# Deep Supramolecular Language Processing for Co-crystal Prediction

**Rebecca Birolo,** [1 2] **Rıza Özçelik,** [1 3] **Andrea Aramini,** [4] **Roberto Gobetto,** [2] **Michele R. Chierotti,** [2] **Francesca Grisoni*** [1 3]

## Abstract

Approximately 40% of marketed drugs exhibit suboptimal pharmacokinetic profiles. Co-crystallization, where pairs of molecules form a multicomponent crystal, constitutes a promising strategy to enhance physicochemical properties without compromising the pharmacological activity. However, finding promising co-crystal pairs is resource-intensive, due to the vast number of possible combinations. We present DeepCocrystal, a novel deep learning approach designed to predict co-crystal formation by processing the 'chemical language' from a supramolecular vantage point. Rigorous validation of DeepCocrystal showed a balanced accuracy of 78% in realistic scenarios, outperforming existing models. By leveraging properties of molecular string representations, DeepCocrystal can also estimate the uncertainty of its predictions. We harness this capability in a challenging prospective study, and successfully discovered two novel co-crystal of diflunisal, an anti-inflammatory drug. This study underscores the potential of deep learning – and in particular of chemical language processing – to accelerate co-crystallization, and ultimately drug development, in both academic and industrial contexts.

## 1. Introduction

Co-crystallization enables the optimization of the pharmacokinetic properties of active pharmaceutical ingredients (APIs)[1,2]. Via co-crystalization, supramolecular interactions between the API and another molecule (coformer) are established to form a multicomponent crystal[3] (Fig. 1a). The resulting co-crystal preserves the bioactivity of the lead molecule while enhancing desirable properties, such as solubility, and stability. Owed to the high number of possible combinations, finding the optimal coformer for a given API is far from trivial, and ultimately relies on a labor- and time-intensive process based on trial and error[4,5].

Machine learning – which extracts relevant information from chemical datasets[6] – can aid in prioritizing API-coformer pairs for co-crystallization[7–11]. Current methods, however, might struggle to generalize to previously unseen molecules[12]. This is in part due to limitations of training datasets, which are unrealistically imbalanced towards existing co-crystals[13]. Therefore, there is a need for approaches that are more robust to data imbalance and demonstrate stronger generalizability to previously unseen molecules.

Here we introduce DeepCocrystal, a novel deep learning approach designed to learn the "supramolecular language" of co-crystallization. Supramolecular chemistry can be viewed as a language[14–16]: atoms ('letters') form molecules ('words'), whose combinations give rise to supramolecular interactions ('sentences'). Building on this analogy, we extend current chemical language processing techniques[17–20] — which predict molecular properties from single string representations[21,22] — to predicting supramolecular interactions between pairs of molecules (*i.e.*, co-crystallization).

DeepCocrystal represents single molecules (API and coformer) as SMILES (Simplified Molecular Input Line Entry Systems[21]) strings (Fig. 1b), whose chemical information is combined to predict whether they form co-crystals. Thanks to intriguing properties of the SMILES language[23], DeepCocrystal addresses the data imbalance and estimates prediction uncertainty, pivotal for prospective applications.

In this work, DeepCocrystal shows superior performance and generalization capacity than existing approaches[24–27]. When applied prospectively to identify coformer candidates, all high-certainty predictions of DeepCocrystals were confirmed experimentally – leading to the identification of two previously unreported diflunisal co-crystals. To the best of our knowledge, this is the first application of "supramolecular language" processing to predict co-crystallization – opening novel opportunities in supramolecular chemistry.

[1] Institute for Complex Molecular Systems, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. [2] Department of Chemistry and NIS Centre, University of Torino, Torino, Italy. [3] Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Utrecht, The Netherlands. [4] Research and Early Development, Dompé Farmaceutici S.p.A, L'Aquila, Italy. Correspondence to: F. Grisoni <f.grisoni@tue.nl>.

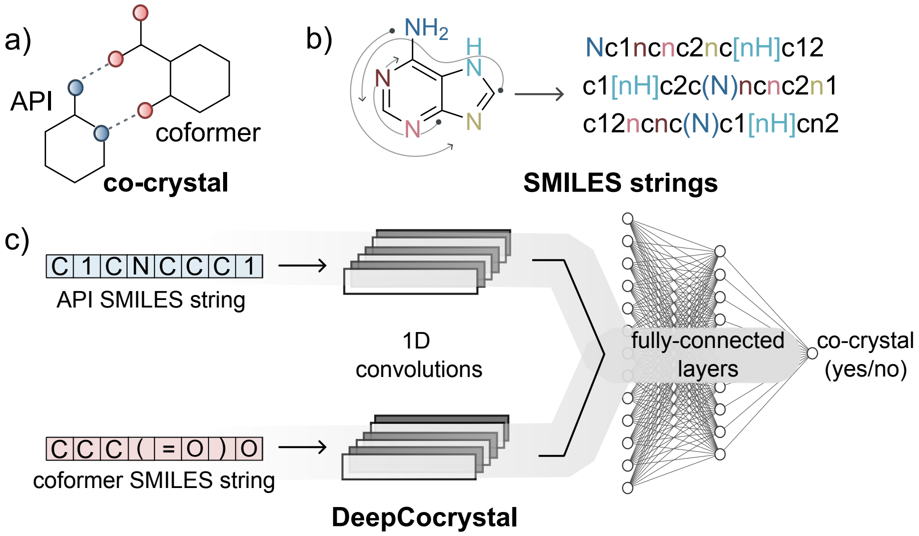

*Figure 1. Overview of key elements of DeepCocrystal for co-crystal prediction.* **a)** The co-crystallization between an active pharmaceutical ingredient (API) and a coformer involves the formation of a multicomponent crystalline structure (co-crystal), in which the API and coformer are held together by non-covalent interactions. **b)** SMILES strings, which convert a molecular graph into one string. One molecule can be represented by many different SMILES strings, based on the starting (non-hydrogen) atom and the chosen direction for graph traversal. **c)** DeepCocrystal represents API and coformers via SMILES strings and passes them through 1-dimensional (1D) convolutions. Fully-connected layers are then used to predict the co-crystallization output as a continuous number between 0 and 1, which can be then discretized (with a cut-off of 0.5) to perform a prediction ("negative" pair if below, and "positive" pair otherwise).

## 2. Results and Discussion

### 2.1. DeepCocrystal architecture

DeepCocrystal has at its core Convolutional Neural Networks (CNNs)[28] for 'chemical language' processing. CNNs are a class of deep learning models commonly used for processing sequences of text[29]. Via convolution – which involves sliding a filter (kernel) over the input text – CNNs can capture information and features at different levels of abstraction, and progressively aggregate it to provide a prediction. DeepCocrystal leverages SMILES[21] strings as an input, which are derived from traversing a molecular graph from a non-hydrogen atom, and annotating atoms and bonds with specific symbols (Fig. 1b). CNNs have been previously applied to predict the properties of single molecules from their SMILES strings[17–19].

DeepCocrystal extends traditional chemical language processing approaches beyond the 'one-molecule-one-property' paradigm, to learn simultaneously from the SMILES strings of *pairs* of molecules (*i.e.*, API-coformer pairs). In particular, DeepCocrystal uses two separate CNNs to learn 'latent representations' of the input molecular structures (of each API and coformer), and then aggregates this information via a fully-connected neural network, to predict the potential co-crystallization of the input pair (Fig. 1c). Via the DeepCocrystal architecture, the co-crystalization potential of any molecular pair is predicted as a number between 0 (negative) and 1 (positive).

In this work, every API-coformer pair was presented to the network twice, once per every separate CNN, as previously suggested[11,27]. This strategy allows artificially increasing the number of inputs available for model training. Moreover, we experimented with different SMILES string variations, to serve as input for DeepCocrystal. In particular, we experimented with (a) *canonical SMILES*, which provide a univocal string per every molecular structure via standardization algorithm[30], and (b) *'randomized' SMILES*, which can provide a different SMILES string based on the chosen starting atom and the graph traversal route (Fig. 1b). Randomized SMILES strings were used to perform 'data augmentation'[23], *i.e.*, to artificially inflate the number of data available for training by using multiple SMILES for a single molecule.

### 2.2. DeepCocrystal training and validation

To train and validate DeepCocrystal, we collected and manually curated a dataset of experimentally-determined co-crystal structures, from (a) the Cambridge Structural Database[31] and (b) existing co-crystal literature[27,32–36]. Moreover, a set of in-house experiments was conducted to measure the co-crystalization of additional molecular pairs. The collected dataset comprises a total of 6632 API-coformer pairs, of which 5240 (79%) are co-crystals ("positive") and 1392 (21%) are physical mixtures ("negative", *i.e.,* no observed co-crystallization).

The training, validation and internal test sets were created by stratified splits of this dataset (10 randomly sampled subsets with 10% molecules in validation and test folds). In addition to using canonical SMILES as input, we also experimented with different levels of augmentation: (a) [positive:negative = 1:4], where one randomized SMILES string is used for every molecule in a "positive" pair, and four SMILES are used for molecules in "negative" pairs, and (b) [positive:negative = 2:7], where a two-fold and a seven-fold augmentation are used for the SMILES strings of positive and negative pairs, respectively. Each model variant was evaluated for its classification performance[37] (Table 1), *i.e.*, via Recall (ability to correctly classify positive pairs, Eq. 1), Specificity (ability to correctly classify negative pairs, Eq. 2) and Balanced Accuracy (overall performance, Eq. 3). These metrics were computed by considering predictions lower than 0.5 as a "negative", or "positive" otherwise.

All DeepCocrystal variants reached a Balanced Accuracy above 88%, with the 2:7 augmentation performing the best. When looking at class performance, different trends can be observed. In identifying "positive" pairs, canonical SMILES lead to the best performance (up to 5% increase in recall). All DeepCocrystal variants have a good capacity to recognize "positive" pairs, with 1:4 and 2:7 augmentations showing comparable performance. DeepCocrystal trained on canonical SMILES showed a significantly higher Recall than the two augmented models (Wilcoxon signed-rank test, $p < 0.05$). On the contrary, the 2:7 SMILES augmentation significantly improves the ability to identify negative pairs (Wilcoxon signed-rank test, $p < 0.05$), resulting in an 8% increase in specificity compared to the canonical version. This evidence highlights how SMILES augmentation on the negative class, can aid in mitigating the data unbalance.

## 2.3. Model benchmarking

The predictive performance of DeepCocrystal was then evaluated on an external test set, which was manually curated by combining public data with in-house experimental co-crystallization results of selected APIs (*see* Materials and Methods). This external set contained 364 pairs (129 are co-crystals and 235 non-co-crystals), with a lower substructure similarity[38] to the training set than the internal test set (Supporting Fig. S1) – constituting a more challenging validation set.

DeepCocrystal was benchmarked with four existing approaches: (i) CCGNet[27], which relies on graph neural networks to perform a prediction; (ii) CC-Descriptor ML, which relies on an array of 'classical' machine learning models trained on co-crystal descriptors[26]; (iii) Descriptor-DNN, based on a fully-connected neural network trained on molecular descriptors[24]; and (iv) Fingerprint-DNN, a fully-connected neural network trained on extended connectivity fingerprints[25,39]. To ensure comparability and account for the lack of provided code, data, and/or hyperparameters, we re-implemented and trained Descriptor-DNN and Fingerprint-DNN, using the same dataset as DeepCocrystal (*see* Materials and Methods).

DeepCocrystal consistently outperformed the benchmarks (Table 1). DeepCocrystal, in its augmented 2:7 configuration, achieved 15%-21% higher balanced accuracy and 12%-56% higher specificity than the benchmarks, albeit with a moderate recall reduction (of up to 15% lower). These results indicate that DeepCocrystal finds a better trade-off between positive and negative prediction power than the benchmarks, which are unbalanced toward the positives. Furthermore, the SMILES augmentation increased the balanced accuracy by 10% and 19%, respectively for 1:4 and 2:7 augmentation levels, compared to using canonical SMILES strings, indicating a higher generalization potential provided by learning from different SMILES versions of the same molecule.

## 2.4. Uncertainty estimation

To extend the applicability of DeepCocrystal to real-world scenarios, we equipped it with an estimate of its (un)certainty. We represented each molecular pair with ten different (pairs of) SMILES strings, and used DeepCocrystal (2:7) predictions to estimate uncertainty. Considering the predictions on SMILES ensembles (*i.e.*, by average prediction, Fig. 2), allows detecting some of the model errors.

We tested two ways of estimating the DeepCocrystal's uncertainty starting from its predictions on the 'molecular-pair ensemble' (*i.e.*, 10-fold SMILES repetitions for each molecular pair): (a) *Majority voting*, whereby the number of agreements in the predicted class per each molecular pair is used as a measure of confidence (the higher, the better); and (b) *Standard deviation-based estimation*, whereby the standard deviation across augmented SMILES (per each pair) is computed (the lower, the better). For each approach, several thresholds of uncertainty (*i.e.*, on standard deviation or on number of agreeing predictions) were used to analyse their effect on performance, in terms of classification accuracy and number of molecules retained for prediction (Table 2).

For both uncertainty estimation strategies, DeepCocrystal performance consistently increases when using stricter thresholds (up to 10% improvement across metrics), with a progressively smaller number of predicted pairs (Table 2). Both approaches have their merits and drawbacks. Standard deviation outperforms majority voting in classification performance (up to 2% improvement), at the expanses of the number of predicted molecular pairs (57 fewer pairs). The approach to use should be chosen on a case-by-case basis, and here, we used a threshold on standard deviation equal to 0.10, to maximize prediction performance.

*Table 1. Performance of DeepCocrystal.* DeepCocrystal was tested on two test sets, one internal and one external. The internal test sets was composed of 664 molecular pairs, which were sampled by stratified splits of the collected dataset. The external set was composed of 364 pairs collected in a second phase of the project, and containing more structurally diverse molecular pairs. The external test set was used to benchmark DeepCocrystal with existing literature models (*i.e.*, Fingerprint-DNN, Descriptor-DNN, CC-Descriptor-ML, and CCGNet[24–27]). Balanced accuracy (global performance), recall (performance on "positive" pairs), and specificity (performance on "negative" pairs) are reported for each set and each model (the closer to 100%, the better). The best performance per metric is highlighted in boldface for each considered test set.

| Test set | Model | BAcc | Recall | Specificity |
|---|---|---|---|---|
| Internal | DeepCocrystal - canonical | 88% ± 2% | **96% ± 1%** | 79% ± 6% |
| | DeepCocrystal - augmented (1:4) | 88% ± 2% | 91% ± 2% | 86% ± 3% |
| | DeepCocrystal - augmented (2:7) | **89% ± 2%** | 92% ± 2% | **87% ± 3%** |
| External | DeepCocrystal - canonical | 59% | **93%** | 26% |
| | DeepCocrystal - augmented (1:4) | 69% | 71% | 66% |
| | DeepCocrystal - augmented (2:7) | **78%** | 75% | **81%** |
| | CCGNet[27] | 60% | 51% | 69% |
| | CC-Descriptor-ML[a] [26] | 63% | 79% | 48% |
| | Descriptor-DNN[24] | 63% | 84% | 41% |
| | Fingerprint-DNN[25] | 57% | 90% | 25% |

[a] Performance computed by excluding five molecular pairs that were used for model training.

*Table 2. Uncertainty estimation with DeepCocrystal.* External test set molecules were represented as 10 SMILES strings each before prediction (using DeepCocrystal 2:7). Two approaches were considered to estimate uncertainty, *i.e.*, majority voting, which picks the most frequent class among the predictions (per molecular pair), and standard deviation computed on the individual model predictions per each pair. Different uncertainty thresholds on each approach were analyzed for their effect on the model performance, as well as on the number of molecular pairs predicted. The number and percentage of predicted pairs (*i.e.*, predictions below the considered thresholds), balanced accuracy (BAcc), recall, and specificity are reported. DeepCocrystal on canonical SMILES (which is invariant to augmentation and cannot be used for uncertainty estimation) was used as a performance baseline. The best performing models per metric are highlighted in boldface.

| SMILES input | Method | Thr. | No. Pairs (%) | BAcc | Recall | Specificity |
|---|---|---|---|---|---|---|
| Canonical | - | - | 364 (100%) | 78% | 75% | 81% |
| Augmented (10-fold) | Major. | $\geq$ 50% | 364 (100%) | 76% | 75% | 77% |
| | Major. | $\geq$ 60% | 348 (96%) | 77% | 75% | 79% |
| | Major. | $\geq$ 70% | 313 (86%) | 79% | 77% | 82% |
| | Major. | $\geq$ 80% | 287 (79%) | 82% | 79% | 84% |
| | Major. | $\geq$ 90% | 254 (70%) | 84% | 82% | 86% |
| | Major. | = 100% | 218 (60%) | 87% | **86%** | 89% |
| | St. dev. | $\leq$ 0.50 | 364 (100%) | 76% | 75% | 77% |
| | St. dev. | $\leq$ 0.40 | 351 (96%) | 77% | 76% | 78% |
| | St. dev. | $\leq$ 0.30 | 275 (76%) | 82% | 80% | 83% |
| | St. dev. | $\leq$ 0.20 | 227 (62%) | 86% | 85% | 87% |
| | St. dev. | $\leq$ 0.10 | 191 (52%) | **88%** | **86%** | 90% |
| | St. dev. | $\leq$ 0.05 | 161 (44%) | **88%** | 84% | **91%** |

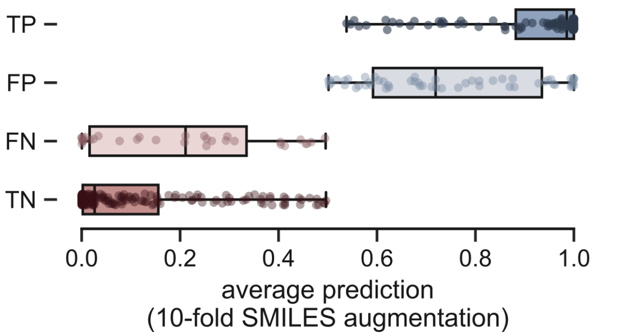

*Figure 2. Relationship between DeepCocrystal predictions and classification performance.* The SMILES of external test set samples were augmented 10 times and the average prediction was computed per API-coformer pair. Such average prediction was used to classify the molecular pairs based on a cut-off of 0.5 (negative if below, and positive otherwise). Molecular pairs were by comparing their true class with the predicted class: TP = True Positive; FP = False Positive; FN = False Negative; TN = True Negative. Box plots depict the distribution of DeepCocrystal's predictions for each group (central line: median; box: inter-quartile range; whiskers: minimum and maximum values). The median predictions of DeepCocrystal were significantly different between true and false classifications (*i.e.*, TP *vs.* FP, and TN *vs.* FN; Kruskal-Wallis H-test, $p < 0.05$).

## 2.5. Prospective experimental application

We applied DeepCocrystal prospectively, to previously unseen molecular pairs. Diflunisal, an anti-inflammatory drug[40] (Fig. 3), was selected as API, since its poor water solubility renders co-crystallization a viable strategy to enhance its bioavailability[41]. As potential coformers, we selected 12 natural products containing polyphenolic or purine moieties (Supporting Table S11), due to their co-administrability and health benefits such as central nervous system stimulation, reduced risk of neurodegenerative diseases, and anti-inflammatory properties[42–45].

10-fold augmentation was performed on each SMILES strings, and the co-crystalization potential of the respective 12 API-coformer pairs was predicted with DeepCocrystal (Supporting Table S11). For experimental validation, three categories of predictions were considered (Table 3): (a) top-two high-certainty, positive prediction (adenine and caffeine), (b) top-two high-certainty, negative predictions (rosmarinic acid and riboflavin), and (c) two most uncertain predictions (theobromine and xanthine). Each selected pair was tested in the lab via well-established protocols, *i.e.*, via grinding, liquid-assisted grinding, and slurry methods[46]. The co-crystalization outcome was determined on the obtained powder samples, via infrared spectroscopy and solid-state nuclear magnetic resonance (*see* Materials and Methods).

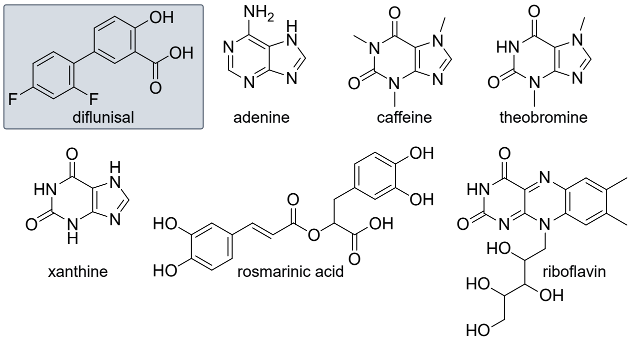

*Figure 3. Coformer candidates for diflunisal (API), selected for the prospective experimental validation.* DeepCocrystal was used to select two 'positive' predictions (adenine and caffeine), two 'negative' predictions (rosmarinic acid and riboflavin), and two high-uncertainty predictions (theobromine and xanthine) for experimental testing. The experimental tests confirmed DeepCocrystal predictions (Table 3).

*Table 3. Results of the prospective experiments guided by DeepCocrystal.* DeepCocrystal (2:7 augmentation) was used to predict the co-crystalization potential with diflunisal, among a list of 12 candidates. Mean and standard deviation of the predictions are reported (as computed on 10-fold SMILES augmentation), and a threshold on the standard deviation . The experimental outcome after lab validation is reported for six selected molecules. Symbols indicate the outcome of the predictions and experimental validation (× = negative outcome; ? = uncertain outcome; ✓= positive outcome).

| **Tested** | **DeepCocrystal** | | **Experimental** |
|---|---|---|---|
| **coformer** | **Prediction** | **Outcome** | **Outcome** |
| Adenine | $0.99 \pm 0.00$ | ✓ | ✓ |
| Caffeine | $0.99 \pm 0.01$ | ✓ | ✓ |
| Theobromine | $0.66 \pm 0.35$ | ? | × |
| Xanthine | $0.63 \pm 0.38$ | ? | × |
| Rosmarinic acid | $0.02 \pm 0.02$ | × | × |
| Riboflavin | $0.00 \pm 0.00$ | × | × |

All four high-certainty predictions of DeepCocrystal (adenine and caffeine as 'positive' predictions, and rosmarinic acid and riboflavin as 'negative' predictions) were confirmed experimentally (Table 3). To the best of our knowledge, the use of adenine and caffeine as coformers for diflunisal has not been previously reported. Future dissolution studies and activity assays will be needed to investigate whether this co-crystal leads to improvement in the solubility and pharmacokinetic profile of diflunisal, as observed in other caffeine-based systems[47–49]. Furthermore, both selected high-uncertainty pairs (theobromine and xanthine) did not form co-crystals (Table 3), indicating the usefulness of our uncertainty estimation approach to rule out false predictions. This experimental validation confirms the potential of DeepCocrystal to accelerate the discovery of novel co-crystal pairs, even with the structurally-similar selection of potential coformers selected in this study.

SMILES augmentation seemed pivotal to achieve these results. DeepCocrystal trained on canonical SMILES, in fact, predicted all purine derivate coformers as 'positive' for co-crystallization with high scores (*see* Supporting Table S11). These findings indicated that chemical language processing and SMILES augmentation allowed DeepCocrystal to capture small structural changes that might be relevant for co-crystallization. DeepCocrystal's capacity to correctly recognize both negative and positive pairs with high certainty underscores its potential to reduce experimental efforts in co-crystal screening and discovery.

## 3. Conclusions

Optimizing the pharmacokinetic properties of active compounds is an ever-lasting challenge in drug discovery, and co-crystallization is an attractive strategy to address this issue. However, identifying suitable co-crystallization partners for active compounds is both resource- and time-intensive. To accelerate this process, we developed DeepCocrystal, a deep chemical language processing approach designed to predict the co-crystallization of any selected molecular pairs.

This study shows the potential of DeepCocrystal to advance the state-of-the-art. DeepCocrystal owes its performance to the intriguing properties of the SMILES language, which allowed mitigating data imbalance and estimating uncertainty. By learning (and then combining) single-molecule information, DeepCocrystal learns elements of the "supramolecular language"[14–16] of co-crystal formation. The experimental validation of DeepCocrystal further corroborated its potential and identified adenine and caffeine as two previously unreported coformers of diflunisal. These results, taken together, underscore the potential of DeepCocrystal to accelerate the discovery of co-crystallization partners.

This first-in-time adoption of the "supramolecular language" perspective with SMILES strings shows its potential for co-crystalization prediction. While this study only focused on 'two-word sentences' (*i.e.*, molecule *pairs*), our approach could be extended to supramolecular interactions among multiple molecular partners. Ultimately, extensions of DeepCocrystal might open unexplored opportunities in supramolecular chemistry, *e.g.* for drug development[50], materials discovery,[51] and beyond.

## 4. Materials and methods

### Dataset creation and curation

#### Internal data

Co-crystal data were collected from the Cambridge Structural Database[31], by searching for all the structures with no more than two different interacting organic molecules per asymmetric unit, using ConQuest 2021.2.0[52]. Hydrates, solvates, metal-organic systems, and duplicates were eliminated, leading to 9647 co-crystals ("positive pairs"). Non-co-crystal structures, *i.e.*, physical mixtures of two materials that do not form a co-crystal, were sourced from literature (1274 pairs, 88%) and in-house experiments (174 pairs, 12%), resulting in 1448 "negative" molecular pairs. All molecules were represented by canonical SMILES strings and entries longer than 80 characters were removed. Salts were deleted, the stereochemistry annotations were omitted, and molecular salts were converted into neutral molecules by uncharging the components, using RDKit (v. 2023.03.3). Molecules that only contributed to one class (*i.e.*, either exclusively partaking in ”positive” pairs or in ”negative” pairs) were removed. The resulting dataset contained 5240 co-crystal structures and 1392 non-co-crystal pairs, spanning different subclasses, *i.e.*, pharmaceutical, $\pi$–$\pi$, and energetic co-crystals.

#### External test set

The external test set was built to contain co-crystallization data of pharmaceutical co-crystal of anti-inflammatory, antitubercular, nootropic, and anti-depressant drugs from both the scientific literature[53–65] and in-house experimental co-crystallization screening. 134 co-crystalization data were collected in-house, via co-crystallization lab experiments (Supporting Tables S6, S5, S9, S8, S10, S7). Co-crystal formation was tested via mechanochemical (grinding and liquid-assisted grinding) and solution-based synthetic techniques (slurry and slow evaporation)[5,66,67], with at least three replicates and by changing the polarity of the solvent in syntheses involving its use. The co-crystal formation was investigated comparing the Fourier transform infrared (FT-IR) spectrum of the powder samples obtained with those of the starting materials, and confirmed by solid-state NMR. We report the FT-IR spectrum of the novel co-crystals in the Supporting Information (Supporting Figs. S2-S31). The resulting external test set collect 364 data of which 129 are co-crystal and 235 non-co-crystal.

## Model training and optimization

### Data preparation

For training and validating DeepCocrystal, and the re-trained benchmarks (*i.e.*, Fingerprint-DNN, and Descriptor-DNN) the internal dataset was split into training, validation, and test folds (80%, 10%, 10%, respectively) using stratified splitting. The splitting was repeated ten times with different random seeds.

### DeepCocrystal

DeepCocrystal was implemented in Tensorflow (v. 2.7.1). Various SMILES augmentation ratios were used for the training set (1:4 and 2:7 positive:negative augmentation ratios), in addition to canonical SMILES. The SMILES strings were label-encoded and padded to a length of 80 characters. Random search was used for hyperparameter optimization (Supporting Table S1), by: (a) running 3000 hyper-parameter combinations, (b) using early stopping on validation accuracy, with a patience of five epochs, and a tolerance of $10^{-5}$, and (c) selecting the combination with the best mean validation accuracy across the 10 dataset splits. The top-performing model is used to predict the test set.

### Benchmarks

CCGNet. The model was applied by using all the data and source code available in the original repository: https://github.com/Saoge123/CCGNet. After testing the code for reproducibility, it was applied, without modification, to predict the co-crystallization data of the external test set. The absence of duplicates, between the external test set of this work and the training set of CCGNet, was ensured. CCGNet associates each API-coformer pair with a score; positive scores are assigned to the co-crystal class, while negatives are to the non-co-crystals.

CC-Descriptors-ML. The model was re-implemented as described in the original paper[26]. Each of the models of this approach (*i.e.*, an ensemble of seven models developed using Support Vector Machine, XGBoost, Light Gradient-Boosting Machine, and Random Forest algorithms) was trained on 8, 10, or 14 selected features from 16 descriptors that show correlation with co-crystallization[26], including fingerprints, molecular radius, RDkit molecular descriptors and Hansen solubility parameters[68]. Training set preparation followed the same procedure as before. The external test set performance was reported considering instances where at least four models produced positive results for co-crystals and negative results for non-co-crystals, as suggested by the authors. Duplicate entries between our external test set and the dataset shared for this model were checked and removed from the test set (five molecule pairs).

Descriptor-DNN model. 0D, 1D, and 2D 'classical' molecular descriptors (in total 1056 descriptors per molecule) were calculated by Mordred[69]. The descriptors of API and coformers were concatenated, obtaining 2112 descriptors for each API-coformer pair. The features were standardized using standard scaling and passed through fully connected layers. We conducted hyperparameter optimization using the same strategy applied for DeepCocrystal (see Supporting Table S1).

Fingerprint-DNN model. The model was trained on extended connectivity fingerprints (radius=2 and nBits=1024), computed by RDKit (v. 2023.03.3)[70]; the fingerprints of the APIs and coformers were concatenated and passed through fully connected layers. We conducted hyperparameter optimization using the same strategy applied for DeepCocrystal (see Supporting Table S1).

### Classification Performance

Model performance was quantified via Recall ($Rec$), Specificity ($Sp$), and Balanced Accuracy ($BAcc$), computed as follows[37]:

$$Rec = \frac{TP}{TP + FN} \times 100, \tag{1}$$

$$Sp = \frac{TN}{TN + FP} \times 100, \tag{2}$$

$$BAcc = \frac{Sp + Rec}{2} \times 100. \tag{3}$$

$TP$, $TF$, $FP$, and $FN$ are the number of true positives (*i.e.*, correctly predicted co-crystals), true negatives (*i.e.*, correctly predicted non-co-crystals), false positives and false negatives, respectively. Recall (*Rec*), quantifies the ability to accurately predict co-crystals, while Specificity (*Sp*) captures the ability of a model to accurately predict negative API-conformer pairs. Balanced accuracy is a global measure of overall classification performance. For all metrics, the closer to 100%, the better the model performance[37].

## Experimental laboratory validation

### Diflunisal-purines experiments

Diflunisal (Apollo Scientific Ltd, 98%), adenine (TCI, 98%), caffeine (TCI, 98%), theobromine (Sigma-Aldrich, 98%), xanthine (Sigma-Aldrich, 99%), rosmarinic acid (Thermo Scientific, >97%), and riboflavin (Sigma-Aldrich, 99%) were used as received. Acetone, ethanol, ethyl acetate and dichloromethane were selected as solvents for the liquid-assisted grinding tests. Slurry experiments were conducted using 1 mL of ethanol, hexane or acetonitrile. *Diflunisal-caffeine* co-crystal in the form of a white microcrystalline powder was obtained by grinding 100 mg (0.4 mmol) of diflunisal and 78 mg (0.4 mmol) of caffeine for 20 minutes, after which drops of ethanol were added, continuing the liquid-assisted grinding for a further 20 minutes and repeating the adding of ethanol every 5 minutes. *Diflunisal-adenine* was obtained by slurry in ethanol: 100 mg of diflunisal (0.4 mmol) and 54 mg (0.4 mmol) of adenine were mixed in a 10 mL beaker, adding 1 mL of ethanol. The suspension was left under stirring for two days at room temperature. After solvent evaporation, the sample was obtained in the form of a microcrystalline white powder. The powder samples were characterized by FT-IR ATR (Supporting Information, Figs. S34 and S33) and solid-state NMR (Supporting Fig. S35 and S36).

### Fourier transform infrared spectroscopy

FT-IR spectra were reordered on an Equinox 55 (Bruker, Milan, Italy) spectrometer with an ATR reflectance attachment. Spectra were collected in the 400-3800 cm$^{-1}$ range with a resolution of 2 cm$^{-1}$ and 16 scans.

### 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, for $^{1}$H, $^{13}$C and $^{15}$N nuclei, respectively. *Diflunisal-caffeine* and *diflunisal-adenine* powdered sample was packed into cylindrical zirconia rotors with a 4 mm o.d. and a 90 mL volume. $^{13}$C and $^{15}$N CPMAS spectra were acquired using a ramp cross-polarisation pulse sequence with a 90° $^{1}$H pulse of 3.60 ms, a contact time of 3 ($^{13}$C) or 4 ($^{15}$N) ms, an optimized recycle delays, and a spinning speed of 12 kHz and 9 kHz, respectively. The $^{13}$C CPMAS spectra were registered for 140 scans for *diflunisal-caffeine* and 880 scans for *diflunisal-adenine*, while the $^{15}$N spectra for 13064 and 26400 scans, respectively. For every spectrum, a two-pulse phase modulation (TPPM) decoupling scheme was used, with a radiofrequency field of 69.4 kHz. The $^{13}$C and $^{15}$N chemical shift scales were calibrated through the signals of $\gamma$-glycine ($^{13}$C methylenic peak at 43.7 ppm and $^{15}$N peak at 33.4 ppm with reference to $NH_3$) as an external standard.

## Author Contributions

*Conceptualization*: FG, RÖ and RB; *Methodology*: RÖ, FG, RB; *Software*: RÖ, RB; *Validation*: RB; *Formal analysis and Investigation*: RB, RÖ, FG; *Resources*: MRC, FG, AA; *Data Curation*: RB; *Supervision*: FG, with support from MRC, RG; *Writing - Original Draft*: RB, RÖ, FG; *Writing - Review and Editing*: RB, RÖ, FG with contributions from RG and MRC. All authors approved the content of this manuscript before submission.

## Acknowledgements

FG and RÖ acknowledge the support from the Centre for Living Technologies (CLT) and from the Irène Curie Fellowship. This research was partially supported by the Nederlandse Organisatie voor Wetenschappelijk (NWO, grant no. EINF-7239 to FG). MRC, RG and RB acknowledge the support from Project CH4.0 under the MUR program "Dipartimenti di Eccellenza 2023–2027" (CUP: D13C22003520001) and Project "Predire" under the program NODES (CUP: D17G22000150001). The authors acknowledge Roberta Beccaria, Filippo Turchi, Natalia Velonia Bellonia, Elena Amadio, Chiara Sabena, Chiara Rosso and Rossella Zeolla for their lab experimental contributions to the external dataset. The authors also acknowledge Derek van Tilborg and Luke Rossen for technical suggestions.

## Code availability

The code to apply DeepCocrystal to any API-coformer pair, along with all publicly available data that took part in its training and validation will be available upon paper acceptance at the following URL: https://github.com/molML/deep-cocrystal.

## References

[1] N. K. Duggirala, M. L. Perry, Ö. Almarsson, and M. J. Zaworotko, "Pharmaceutical cocrystals: along the path to improved medicines," *Chemical communications*, vol. 52, no. 4, pp. 640–655, 2016.

[2] A. R. Thayyil, T. Juturu, S. Nayak, and S. Kamath, "Pharmaceutical co-crystallization: Regulatory aspects, design, characterization, and applications," *Advanced Pharmaceutical Bulletin*, vol. 10, no. 2, p. 203, 2020.

[3] G. R. Desiraju, "Supramolecular synthons in crystal engineering—a new organic synthesis," *Angewandte Chemie International Edition in English*, vol. 34, no. 21, pp. 2311–2327, 1995.

[4] J. B. Ngilirabanga and H. Samsodien, "Pharmaceutical co-crystal: An alternative strategy for enhanced physicochemical properties and drug synergy," *Nano Select*, vol. 2, no. 3, pp. 512–526, 2021.

[5] C. Cappuccino, D. Cusack, J. Flanagan, C. Harrison, C. Holohan, M. Lestari, G. Walsh, and M. Lusi, "How many cocrystals are we missing? assessing two crystal engineering approaches to pharmaceutical cocrystal

screening," *Crystal Growth & Design*, vol. 22, no. 2, pp. 1390–1397, 2022.

[6] N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, and A. Walsh, "Best practices in machine learning for chemistry," *Nature chemistry*, vol. 13, no. 6, pp. 505–508, 2021.

[7] N. Sarkar, N. C. Gonnella, M. Krawiec, D. Xin, and C. B. Aakeröy, "Evaluating the predictive abilities of protocols based on hydrogen-bond propensity, molecular complementarity, and hydrogen-bond energy for cocrystal screening," *Crystal Growth & Design*, vol. 20, no. 11, pp. 7320–7327, 2020.

[8] F. Molajafari, T. Li, M. Abbasichaleshtori, M. H. ZD, A. F. Cozzolino, D. R. Fandrick, and J. D. Howe, "Computational screening for prediction of co-crystals: method comparison and experimental validation," *CrystEngComm*, 2024.

[9] D. Wang, Z. Yang, B. Zhu, X. Mei, and X. Luo, "Machine-learning-guided cocrystal prediction based on large data base," *Crystal Growth & Design*, vol. 20, no. 10, pp. 6610–6621, 2020.

[10] D. Yang, L. Wang, P. Yuan, Q. An, B. Su, M. Yu, T. Chen, K. Hu, L. Zhang, Y. Lu, *et al.*, "Cocrystal virtual screening based on the xgboost machine learning model," *Chinese Chemical Letters*, vol. 34, no. 8, p. 107964, 2023.

[11] Y. Kang, J. Chen, X. Hu, Y. Jiang, and Z. Li, "A cocrystal prediction method of graph neural networks based on molecular spatial information and global attention," *CrystEngComm*, vol. 25, no. 46, pp. 6405–6415, 2023.

[12] C. von Essen and D. Luedeker, "In silico co-crystal design: assessment of the latest advances," *Drug Discovery Today*, p. 103763, 2023.

[13] T. Heng, D. Yang, R. Wang, L. Zhang, Y. Lu, and G. Du, "Progress in research on artificial intelligence applied to polymorphism and cocrystal prediction," *ACS omega*, vol. 6, no. 24, pp. 15543–15550, 2021.

[14] P. J. Cragg and P. J. Cragg, *An introduction to supramolecular chemistry*. Springer, 2010.

[15] J.-M. Lehn, "Supramolecular chemistry—scope and perspectives molecules, supermolecules, and molecular devices (nobel lecture)," *Angewandte Chemie International Edition in English*, vol. 27, no. 1, pp. 89–112, 1988.

[16] C. P. Brock and J. D. Dunitz, "Towards a grammar of crystal packing," *Chemistry of materials*, vol. 6, no. 8, pp. 1118–1127, 1994.

[17] M. Hirohara, Y. Saito, Y. Koda, K. Sato, and Y. Sakakibara, "Convolutional neural network based on smiles representation of compounds for detecting chemical motif," *BMC bioinformatics*, vol. 19, pp. 83–94, 2018.

[18] T. B. Kimber, S. Engelke, I. V. Tetko, E. Bruno, and G. Godin, "Synergy effect between convolutional neural networks and the multiplicity of smiles for improvement of molecular prediction," *arXiv preprint arXiv:1812.04439*, 2018.

[19] D. van Tilborg, A. Alenicheva, and F. Grisoni, "Exposing the limitations of molecular machine learning with activity cliffs," *Journal of chemical information and modeling*, vol. 62, no. 23, pp. 5938–5951, 2022.

[20] H. Öztürk, A. Özgür, P. Schwaller, T. Laino, and E. Ozkirimli, "Exploring chemical space using natural language processing methodologies for drug discovery," *Drug Discovery Today*, vol. 25, no. 4, pp. 689–705, 2020.

[21] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[22] M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle, K. M. Jablonka, *et al.*, "Selfies and the future of molecular string representations," *Patterns*, vol. 3, no. 10, 2022.

[23] E. J. Bjerrum, "Smiles enumeration as data augmentation for neural network modeling of molecules," *arXiv preprint arXiv:1703.07076*, 2017.

[24] M. E. Mswahili, M.-J. Lee, G. L. Martin, J. Kim, P. Kim, G. J. Choi, and Y.-S. Jeong, "Cocrystal prediction using machine learning models and descriptors," *Applied Sciences*, vol. 11, no. 3, p. 1323, 2021.

[25] J.-J. Devogelaer, H. Meekes, P. Tinnemans, E. Vlieg, and R. De Gelder, "Co-crystal prediction by artificial neural networks," *Angewandte Chemie International Edition*, vol. 59, no. 48, pp. 21711–21718, 2020.

[26] X. Liang, S. Liu, Z. Li, Y. Deng, Y. Jiang, and H. Yang, "Efficient cocrystal coformer screening based on a machine learning strategy: A case study for the preparation of imatinib cocrystal with enhanced physicochemical properties," *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 196, p. 114201, 2024.

[27] Y. Jiang, Z. Yang, J. Guo, H. Li, Y. Liu, Y. Guo, M. Li, and X. Pu, "Coupling complementary strategy to flexible graph neural network for quick discovery of coformer in diverse co-crystal materials," *Nature Communications*, vol. 12, no. 1, p. 5950, 2021.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[29] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of cnn and rnn for natural language processing," *arXiv preprint arXiv:1702.01923*, 2017.

[30] N. Schneider, R. A. Sayle, and G. A. Landrum, "Get your atoms in order – an open-source implementation of a novel and robust molecular canonicalization algorithm," *Journal of chemical information and modeling*, vol. 55, no. 10, pp. 2111–2120, 2015.

[31] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, "The cambridge structural database," *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, vol. 72, no. 2, pp. 171–179, 2016.

[32] T. Shen, "Chemical and pharmacological properties of diflunisal," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 3, no. 2P2, pp. 3S–8S, 1983.

[33] C. B. Aakeröy, A. B. Grommet, and J. Desper, "Cocrystal screening of diclofenac," *Pharmaceutics*, vol. 3, no. 3, pp. 601–614, 2011.

[34] T. Grecu, C. A. Hunter, E. J. Gardiner, and J. F. McCabe, "Validation of a computational cocrystal prediction tool: comparison of virtual and experimental cocrystal screening results," *Crystal growth & design*, vol. 14, no. 1, pp. 165–171, 2014.

[35] T. Grecu, H. Adams, C. A. Hunter, J. F. McCabe, A. Portell, and R. Prohens, "Virtual screening identifies new cocrystals of nalidixic acid," *Crystal growth & design*, vol. 14, no. 4, pp. 1749–1755, 2014.

[36] L. Roca-Paixão, N. T. Correia, and F. Affouard, "Affinity prediction computations and mechanosynthesis of carbamazepine based cocrystals," *CrystEngComm*, vol. 21, no. 45, pp. 6991–7001, 2019.

[37] D. Ballabio, F. Grisoni, and R. Todeschini, "Multivariate comparison of classification performance measures," *Chemometrics and Intelligent Laboratory Systems*, vol. 174, pp. 33–44, 2018.

[38] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[39] J. Chen, Z. Li, Y. Kang, and Z. Li, "Cocrystal prediction based on deep forest model—a case study of febuxostat," *Crystals*, vol. 14, no. 4, p. 313, 2024.

[40] T. Shen, "Chemical and pharmacological properties of diflunisal," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 3, no. 2P2, pp. 3S–8S, 1983.

[41] P. Snetkov, S. Morozkina, R. Olekhnovich, and M. Uspenskaya, "Diflunisal targeted delivery systems: A review," *Materials*, vol. 14, no. 21, p. 6687, 2021.

[42] E. Martínez-Pinilla, A. Oñatibia-Astibia, and R. Franco, "The relevance of theobromine for the beneficial effects of cocoa consumption," *Frontiers in pharmacology*, vol. 6, p. 126866, 2015.

[43] N. Yahfoufi, N. Alsadi, M. Jambi, and C. Matar, "The immunomodulatory and anti-inflammatory role of polyphenols," *Nutrients*, vol. 10, no. 11, p. 1618, 2018.

[44] C. Luo, L. Zou, H. Sun, J. Peng, C. Gao, L. Bao, R. Ji, Y. Jin, and S. Sun, "A review of the anti-inflammatory effects of rosmarinic acid on inflammatory diseases," *Frontiers in pharmacology*, vol. 11, p. 153, 2020.

[45] K. Rodak, I. Kokot, and E. M. Kratz, "Caffeine as a factor influencing the functioning of the human body—friend or foe?," *Nutrients*, vol. 13, no. 9, p. 3088, 2021.

[46] M. Guo, X. Sun, J. Chen, and T. Cai, "Pharmaceutical cocrystals: A review of preparations, physicochemical properties and applications," *Acta Pharmaceutica Sinica B*, vol. 11, no. 8, pp. 2537–2564, 2021.

[47] S. Bordignon, P. Cerreia Vioglio, E. Priola, D. Voinovich, R. Gobetto, Y. Nishiyama, and M. R. Chierotti, "Engineering codrug solid forms: mechanochemical synthesis of an indomethacin–caffeine system," *Crystal Growth & Design*, vol. 17, no. 11, pp. 5744–5752, 2017.

[48] G. S. Kumar, P. Seethalakshmi, N. Bhuvanesh, and S. Kumaresan, "Studies on the syntheses, structural characterization, antimicrobial-, and dpph radical scavenging activity of the cocrystals caffeine: cinnamic acid and caffeine: eosin dihydrate," *Journal of Molecular Structure*, vol. 1050, pp. 88–96, 2013.

[49] N. R. Goud, S. Gangavaram, K. Suresh, S. Pal, S. G. Manjunatha, S. Nambiar, and A. Nangia, "Novel furosemide cocrystals and selection of high solubility drug forms," *Journal of pharmaceutical sciences*, vol. 101, no. 2, pp. 664–680, 2012.

[50] K. Kawakami, M. Ebara, H. Izawa, N. M Sanchez-Ballester, J. P Hill, and K. Ariga, "Supramolecular approaches for drug development," *Current medicinal chemistry*, vol. 19, no. 15, pp. 2388–2398, 2012.

[51] S. I. Stupp and L. C. Palmer, "Supramolecular chemistry and self-assembly in organic materials design," *Chemistry of Materials*, vol. 26, no. 1, pp. 507–518, 2014.

[52] I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson, and R. Taylor, "New software for searching the cambridge structural database and visualizing crystal structures," *Acta Crystallographica Section B: Structural Science*, vol. 58, no. 3, pp. 389–397, 2002.

[53] X. Buol, K. Robeyns, C. Caro Garrido, N. Tumanov, L. Collard, J. Wouters, and T. Leyssens, "Improving nefiracetam dissolution and solubility behavior using a cocrystallization approach," *Pharmaceutics*, vol. 12, no. 7, p. 653, 2020.

[54] S. G. Dash and T. S. Thakur, "Computational screening of multicomponent solid forms of 2-aryl-propionate class of nsaid, zaltoprofen, and their experimental validation," *Crystal Growth & Design*, vol. 21, no. 1, pp. 449–461, 2020.

[55] X. Ma, X. Chen, Y. Zhen, X. Zheng, C. Shi, S. Jin, B. Liu, B. Chen, and D. Wang, "Molecular structures of eight hydrogen bond-mediated minoxidil adducts from different aryl acids," *Journal of Molecular Structure*, vol. 1298, p. 136942, 2024.

[56] A. G. O. Pontes, L. M. T. Vidal, Y. S. de Oliveira, B. P. Bezerra, S. B. H. Girão, and A. P. Ayala, "Exploring the formation and diversity of secnidazole cocrystals," *Journal of Molecular Structure*, vol. 1311, p. 138374, 2024.

[57] J. Ouyang, L. Liu, Y. Li, M. Chen, L. Zhou, Z. Liu, L. Xu, and H. Shehzad, "Cocrystals of carbamazepine: Structure, mechanical properties, fluorescence properties, solubility, and dissolution rate," *Particuology*, vol. 90, pp. 20–30, 2024.

[58] W.-J. Ji, J.-Y. Jiang, M. Hong, B. Zhu, G.-B. Ren, and M.-H. Qi, "Colorful prothionamide salt forms with enhancement in water solubility and dissolution behavior," *Crystal Growth & Design*, vol. 23, no. 8, pp. 5770–5784, 2023.

[59] F. Rossi, P. Cerreia Vioglio, S. Bordignon, V. Giorgio, C. Nervi, E. Priola, R. Gobetto, K. Yazawa, and M. R. Chierotti, "Unraveling the hydrogen bond network in a theophylline–pyridoxine salt cocrystal by a combined x-ray diffraction, solid-state nmr, and computational approach," *Crystal Growth & Design*, vol. 18, no. 4, pp. 2225–2233, 2018.

[60] S. Bordignon, P. Cerreia Vioglio, E. Amadio, F. Rossi, E. Priola, D. Voinovich, R. Gobetto, and M. R. Chierotti, "Molecular crystal forms of antitubercular ethionamide with dicarboxylic acids: solid-state properties and a combined structural and spectroscopic study," *Pharmaceutics*, vol. 12, no. 9, p. 818, 2020.

[61] F. Liu, L.-Y. Wang, M.-C. Yu, Y.-T. Li, Z.-Y. Wu, and C.-W. Yan, "A new cocrystal of isoniazid-quercetin with hepatoprotective effect: The design, structure, and in vitro/in vivo performance evaluation," *European Journal of Pharmaceutical Sciences*, vol. 144, p. 105216, 2020.

[62] S. Bordignon, P. Cerreia Vioglio, C. Bertoncini, E. Priola, R. Gobetto, and M. R. Chierotti, "Pseudopolymorphism driven by stoichiometry and hydrated/anhydrous reagents: The riveting case of methyl gallate· l-proline," *Crystal Growth & Design*, vol. 21, no. 12, pp. 6776–6785, 2021.

[63] I. D'Abbrunzo, E. Bianco, L. Gigli, N. Demitri, R. Birolo, M. R. Chierotti, I. Škorić, J. Keiser, C. Häberli, D. Voinovich, *et al.*, "Praziquantel meets niclosamide: A dual-drug antiparasitic cocrystal," *International journal of pharmaceutics*, vol. 644, p. 123315, 2023.

[64] I. D'Abbrunzo, R. Birolo, M. R. Chierotti, D.-K. Bučar, D. Voinovich, B. Perissutti, and D. Hasa, "Enantiospecific crystallisation behaviour of malic acid in mechanochemical reactions with vinpocetine," *European Journal of Pharmaceutics and Biopharmaceutics*, p. 114344, 2024.

[65] R. Birolo, F. Bravetti, E. Alladio, E. Priola, G. Bianchini, R. Novelli, A. Aramini, R. Gobetto, and M. R. Chierotti, "Speeding up the cocrystallization process: Machine learning-combined methods for the prediction of multicomponent systems," *Crystal Growth & Design*, vol. 23, no. 11, pp. 7898–7911, 2023.

[66] D. R. Weyna, T. Shattock, P. Vishweshwar, and M. J. Zaworotko, "Synthesis and structural characterization of cocrystals and pharmaceutical cocrystals: mechanochemistry vs slow evaporation from solution," *Crystal Growth and Design*, vol. 9, no. 2, pp. 1106–1123, 2009.

[67] M. D. Charpentier, J.-J. Devogelaer, A. Tijink, H. Meekes, P. Tinnemans, E. Vlieg, R. de Gelder, K. Johnston, and J. H. Ter Horst, "Comparing and quantifying the efficiency of cocrystal screening methods for praziquantel," *Crystal Growth & Design*, vol. 22, no. 9, pp. 5511–5525, 2022.

[68] M. A. Mohammad, A. Alhalaweh, and S. P. Velaga, "Hansen solubility parameter as a tool to predict cocrystal formation," *International journal of pharmaceutics*, vol. 407, no. 1-2, pp. 63–71, 2011.

[69] H. Moriwaki, Y.-S. Tian, N. Kawashita, and T. Takagi, "Mordred: a molecular descriptor calculator," *Journal of cheminformatics*, vol. 10, pp. 1–14, 2018.

[70] G. Landrum, "Rdkit documentation," *Release*, vol. 1, no. 1-79, p. 4, 2013.

[71] R. Birolo, F. Bravetti, S. Bordignon, I. D'Abbrunzo, P. P. Mazzeo, B. Perissutti, A. Bacchi, M. R. Chierotti, and R. Gobetto, "Overcoming the drawbacks of sulpiride by means of new crystal forms," *Pharmaceutics*, vol. 14, no. 9, p. 1754, 2022.

# Supporting Information

## Models hyper-parameters

*Table S1. Hyper-parameters optimization.* The best hyper-parameters found for each model implemented in this work are reported.

| Hyper-parameter | Values | DeepCocrystal canonical | DeepCocrystal (1:4) | DeepCocrystal (2:7) | Fingerprint DNN | Descriptor DNN |
|---|---|---|---|---|---|---|
| No. convolutions | 1, 2 , 3, 4 | 1 | 1 | 2 | | |
| No. filters | 16, 32, 64, 128, 256 | 256 | 256 | 256 | | |
| Kernel sizes | 3, 4, 5, 7 | 4 | 3 | 4 | | |
| Convol. activation | relu, selu | relu | selu | selu | | |
| Number dense | 1, 2, 3, 4, 5 | 5 | 4 | 3 | 1 | 1 |
| Dense layer size | 256, 512, 1024 | 512 | 1024 | 1024 | 1024 | 512 |
| Dense activation | relu, selu | relu | relu | relu | relu | relu |
| Embedding dim. | 32, 64, 128, 256 | 256 | 64 | 32 | | |
| Dropout rate | 0.0, 0.1, 0.25 | 0.0 | 0.25 | 0.1 | 0.1 | 0.1 |
| Optimizer | adam, rmsprop | adam | adam | adam | rmsprop | adam |
| Learning rate | 5e-2, 1e-3, 5e-3, 5e-4 | 5e-2 | 5e-2 | 5e-2 | 5e-2 | 5e-2 |
| Batch size | 64, 256, 512, 1024 | 512 | 64 | 512 | 64 | 64 |

## Internal test

*Table S2. Dataset sampling:* training set = 80% (5306 data), validation set = 10% (663 data), test set = 10% (664 data). After the stritified splitting, the data in training and in validation sets are augmentated (1:4 ratio – 1 times positive data and 4 times negatives; 2:7 ratio – 2 times positive data and 7 times negatives) by SMILES augmentation, duplicates are checked and removed. Model performances are evaluated by predicting the co-crystallization of pairs in the 10 test sets using canonical SMILES as molecule inputs.

| | training set | | | validation set | | | test set | | |
|---|---|---|---|---|---|---|---|---|---|
| setup | YES | NO | total | YES | NO | total | YES | NO | total |
| 0 (1:4) | 8382 | 8887 | 17269 | 1048 | 1110 | 2158 | 525 | 139 | 664 |
| 1 (1:4) | 8380 | 8881 | 17261 | 1047 | 1106 | 2153 | 525 | 139 | 664 |
| 2 (1:4) | 8377 | 8889 | 17266 | 1048 | 1109 | 2157 | 525 | 139 | 664 |
| 3 (1:4) | 8382 | 8887 | 17269 | 1044 | 1108 | 2152 | 525 | 139 | 664 |
| 4 (1:4) | 8380 | 8882 | 17262 | 1048 | 1110 | 2158 | 525 | 139 | 664 |
| 5 (1:4) | 8380 | 8887 | 17267 | 1048 | 1109 | 2157 | 525 | 139 | 664 |
| 6 (1:4) | 8382 | 8882 | 17264 | 1048 | 1112 | 2160 | 525 | 139 | 664 |
| 7 (1:4) | 8379 | 8879 | 17258 | 1048 | 1108 | 2156 | 525 | 139 | 664 |
| 8 (1:4) | 8381 | 8890 | 17271 | 1048 | 1108 | 2156 | 525 | 139 | 664 |
| 9 (1:4) | 8378 | 8885 | 17263 | 1048 | 1109 | 2157 | 525 | 139 | 664 |
| 0 (2:7) | 16743 | 15508 | 32251 | 2092 | 1937 | 4029 | 525 | 139 | 664 |
| 1 (2:7) | 16795 | 15425 | 32220 | 2097 | 1935 | 4032 | 525 | 139 | 664 |
| 2 (2:7) | 16737 | 15491 | 32228 | 2093 | 1939 | 4032 | 525 | 139 | 664 |
| 3 (2:7) | 16748 | 15494 | 32242 | 2085 | 1941 | 4026 | 525 | 139 | 664 |
| 4 (2:7) | 16732 | 15510 | 32242 | 2097 | 1930 | 4027 | 525 | 139 | 664 |
| 5 (2:7) | 16731 | 15500 | 32231 | 2093 | 1934 | 4027 | 525 | 139 | 664 |
| 6 (2:7) | 16736 | 15496 | 32232 | 2090 | 1935 | 4025 | 525 | 139 | 664 |
| 7 (2:7) | 16736 | 15506 | 32242 | 2093 | 1942 | 4035 | 525 | 139 | 664 |
| 8 (2:7) | 16743 | 15489 | 32232 | 2093 | 1931 | 4024 | 525 | 139 | 664 |
| 9 (2:7) | 16728 | 15495 | 32223 | 2091 | 1928 | 4019 | 525 | 139 | 664 |

*Table S3. DeepCocrystal performance on internal test per setups.* Balanced accuracy (BAcc), Recall and Specificity are computed and reported for DeepCocrystal - canonical and DeepCocrystal - augmentated (1:4 and 2:7 configurations).

| | **canonical** | | | **1:4 augmentation** | | | **2:7 augmentation** | | |
|---|---|---|---|---|---|---|---|---|---|
| **setup** | **BAcc** | **Recall** | **Spec.** | **BAcc** | **Recall** | **Spec.** | **BAcc** | **Recall** | **Spec.** |
| 0 | 0.88 | 0.96 | 0.79 | 0.89 | 0.91 | 0.86 | 0.87 | 0.91 | 0.83 |
| 1 | 0.88 | 0.97 | 0.79 | 0.86 | 0.96 | 0.76 | 0.88 | 0.90 | 0.87 |
| 2 | 0.85 | 0.99 | 0.71 | 0.89 | 0.91 | 0.87 | 0.92 | 0.93 | 0.91 |
| 3 | 0.89 | 0.97 | 0.81 | 0.90 | 0.87 | 0.94 | 0.88 | 0.96 | 0.80 |
| 4 | 0.88 | 0.94 | 0.81 | 0.88 | 0.91 | 0.85 | 0.91 | 0.93 | 0.88 |
| 5 | 0.83 | 0.98 | 0.68 | 0.87 | 0.92 | 0.81 | 0.90 | 0.91 | 0.89 |
| 6 | 0.88 | 0.96 | 0.81 | 0.87 | 0.88 | 0.86 | 0.90 | 0.94 | 0.86 |
| 7 | 0.90 | 0.95 | 0.85 | 0.89 | 0.92 | 0.86 | 0.90 | 0.90 | 0.89 |
| 8 | 0.88 | 0.97 | 0.80 | 0.88 | 0.89 | 0.88 | 0.86 | 0.88 | 0.85 |
| 9 | 0.92 | 0.95 | 0.88 | 0.90 | 0.90 | 0.91 | 0.91 | 0.93 | 0.88 |
| **average** | 0.88 | 0.96 | 0.79 | 0.88 | 0.91 | 0.86 | **0.89** | **0.92** | **0.87** |
| **STD** | 0.02 | 0.01 | 0.06 | 0.01 | 0.02 | 0.05 | 0.02 | 0.02 | 0.03 |

*Table S4. Performance comparison on internal test.* The developed model trained on extended connectivity fingerprints and descriptors metric were compared to similar models reported in literature.

| | **This work** | | | | **Literature** |
|---|---|---|---|---|---|
| **Model** | **Accuracy** | **BAcc** | **Recall** | **Specificity** | **Accuracy** |
| Fingerprint-DNN | 0.94 ± 0.01 | 0.91 ± 0.01 | 0.96 ± 0.01 | 0.85 ± 0.02 | 0.97 ± 0.01 [25] |
| Descriptors-DNN | 0.93 ± 0.01 | 0.89 ± 0.01 | 0.96 ± 0.01 | 0.83 ± 0.03 | 0.83 [24] |

## External test

In-house experimental screening are performed for prothionamide, pyrazinamide, p-aminosalicylic acid, sulpiride, lamotrigine, flurbiprofen, ketoprofen, naproxen and gentisic acid. The external test set includes also 230 published co-crystallization data of structurally similar molecules. The not pubblished data are reported in the following tables.

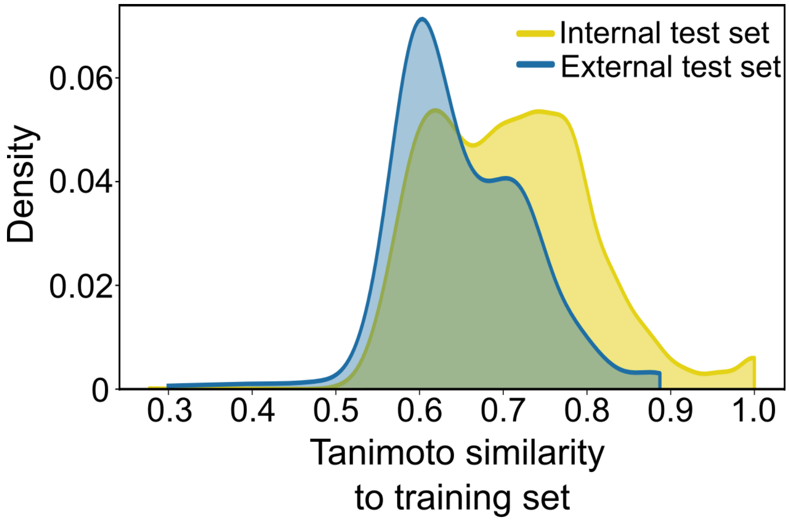

*Figure S1. Similarity of the test sets to the training set.* Maximum Tanimoto similarity of test set compounds to the training set is computed using fingerprints (nBits=1024, radius=2). The distribution across test set samples is visualized. The distributions show that the curated external test set contains structurally more novel compounds than the internal test set and presents a more challenging evaluation setting.

*Table S5. Pyrazinamide* experimental co-crystal screening results (× = no-co-crystal; ✓= co-crystal).

| **Coformer** | **Grinding** | **LAG** | **Slurry** | **Slow evap.** | **EXP** | **IR spectrum** |
|---|---|---|---|---|---|---|
| **2,5-dihydroxyterephtalic acid** | × | ✓ | × | × | 1 | Figure S2 |
| **pyridine-2,6-dicarboxylic acid** | × | × | ✓ | × | 1 | Figure S3 |
| **trimesic acid** | × | ✓ | × | ✓ | 1 | Figure S4 |
| phtalic acid | × | × | × | × | 0 | |
| **mandelic acid** | × | × | ✓ | × | 1 | Figure S5 |
| caffeine | × | × | × | × | 0 | |
| prothionamide | × | × | × | × | 0 | |

*Table S6. Prothionamide* experimental co-crystal screening results (× = no-co-crystal; ✓= co-crystal).

| **Coformer** | **Grinding** | **LAG** | **Slurry** | **Slow evap.** | **EXP** | **IR spectrum** |
|---|---|---|---|---|---|---|
| adipic acid | × | × | × | × | 0 | |
| **tartaric acid** | × | × | ✓ | ✓ | 1 | Figure S6 |
| ibuprofen | × | × | × | × | 0 | |
| indomethacin | × | × | × | × | 0 | |
| glutamic acid | × | × | × | × | 0 | |
| isoniazide | × | × | × | × | 0 | |
| proline | × | × | × | × | 0 | |
| 4-acetamidobenzoic acid | × | × | × | × | 0 | |
| **ketoglutaric acid** | × | × | ✓ | × | 1 | Figure S7 |
| nicotinic acid | × | × | × | × | 0 | |
| hydroquinone | × | × | × | × | 0 | |
| caffeic acid | × | × | × | × | 0 | |
| benzoic acid | × | × | × | × | 0 | |
| GABA | × | × | × | × | 0 | |
| **trimesic acid** | × | × | ✓ | ✓ | 1 | Figure S8 |
| **pyridine-2,6-dicarboxylic acid** | × | × | ✓ | ✓ | 1 | Figure S9 |
| vanillic acid | × | × | × | × | 0 | |
| **phthalic acid** | × | × | ✓ | ✓ | 1 | Figure S10 |
| mandelic acid | × | × | × | × | 0 | |
| **diflunisal** | × | × | ✓ | × | 1 | Figure S11 |
| nalidixic acid | × | × | × | × | 0 | |
| **2,3-pyrazinedicarboxylic acid** | × | × | ✓ | × | 1 | Figure S12 |
| pyridine-2,3-dicarboxylic acid | × | × | × | × | 0 | |
| naproxen | × | × | × | × | 0 | |
| gentisic acid | × | × | × | × | 0 | |
| caffeine | × | × | × | × | 0 | |
| theophylline | × | × | × | × | 0 | |
| pyrogallol | × | × | × | × | 0 | |

*Table S7. Lamotrigine* experimental co-crystal screening results (× = no-co-crystal; ✓= co-crystal).

| **Coformer** | **Grinding** | **LAG** | **Slurry** | **Slow evap.** | **EXP** | **IR spectrum** |
|---|---|---|---|---|---|---|
| **2,6-dihydroxybenzoic acid** | × | ✓ | × | ✓ | 1 | Figure S13 |
| **2,6-dimethylbenzoic acid** | × | ✓ | × | × | 1 | Figure S14 |
| **2,6-(trifluoromethyl)benzoic acid** | × | ✓ | × | × | 1 | Figure S15 |
| **salicylic acid** | × | ✓ | × | × | 1 | Figure S16 |
| **benzoic acid** | × | ✓ | × | × | 1 | Figure S17 |

*Table S8. p-aminosalicylic acid* experimental co-crystal screening results (× = no-co-crystal; ✓= co-crystal).

| Coformer | Grinding | LAG | Slurry | Slow evap. | EXP | IR spectrum |
|---|---|---|---|---|---|---|
| adipic acid | × | × | × | × | 0 | |
| ascorbic acid | × | × | × | × | 0 | |
| caffeic acid | × | × | × | × | 0 | |
| citric acid | × | × | × | × | 0 | |
| nicotinic acid | × | × | × | × | 0 | |
| quercetin | × | × | × | × | 0 | |
| pimelic acid | × | × | × | × | 0 | |
| hippuric acid | × | × | × | × | 0 | |
| 4-aminobenzoic acid | × | × | × | × | 0 | |
| **lysine** | ✓ | × | ✓ | × | 1 | Figure S18 |
| lactic acid | × | × | × | × | 0 | |
| aspartic acid | × | × | × | × | 0 | |
| glycolic acid | × | × | × | × | 0 | |
| thymine | × | × | × | × | 0 | |
| **adenine** | × | ✓ | ✓ | ✓ | 1 | Figure S19 |
| histidine | × | × | × | × | 0 | |
| glutamine | × | × | × | × | 0 | |
| acetylsalicylic acid | × | × | × | × | 0 | |
| ketoglutaric acid | ✓ | × | × | × | 1 | Figure S20 |
| theobromine | × | × | × | × | 0 | |
| phenylalanine | × | × | × | × | 0 | |
| N-acetylcysteine | × | × | × | × | 0 | |
| prothionamide | × | × | × | × | 0 | |
| 5-aminouracil | × | × | × | × | 0 | |
| riboflavin | × | × | × | × | 0 | |
| rosmarinic acid | × | × | × | × | 0 | |
| ethionamide | × | × | × | × | 0 | |
| malonamide | × | × | × | × | 0 | |
| maleamic acid | × | × | × | × | 0 | |
| barbituric acid | × | × | × | × | 0 | |
| **2-thiobarbituric acid** | × | ✓ | × | ✓ | 1 | Figure S21 |
| carbocysteine | × | × | × | × | 0 | |
| **oxalic acid** | × | × | ✓ | ✓ | 1 | Figure S22 |
| **tartaric acid** | × | × | × | ✓ | 1 | Figure S23 |
| **malonic acid** | × | ✓ | × | ✓ | 1 | Figure S24 |
| **maleic acid** | × | ✓ | × | × | 1 | Figure S25 |
| fumaric acid | × | × | × | × | 0 | |
| malic acid | × | × | × | × | 0 | |
| **glutaric acid** | × | × | ✓ | ✓ | 1 | Figure S26 |
| succinic acid | × | × | × | × | 0 | |

*Table S9. Non-steroidal ant-inflammatory drugs and gentisic acid* experimental co-crystal screening results (× = no-co-crystal; ✓= co-crystal).

| Coformer | Grinding | LAG | Slurry | Slow evap. | EXP | IR spectrum |
|---|---|---|---|---|---|---|
| **gentisic acid - adenine** | | ✓ | ✓ | | 1 | Figure S27 |
| **gentisic acid - guanine** | ✓ | ✓ | ✓ | | 1 | Figure S28 |
| **KET - tyramine** | × | ✓ | × | ✓ | 1 | Figure S29 |
| **FLU - tyramine** | × | ✓ | × | ✓ | 1 | Figure S30 |
| **NAP - tyramine** | × | ✓ | × | ✓ | 1 | Figure S31 |

*Table S10. Sulpiride* experimental co-crystal screening results (× = no-co-crystal; ✓= co-crystal). The co-crystal characterization is reported in a separate pubblication[71] .

| Coformer | Grinding | LAG | Slurry | Slow evap. | EXP |
|---|---|---|---|---|---|
| **adipic acid** | × | × | ✓ | ✓ | 1 |
| **4-aminobenzoic acid** | × | × | ✓ | × | 1 |
| **caffeic acid** | × | × | ✓ | × | 1 |
| **fumaric acid** | × | × | ✓ | ✓ | 1 |
| **maleic acid** | × | × | ✓ | × | 1 |
| **malic acid** | × | × | ✓ | ✓ | 1 |
| **malonic acid** | × | × | ✓ | × | 1 |
| **nicotinic acid** | × | × | ✓ | ✓ | 1 |
| **succinic acid** | × | × | ✓ | × | 1 |
| **acetazolamide** | × | × | ✓ | ✓ | 1 |
| **ibuprofen** | × | × | ✓ | × | 1 |
| **indomethacin** | × | × | ✓ | ✓ | 1 |
| quercetin | × | × | × | × | 0 |
| hippuric acid | × | × | × | × | 0 |
| lactose | × | × | × | × | 0 |
| d-mannitol | × | × | × | × | 0 |
| caffeine | × | × | × | × | 0 |
| cytosine | × | × | × | × | 0 |
| thymine | × | × | × | × | 0 |
| trimesic acid | × | × | × | × | 0 |
| piracetam | × | × | × | × | 0 |
| N-acetylcysteine | × | × | × | × | 0 |
| ketoglutaric acid | × | × | × | × | 0 |
| tyrosine | × | × | × | × | 0 |
| GABA | × | × | × | × | 0 |
| proline | × | × | × | × | 0 |
| lysine | × | × | × | × | 0 |
| theophylline | × | × | × | × | 0 |
| ascorbic acid | × | × | × | × | 0 |
| n-propyl gallate | × | × | × | × | 0 |
| glutamic acid | × | × | × | × | 0 |
| L-phenylalanine | × | × | × | × | 0 |
| melatonine | × | × | × | × | 0 |

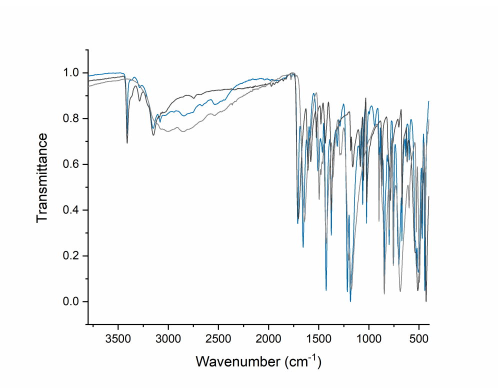

*Figure S2. Pyrazinamide - 2,5-dihydroxyterephthalic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = 2,5-dihydroxyterephthalic acid, black = pyrazinamide).

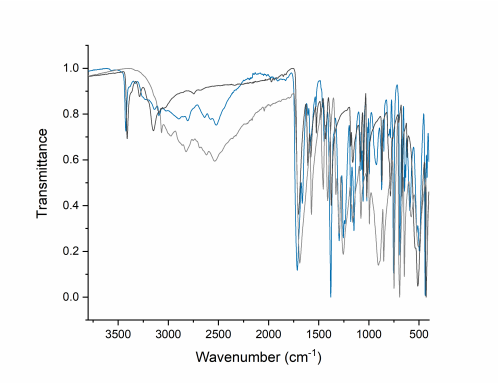

*Figure S3. Pyrazinamide - pyridine-2,6-dicarboxylic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = pyridine-2,6-dicarboxylic acid, black = pyrazinamide).

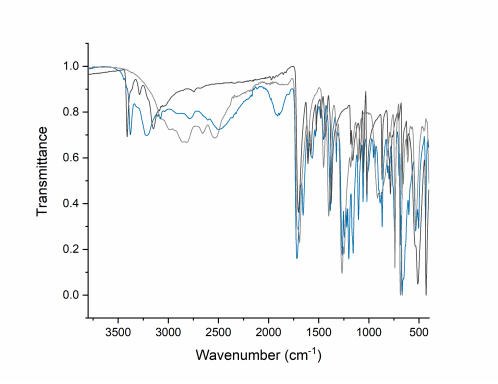

*Figure S4. Pyrazinamide - trimesic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = trimesic acid, black = pyrazinamide).

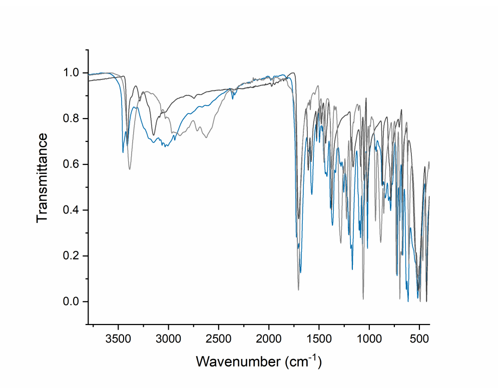

*Figure S5. Pyrazinamide - mandelic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = mandelic acid, black = pyrazinamide).

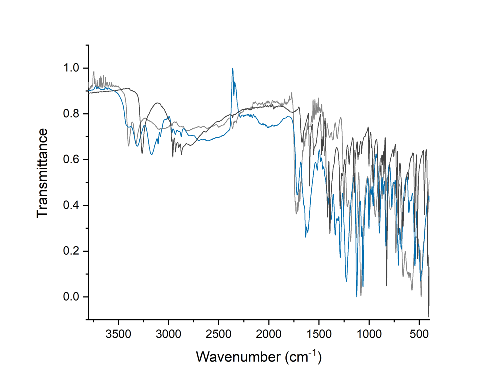

*Figure S6. Prothionamide - tartaric acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = tartaric acid, black = prothionamide).

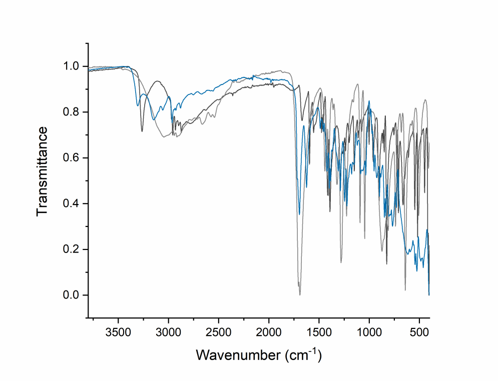

*Figure S7. Prothionamide - ketoglutaric acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = ketoglutaric acid, black = prothionamide).

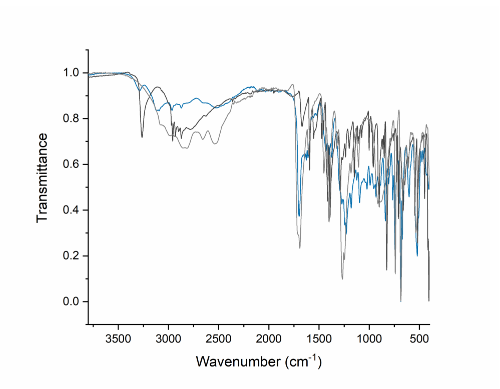

*Figure S8. Prothionamide - trimesic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = trimesic acid, black = prothionamide).

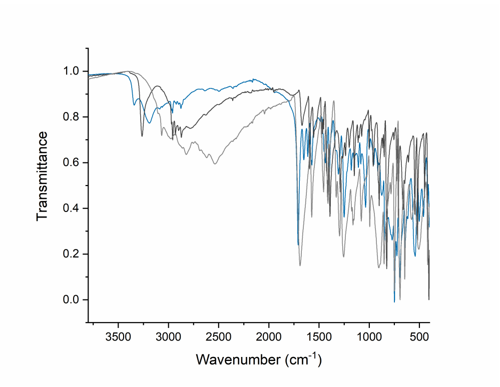

*Figure S9. Prothionamide - pyridine-2,6-dicarboxylic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = pyridine-2,6-dicarboxylic acid, black = prothionamide).

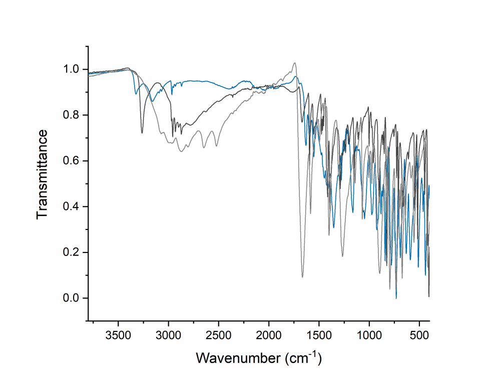

*Figure S10. Prothionamide - phthalic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = phthalic acid, black = prothionamide).

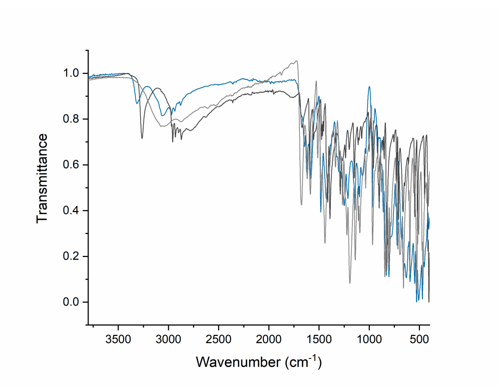

*Figure S11. Prothionamide - diflunisal.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = diflunisal, black = prothionamide).

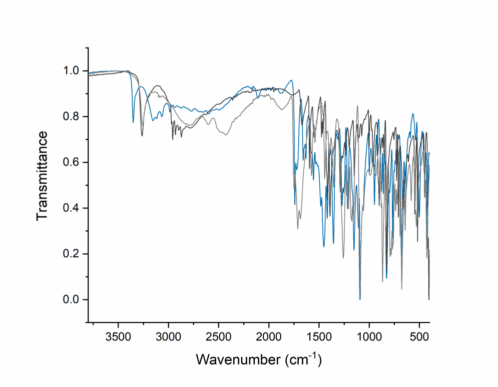

*Figure S12. Prothionamide - 2,3-pyrazinedicarboxylic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = 2,3-pyrazinedicarboxylic acid, black = prothionamide).

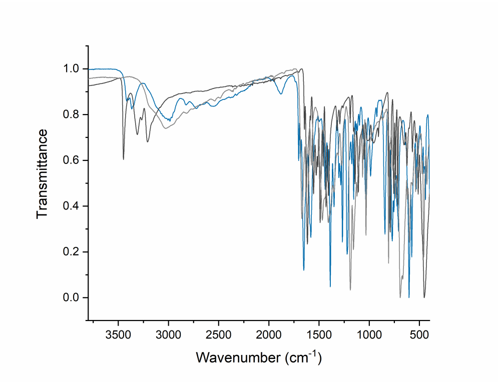

*Figure S13. Lamotrigine - 2,6-dihydroxybenzoic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = 2,6-dihydroxybenzoic acid, black = lamotrigine).

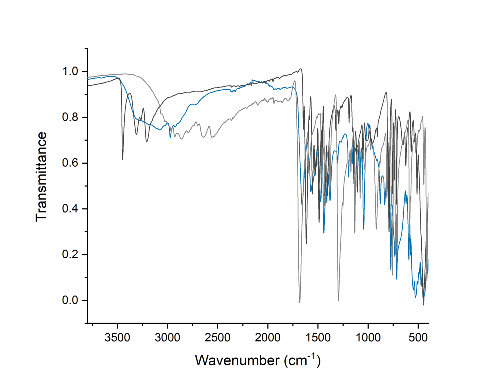

*Figure S14. Lamotrigine - 2,6-dimethylbenzoic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = 2,6-dimethylbenzoic acid, black = lamotrigine).

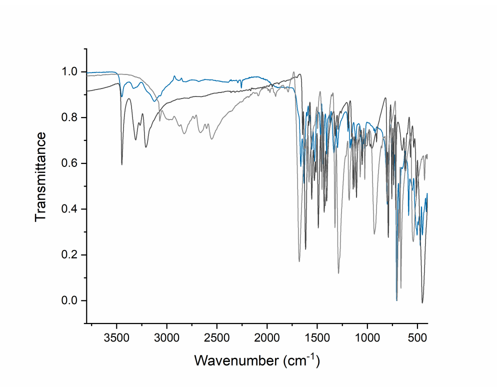

*Figure S17. Lamotrigine - benzoic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = benzoic acid, black = lamotrigine).

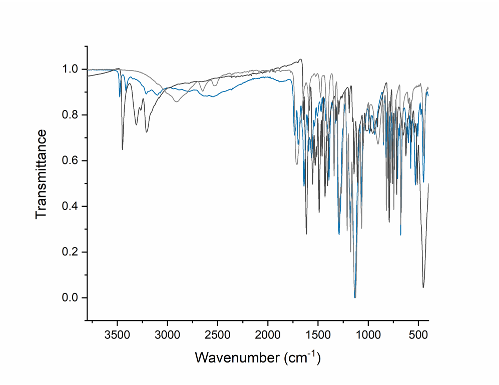

*Figure S15. Lamotrigine - 2,6-(trifluoromethyl)benzoic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = 2,6-(trifluoromethyl)benzoic acid, black = lamotrigine).

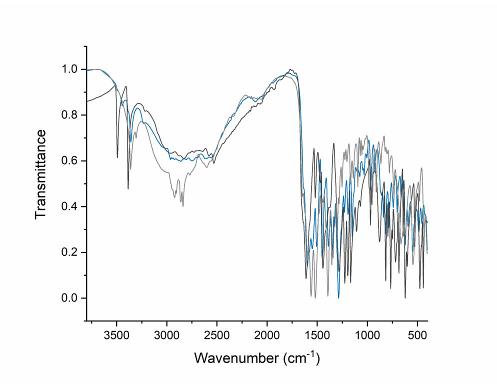

*Figure S18. p-aminosalicylic acid - lysine.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = lysine, black = p-aminosalicylic acid).

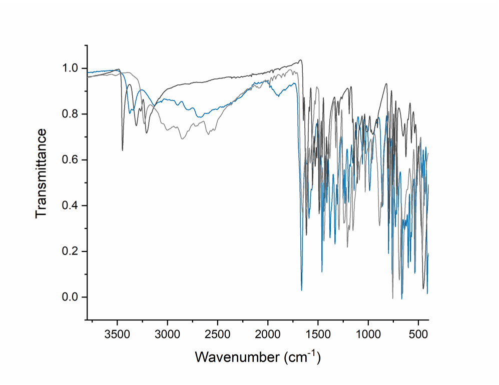

*Figure S16. Lamotrigine - salicylic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = salicylic acid, black = lamotrigine).

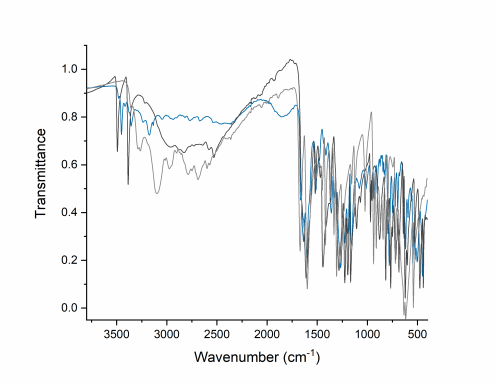

*Figure S19. p-aminosalicylic acid - adenine.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = adenine, black = p-aminosalicylic acid).

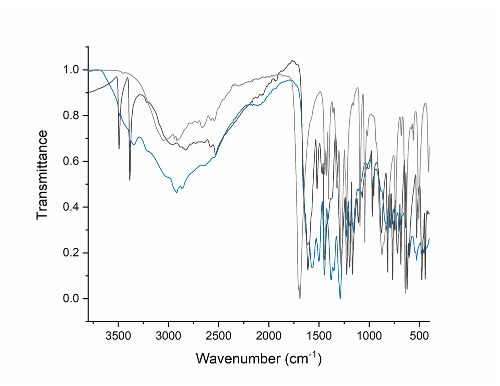

*Figure S20. p-aminosalicylic acid - ketoglutaric acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = ketoglutaric acid, black = p-aminosalicylic acid).

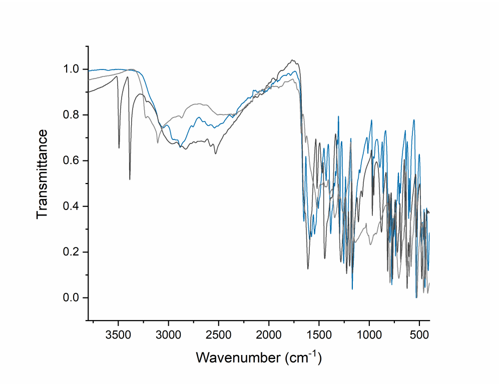

*Figure S21. p-aminosalicylic acid - 2-thiobarbituric acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = 2-thiobarbituric acid, black = p-aminosalicylic acid).

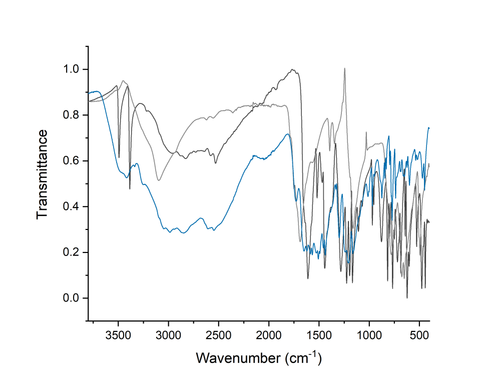

*Figure S22. p-aminosalicylic acid - oxalic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = oxalic acid, black = p-aminosalicylic acid).

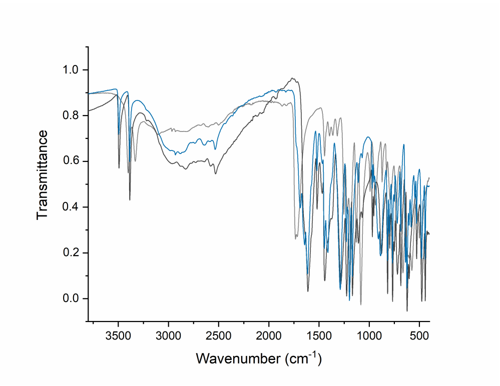

*Figure S23. p-aminosalicylic acid - tartaric acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = tartaric acid, black = p-aminosalicylic acid).

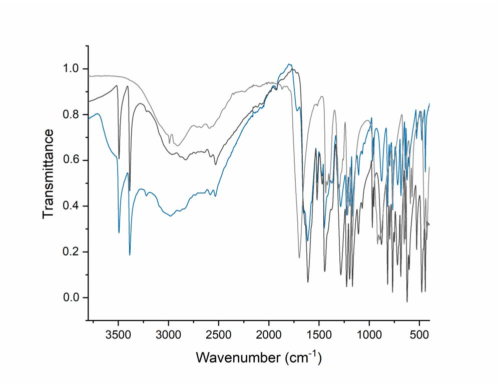

*Figure S24. p-aminosalicylic acid - malonic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = malonic acid, black = p-aminosalicylic acid).

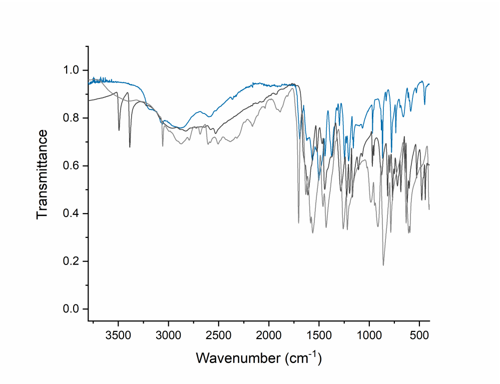

*Figure S25. p-aminosalicylic acid - maleic acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = maleic acid, black = p-aminosalicylic acid).

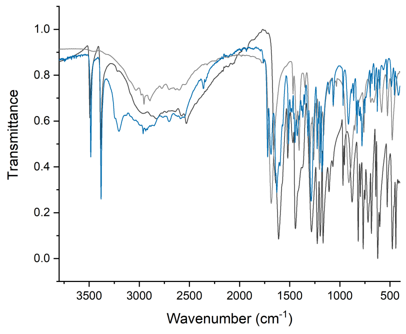

*Figure S26. p-aminosalicylic acid - glutaric acid.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = glutaric acid, black = p-aminosalicylic acid).

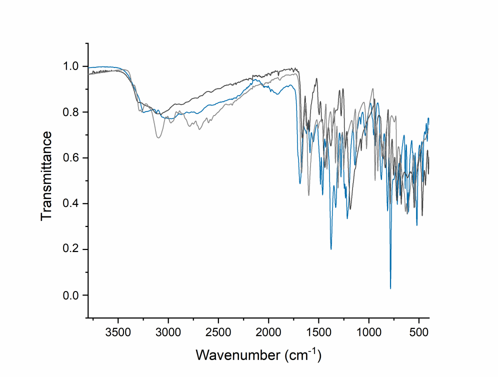

*Figure S27. gentisic acid - adenine.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = adenine, black = gentisic acid).

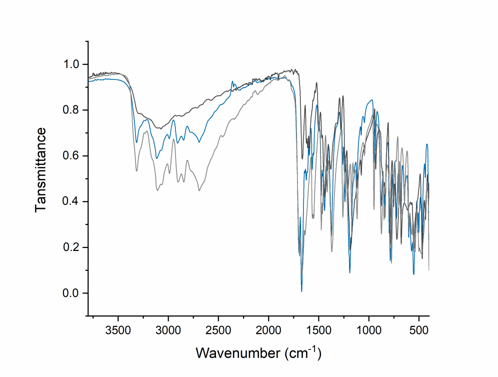

*Figure S28. gentisic acid - guanine.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = guanine, black = gentisic acid).

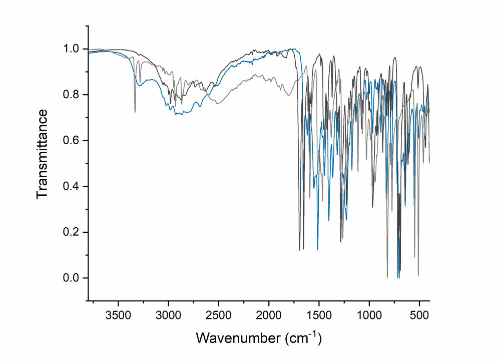

*Figure S29. ketoprofen - tyramine.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = tyramine, black = ketoprofen).

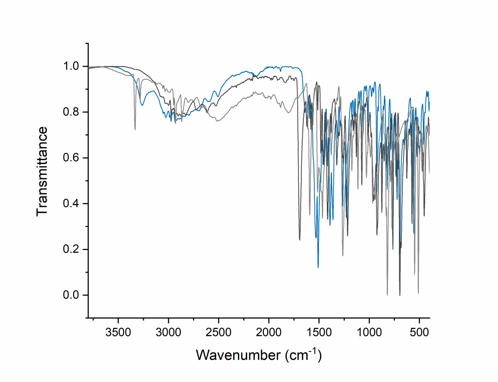

*Figure S30. flurbiprofen - tyramine.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = tyramine, black = flurbiprofen).

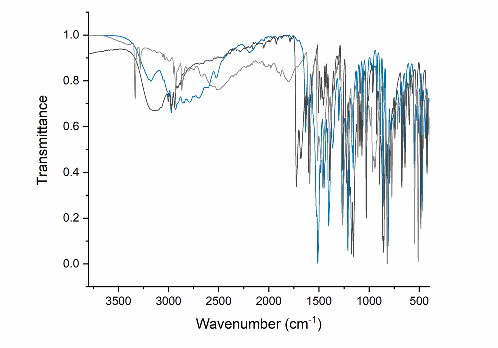

*Figure S31. naproxen - tyramine.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = tyramine, black = naproxen).

## Diflunisal-purine

### DeepCocrystal prediction

*Table S11. Results of the Prospective Study and of DeepCoCrystal prediction.* The model predictions (with the canonical and with the augmented SMILES) are reported, along with the results of majority voting. Predictions are compared with the experimental validation (*n.a.* = not tested; ✓= positive; × = negative).

| Tested coformer | DeepCocrystal Canonical | DeepCocrystal Augmented (2:7) | Majority vote | Lab validation |
|---|---|---|---|---|
| Caffeine | 0.99 | 0.99 ± 0.01 | 100% | ✓ |
| Adenine | 0.99 | 0.99 ± 0.00 | 100% | ✓ |
| Theobromine | 0.99 | 0.66 ± 0.35 | 60% | × |
| Xanthine | 0.99 | 0.63 ± 0.38 | 60% | × |
| 7-hydroxycoumarin | 0.99 | 0.26 ± 0.29 | 80% | *n.a.* |
| Ternatin | 0.99 | 0.26 ± 0.27 | 80% | *n.a.* |
| Resveratrol | 0.61 | 0.21 ± 0.28 | 90% | *n.a.* |
| Quercitin | 0.94 | 0.06 ± 0.09 | 100% | *n.a.* |
| Cinnamic acid | 0.21 | 0.05 ± 0.07 | 100% | *n.a.* |
| Curcumin | 0.09 | 0.03 ± 0.05 | 100% | *n.a.* |
| Rosmarinic acid | 0.94 | 0.02 ± 0.02 | 100% | × |
| Riboflavin | 0.05 | 0.00 ± 0.00 | 100% | × |

### Co-crystal characterization (solid-state NMR and IR spectroscopy)

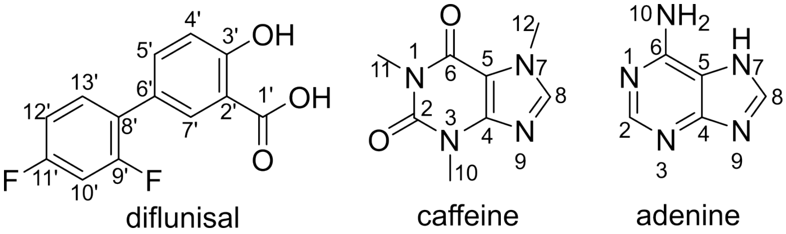

*Figure S32. Chemical structure and atom numbering* of diflunisal (DIF), caffeine (CAF) and adenine (ADE).

$^{13}$C (Figure S35) and $^{15}$N (Figure S36) CPMAS SSNMR spectra of diflunisal-caffeine (DIF-CAF) and of diflunisal-adenine (DIF-ADE) are recorded to investigate structural information, such as the number of independent molecules in the unit cell and the proton position along hydrogen-bond axes. The spectra confirm the formation of two novel crystal forms with a discrete degree of crystallinity (average full width at half maximum value of about 120 Hz). The structures of both supramolecular systems are characterized by the presence of one DIF molecule and one molecule of the respective coformer per asymmetric unit, since only a single signal pattern is detectable for each compound. The shift toward lower frequencies of the signal of the diflunisal carboxylic group (C1', *see* Fig. S32) in the DIF-CAF sample (from 175.2 to 172.9 ppm), combined with the shift of the unsaturated nitrogen of the imidazole ring (N9) signal (from 230.0 ppm for pure caffeine to 221.1 ppm for DIF-CAF), suggests the formation of the COOH$\cdots$N$_{ar}$ H-bond interaction, involving these two groups. Determining the main supramolecular interactions in DIF-ADE is challenging due to the high number of donor and acceptor groups in adenine. The most probable hypothesis is the formation of the COO$^{-}\cdots{}^{+}$NH$_3$ hydrogen bond interaction, involving a proton transfer from the carboxylic group of DIF to the most basic group (N10) of ADE. This interaction is confirmed by the C1' signal at 175.4 ppm and the N10 signal at 83.7 ppm in DIF-ADE, showing a shift of 7 ppm compared to pure ADE. The supramolecular interaction involving the N1, N3 and N7 atoms are neutral as detected from the maximum shift of the signals of 31 ppm, consistent with neutral XH$\cdots$N$_{ar}$ hydrogen bond interactions, where X could be the hydroxylic group of DIF as well as the NH or NH$_3^+$ groups of the ADE. Nonetheless, the recorded chemical shifts are consistent and unequivocally indicate the formation of a novel multicomponent crystal form.

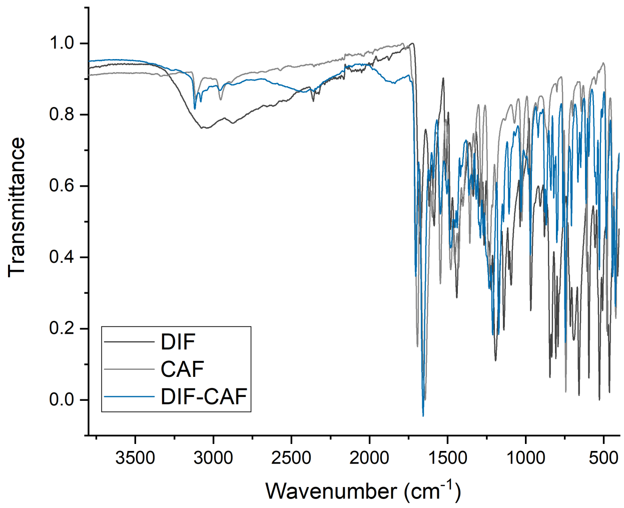

*Figure S33. Diflunisal - caffeine.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = caffeine (CAF), black = diflunisal(DIF)).

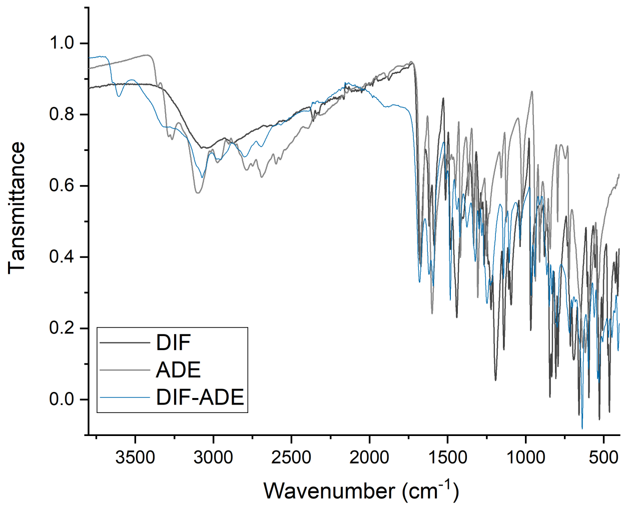

*Figure S34. Diflunisal - adenine.* FT-IR ATR spectrum of the novel multicomponent crystal form (blue) compared with those of the starting materials (light grey = adenine (ADE), black = diflunisal(DIF)).

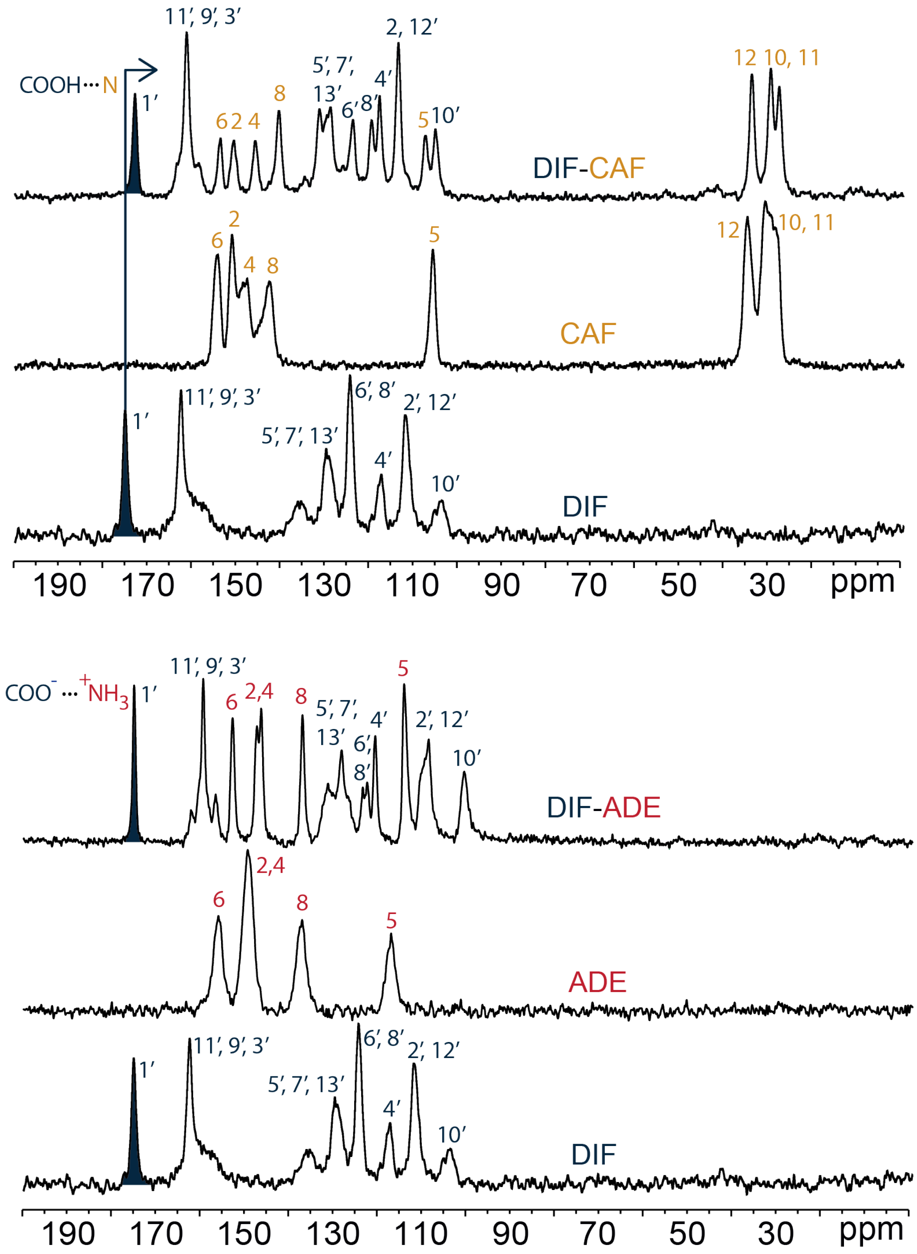

*Figure S35.* $^{13}$*C (100 MHz) CPMAS* spectra of DIF-CAF and DIF-ADE, compared with the respective starting materials, acquired with a spinning speed of 12 kHz at room temperature. The numbering of DIF, CAF and ADE atoms refers to Figure S32. The assignment of $^{13}$C signals for DIF-CAF was made possible through the non-quaternary suppression experiment. Focusing on the C1' signal, its chemical shift in the DIF-CAF spectrum suggests the formation of the COOH$\cdots$N$_{ar}$ synthon, differently for DIF-ADE where the COO$^{-}\cdots{}^{+}$NH$_3$ hydrogen bond interaction is detected.

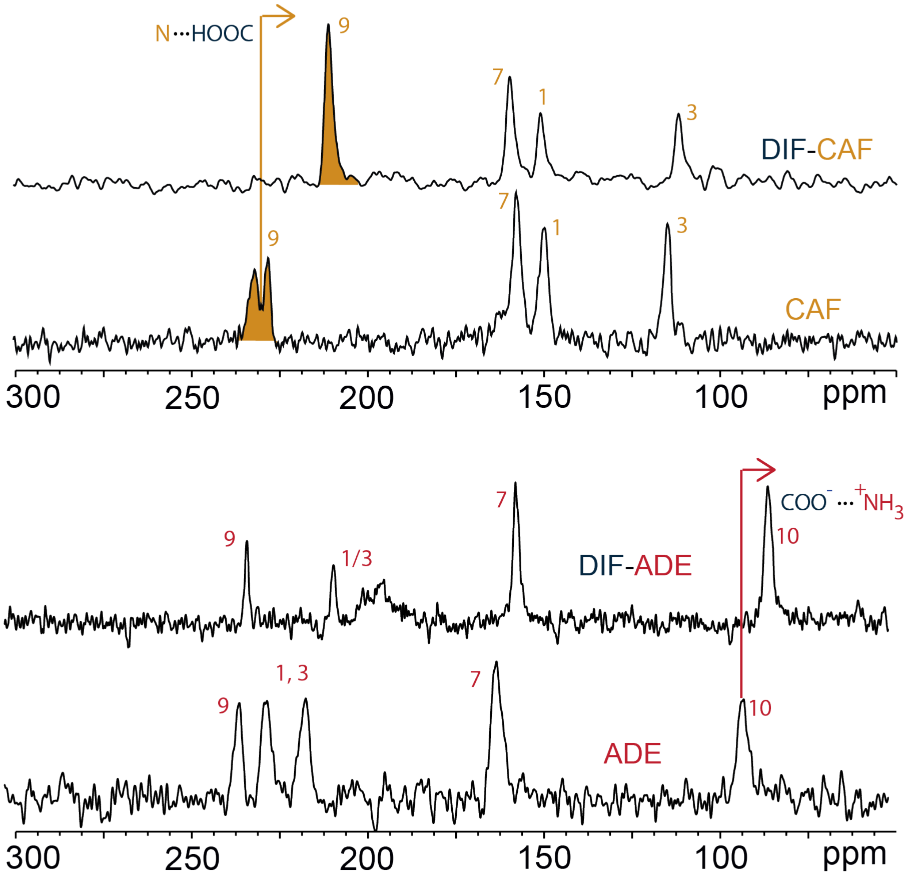

*Figure S36.* $^{15}$*N (40.56 MHz) CPMAS* spectra of DIF-CAF (top), DIF-ADE (bottom) compared with pure CAF and pure ADE, acquired with a spinning speed of 9 kHz at room temperature. The numbering refers to Figure S32.