Electronic Supplementary Material (ESI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2022

Supporting Information

Aikaterini Vriza,[a,b] *Ioana Sovago,*[c] *Daniel Widdowson,*[b,d] *Vitaliy Kurlin,*[b,d] *Peter A. Wood,*[c]

*Matthew S. Dyer**[a,b]

a. Department of Chemistry and Materials Innovation Factory, University of Liverpool, 51 Oxford Street, Liverpool L7 3NY, UK.

b. Leverhulme Research Centre for Functional Materials Design, University of Liverpool, Oxford Street, Liverpool L7 3NY, UK.

c. Materials Innovation Factory and Computer Science department, University of Liverpool, Liverpool, L69 3BX UK.Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK.

Email: M.S.Dyer@liverpool.ac.uk

Table of contents

1 Co-crystals in the CSD

1.1 Creating the CSD co-crystals map

Table S1. Co-crystals categorized based on the types of bonding. In the cases where more than one type of bonding was identified, the co-crystals were categorized considering *i*) the type of the interaction, e.g., if H-bonding and π-π interactions where both found in a pair then the pair is categorized as H-bonded as this is stronger than π-π stacking *ii*) based on the distance between the bonds according to Mercury software, e.g., if both H-bonding and halogen bonding were found in a pair then the pair is categorized according to the shorter bond as it is the strongest.

Type of bonding	Functional groups	Comments
Hydrogen bonding	Both molecules have OH or NH or SH	the donor atom D is any of N, O, or S, and the acceptor atom A is any of N, O, or S
Halogen bonded	One molecule should have a halogen and the other a heteroatom	D···X-A, where D is one of N, O, S, or Cl; X is either Br or I
Weakly bound (π-π stacking)	At least one molecule of the pair has one aromatic ring without heteroatoms	interactions that do not belong to any other category, mainly π-π interconnected

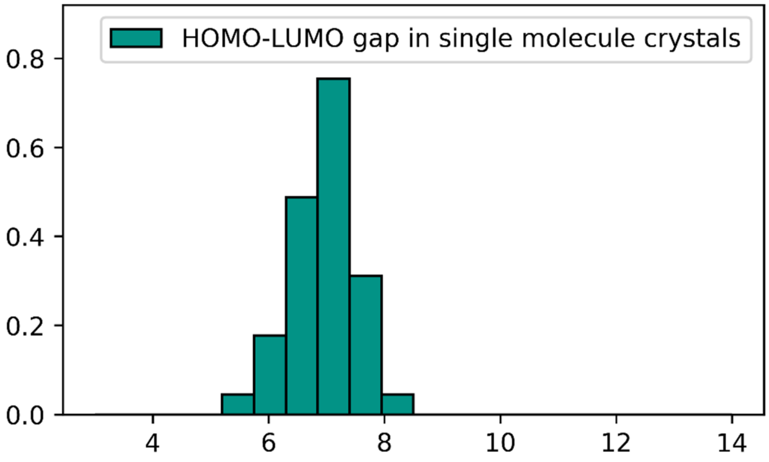


Figure S1. HOMO-LUMO gap in single molecule semiconductors. The orbital energies using PM6 were calculated for the list of the top 40 molecules reported in the SI of Nematiaram *et al*.[2]

Table S2. Solvents and single atoms that were excluded from the molecular pairs during the co-crystal extraction.

CC(Cl)(Cl)Cl	NC=O	CCNCC
OCC(F)(F)F	OC=O	F
ClC=C(Cl)Cl	CCCCCCC	Br
ClC(Cl)=C	CCCCCC	BrBr
CCOC(CC)OCC	CC(C)COC(C)=O	[F]
COCOC	CCCCCC(C)C	[O]
ClCCCl	CC(C)O	[C]
ClC=CCl	CC(C)OC(C)=O	[Cl]
COCCOC	CC(C)OC(C)C	[Br]
C1COCCO1	Cc1cccc(C)c1	[Xe]
CCCCO	CO	[N]
CCCCCO	COc1ccccc1	[H]
CCCO	COC(C)=O	[I]
COC(C)(C)OC	CCCCC(C)=O	[He]
CCC(C)O	CC1CCCCC1	Cl
CCOCCO	CCC(C)=O	ClCl
COCCO	CC(C)CC(C)=O	I
CC(C)CO	CC(C)C(C)=O	II
CC1CCCO1	C1COCCN1	IIII
CC(C)CCO	CN(C)C(C)=O	IC(I)I
CC(O)=O	CN1CCCC1=O	ICl
CC(C)=O	CN([O])=O	C=O
CC#N	Cc1ccccc1C	C#C
c1ccccc1	Cc1ccc(C)cc1	ClCl
CCCCOC(C)=O	CCCCC	ClI
ClC(Cl)(Cl)Cl	CCCOC(C)=O	COC
Clc1ccccc1	c1ccncc1	OB(O)O
ClC(Cl)Cl	O=S1(=O)CCCC1	S=C=S
CC(C)c1ccccc1	COC(C)(C)C	O=S=O
C1CCCCC1	C1CCc2ccccc2C1	O=C=O
ClCCl	C1CCOC1	N#N
CCOCC	Cc1ccccc1	C#C
CC(C)NC(C)C	OC(=O)C(Cl)(Cl)Cl	CC#CC
CN(C)C=O	OC(=O)C(F)(F)F	I[As](I)I
CS(C)=O	O	NCCN
CCO	OO	IC#CI
CCOC(C)=O	C	CBr
OCCO	S	BrI
CCOC=O	N	

1.2 New Invariants of Crystal Structures: Pointwise Distance Distributions (PDD)

To visualize the co-crystal space as a Minimum Spanning Tree in Figure 3, we applied new isometry invariants, which continuously quantify the similarity of any crystals using geometry.

An *isometry* is any composition of translations, rotations or reflections. An isometry *invariant* of a crystal does not change under isometries applied to the input, and so is independent of transformations that do not affect the rigid structure of the crystal, as well as superfluous changes in representation like extensions of the unit cell.[3]

The new isometry invariants are defined for any *periodic sets*, given by a finite motif of points which repeats periodically according to a lattice. A crystal structure can give rise to a periodic set either by taking a point in the center of each atom, or in the center of mass of each molecule.

To construct the *Pointwise Distance Distribution* (PDD) invariant of a periodic set S with points p_1, ..., p_m in a unit cell, we first find for each motif point p_i the row of ordered distances $d_{i1} \leq d_{i2} \leq \ldots \leq d_{ik}$ to the first k nearest neighbors of p_i in the infinite periodic set S.

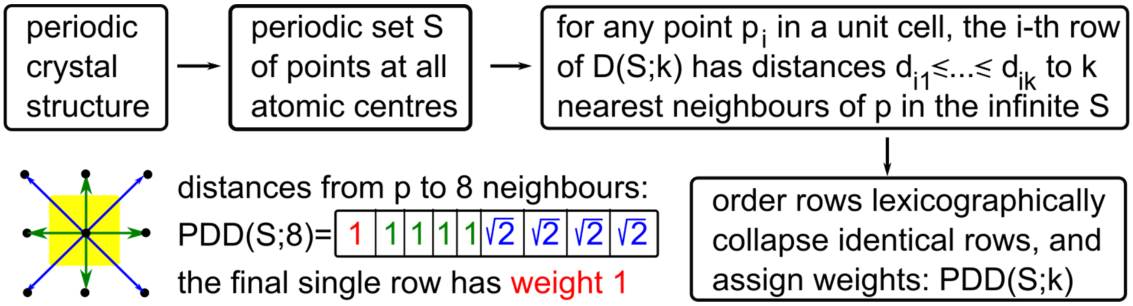

periodic crystal structure
periodic set S of points at all atomic centres
for any point p_i in a unit cell, the i-th row of D(S;k) has distances $d_{i1} \leq \ldots \leq d_{ik}$ to k nearest neighbours of p in the infinite S
order rows lexicographically collapse identical rows, and assign weights: PDD(S;k)
distances from p to 8 neighbours:
PDD(S;8)= 1 | 1 | 1 | 1 | 1 | √2 | √2 | √2 | √2
the final single row has weight 1

Figure S2. Computing the Pointwise Distance Distribution (PDD) for the unit square lattice.

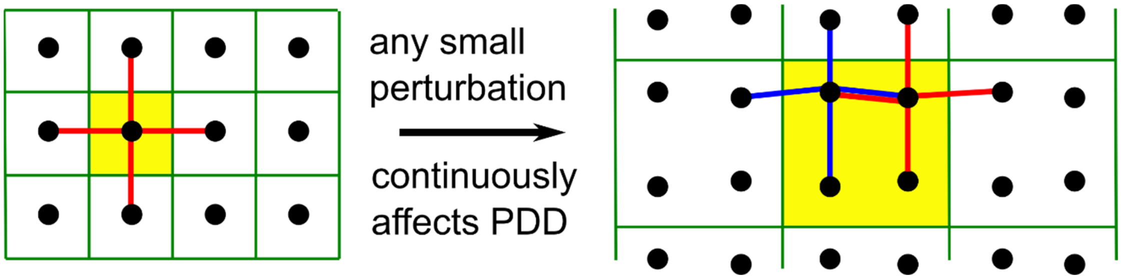
Two ordered rows of distances can be *lexicographically* compared as in a dictionary: a row $(d_{i1},d_{i2},\ldots,d_{ik})$ is less than another row $(d'_{i1},d'_{i2},\ldots,d'_{ik})$ if, comparing two rows coordinate-wise, we find a strictly smaller distance in the former row, so $d_{ij}<d'_{ij}$ for some index j from {1,...,k}.

After lexicographically sorting the rows, if any group of w rows are identical, replace them with one row and give the weight w/m (so unique rows have weight 1/m). Weights are canonically placed in the extra first column, giving a matrix with k+1 columns and no more than m rows.

The resulting matrix is an isometry invariant [Ref [4], Thm 5], which was also proved to be complete at least in a general position, meaning that almost any two non-isometric periodic sets have different PDDs for large enough k [Ref [4], Thm 9]. We have no counter-examples to completeness and conjecture that PDD is complete for all periodic sets of points.

To define a proper distance between PDDs, a base distance between rows of two PDDs (without weights) is required. This could be any metric between vectors. In the experiments we used the L_infinity, the maximum absolute difference between any corresponding elements of rows. Once a distance between rows is chosen, the Earth Mover's Distance compares the weighted distributions of rows, finding an optimal matching between rows while respecting the weights, and the 'cost' for this optimal matching is a proper EMD metric satisfying all metric axioms and also the continuity under perturbations of points [Ref [4], Thm 7]. This distance has units of Angstroms, being a weighted sum of differences of inter-point distances.

An invariant of periodic sets is continuous if small perturbations to the points of the input always result in a small distance between outputs. This continuity is needed to capture the notion of similarity between non-isometric structures but is not satisfied by other invariants such as the reduced cell and is impossible with discrete invariants such as space groups, as in Figure S3 below.



PDD(S;4) =

1	1	1	1	1	1

PDD(S';4) =

0.5	0.8	1.005	1.005	1.2
0.5	1	1	1.02	1.02

EMD(PDD(S;4), PDD(S';4)) = 0.102

Figure S3. The Earth Mover's Distance (EMD) between PDD of close structures is small [39b].

Figure S3 shows a unit square lattice with a point at (0.5,0.5) whose PDD consists of one row with the first four equal distances (1,1,1,1) shown in red, and a perturbed set obtained by extending the unit cell to 2x2 and perturbing two points, giving the motif (0.5,0.4), (0.5,1.6), (1.5,0.5), (1.5,1.5). The initial matrix has two pairs of identical rows of distances: (0.8,1.005,1.005, 1.2) in blue and (1,1,1.005,1.005) in red. The final

PDD has the two above rows with weight 0.5. In this case, the Earth Mover's distance (with k=4) is equal to the average L_infinity distance between the initial row and two perturbed rows, so EMD=0.5(|0.8-1|+|1.005-1|) = 0.102.

$$\begin{pmatrix} 0.8 & 1.005 & 1.005 & 1.2 \\ 0.8 & 1.005 & 1.005 & 1.2 \\ 1 & 1 & 1.005 & 1.005 \\ 1 & 1 & 1.005 & 1.005 \end{pmatrix} \longrightarrow \left(\begin{array}{c|cccc} 1/2 & 0.8 & 1.005 & 1.005 & 1.2 \\ 1/2 & 1 & 1 & 1.005 & 1.005 \end{array}\right)$$

Figure S4. The four motif points of the set in Figure S3 give two repeated sets of distances because of the reflectional symmetry in the horizontal line cutting the unit cell in half. Since both rows appear twice out of four total rows, they both have the weight 2/4=1/2.

1.3 Creating the external validation datasets

The external validation dataset was created from published work on experimental co-crystal screening. The SMILES strings were then canonicalized to be in same format as the CCDC canonical SMILES. The publicly available sources alongside with more details regarding the type of the reported co-crystals and the experimental methods tested are discussed below in chronological order:

1. Karki *et al*, 2010 (Artemisinin dataset): An experimental screening was performed using liquid-assisted grinding (LAG) was performed using a short-list of 75 chemically diverse co-formers which are reported in the Supporting Information (Figure S1-S75).[5] Only 2 out of the 75 co-formers resulted in observation of a co-crystal, forming a dataset of **73** negatives and **2** positive pairs.

2. Grecu *et al*, 2014 (MEPS dataset): This work presents an *ab initio* co-crystal screening approach, namely molecular electrostatic potential surfaces (MEPS), which is validated using experimental co-crystal screens reported in literature. These screens involve 18 APIs tested with a wide range of co-formers.[6] The names of the tested co-formers as well as the positive or negative outcome of the co-crystal screen were extracted from the Supporting Information of the paper. The MEPS paper reports 303 negative and 129 positive molecular pairs. However, as Wang *et al* stated,[7] five pairs that have been reported as negatives which were experimentally proven to be positive in later publications. These pairs involve pyrazinecarboxamide with 3,5-dihydroxybenzoic acid (ACOPOA), oxalic acid (UZODUK), malonic acid (SIHRAE), adipic acid (KOVSAR), and glutaric acid (SIHQOR). Moreover, Grecu *et al* report two pairs molecular pairs which form salts instead of co-crystals *i.e.*, indomethacin and N-methyl-D-glucamine, indomethacin and tromethamine. These two pairs were labelled as negative. Taking into consideration these corrections, the final MEPS dataset consists of **300** negatives and **132** positive pairs.

3. Wicker *et al*, 2017 (H-bond synthons dataset): A set of 20 target molecules was screened for co-crystallization against 34 substituted aromatic acid and amide co-formers. Major consideration was given in the incorporation of the four main hydrogen-bond supramolecular synthons between the molecular pairs.[8] The experimental screening involves solid state grinding and the co-crystal formation assessment was based upon changes in the PXRD pattern when accompanied in IR by a shift of the characteristic peaks traditionally involved in hydrogen bonding. In some cases, Differential Scanning Calorimetry (DSC) was also used to assess co-crystal formation. The whole process revealed a dataset of **408** negative and **272** positive molecular pairs.

4. Mapp *et al*, 2017 (Propyphenazone dataset): Propyphenazone, an analgesic drug with limited or no hydrogen bonding functionality, was screened against 89 co-formers.[9] The experimental methods mainly used were solvent drop and neat grinding. Solution crystallization experiments were also performed for some of the combinations that were difficult to characterize after the grinding experiments. The co-crystal formation assessment was performed based on the PXRD patterns which are different from those of the parent materials. This process resulted in a dataset of **81** negative and **8** positive molecular pairs.

5. Przybyłek *et al*, 2018 (Phenolic acids dataset): This work is related to the development of a theoretical co-crystal screening model based on 1D and 2D molecular descriptors for phenolic acid co-crystals.[10] A dataset containing both phenolic acid co-crystals and eutectics was created from the authors for validating their approach. The reported molecular pairs were extracted from the Supporting Information Table S1. A duplicate pair was found, namely paliperidone- hydroxybenzoic acid, which was removed, resulting in a dataset of **58** negative and **167** positive molecular pairs.

6. Przybyłek *et al*, 2019 (Dicarboxylic acids dataset): This work is related to the testing of a theoretical co-crystal screening model based on 1D and 2D molecular descriptors for dicarboxylic acid co-crystals.[11] A dataset containing both dicarboxylic acid co-crystals and eutectics was created from the authors. The reported molecular pairs were extracted from the Supporting Information Table S1. Two duplicate pairs were found, namely 2-pyridone-adipic acid and exemestane-maleic acid, which were removed, resulting in a dataset of **104** negative and **606** positive molecular pairs.

7. Sarkar *et al*, 2020 ((des)loratadine dataset): This work involves the experimental screening of two APIs, namely loratadine and desloratadine, against 41 potential co-formers.[12] The experimental method used was solvent-assisted grinding and the co-crystallization assessment was based on the IR spectrum. If IR spectrum of the molecular mixture had a shift greater than 3 cm-1 in several modes, the mixture was characterized as a successful co-crystal, whereas consistent un-changed peaks were characterized as unsuccessful co-crystal. This process resulted in a dataset of **17** negative and **65** positive molecular pairs.

8. Khalaji *et al*, 2021 (Linezolid dataset): Linezolid, an antibacterial drug, was experimentally screened against 19 different co-formers. The experimental technique used was liquid-assisted grinding (LAG) testing three different solvents, *i.e.,* methanol, toluene and water. The co-crystal formation was assessed from the PXRD patterns.[13] This process resulted in **9** negative and **10** positive molecular pairs.

9. Vriza *et al*, 2021 (Pyrene dataset): Pyrene was screened against 6 polyaromatic hydrocarbons for the formation of π-π co-crystals. The co-crystallization reactions were performed after dissolving the molecular mixtures in dicloromethane under 45°C with continuous stirring. The co-crystals started forming after slow evaporation in open air. The co-crystal formation was verified by the PXRD patterns.[14] This process resulted in **4** negative and **2** positive molecular pairs.

10. Devogelaer *et al*, 2021 (Praziquantel dataset): Praziquantel, an anthelmintic drug, was screened against 30 co-formers.[15] The experimental techniques used were liquid assisted grinding and solvent evaporation. The co-crystal formation was verified by the PXRD patterns. This process resulted in **18** negative and **12** positive molecular pairs.

11. Wu *et al*, 2021 (Mop dataset): In this work, 2-amino-4,6-dimethoxypyrimidine (MOP) was experimentally screened with 63 co-formers.[16] The experimental techniques used were liquid assisted grinding and the resulted powders were characterized by powder X-ray diffraction (PXRD) and differential scanning calorimetry (DSC) to identify possible solid forms. This process resulted in **22** negative and **41** positive molecular pairs.

All the datasets were further combined and the duplicate molecular pairs were removed to create the final external validation database.

1.4 Molecular Pair representations using pretrained models

The two pretrained with self-supervised learning networks which have been used to learn the molecular fingerprints of the molecular pairs are analysed below:

1.4.1 GNN pretrained model

The GNN model is pretrained with Attribute Masking, where the input node/edge attributes, *i.e.,* the atom type in the molecular graph, are randomly masked with special masked indicators. The training task of the GNN model is to predict the masked nodes based on the neighboring structure (Figure S5).[17] The pretrained GNN was acquired from the following repository: http://snap.stanford.edu/gnn-pretrain.

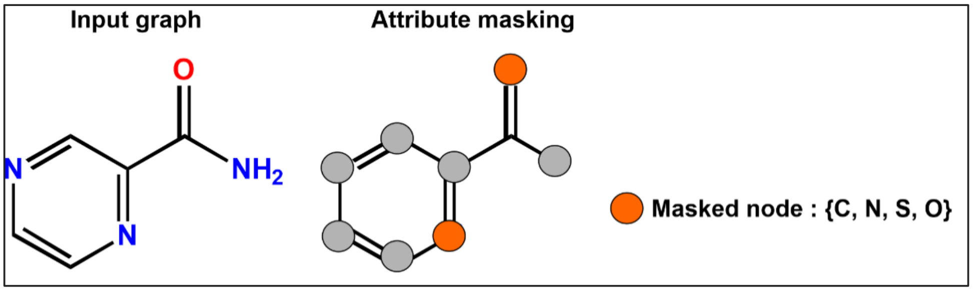


Figure S5. Visualization of the attribute masking technique

For the node level self-supervised pretraining Hu *et al* used 2M unlabelled molecules from ZINC15 database. The model was trained with Adam optimizer with a learning rate of 0.001 for 100 epochs using a batch size of 256.

1.4.2 ChemBERTa pretrained model

ChemBERTA uses the Byte-Pair Encoder (BPE) tokenization strategy as provided from the HuggingFace tokenizers library.[18] BPE is a hybrid between character and word-level representations, which allows for the handling of large vocabularies in natural language corpora. This tokenization strategy finds the best word segmentation by iteratively and greedily merging frequent pairs of characters.

2 Machine learning procedures

Table S3. Hyperparameters optimization.

Hyperparameters	Values range
Learning rate	[10^{-2} ,10^{-3}, 10^{-4}, $3*10^{-4}$, 10^{-5}]
Batch size	[32, 64, 128, 256]
Number of epochs	[10, 50, 100]
Weight decay	[10^{-3}, 10^{-4} , 10^{-5} , 10^{-6}]
Dropout	[0.1, 0.2, 0.3, 0.4]

2.1 Hyperparameters selection with wandb

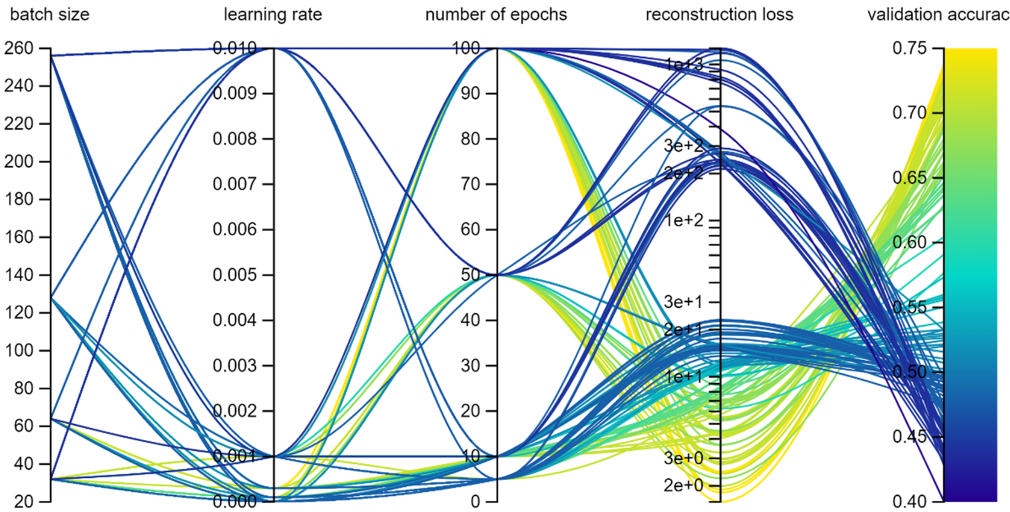


Figure S6. Hyperparameter optimization using the Mordred library for extracting the molecular features. The optimal hyperparameters identified with the wandb library screening are the following: learning rate = 1e^{-04}, number of epochs = 100, batch size =128, dropout=0.1 and weight decay =1e^{-05}.

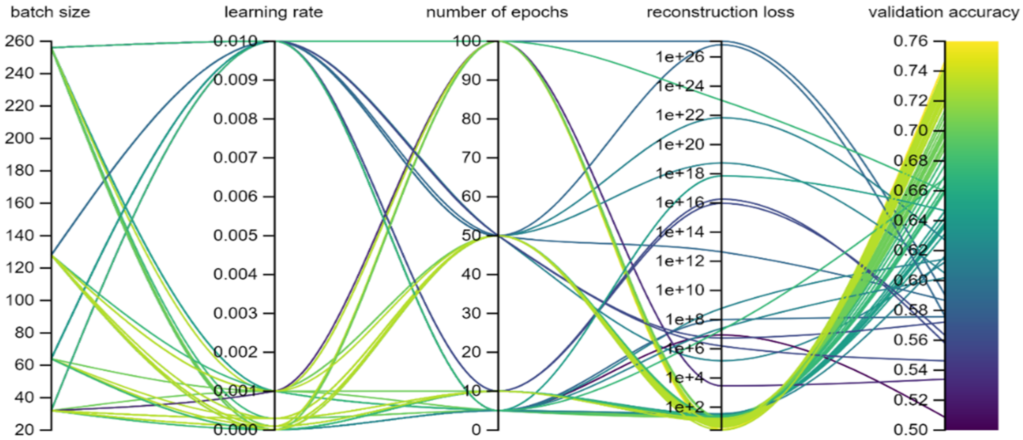


Figure S7. Hyperparameter optimization using the Morgan Fingerprint. The optimal hyperparameters identified with the wandb library screening are the following: learning rate = 1e^{-03}, number of epochs = 100, batch size =64, dropout=0.1 and weight decay =1e^{-05}.

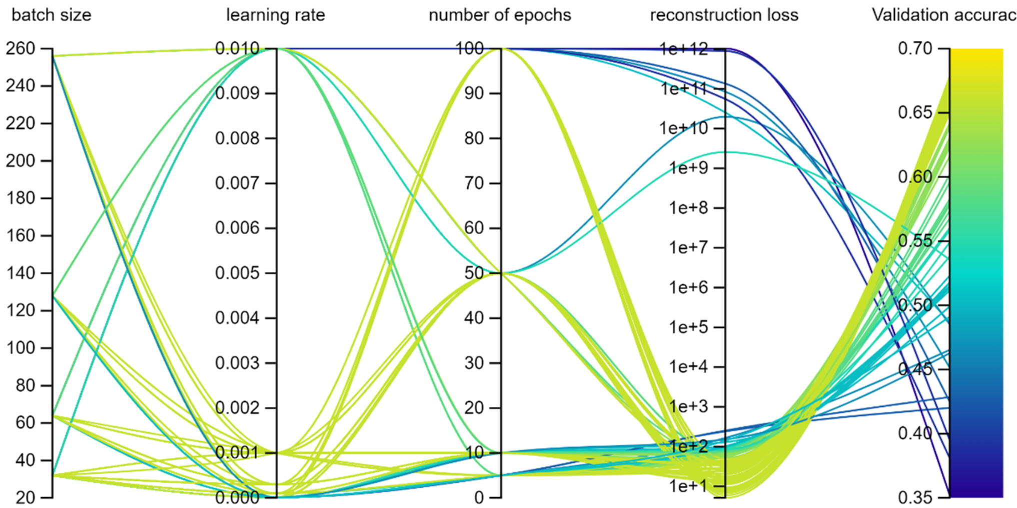


Figure S8. Hyperparameter optimization using the NLP (ChemBERTa) fingerprint. The optimal hyperparameters identified with the wandb library screening were : learning rate = 1e^{-05}, number of epochs = 100, batch size =64, dropout=0.1 and weight decay =1e^{-04}.

2.2 Algorithm performance

2.2.1 Algorithms performance per dataset

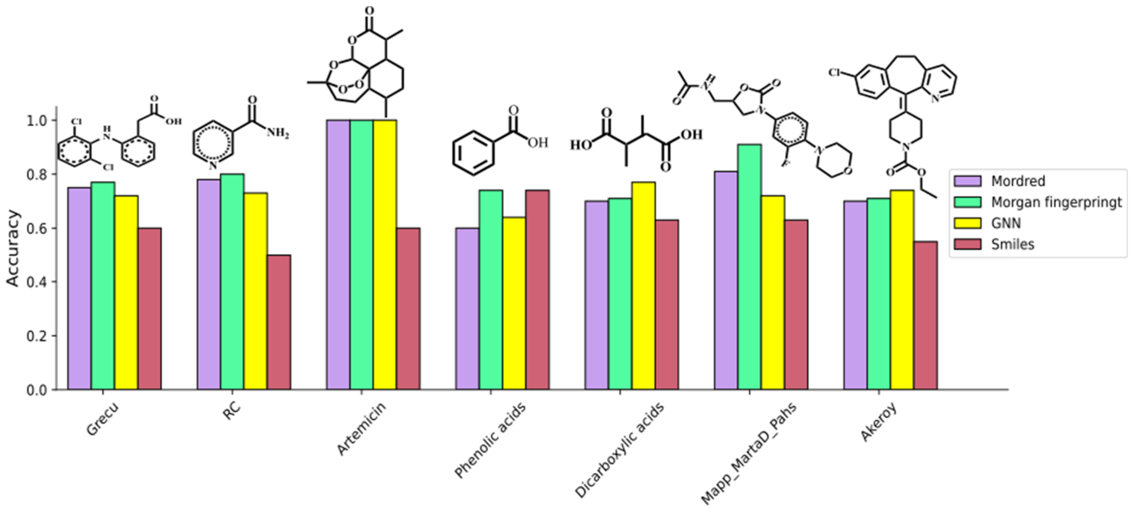


Figure S9. Models accuracy per dataset

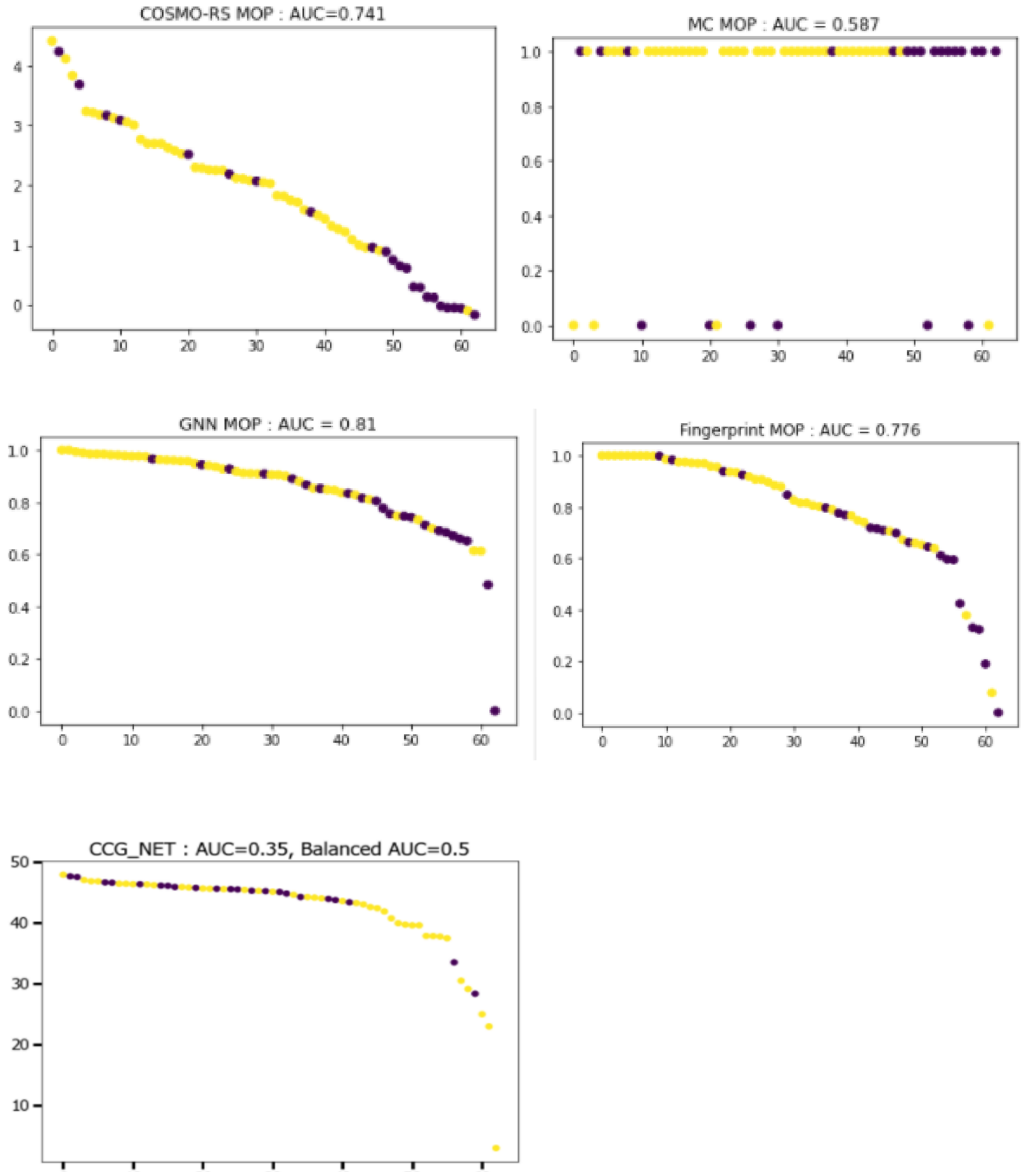


Figure S10. Models accuracy on an external dataset (MOP dataset)

2.3 Final models

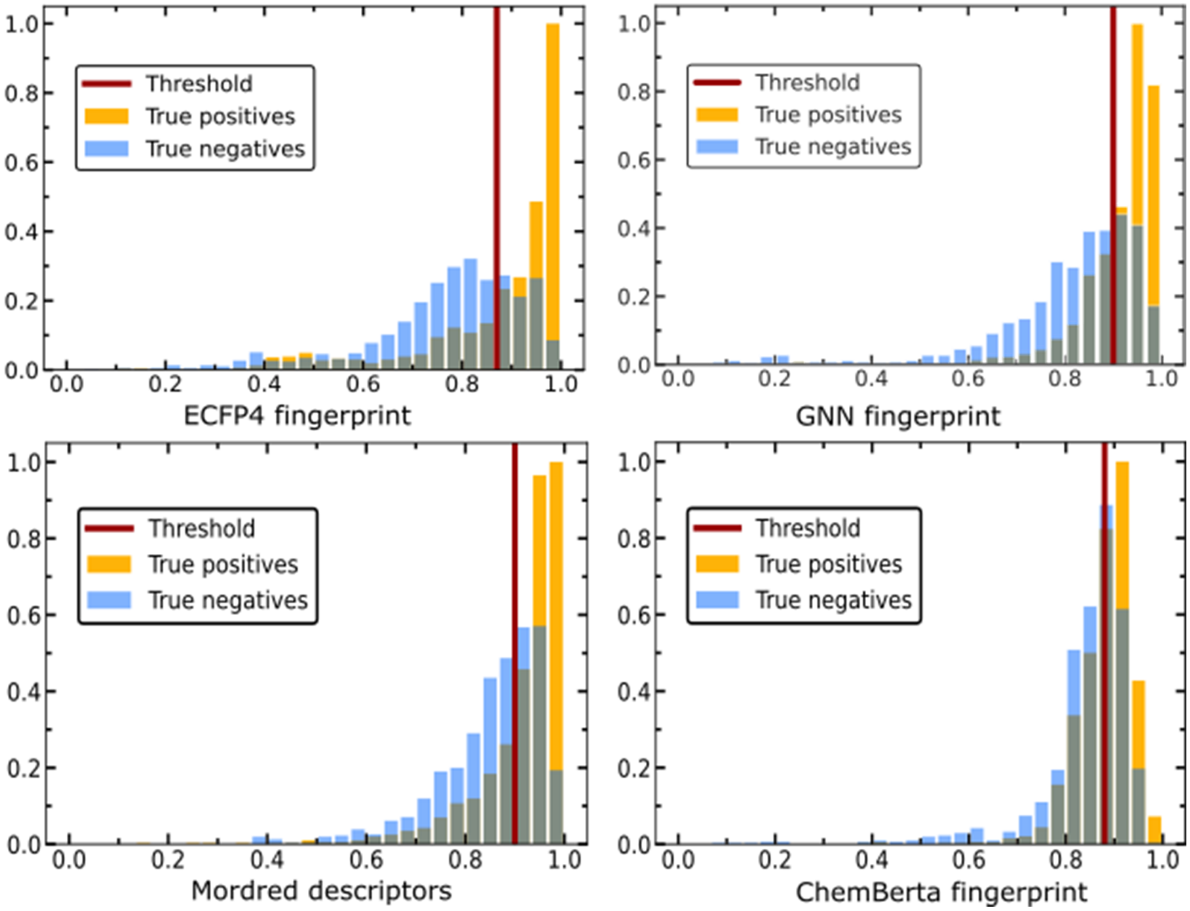
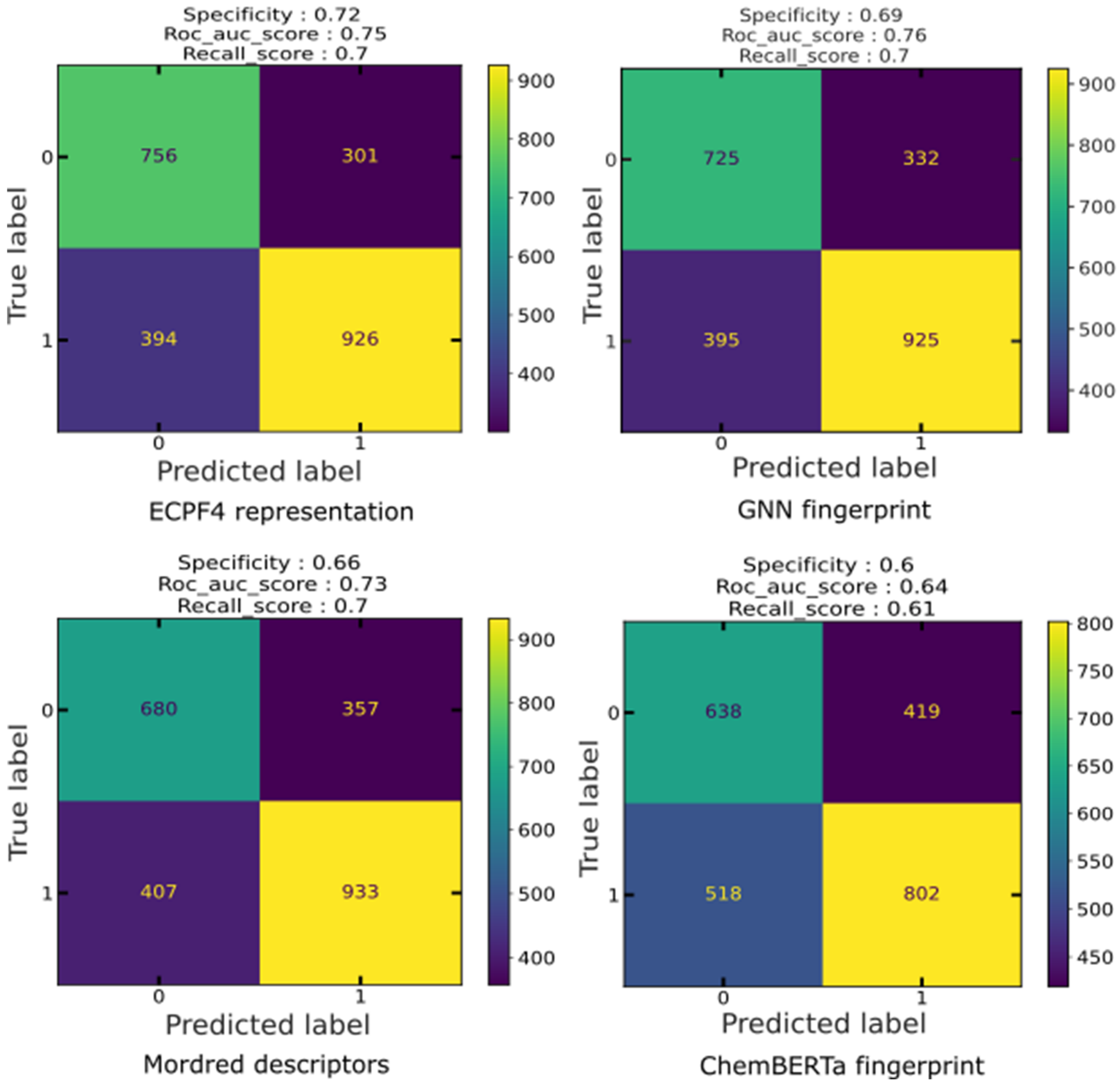


Figure S11. Scores distribution of the different models on the external validation sets. The real positives (orange bars) have higher scores than the true negatives (blue bars) for all four models. A better discrimination between the two classes is achieved for the ECFP4 and GNN models.

```
from sklearn.metrics import roc_curve
def get_threshold(labels, scores):
  fpr, tpr, thresholds = roc_curve(labels, scores)
  gmeans = np.sqrt(tpr * (1 - fpr))
  ix = np.argmax(gmeans)
  threshold = thresholds[ix]
  return threshold
```

Figure S12. Code explaining the threshold selection using the Receiver Operating Characteristic (ROC) curve from the scikit-learn library.[19] The ROC curve features the relationship between the true positive and the false positive rate. The optimal threshold is the point where the true positive rate is maximized whilst the false positive rate is minimized.



ECPF4 representation

GNN fingerprint

Mordred descriptors

ChemBERTa fingerprint

Figure S13. Confusion matrices of the four different models based on the representation techniques.

2.4 Comparison with other methods per API

Our two best models were compared to other methods for co-crystal screening which report their accuracy on the external validation data. The reported labelled from the other methods are shown in Table S3.

Table S3. Reported AUC from other methods tested on the APIs

API	MEPS*	Wang**	COSMO-RS***
Piracetam	0.93	0.895	
Paracetamol	0.7676	0.543	0.6
Diclofenac	0.75	0.857	
Pyrazinamide	0.841	0.876	
Acetazolamide	0.794	0.311	
Indomethacin	0.68	0.658	0.54
Drug_candidate	0.792	0.986	
Furosemide	0.65	0.894	
Nalidixicacid	1	0.948	
3-Cyanophenol	0.76	0.911	0.98
4-Cyanophenol	0.89	0.964	1
3-Cyanopyridine	0.94	1	
4-Cyanopyridine	0.95	0.911	0.96
Benzamide	0.4	0.738	0.71
Itraconazole	0.8125	0.844	1
Bicalutamide	0.68	0.688	0.94
Meloxicam	0.633	0.733	0.67
Nicotinamide	1	1	0.92

*Reported in the paper: https://pubs.acs.org/doi/abs/10.1021/cg401339v

**Reported in the paper: https://pubs.acs.org/doi/10.1021/acs.cgd.0c00767

***Reported in the paper: https://europepmc.org/article/med/32969658

For direct comparison with the CCGnet model the github repository https://github.com/Saoge123/ccgnet was used only for those APIs that are not contained in the training set of CCGnet.

2.5 Interpretability with Shapley

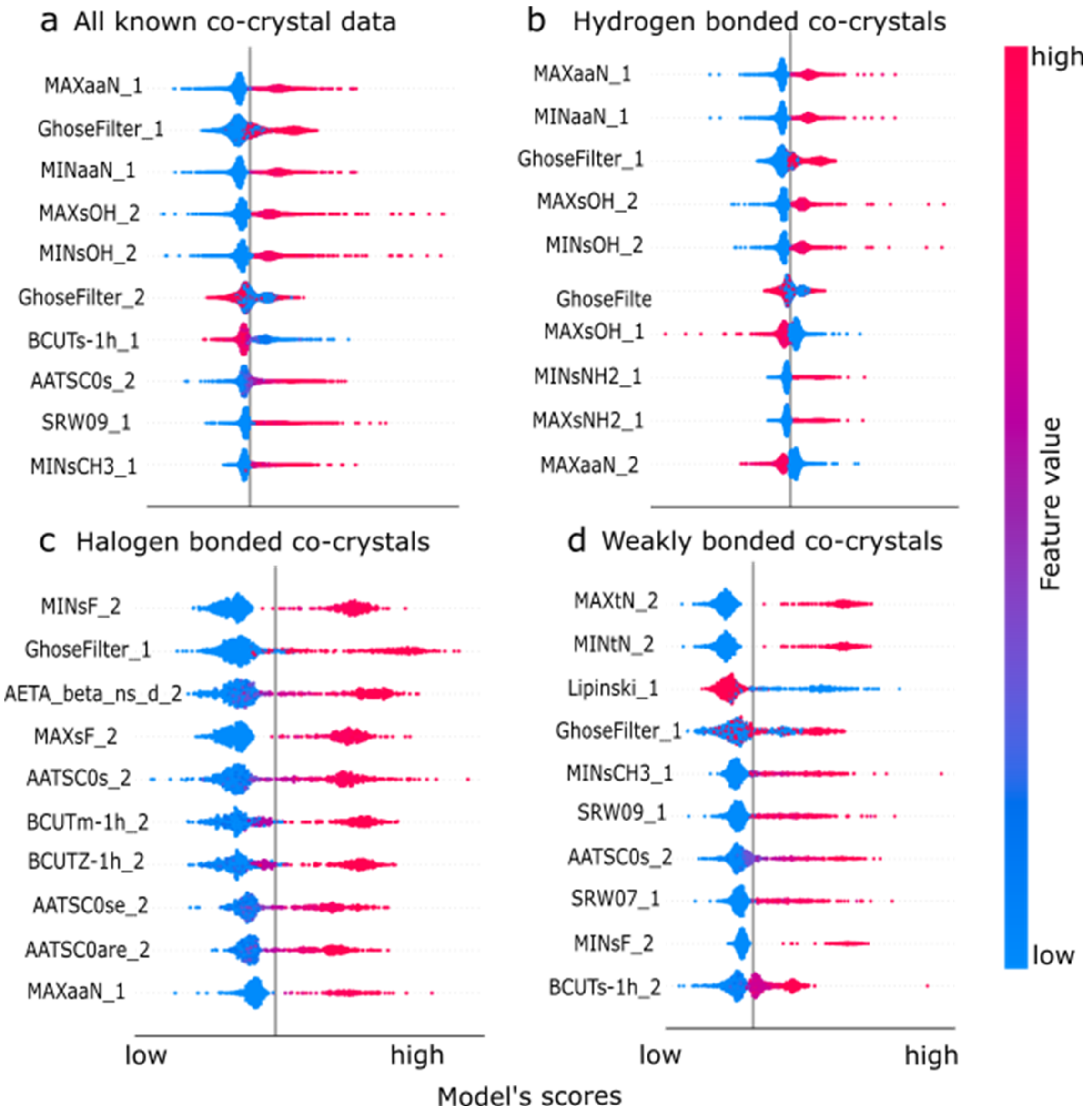
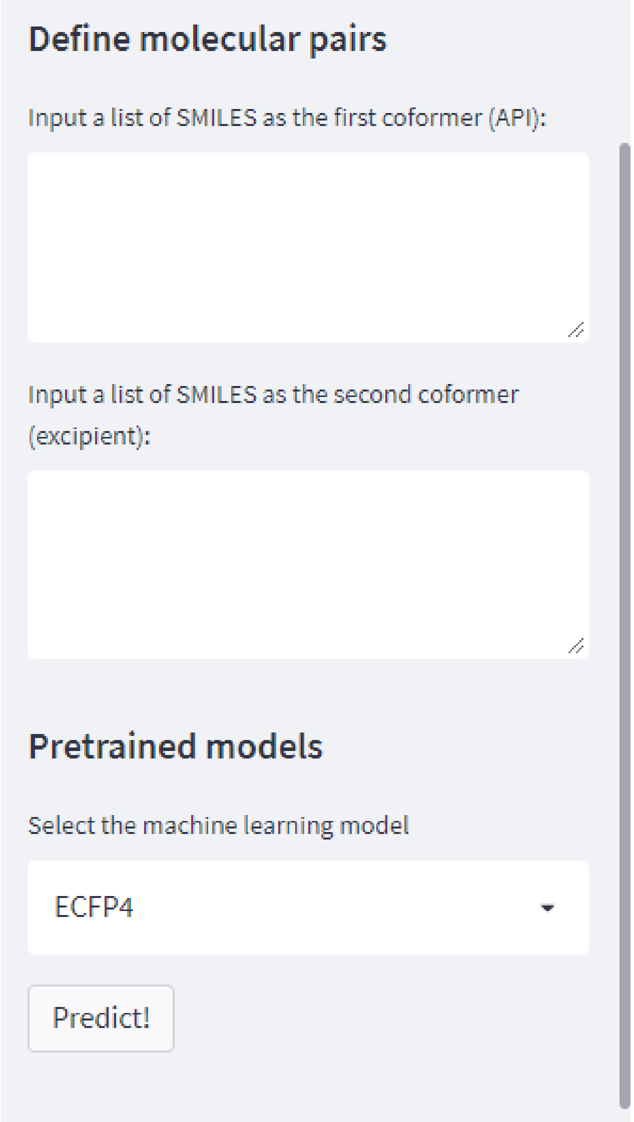


Figure S14. Shapley additive explanations categorized according to the type of interactions between the molecular pairs. Each molecule is represented as a vector containing the Mordred descriptors. The notation _1 and _2 indicate the first or second molecule in the pair. a) global interpretation of the whole co-crystals dataset and local interpretations of b) hydrogen bonded pairs, c) halogen bonded pair, d) weakly bonded molecular pairs. The pink colour refers to high values of the molecular features and the blue to low values, whereas the x axis refers to the model's scores being high or the right and low on the left.

As for the co-crystal formation the type of interactions among the molecular pairs plays a crucial role, we got insights for what affected co-crystallization based on which bonding group the pairs belong to. Molecular descriptors representation is straight-forward and the weight of each feature can be directly extracted with SHAP. According to Figure S11a representing the Shapley global interpretation, we can observe that as the dataset is dominated by H-bond interactions the most important features are related to the OH group (MAXsOH,MINsOH) and the N group (MAXaaN, MinaaN). According to the Shapley local interpretations we can derive i) the important features for hydrogen bonded pairs (Figure S11b) where the existence of OH (MAXsOH,MINsOH), NH2 (MAXsNH2, MINsNH2) and N groups (MAXaaN, MinaaN) are highlighted as the most important contributing factors, ii) the dominating features for the halogen bonded co-crystals (Figure S11) are those related to the existence of F groups (MINsF, MAXsF) and iii) in the case of weak interactions the existence of electronegative groups such as terminal triple bonded N (≡N) groups (MAXtN, MINtN) or F groups (MINsF) was found to be the most important for the formation of that type of weak bonding. It can be concluded that the top important descriptors of each category are mostly related to the existence of some functional groups in the molecules that form the pairs and not a physical property. That could be the reason why using the molecular fingerprints for the co-crystallization prediction shown a good performance in the tested systems.

2.6 Using the GUI

The code is publicly available at https://github.com/lrcfmd/MolecularSetTransformer, which also includes a web interface for enabling the wider use of our pretrained models. See Figures S14-S16 for screenshots demonstrating the use of the wed interface.



Define molecular pairs

Input a list of SMILES as the first coformer (API):

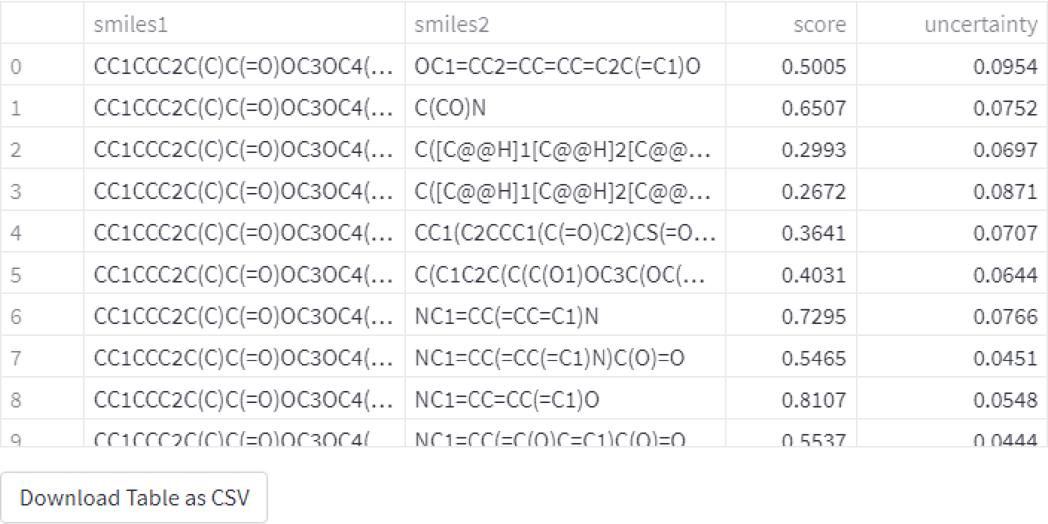
Input a list of SMILES as the second coformer (excipient):

Pretrained models

Select the machine learning model

ECFP4

Figure S15. The user can insert the SMILES strings of the molecular pairs on the empty boxes on the left of the main screen. Then the user can select the desired pretrained model between the ECP4 and GNN representation. By pressing the predict button, the selected model will calculate the score and uncertainty of each given pair and print the results on the main page.



	smiles1	smiles2	score	uncertainty
0	CC1CCC2C(C)C(=O)OC3OC4(…	OC1=CC2=CC=CC=C2C(=C1)O	0.5005	0.0954
1	CC1CCC2C(C)C(=O)OC3OC4(…	C(CO)N	0.6507	0.0752
2	CC1CCC2C(C)C(=O)OC3OC4(…	C([C@@H]1[C@@H]2[C@@…	0.2993	0.0697
3	CC1CCC2C(C)C(=O)OC3OC4(…	C([C@@H]1[C@@H]2[C@@…	0.2672	0.0871
4	CC1CCC2C(C)C(=O)OC3OC4(…	CC1(C2CCC1(C(=O)C2)CS(=O…	0.3641	0.0707
5	CC1CCC2C(C)C(=O)OC3OC4(…	C(C1C2C(C(C(O1)OC3C(OC(…	0.4031	0.0644
6	CC1CCC2C(C)C(=O)OC3OC4(…	NC1=CC(=CC=C1)N	0.7295	0.0766
7	CC1CCC2C(C)C(=O)OC3OC4(…	NC1=CC(=CC(=C1)N)C(O)=O	0.5465	0.0451
8	CC1CCC2C(C)C(=O)OC3OC4(…	NC1=CC=CC(=C1)O	0.8107	0.0548
9	CC1CCC2C(C)C(=O)OC3OC4(	NC1=CC(=C(O)C=C1)C(O)=O	0.5537	0.0444

Figure S16. A downloadable table is printed reporting the given SMILES strings of the two molecules, the score and the uncertainty of each pair.

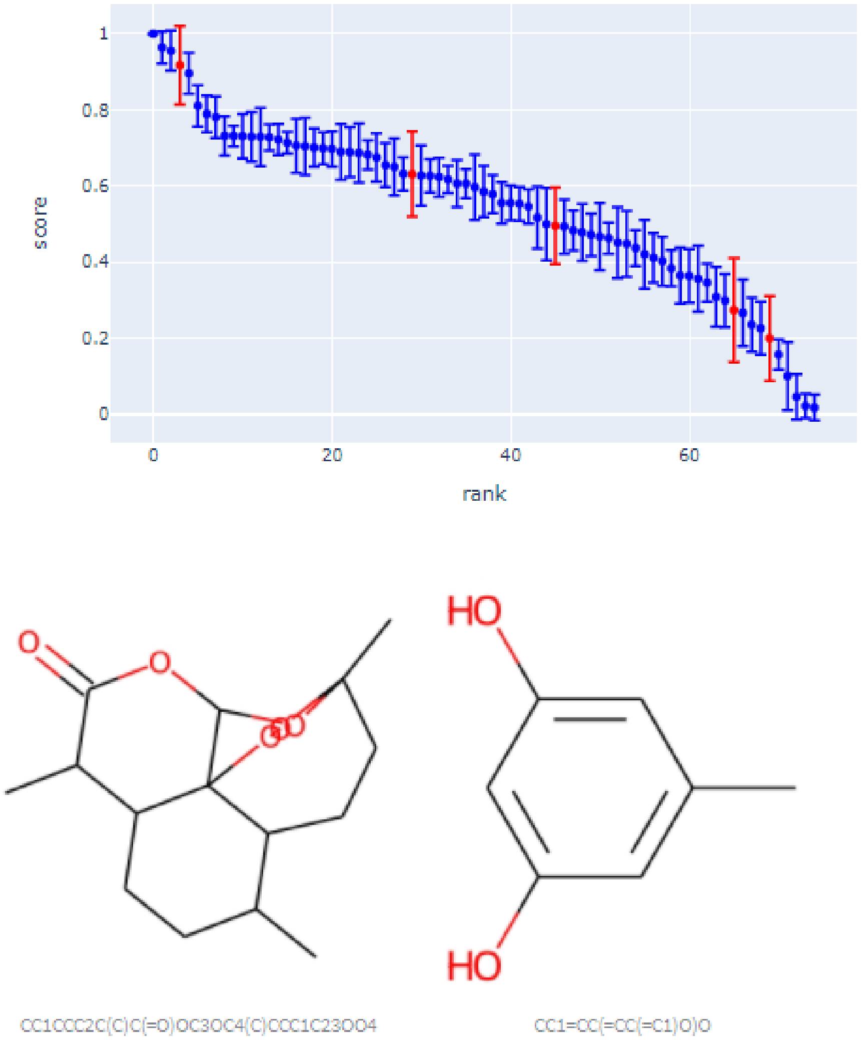


Figure S17. The ranking plot is shown for the user given molecular pairs (default dataset: artemisinin). The points with high uncertainty are colored in red. After clicking on each point the image of the molecules is printed on the main page.

References:

1 C. F. Macrae, I. Sovago, S. J. Cottrell, P. T. A. Galek, P. McCabe, E. Pidcock, M. Platings, G. P. Shields, J. S. Stevens, M. Towler and P. A. Wood, *Journal of Applied Crystallography*, 2020, **53**, 226–235.

2 T. Nematiaram, D. Padula, A. Landi and A. Troisi, *Advanced Functional Materials*, 2020, **30**, 2001906.

3 D. Widdowson, M. M. Mosca, A. Pulido, A. I. Cooper and V. Kurlin, *Match*, 2021, **87**, 529–559.

4 D. Widdowson and V. Kurlin, *arXiv:2108.04798*.

5 S. Karki, T. Friić, L. Fábián and W. Jones, *CrystEngComm*, 2010, **12**, 4038–4041.

6 T. Grecu, C. A. Hunter, E. J. Gardiner and J. F. McCabe, *Crystal Growth and Design*, 2014, **14**, 165–171.

7 D. Wang, Z. Yang, B. Zhu, X. Mei and X. Luo, *Crystal Growth and Design*, 2020, **20**, 6610–6621.

8 J. G. P. Wicker, L. M. Crowley, O. Robshaw, E. J. Little, S. P. Stokes, R. I. Cooper and S. E. Lawrence, *CrystEngComm*, 2017, **19**, 5336–5340.

9 L. K. Mapp, S. J. Coles and S. Aitipamula, *Crystal Growth and Design*, 2017, **17**, 163–174.

10 M. Przybyłek and P. Cysewski, *Crystal Growth and Design*, 2018, **18**, 3524–3534.

11 M. Przybyłek, T. Jeliński, J. Słabuszewska, D. Ziółkowska, K. Mroczyńska and P. Cysewski, *Crystal Growth and Design*, 2019, **19**, 3876–3887.

12 N. Sarkar, J. Mitra, M. Vittengl, L. Berndt and C. B. Aakeröy, *CrystEngComm*, 2020, **22**, 6776–6779.

13 M. Khalaji, M. J. Potrzebowski and M. K. Dudek, *Crystal Growth and Design*, 2021, **21**, 2301–2314.

14 A. Vriza, A. B. Canaj, R. Vismara, L. J. Kershaw Cook, T. D. Manning, M. W. Gaultois, P. A. Wood, V. Kurlin, N. Berry, M. S. Dyer and M. J. Rosseinsky, *Chemical Science*, 2021, **12**, 1702–1719.

15 J.-J. Devogelaer, M. D. Charpentier, A. Tijink, V. Dupray, G. Coquerel, K. Johnston, H. Meekes, P. Tinnemans, E. Vlieg, J. H. ter Horst and R. de Gelder, *Crystal Growth & Design*, 2021, **21**, 3428–3437.

16 D. Wu, B. Zhang, Q. Yao, B. Hou, L. Zhou, C. Xie, J. Gong, H. Hao and W. Chen, *Crystal Growth and Design*, 2021, **21**, 4531–4546.

17 W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande and J. Leskovec, *arXiv:1905.12265*.

18 T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest and A. Rush, in *arXiv preprint arXiv:1910.03771*, 2020, pp. 38–45.

19 M. P. and É. D. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, *Scikit-learn: Machine Learning in Python*, 2011, vol. 12.