Leveraging Deep Chemical Language Processing for Co-crystal Prediction

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

[1]Institute for Complex Molecular Systems (ICMS), Department of Biomedical Engineering, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands.
[2]Department of Chemistry and NIS Centre, University of Torino, Via P. Giuria 7, 10125 Torino, Italy.
[3]Research and Early Development, Dompé Farmaceutici S.p.A, Via Campo di Pile, s.n.c., 67100, L'Aquila, Italy
[4]Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Princetonlaan 6, 3584 CB, Utrecht, The Netherlands.
[*]e-mail: f.grisoni@tue.nl

Abstract

More than 40% of marketed drugs, including new chemical entities, suffer from low aqueous solubility. Enhancing the solubility profile of these molecules, without altering their chemical identity or pharmacological activity, is achievable through co-crystallization, a process wherein the drug and another organic molecule coexist in the crystal structure. However, finding the most promising combination of molecules for co-crystallization is challenging, as well as time-consuming and source-intensive, due to the vast search space. To overcome this limitation and rationally design experimental trials, we propose DeepCrystal, a deep learning model based on chemical language. DeepCrystal is rigorously validated, achieving a balanced accuracy of 78% on the external test set, as well as superior performance to existing models. In addition, thanks to chemical language for molecule representation, we estimate the uncertainty of the model to gauge its reliability for future applications. Finally, DeepCrystal is successfully employed to discover novel diflunisal co-crystals, highlighting its potential, in both academic and industrial settings, for the design of new pharmaceutical formulations.

1 Introduction

Co-crystallization [5] is a well-established technique to enhance the solubility, stability, and processability of active pharmaceutical ingredients (APIs) [2]. In this approach, another molecule, called coformer, is searched to form a multicomponent crystal based on hydrogen bonding interacrions with the API, so that physicochemical properties of the API are optimized while its activity is preserved [7, 20]. However, due to the thousands of potentially available molecules, finding the optimal coformer is a labor-intensive and time-consuming process based on trial and error [3,8].

Machine learning models have been developed to

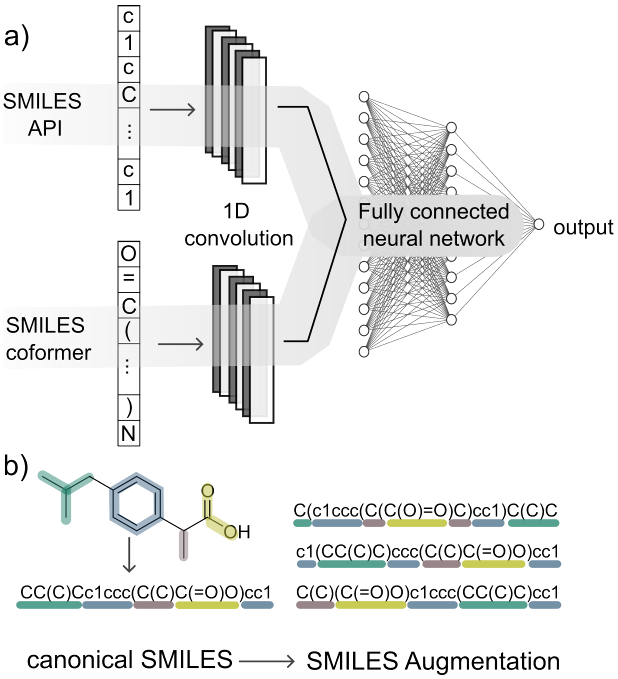
a)
SMILES API
c
1
c
C
⋮
c
1
1D convolution
Fully connected neural network
output
SMILES coformer
O
=
C
(
⋮
)
N
b)
OH
CC(C)Cc1ccc(C(C)C(=O)O)cc1
C(c1ccc(C(C(O)=O)C)cc1)C(C)C
c1(CC(C)C)ccc(C(C)C(=O)O)cc1
C(C)(C(=O)O)c1ccc(CC(C)C)cc1
canonical SMILES ⟶ SMILES Augmentation

Figure 1: *DeepCrystal* **a)** architecture: API and coformer are represented by SMILES strings; after passing via the convolutional layers, the learned representations are then concatenated and passed through fully connected layers for the prediction of the outputs. b) SMILES augmentation: the same molecule is represented by n diverse SMILES strings, randomly generated.

rapidly identify promising API-coformer pairs in the vast search space [13,17]. The early approaches adopt either fully connected neural networks or tree-based models on molecular descriptors to predict the co-crystallization of molecule pairs [21, 22, 24]. More recent works introduced deep learning to the task and utilized convolutions over molecular graphs [10, 23] to improve the predictive performance. Common to those models, they are trained on unbalanced datasets with folds of more API-coformer co-crystals than negative pairs, due to limited data availability, and struggle to generalize unseen data [21]. Random sampling of molecule pairs as negative API-coformer mixtures are introduced to target this problem [6], however, this approach comes at the cost of potentially mislabeled data. Therefore new approaches that are more robust to class imbalance and have stronger generalizability to unseen data are needed.

Here we propose DeepCrystal, a novel deep learning approach to predict co-crystallization with higher generalizability. DeepCrystal utilizes string representation of molecules, the so-called chemical language [15]. Leveraging the chemical language, DeepCrystal can use “SMILES augmentation” – the technique of representing the same molecule with multiple strings – to target class imbalance and to improve its generalizability. The evaluation of DeepCrystal via internal and external test sets shows its superior performance to the models in the literature and its stronger generalizability. Ablation studies reveal that the performance boost is thanks to using chemical language and SMILES augmentation adopted by DeepCrystal. Furthermore, SMILES augmentation equips DeepCrystal with a simple and effective uncertainty estimation approach.

Last, we use DeepCrystal challenging prospective study: we screened four structurally similar coformer candidates for diflunisal, a widely used anti-inflammatory drug. Despite the high structure similarity of the coformers, DeepCrystal accurately identified the co-cocrystals thanks to the uncertainty estimation module, as validated by the experimental screening. This prospective study underlines DeepCrystal applicability in a realistic setting to accelerate the coformer search in the co-crystallization workflow.

2 Results and Discussion

2.1 Overview of DeepCrystal

DeepCrystal is a chemical language-based deep learning architecture to predict the co-crystallization of API-coformer pairs. Specifically, DeepCrystal consists of three blocks: (a) it represents the input molecules (API and coformers) via SMILES strings, (b) it learns 'latent representations' of the molecular structures via a convolutional neural network (CNN), and (c) a fully-connected neural network is used to predict the potential co-crystallization of the input pair (Figure 1a). DeepCrystal was trained on a dataset collected from the Cambridge Structural Database, literature, and in-house experiments. The dataset contains different types of co-crystal systems (pharmaceutical, π–π, and energetic), collecting 5240 co-crystals ("positives") and 1392 physical mixtures ("negatives", *i.e.,* no observed co-crystallization), thereby presenting the class imbalance typical to co-crystallization prediction.

DeepCrystal is trained, validated, and tested on stratified splits of this dataset (10 randomly sampled subsets with 10% molecules). As in existing studies [9, 10], every API-coformer pair is presented to the network twice, by switching the order of the inputs. The number of negative and positive pairs is balanced via SMILES augmentation, leveraging the power of chemical language (Figure 1b).

2.2 Performance of DeepCrystal

We first analyze the effect of SMILES augmentation by comparing the model trained on augmented data to the model trained only on canonical SMILES (Table 1). Both models reached an average balanced accuracy (BAcc) [1] above 88%. Remarkably, increasing the ratio of negative samples in the dataset via SMILES augmentation improves the capacity of DeepCrystal to identify negative pairs by 8% on average, as quantified by the specificity. A Wilcoxon signed-rank test (p-value = 0.0108) also validates the significance of the increase in the metric. This comparison suggests that SMILES augmentation improves predictive performance for the negative data class, appreciable even with a small number of negative samples (525 positive/139 negative) resulting from stratified partitioning of the entire dataset.

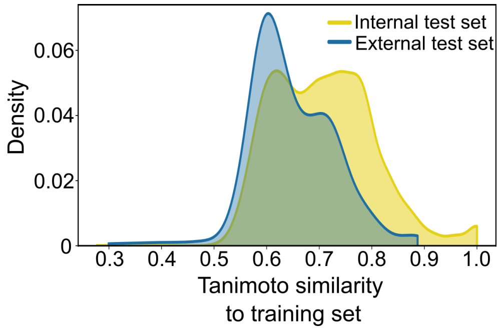

Figure 2: *Tanimoto similarity* to the training set computed for internal (yellow) and external (blue) test sets.

Having a particular interest in pharmaceutical co-crystals (mainly anti-inflammatory, anti-tubercular, nootropic, and anti-depressant drugs), we generate a new set of API-coformer pairs and evaluate the applicability domain of DeepCrystal. This external test set is composed of 364 pairs, of which 129 are co-crystals and 235 non-co-crystals. The lower similarity of molecule pairs in this set compared to the previous one (Tanimoto similarity computed over extended connectivity fingerprints with radius=2 and nBits=1024) suggests that the external test presents a more challenging and realistic setting to evaluate generalizability (Figure 2).

Using the external test set, we compare DeepCrystal to three existing approaches in the literature: (i) a fully-connected neural network trained on extended connectivity fingerprints [4, 6]; (ii) a fully-connected neural network trained on molecular descriptors [14]; and (iii) CCGNet, a graph neural network available in the literature [9]. CCGNet is trained on an unbalanced and non-augmented dataset (6819 positive and 1052 negative) and publicly available, while the first two models are trained on the same DeepCrystal dataset. The released model of CCGNet has a reported 100% accuracy for independent test sets of nicotinamide, carbamazepine, and paracetamol, but is suspected of test-set data leakage [11].

Table 1: *Performance of DeepCrystal* on the internal test sets (664 molecular pairs obtained from stratified split). DeepCrystal trained on canonical SMILES and augmented SMILES is evaluated. DeepCrystal, CCGNet, as well as DNN models based on ECFPs or classical molecular descriptors were compared on the external test set (364 molecular pairs). Balanced accuracy, recall and specificity are reported. The best performances per metric are highlighted in boldface for each test set.

Test set	Model	BAcc	Recall	Specificity
Internal	DeepCrystal - canonical	0.88 ± 0.02	**0.96 ± 0.01**	0.79 ± 0.06
	DeepCrystal - augmented	**0.89 ± 0.02**	0.92 ± 0.02	**0.87 ± 0.03**
External	CCGNet [9]	0.60	0.51	0.69
	Fingerprint-DNN	0.57	0.90	0.25
	Descriptors-DNN	0.63	0.84	0.41
	DeepCrystal - canonical	0.59	**0.93**	0.26
	DeepCrystal - augmented	**0.78**	0.75	**0.81**

The results (Table 1) indicate the strong performance of DeepCrystal over the benchmarks. DeepCrystal trained with data augmentation achieves 15%-21% higher balanced accuracy than the benchmarks and 12%-56% higher specificity, at the cost of a lower recall by 15%, in the worst case. These results show that DeepCrystal finds a better tradeoff between positive and negative prediction power than the models in the literature. Furthermore, the SMILES augmentation increases the balanced accuracy by 19% (compared to using canonical SMILES strings) in this challenging setting, validating the previous results and demonstrating its benefits for better generalizability.

2.3 Uncertainty Estimation with DeepCrystal

Interested in estimating the uncertainty in model predictions, we develop a pipeline on top of DeepCrystal using test-time SMILES augmentation. Specifically, (i) we represent API-coformer pairs in the test via ten different SMILES strings, (ii) predict the cocrystallization of each representation, and (iii) compute the average and standard deviation across predictions per molecule pair.

To qualitatively evaluate the performance of this uncertainty estimation pipeline on our external test set, we visualize the prediction average per true neg-

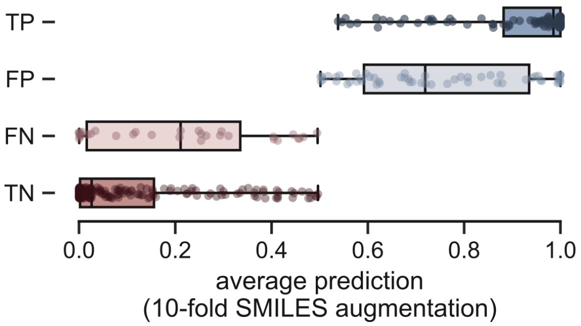

Figure 3: *Uncertainty of the model evaluation.* The box plot depicts the average prediction on external test set values, derived from 10-fold SMILES augmentation results for each API-coformer pair.

Table 2: *Metrics vs standard deviation.* Predictive performance of DeepCrystal on the external test set, calculated decreasing standard deviation across the ten times SMILES augmentation predictions. N° = number of samples on which the metrics are calculated.

Standard dev.	N°	BAcc	Recall	Spec.
≤ 0.50	364	0.76	0.75	0.77
≤ 0.40	351	0.77	0.76	0.78
≤ 0.30	275	0.82	0.80	0.83
≤ 0.20	227	0.86	0.85	0.87
≤ 0.10	191	0.88	0.86	0.90
≤ 0.05	161	**0.88**	**0.84**	**0.91**

atives, false positives, false negatives, and true negatives (Figure 3), by considering mean predictions over 0.5 as a positive prediction. The results show that true positives cluster more toward the high prediction average regions and have a higher median than false positives, suggesting that correct positive predictions have high prediction value across randomized API-coformer representations. Similar results were observed for false negatives and true negatives, where true negatives had a lower median of average prediction. Taken together, these results suggest that mean prediction across ten SMILES representations per molecule pair is an indicator for model accuracy downstream.

As uncertainty parameter, we compute balanced accuracy, recall, and specificity evaluating the standard deviation across the ten predictions for each molecule pair (Table 2). The results show an increasing trend across the metrics, narrowing the analysis to samples with a lower standard deviation. The balance accuracy increases from 76% to 88% as the standard deviation decreases from ± 0.50 to ± 0.05, a remarkable gap can be noted when considering samples showing a standard deviation below or above ± 0.20.

The same trend emerges computing metrics per number of majority class predictions, increasing performance as one prediction dominates the other across ten predictions. Again, the balanced accuracy increases from 76% to 87% as majority class predictions increase from five, a tie between negative and positive, to ten, where models predict the same class across all SMILES representations of the molecule pair. The uncertainty, measured by standard deviation, as well as the consistent prediction values, highlight that using SMILES augmentation at test time unlocks an estimation of prediction accuracy and further advocates for the adoption of DeepCrystal, especially in prospective settings where such an estimate can lower misses in experiments and save costs.

2.4 A Prospective Study with DeepCrystal

Motivated by the success of DeepCrystal in predicting co-crystallization and presenting an uncertainty estimate of its predictions, we ask the following question: *Can DeepCrystal empower a prospective co-crystallization study*?

To answer, we select diflunisal (DIF) as the target API for co-crystallization screening. The anti-inflammatory activity of this molecule is limited by its poor water solubility and co-crystallization can succeed in improving its bioavailability [18]. Coformer are sought among natural compounds based on purines moiety because they are co-administrable substance with several health benefits (central nervous systems stimulants, risk reduction of neurodegenerative diseases, and anti-inflammatory properties [12, 16].

Previous research reports the co-crystallization of DIF with theophylline [19]. For stress-testing the model and estimating whether DeepCrystal can distinguish between subtle chemical structural differences, we create a library of four molecules with the same scaffold as theophylline: theobromine (TBR), xanthine (XAN), caffeine (CAF), and adenine (ADE) (Figure 4).

Next, we compared DeepCrystal trained on canonical SMILES with DeepCrystal trained on augmented SMILES to predict the co-crystallization between DIF and the four selected molecules. Employing DeepCrystal augmentated, we use ten-fold test-time augmentation for each pair and compute mean pre-

Table 3: *DeepCrystal prediction* for DIF-CAF, DIF-ADE, DIF-TBR, and DIF-XAN. The columns Deep-Crystal (canonical) and Deep-Crystal (augmentated) refer to the respectively models' prediction. The 'Deep-Crystal augmented' outputs are reported consider the average and standard deviation across ten different prediction computed via test time SMILES augmentation.

Sample	DeepCrystal (canonical)	DeepCrystal (augmented)	Consistent predictions	lab result
DIF-CAF	0.99	0.99 ± 0.01	100%	co-crystal
DIF-ADE	0.99	0.99 ± 0.00	100%	co-crystal
DIF-TBR	0.99	0.66 ± 0.35	60%	physical mixture
DIF-XAN	0.99	0.63 ± 0.38	60%	physical mixture

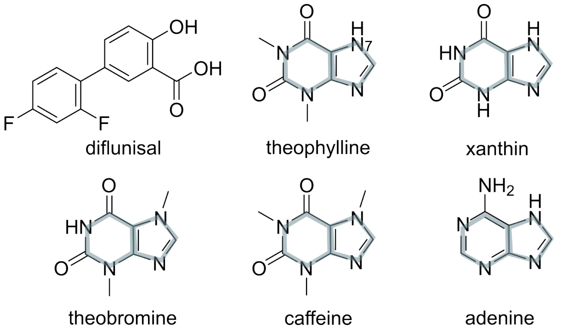

Figure 4: *Case study DIF-purines.* Chemical structure of diflunisal and purines investigated as possible coformers for cocrystallization.

diction and standard deviation across representations (Table 3). DeepCrystal canonical is unable to discriminate among the four candidates, predicting each system as a co-crystal. Differently, DeepCrystal augmentated predicts DIF-CAF and DIF-ADE as co-crystals across all ten representations (with an average probability of 99while TBR and XAN are predicted to co-crystalize six times with an average probability below 66% and a high standard deviation. These results indicate that DeepCrystal is "certain" that CAF and ADE would co-crystallize with DIF, while predictions for TBR and XAN carry more uncertainty.

Moving to the lab, a comprehensive experimental screening is conducted, involving grinding, liquid-assisted grinding, and slurry methods, as these are the most commonly employed techniques for achieving co-crystallization. Each DIF-coformer pair is tested using all three techniques, with variations in experimental conditions such as time, quantity, and the polarity of the solvent added during the liquid-assisted grinding and slurry procedures. Co-crystal formation occurs only for DIF-CAF and DIF-ADE systems, achieved through liquid-assisted grinding in ethanol and slurry in ethanol, respectively. Conversely, TBR and XAN results in physical mixtures with DIF in all trials. To distinguish between co-crystal and non-co-crystal formations, the obtained powder samples are analyzed using infrared spectroscopy, powder X-ray diffraction, and solid-state nuclear magnetic resonance.

The lab results validate the importance and accuracy of uncertainty estimations by DeepCrystal. While CAF and ADE, the compounds predicted to co-crystallize with high certainty, form co-crystals with DIF, whereas TBR and XAN do not. Thanks to using chemical language and test-time augmentation, DeepCrystal prioritized the correct coformers for DIF among four structurally similar candidates, demonstrating the applicability of DeepCrystal in a prospective setting.

3 Conclusions

Optimizing pharmacokinetic properties of active compounds is an ever-lasting challenge in drug discovery and co-crystallization is a useful tool to target this problem. Yet, finding co-crystallization partners to active compounds is resource-intensive, and *in* *silico* approaches to accelerate the search are called for. Here we developed DeepCrystal, a deep learning model that can predict the co-crystallization of any two compounds with unmatched accuracy.

DeepCrystal owes its state-of-the-art performance to the first-time adoption of the chemical language to represent molecules. Chemical language unlocks the use of SMILES augmentation – the technique of representing the same molecule with different strings – to target class imbalance in the training dataset, a pertinent challenge in training co-crystallization models. Our experiments show that SMILES augmentation equips DeepCrystal with stronger generalizability to molecules dissimilar to the training coumpounds.

Chemical language and SMILES augmentation add another capability to DeepCrystal, that is estimating the uncertainty of its predictions. Predicting the co-crystallization of test set pairs across multiple representations, DeepCrystal presents an uncertainty estimate per prediction, which we show to correlate with the performance downstream. We then use DeepCrystal prospectively and find two novel co-crystals of diflunisal among structurally similar coformer candidates.

We see DeepCrystal as a step to accelerate the co-crystallization workflows, reducing time and lab experiments to develop promising new drug formulations. The first-time adoption of chemical language for the task has demonstrated non-disputable benefits and is likely to trigger further research in the same direction. We believe that DeepCrystal will be a stepping stone for such research, in the big goal of speeding up drug discovery.

References

[1] Davide Ballabio, Francesca Grisoni, and Roberto Todeschini. Multivariate comparison of classification performance measures. *Chemometrics and Intelligent Laboratory Systems*, 174:33–44, 2018.

[2] Meenakshi Bhatia and Sunita Devi. Co-crystallization: a green approach for the solubility enhancement of poorly soluble drugs. *CrystEngComm*, 26(3):293–311, 2024.

[3] Chiara Cappuccino, David Cusack, James Flanagan, Carl Harrison, Cillian Holohan, Monica Lestari, Gareth Walsh, and Matteo Lusi. How many cocrystals are we missing? assessing two crystal engineering approaches to pharmaceutical cocrystal screening. *Crystal Growth & Design*, 22(2):1390–1397, 2022.

[4] Jiahui Chen, Zhihui Li, Yanlei Kang, and Zhong Li. Cocrystal prediction based on deep forest model—a case study of febuxostat. *Crystals*, 14(4):313, 2024.

[5] Gautam R Desiraju. Supramolecular synthons in crystal engineering—a new organic synthesis. *Angewandte Chemie International Edition in English*, 34(21):2311–2327, 1995.

[6] Jan-Joris Devogelaer, Hugo Meekes, Paul Tinnemans, Elias Vlieg, and Rene De Gelder. Co-crystal prediction by artificial neural networks. *Angewandte Chemie International Edition*, 59(48):21711–21718, 2020.

[7] Naga K Duggirala, Miranda L Perry, Örn Almarsson, and Michael J Zaworotko. Pharmaceutical cocrystals: along the path to improved medicines. *Chemical communications*, 52(4):640–655, 2016.

[8] Molly M Haskins and Michael J Zaworotko. Screening and preparation of cocrystals: A comparative study of mechanochemistry vs slurry methods. *Crystal Growth & Design*, 21(7):4141–4150, 2021.

[9] Yuanyuan Jiang, Zongwei Yang, Jiali Guo, Hongzhen Li, Yijing Liu, Yanzhi Guo, Menglong Li, and Xuemei Pu. Coupling complementary strategy to flexible graph neural network for quick discovery of coformer in diverse co-crystal materials. *Nature Communications*, 12(1):5950, 2021.

[10] Yanlei Kang, Jiahui Chen, Xiurong Hu, Yunliang Jiang, and Zhong Li. A cocrystal prediction method of graph neural networks based on molecular spatial information and global attention. *CrystEngComm*, 25(46):6405–6415, 2023.

[11] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 2023.

[12] Eva Martínez-Pinilla, Ainhoa Oñatibia-Astibia, and Rafael Franco. The relevance of theobromine for the beneficial effects of cocoa consumption. *Frontiers in pharmacology*, 6:126866, 2015.

[13] Fateme Molajafari, Tianrui Li, Mehrnaz Abbasichaleshtori, Moein Hajian ZD, Anthony F Cozzolino, Daniel R Fandrick, and Joshua D Howe. Computational screening for prediction of co-crystals: method comparison and experimental validation. *CrystEngComm*, 2024.

[14] Medard Edmund Mswahili, Min-Jeong Lee, Gati Lother Martin, Junghyun Kim, Paul Kim, Guang J Choi, and Young-Seob Jeong. Cocrystal prediction using machine learning models and descriptors. *Applied Sciences*, 11(3):1323, 2021.

[15] Hakime Öztürk, Arzucan Özgür, Philippe Schwaller, Teodoro Laino, and Elif Ozkirimli. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug Discovery Today*, 25(4):689–705, 2020.

[16] Kamil Rodak, Izabela Kokot, and Ewa Maria Kratz. Caffeine as a factor influencing the functioning of the human body—friend or foe? *Nutrients*, 13(9):3088, 2021.

[17] Nandini Sarkar, Nina C Gonnella, Mariusz Krawiec, Dongyue Xin, and Christer 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*, 20(11):7320–7327, 2020.

[18] Petr Snetkov, Svetlana Morozkina, Roman Olekhnovich, and Mayya Uspenskaya. Diflunisal targeted delivery systems: A review. *Materials*, 14(21):6687, 2021.

[19] Artem O Surov, Alexander P Voronin, Alex N Manin, Nikolay G Manin, Lyudmila G Kuzmina, Andrei V Churakov, and German L Perlovich. Pharmaceutical cocrystals of diflunisal and diclofenac with theophylline. *Molecular pharmaceutics*, 11(10):3707–3715, 2014.

[20] Abdul Raheem Thayyil, Thimmasetty Juturu, Shashank Nayak, and Shwetha Kamath. Pharmaceutical co-crystallization: Regulatory aspects, design, characterization, and applications. *Advanced Pharmaceutical Bulletin*, 10(2):203, 2020.

[21] Carolina von Essen and David Luedeker. In silico co-crystal design: assessment of the latest advances. *Drug Discovery Today*, page 103763, 2023.

[22] Dingyan Wang, Zeen Yang, Bingqing Zhu, Xuefeng Mei, and Xiaomin Luo. Machine-learning-guided cocrystal prediction based on large data base. *Crystal Growth & Design*, 20(10):6610–6621, 2020.

[23] Fu Xiao, Yinxiang Cheng, Jian-Rong Wang, Dingyan Wang, Yuanyuan Zhang, Kaixian Chen, Xuefeng Mei, and Xiaomin Luo. Cocrystal prediction of bexarotene by graph convolution network and bioavailability improvement. *Pharmaceutics*, 14(10):2198, 2022.

[24] Dezhi Yang, Li Wang, Penghui Yuan, Qi An, Bin Su, Mingchao Yu, Ting Chen, Kun Hu, Li Zhang, Yang Lu, et al. Cocrystal virtual screening based on the xgboost machine learning model. *Chinese Chemical Letters*, 34(8):107964, 2023.