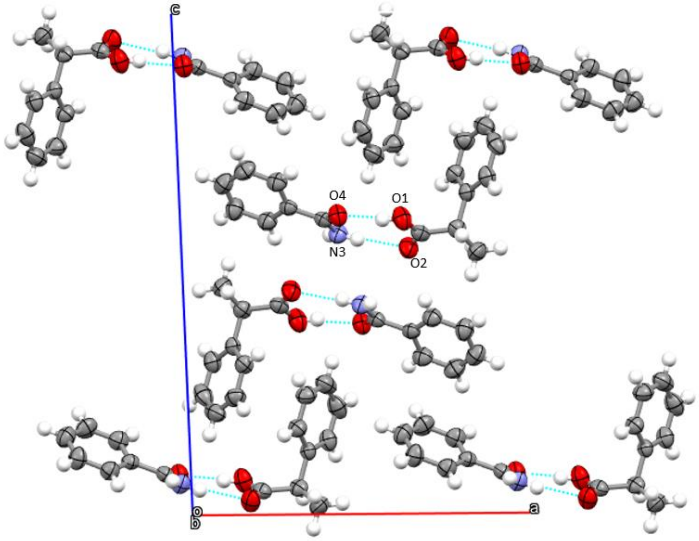


Figure S36. Unit cell of the PPA-BA adduct.

Table S31. Cut-off values of Hansen's parameters reported in the literature.

	$\Delta\delta$	$\Delta\delta t$	Ra	reference
cut-off ($MPa^{0.5}$)	5.0	7.0	5.6	[25]
cut-off ($MPa^{0.5}$)	8.18	16.87	17.64	[13]

Table S32. Comparison of sensitivity, specificity, and accuracy of the HSP, HBE, MC methods for each API in the dataset.

	HSP				HBE			MC		
API	**Cut-off**	**Sensibility**	**Specificity**	**Accuracy**	**Sensibility**	**Specificity**	**Accuracy**	**Sensibility**	**Specificity**	**Accuracy**
riluzole	7	40%	60%	50%	44%	100%	67%	73%	21%	53%
diclofenac	7	50%	60%	50%	20%	100%	42%	29%	60%	37%
indomethacin	11	60%	80%	80%	33%	89%	78%	92%	51%	59%
nalidixic acid	11	80%	50%	60%	83%	33%	52%	100%	26%	51%
caffeine	8	70%	80%	80%	65%	45%	61%	80%	55%	76%
carbamazepine	14	40%	60%	50%	67%	25%	58%	37%	50%	40%
paracetamol	11	60%	40%	50%	11%	95%	55%	68%	25%	46%
piroxicam	13	50%	50%	50%	52%	46%	51%	43%	58%	46%
piracetam	11	40%	40%	40%	54%	72%	65%	100%	5%	45%
pyrazinamide	12	80%	40%	60%	77%	39%	60%	96%	13%	58%
acetazolamide	13	50%	80%	70%	0%	96%	68%	67%	19%	31%
furosemide	8	70%	70%	70%	23%	80%	54%	94%	5%	46%
sulpiride	10	29%	40%	35%	43%	55%	50%	57%	40%	47%
AVERAGE	**10**	**55%**	**58%**	**57%**	**44%**	**67%**	**59%**	**72%**	**33%**	**49%**

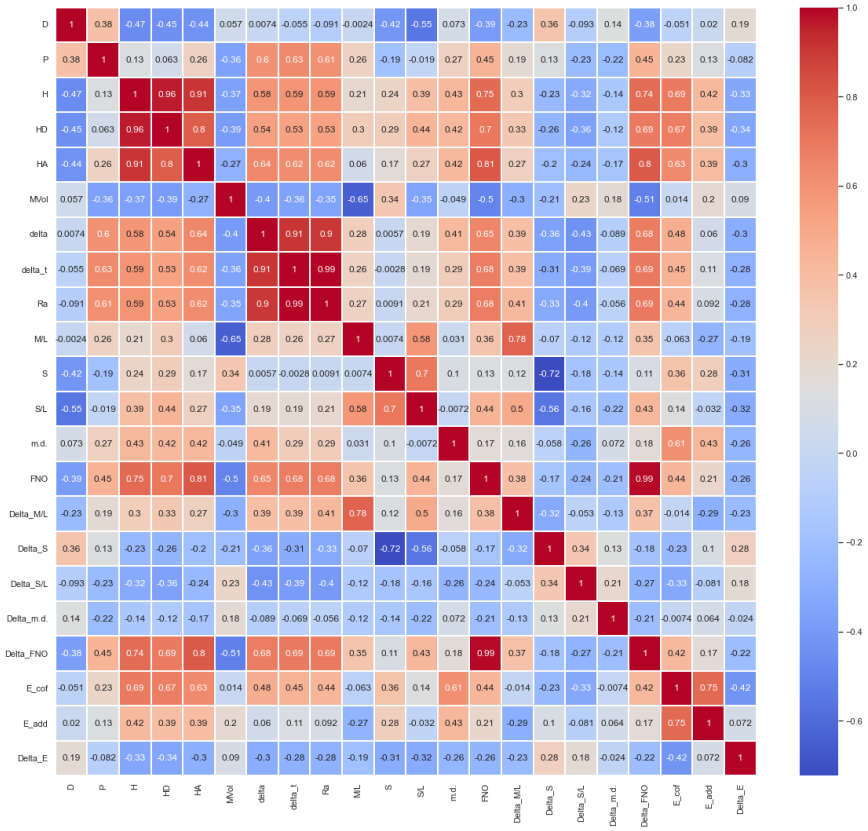


Figure S37. Correlation matrix in terms of Pearson correlation (D = δ_D, H = δ_H, P = δ_p, delta = $\Delta\delta$, delta_t = $\Delta\delta t$).

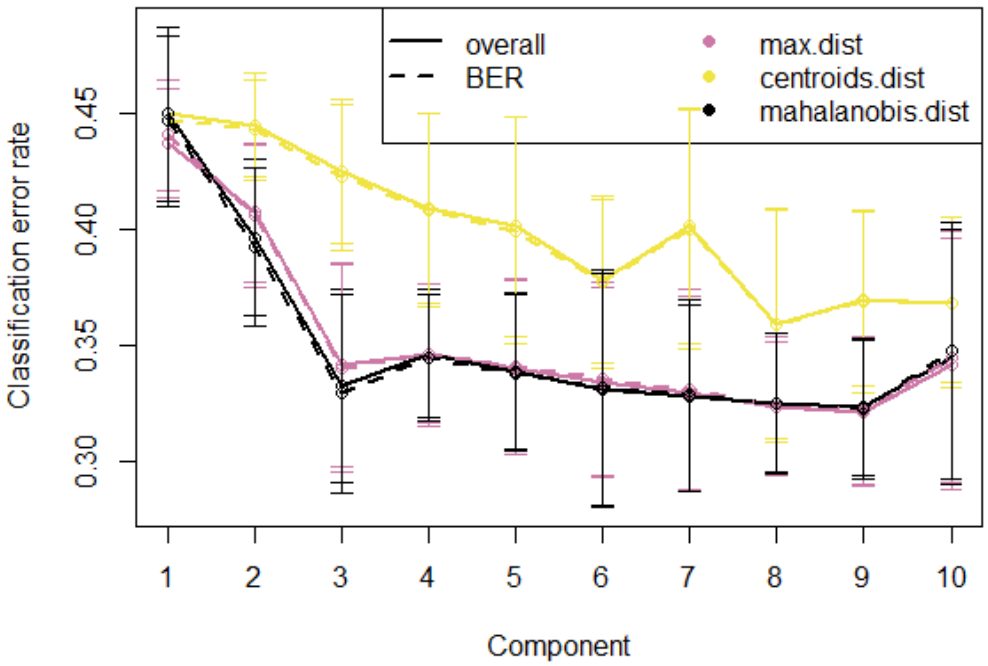


Figure S38. RMSECV for determine the number of components to use in the PLS-DA model for PPA co-crystal prediction.

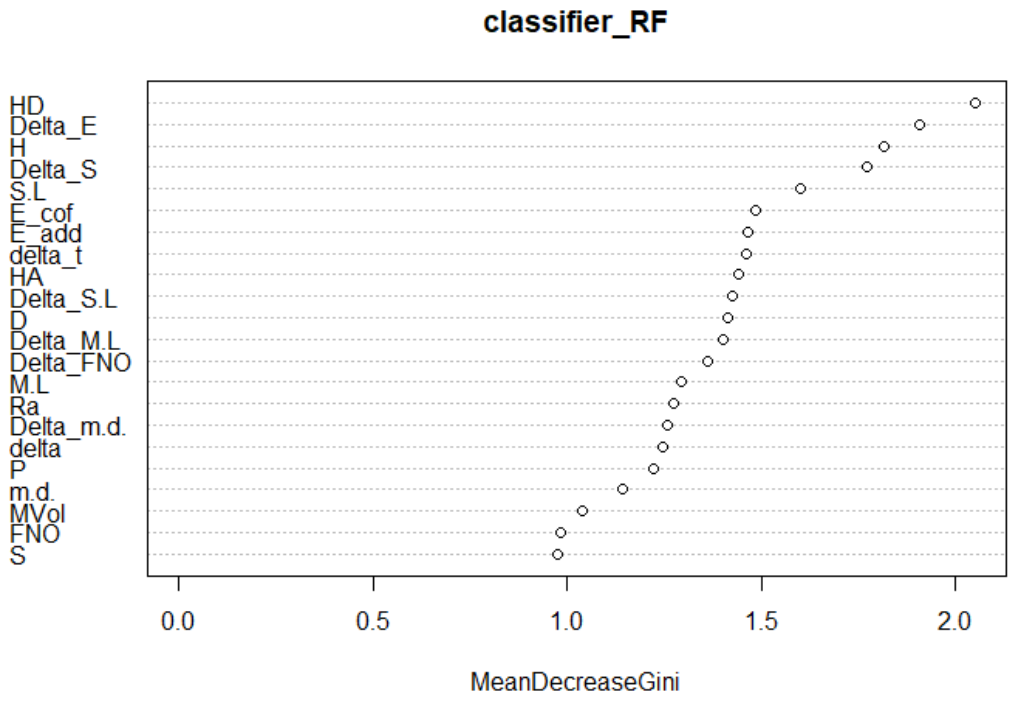


Figure S39. Feature importance in the design of the RF model for the PPA co-crystals prediction.

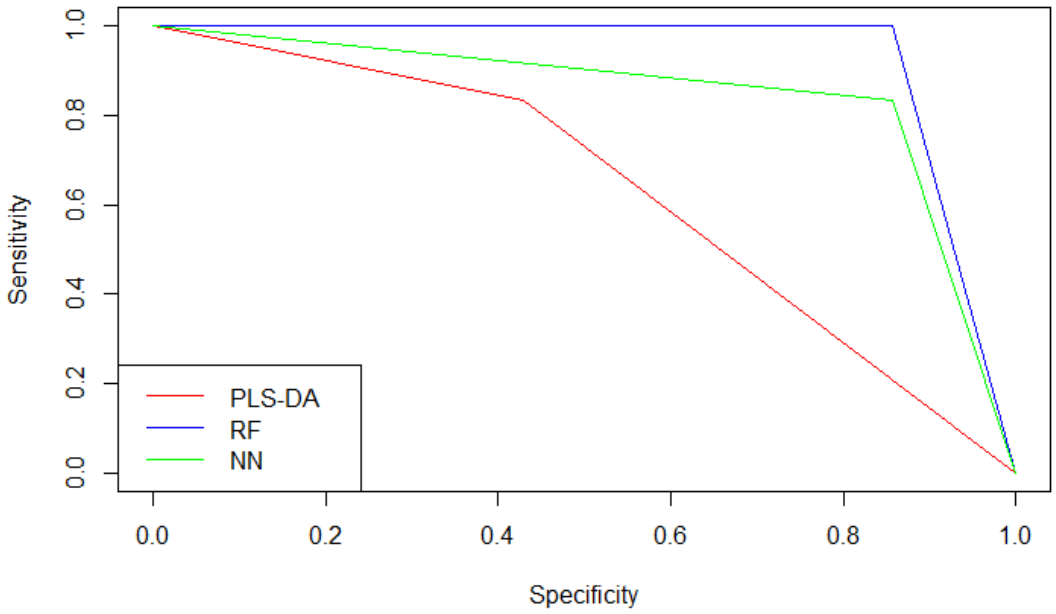


Figure S40. ROC curves to evaluate the sensitivity and specificity of the PPA model.

Table S33. Confusion matrix and performance measures of the models designed for the PPA co-crystal prediction. The training set is composed of 543 co-crystallization data of 13 APIs (see main text for more details). HSP, HBE and MC descriptors was used as features for these models.

TRAINING MIX, HSP-HBE-MC DESCRIPTORS

PLS-DA

predicted \ Experimental outcome	YES	NO
YES	7	3
NO	1	2

sensitivity = 88%
specificity = 40%
accuracy = 69%

RF

predicted \ Experimental outcome	YES	NO
YES	7	4
NO	1	1

sensitivity = 88%
specificity = 20%
accuracy = 62%

NN

predicted \ Experimental outcome	YES	NO
YES	7	3
NO	1	2

sensitivity = 88%
specificity = 40%
accuracy = 69%

Table S34. Confusion matrix and performance measures of the models designed for the PPA co-crystal prediction. The training set is composed of 64 co-crystallization data of NSAIDs (see main text for more details). QSAR descriptors was used as features for these models.

TRAINING NSAIDs, QSAR DESCRIPTORS

PLS-DA

predicted \ Experimental outcome	YES	NO
YES	8	4
NO	0	1

sensitivity = 100%
specificity = 20%
accuracy = 69%

RF

predicted \ Experimental outcome	YES	NO
YES	8	4
NO	0	1

sensitivity = 100%
specificity = 20%
accuracy = 69%

NN

predicted \ Experimental outcome	YES	NO
YES	8	5
NO	0	0

sensitivity = 100%
specificity = 0%
accuracy = 62%

Table S35. Confusion matrix and performance measures of the models designed for the PPA co-crystal prediction. The training set is composed of 543 co-crystallization data of 13 APIs (see main text for more details). QSAR descriptors was used as features for these models.

TRAINING MIX, QSAR DESCRIPTORS

PLS-DA

predicted \ Experimental outcome	YES	NO
YES	7	3
NO	1	2

sensitivity = 88%
specificity = 40%
accuracy = 69%

RF

predicted \ Experimental outcome	YES	NO
YES	2	1
NO	6	4

sensitivity = 25%
specificity = 80%
accuracy = 46%

NN

predicted \ Experimental outcome	YES	NO
YES	3	2
NO	5	3

sensitivity = 38%
specificity = 60%
accuracy = 46%

Table S36. Confusion matrix and performance measures of different published tools for the PPA co-crystal prediction.

OTHER PREDICTIVE TOOLS

HSP

predicted \ Experimental outcome	YES	NO
YES	8	5
NO	0	0

sensitivity = 100%
specificity = 0%
accuracy = 62%

HBE

predicted \ Experimental outcome	YES	NO
YES	6	1
NO	2	4

sensitivity = 75%
specificity = 80%
accuracy = 77%

MC

predicted \ Experimental outcome	YES	NO
YES	6	4
NO	2	1

sensitivity = 75%
specificity = 20%
accuracy = 54%

CCGNet

predicted \ Experimental outcome	YES	NO
YES	2	1
NO	6	4

sensitivity = 25%
specificity = 80%
accuracy = 46%

References

(1) Yadav, B.; Balasubramanian, S.; Chavan, R. B.; Thipparaboina, R.; Naidu, V. G. M.; Shastri, N. R. Hepatoprotective Cocrystals and Salts of Riluzole: Prediction, Synthesis, Solid State Characterization, and Evaluation. *Cryst. Growth Des.* **2018**, *18* (2), 1047–1061. https://doi.org/10.1021/acs.cgd.7b01514.

(2) Aakeröy, C. B.; Grommet, A. B.; Desper, J. Co-Crystal Screening of Diclofenac. *Pharmaceutics* **2011**, *3* (3), 601–614. https://doi.org/10.3390/pharmaceutics3030601.

(3) Surov, A. O.; Voronin, A. P.; Manin, A. N.; Manin, N. G.; Kuzmina, L. G.; Churakov, A. V.; Perlovich, G. L. Pharmaceutical Cocrystals of Diflunisal and Diclofenac with Theophylline. *Mol. Pharm.* **2014**, *11* (10), 3707–3715. https://doi.org/10.1021/mp5004652.

(4) Grecu, T.; Hunter, C. A.; Gardiner, E. J.; McCabe, J. F. Validation of a Computational Cocrystal Prediction Tool: Comparison of Virtual and Experimental Cocrystal Screening Results. *Cryst. Growth Des.* **2014**, *14* (1), 165–171. https://doi.org/10.1021/cg401339v.

(5) Kojima, T.; Tsutsumi, S.; Yamamoto, K.; Ikeda, Y.; Moriwaki, T. High-Throughput Cocrystal Slurry Screening by Use of in Situ Raman Microscopy and Multi-Well Plate. *Int. J. Pharm.* **2010**, *399* (1–2), 52–59. https://doi.org/10.1016/j.ijpharm.2010.07.055.

(6) Ferretti, V.; Dalpiaz, A.; Bertolasi, V.; Ferraro, L.; Beggiato, S.; Spizzo, F.; Spisni, E.; Pavan, B. Indomethacin Co-Crystals and Their Parent Mixtures: Does the Intestinal Barrier Recognize Them Differently? *Mol. Pharm.* **2015**, *12* (5), 1501–1511. https://doi.org/10.1021/mp500826y.

(7) Grecu, T.; Adams, H.; Hunter, C. A.; McCabe, J. F.; Portell, A.; Prohens, R. Virtual Screening Identifies New Cocrystals of Nalidixic Acid. *Cryst. Growth Des.* **2014**, *14* (4), 1749–1755. https://doi.org/10.1021/cg401889h.

(8) Springuel, G.; Norberg, B.; Robeyns, K.; Wouters, J.; Leyssens, T. Advances in Pharmaceutical Co-Crystal Screening: Effective Co-Crystal Screening through Structural Resemblance. *Cryst. Growth Des.* **2012**, *12* (1), 475–484. https://doi.org/10.1021/cg201291k.
(9) Gong, H.; Wang, Q.; Du, Y. Dynamic Investigation of Cocrystallization between Piracetam and Hydroquinone with Terahertz Time-Domain Spectroscopy. In *2017 42nd International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz)*; IEEE: Cancun, Mexico, 2017; pp 1–2. https://doi.org/10.1109/IRMMW-THz.2017.8066975.
(10) Bian, L.; Zhao, H.; Hao, H.; Yin, Q.; Wu, S.; Gong, J.; Dong, W. Novel Glutaric Acid Cocrystal Formation via Cogrinding and Solution Crystallization. *Chem. Eng. Technol.* **2013**, *36* (8), 1292–1299. https://doi.org/10.1002/ceat.201200720.
(11) Roca-Paixão, L.; Correia, N. T.; Affouard, F. Affinity Prediction Computations and Mechanosynthesis of Carbamazepine Based Cocrystals. *CrystEngComm* **2019**, *21* (45), 6991–7001. https://doi.org/10.1039/C9CE01160A.
(12) Musumeci, D.; Hunter, C. A.; Prohens, R.; Scuderi, S.; McCabe, J. F. Virtual Cocrystal Screening. *Chem. Sci.* **2011**, *2* (5), 883. https://doi.org/10.1039/c0sc00555j.
(13) Salem, A.; Nagy, S.; Pál, S.; Széchenyi, A. Reliability of the Hansen Solubility Parameters as Co-Crystal Formation Prediction Tool. *Int. J. Pharm.* **2019**, *558*, 319–327. https://doi.org/10.1016/j.ijpharm.2019.01.007.
(14) Wang, D.; Yang, Z.; Zhu, B.; Mei, X.; Luo, X. Machine-Learning-Guided Cocrystal Prediction Based on Large Data Base. *Cryst. Growth Des.* **2020**, *20* (10), 6610–6621. https://doi.org/10.1021/acs.cgd.0c00767.
(15) Latif, S.; Abbas, N.; Hussain, A.; Arshad, M. S.; Bukhari, N. I.; Afzal, H.; Riffat, S.; Ahmad, Z. Development of Paracetamol-Caffeine Co-Crystals to Improve Compressional, Formulation and *in Vivo* Performance. *Drug Dev. Ind. Pharm.* **2018**, *44* (7), 1099–1108. https://doi.org/10.1080/03639045.2018.1435687.
(16) Sarmah, K. K.; Rajbongshi, T.; Bhowmick, S.; Thakuria, R. First-Line Antituberculosis Drug, Pyrazinamide, Its Pharmaceutically Relevant Cocrystals and a Salt. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2017**, *73* (5), 1007–1016. https://doi.org/10.1107/S2052520617011477.
(17) Kulla, H.; Greiser, S.; Benemann, S.; Rademann, K.; Emmerling, F. In Situ Investigation of a Self-Accelerated Cocrystal Formation by Grinding Pyrazinamide with Oxalic Acid. *Molecules* **2016**, *21* (7), 917. https://doi.org/10.3390/molecules21070917.
(18) Luo, Y.-H.; Sun, B.-W. Pharmaceutical Co-Crystals of Pyrazinecarboxamide (PZA) with Various Carboxylic Acids: Crystallography, Hirshfeld Surfaces, and Dissolution Study. *Cryst. Growth Des.* **2013**, *13* (5), 2098–2106. https://doi.org/10.1021/cg400167w.
(19) Wang, J.-R.; Ye, C.; Zhu, B.; Zhou, C.; Mei, X. Pharmaceutical Cocrystals of the Anti-Tuberculosis Drug Pyrazinamide with Dicarboxylic and Tricarboxylic Acids. *CrystEngComm* **2015**, *17* (4), 747–752. https://doi.org/10.1039/C4CE02044H.
(20) Wood, P. A.; Feeder, N.; Furlow, M.; Galek, P. T. A.; Groom, C. R.; Pidcock, E. Knowledge-Based Approaches to Co-Crystal Design. *CrystEngComm* **2014**, *16* (26), 5839. https://doi.org/10.1039/c4ce00316k.
(21) Hiendrawan, S.; Veriansyah, B.; Widjojokusumo, E.; Soewandhi, S. N.; Wikarsa, S.; Tjandrawinata, R. R. Physicochemical and Mechanical Properties of Paracetamol Cocrystal with 5-Nitroisophthalic Acid. *Int. J. Pharm.* **2016**, *497* (1–2), 106–113. https://doi.org/10.1016/j.ijpharm.2015.12.001.
(22) Birolo, R.; Bravetti, F.; Bordignon, S.; D'Abbrunzo, I.; Mazzeo, P. P.; Perissutti, B.; Bacchi, A.; Chierotti, M. R.; Gobetto, R. Overcoming the Drawbacks of Sulpiride by Means of New Crystal Forms. *Pharmaceutics* **2022**, *14* (9), 1754. https://doi.org/10.3390/pharmaceutics14091754.
(23) Othman, M. F.; Anuar, N.; Ad Rahman, S.; Ahmad Taifuddin, N. A. Cocrystal Screening of Ibuprofen with Oxalic Acid and Citric Acid via Grinding Method. *IOP Conf. Ser. Mater. Sci. Eng.* **2018**, *358*, 012065. https://doi.org/10.1088/1757-899X/358/1/012065.
(24) Dash, S. G.; Thakur, T. S. Computational Screening of Multicomponent Solid Forms of 2-Aryl-Propionate Class of NSAID, Zaltoprofen, and Their Experimental Validation. *Cryst. Growth Des.* **2021**, *21* (1), 449–461. https://doi.org/10.1021/acs.cgd.0c01278.

(25) Mohammad, M. A.; Alhalaweh, A.; Velaga, S. P. Hansen Solubility Parameter as a Tool to Predict Cocrystal Formation. *Int. J. Pharm.* **2011**, *407* (1–2), 63–71. https://doi.org/10.1016/j.ijpharm.2011.01.030.