SYNTHON-CGR: CO-CRYSTAL FORMATION PREDICTION WITH CRYSTAL-SUPERVISED SYNTHON REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Co-crystallization offers a practical way to tune the physicochemical properties of molecular compounds without changing their covalent structure, but finding compatible molecular pairs still requires substantial experimental effort. Most machine-learning approaches treat this problem as binary classification and are trained only on pair-level formation labels. As a result, they provide limited information about the intermolecular motifs that support a predicted co-crystal. We introduce **Synthon Candidate Graph Reasoner (Synthon-CGR)**, a graph neural network that predicts co-crystal formation together with a ranked set of candidate supramolecular synthons. The model combines molecular graph representations, functional motifs, and crystal-derived supervision so that intermolecular interactions become explicit prediction targets rather than post-hoc explanations. We curate literature and crystallographic data, augment the training pool to 27,718 molecular pairs to address the scarcity of reported negatives, and derive synthon labels from resolved crystal structures. On a pair-disjoint held-out test set, Synthon-CGR reaches F1 = 0.893 for co-crystal formation and HitRate@5 = 0.935 on the synthon-ranking subset. We also evaluate the model on an independent dataset obtained from co-crystallization experiments conducted in our laboratory and compare predicted interaction motifs with expert annotations for experimentally confirmed co-crystals. Together, these experiments show that crystal-derived interaction supervision can extend pair-level co-crystal prediction with explicit and experimentally testable hypotheses about molecular recognition.

1 INTRODUCTION

Co-crystallization provides a way to modify the physicochemical properties of molecular compounds without changing their covalent structure (Bolla et al., 2022). In pharmaceutical development, co-crystals can improve solubility, stability, mechanical properties, and bioavailability (Guo et al., 2021). The practical challenge is that identifying a suitable coformer is still largely an experimental search problem, often requiring substantial screening effort (Sugden et al., 2022; Karimi-Jafari et al., 2018). This has motivated a growing body of data-driven methods that predict co-crystal formation directly from molecular pairs (Wang et al., 2020; Jiang et al., 2021; Birolo et al., 2025).

A persistent limitation of this setting is the form of the available supervision. Most datasets record only whether a pair formed a co-crystal, while unsuccessful crystallization experiments are reported much less frequently (Wicker et al., 2017; Xiao et al., 2022). The resulting data are therefore biased toward positive examples and describe the outcome of co-crystallization more readily than the interactions that produced it. A binary label can tell a model that two molecules form a co-crystal, but it does not identify which functional groups interact or which intermolecular motif is realized in the solid state. This distinction matters if predictions are intended to support chemical reasoning rather than only rank candidate pairs.

Resolved crystal structures provide a natural source of additional supervision. They expose the intermolecular contacts present in successful co-crystals and connect directly to the language of supramolecular synthons used in crystal engineering (Desiraju, 1995; Galek et al., 2007; Sandhu et al., 2021). Existing learning-based predictors have introduced attention maps, fragment rele-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

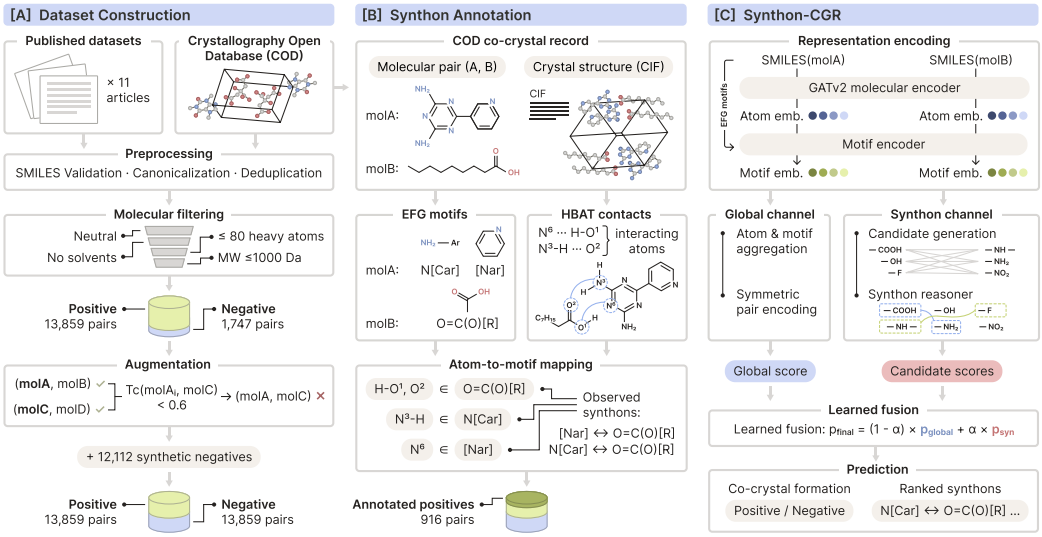

Figure 1: Overview of Synthon-CGR. Pair-level formation data and crystal-derived synthon annotations supervise a model that combines global molecular representations with contextual reasoning over candidate synthons.

vance, and other forms of post-hoc attribution, but these explanations are typically derived after the formation decision. They do not directly train the model to identify an intermolecular motif that can be compared with a crystal structure. We ask whether crystal-derived interaction information can instead be incorporated into the learning objective itself.

We address this question with **Synthon Candidate Graph Reasoner (Synthon-CGR)**. The model represents both coformers as molecular graphs, identifies functional motifs, constructs candidate synthons between the two molecules, and reasons over those candidates in the context of the complete pair. Training combines ordinary pair-level labels with synthon annotations extracted from crystal structures, allowing the same model to return a co-crystal formation probability and a ranked set of interaction hypotheses. Figure 1 summarizes the resulting pipeline.

Our main contributions are:

1. We assemble and curate a co-crystal dataset that combines published formation data with crystallographic records. Similarity-filtered negative augmentation is applied only to the training pool, preserving the held-out evaluation data.

2. We derive crystal-based synthon supervision by mapping observed intermolecular contacts to pairs of functional motifs. This provides an interaction-level learning signal in addition to the usual pair-level formation labels.

3. We introduce Synthon-CGR, which combines a global representation of the molecular pair with contextual reasoning over candidate synthons. The model jointly predicts co-crystal formation and ranks explicit intermolecular interaction hypotheses.

4. We evaluate the method on both a common held-out test set and an independent experimental validation dataset obtained through co-crystallization experiments in our laboratory. For experimentally confirmed co-crystals, we further compare the predicted synthons with expert-annotated interactions in the corresponding crystal structures.

Code and accompanying scripts will be released at `https://anonymous.4open.science/r/Synthon-CGR-D8E3`.

2 Related Work

2.1 Co-crystal Formation Prediction

Co-crystal prediction has developed from knowledge-based criteria (Fábián, 2009; Wood et al., 2014; Sarkar et al., 2020) toward data-driven models based on molecular descriptors and fingerprints (Wicker et al., 2017; Wang et al., 2020), molecular graphs (Jiang et al., 2021; Xiao et al., 2022), three-dimensional representations (Zheng et al., 2023; Kang et al., 2023; Ahmadi et al., 2024), SMILES representations (Birolo et al., 2025), and hybrid graph-descriptor models (Ghanavati & Rohani, 2025). These studies show that molecular representations can support accurate pair-level prediction, but the underlying data remain difficult. Negative crystallization outcomes are sparsely reported (Wicker et al., 2017), synthetic negatives are therefore often introduced during training (Wang et al., 2020; Xiao et al., 2022), and performance can drop when models are evaluated outside the chemical space represented in the training data (Birolo et al., 2025). Synthon-CGR retains the same pair-level prediction task, but augments it with supervision over the intermolecular motifs observed in resolved crystal structures.

2.2 Supramolecular Interaction Modeling

Supramolecular synthons describe recurring intermolecular motifs associated with preferred noncovalent recognition in molecular crystals (Desiraju, 1995). Computational crystal-engineering methods characterize these interactions using CSD-derived propensities and coordination statistics (Galek et al., 2007; 2014), interaction maps (Wood et al., 2012), electrostatic analysis (Wang et al., 2019), and synthon-guided crystal-structure prediction (Sarma & Desiraju, 2002; Momenzadeh Abardeh et al., 2025). Identifying the dominant motif remains difficult when several recognition sites or interaction types compete (Sandhu et al., 2021; Cruz-Cabeza et al., 2024). More recent co-crystal prediction models provide atom-level attention, fragment relevance, or cross-attention analyses (Jiang et al., 2021; Birolo et al., 2025; Liu et al., 2026b;a). These approaches improve interpretability, but they still primarily explain a pair-level decision. Synthon-CGR instead treats the intermolecular motif as a supervised prediction target and evaluates it directly against crystal-derived interaction labels.

3 Data and Crystal-Derived Supervision

3.1 Co-crystal Pair Dataset

Data collection. We aggregate publicly available co-crystal formation data from prior machine-learning studies and supplement them with 1,369 co-crystal records from the Crystallography Open Database (COD). In total, 11 literature sources were used; the full list of sources with the corresponding number of molecular pairs is provided in Table 5 (Appendix A.3).

Curation and standardization. Molecular structures are standardized with RDKit by parsing and sanitizing SMILES, canonicalizing valid structures, and removing records that do not satisfy the molecular and pair-level quality criteria. The curation step excludes invalid or disconnected structures, simple solvents and inorganic ions, charged species, excessively large molecules, identical-component pairs, duplicate pairs, and records with contradictory labels. Additional details on molecular filtering and pair-level standardization are provided in Appendices A.1 and A.2. After curation, the dataset contains 15,606 unique molecular pairs, including 13,859 positives and 1,747 negatives. Molecular property distributions before final filtering are shown in Appendix A.5, while additional composition and stoichiometry statistics are reported in Appendix A.6.

Negative-pair augmentation. The scarcity of reported negative experiments creates a strong imbalance in the curated data. To reduce this imbalance without modifying the held-out evaluation set, we generate synthetic negatives only within the training pool. For a molecule from a known positive pair, candidate partners are sampled from other positive pairs and retained only when their ECFP4 Tanimoto similarity to every known coformer of that molecule is below 0.6 (Wang et al., 2020). The similarity filter is intended to avoid treating close analogues of known coformers as

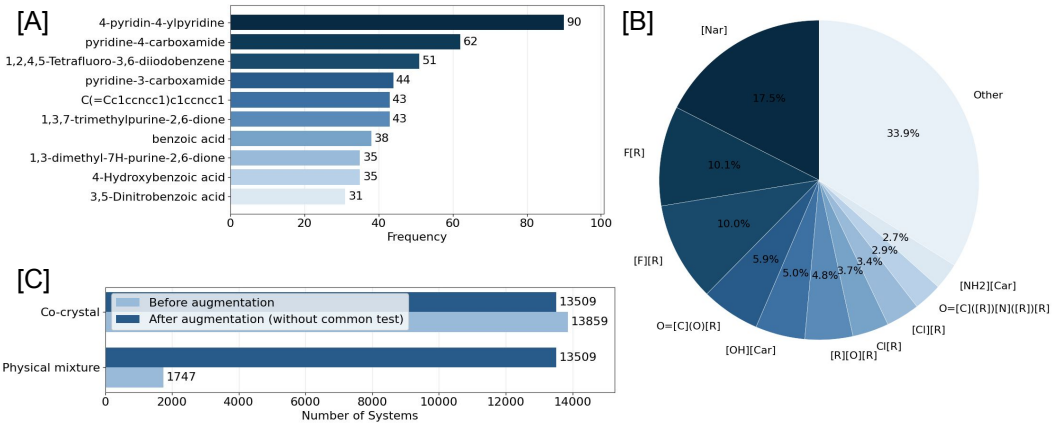

Figure 2: Overview of the curated data: frequent molecular components, functional-motif distribution in the crystal-structure subset, and class balance before and after training-set augmentation.

obvious negatives. The test split is created before augmentation, and the full procedure is described in Appendix A.4.

3.2 Crystal Structures and Synthon Annotation

Crystal-structure retrieval. To obtain interaction-level supervision, co-crystal CIF files are retrieved from COD and matched to molecular pairs using molecular formulae and canonicalized structures. This procedure yields 2,315 crystal structures associated with pairs in the curated dataset. The crystal structures are not used as model inputs at inference time. Their role is to reveal which intermolecular contacts were realized experimentally and to convert those contacts into labels over candidate synthons.

Functional motifs and candidate synthons. Functional motifs are detected with the Ertl algorithm implemented in EFGs (Colmenarejo, 2025). Each motif is represented by its pseudo-SMILES type and the atoms that belong to it. For every molecular pair, candidate synthons are formed by pairing motifs from opposite components. Each candidate is also assigned a statistical interaction prior based on how often the corresponding motif pair occurs in crystal structures. This construction turns an unconstrained atom-level interaction problem into a finite set of chemically defined alternatives that can be ranked by the model. Across the curated data, the two components contain on average 2.9 and 2.6 EFGs motifs, respectively. This produces an average of 7.5 candidate synthons per molecular pair, with a median of 6 and a maximum of 171. No retained candidates are available for 6.3% of pairs. Further statistics are reported in Appendix E.2.

HBAT-derived synthon supervision. Hydrogen and halogen bonds are extracted from the crystal structures with HBAT 2 and mapped from the interacting donor and acceptor atoms to the corresponding EFGs motifs. The resulting motif pairs define binary labels over the candidate set. In this way, a molecular pair can carry two levels of supervision: the pair-level label indicates whether a co-crystal forms, while the synthon mask identifies which candidate interactions are observed in the resolved structure. HBAT analysis of 1,970 usable crystal structures identifies 5,336 intermolecular interactions, including 4,951 hydrogen bonds (92.8%) and 385 halogen bonds (7.2%). On average, 2.71 interactions are detected per molecular pair.

Reliable non-empty masks are available only for a subset of the crystal structures. Out of 1,970 usable crystal structures, 916 molecular pairs yield non-empty synthon masks that can be unambiguously mapped to a unique RDKit atom and EFGs motif. Symmetry-equivalent molecular copies and internally symmetric atoms can make it impossible to map a crystallographic interaction to a unique RDKit atom and EFGs motif. Rather than assign uncertain labels, we exclude these ambiguous cases from direct synthon supervision. The complete mapping procedure, quality controls, and common failure modes are described in Appendix B.3 and Appendix B.4.

4 SYNTHON-CGR

4.1 MOLECULAR AND MOTIF ENCODING

Synthon-CGR takes two molecular graphs as input and encodes them independently with a GATv2 backbone. GATv2 was selected from GCN, GAT, GATv2, GIN, and GINE encoders in the preliminary comparison reported in Appendix E.1. Each EFGs motif is then represented by pooling the embeddings of its constituent atoms and combining the pooled representation with a learnable motif-type embedding, an atom-index embedding, and a scalar motif-size feature. This gives the model two complementary views of each molecule. The pooled graph embedding captures the overall molecular context, while the motif embeddings preserve the local chemical environments that define possible intermolecular interactions.

4.2 CANDIDATE SYNTHON CONSTRUCTION

For a molecular pair (A, B), let h_i and h_j denote motif embeddings from the two components. A candidate synthon $c = (i, j)$ is encoded from the two motif representations, simple symmetric pairwise operations, and a small set of interaction-level scalar features:

$$h_c = \mathrm{MLP}\left([h_i; h_j; |h_i - h_j|; h_i \odot h_j; s_c]\right), \tag{1}$$

where s_c contains the statistical interaction prior, a halogen-bond indicator, donor and acceptor scores for both motifs, and normalized motif sizes. The concatenated motif embeddings preserve the identity of the two groups. The absolute difference and element-wise product provide compact descriptions of their compatibility. Appendix C.1 gives the complete feature definition.

4.3 CONTEXTUAL SYNTHON REASONING

Candidate interactions within the same molecular pair are not independent. Several donor and acceptor sites can compete, and the plausibility of one interaction depends on what alternatives are available elsewhere in the pair. Synthon-CGR therefore processes the full candidate set jointly with the *SynthonReasoner*. For candidate set $\mathcal{C}$, the first contextual update is

$$\tilde{h}_c = h_c + \lambda\, \mathrm{MHA}(h_c, \{h_{c'}\}_{c' \in \mathcal{C}}, \{h_{c'}\}_{c' \in \mathcal{C}}), \tag{2}$$

followed by residual MLP blocks with LayerNorm. Attention is restricted to candidates belonging to the same molecular pair. This is the main distinction from V2, where otherwise similar candidate representations are scored independently. The contextual formulation allows a candidate to be assessed relative to the complete interaction landscape of the pair rather than from local features alone. Architectural details are provided in Appendix C.2.

4.4 JOINT FORMATION AND SYNTHON PREDICTION

A synthon head assigns a score to each contextualized candidate and produces the ranked interaction output used for synthon evaluation. The same contextualized candidate representations are aggregated into a synthon-level formation channel. In parallel, a global channel constructs a symmetric representation of the molecular pair from pooled atom-level and motif-level embeddings using sums, absolute differences, and element-wise products. A learned coefficient α combines the global and synthon channels into the final co-crystal formation score.

The two-channel design is deliberate. The global representation provides a direct route from the complete molecular structures to the pair-level prediction, while the synthon channel introduces explicit interaction hypotheses that can influence the same decision. The model therefore does not require the interaction module to carry the entire classification task. Instead, it can use local interaction evidence where it is informative while retaining global molecular context. At inference time, the output consists of both a formation probability and a ranked list of candidate synthons for the same molecular pair.

4.5 TWO-STAGE TRAINING

Pair-level and synthon-level labels are available at very different scales. Formation labels cover the full curated training set, whereas reliable synthon masks exist only for a subset of positive pairs with

usable crystal structures. Training the complete model only on that subset would discard most of the available formation data. We therefore separate representation learning from interaction-level finetuning.

Stage 1 trains the molecular representation and the co-crystal formation objective on the large augmented training pool without requiring synthon masks. Stage 2 then finetunes the model on the smaller masked subset, where pair-level classification is optimized together with candidate-level synthon supervision and ranking-oriented objectives. This schedule allows the model to first learn broad regularities associated with co-crystal formation and then refine its interaction predictions using the structurally annotated data. The loss functions, optimization procedure, and hyperparameters are reported in Appendix D.1–D.3. CIF structures and HBAT annotations are needed only during training. Prediction for a new pair requires only the two molecular structures.

5 Experimental Setup

5.1 Data Splits and Evaluation Protocol

All reported model variants use the same held-out test set. Exact molecular pairs are disjoint between training and test, although an individual molecule may occur in different splits when paired with different partners. Before any negative augmentation, we reserve a balanced test set of 700 pairs containing 350 positive and 350 negative examples. Among the positive test pairs, 245 have non-empty synthon masks and form the synthon-ranking subset. The remaining 14,906 curated pairs constitute the training pool. Similarity-filtered augmentation then adds 12,112 synthetic negatives to the training data, producing a balanced training pool of 27,018 examples (13,509 positive and 13,509 negative). Together with the held-out test set, the final dataset comprises 27,718 pairs (13,859 positive and 13,859 negative).

For synthon-supervised finetuning, 729 positive pairs with non-empty masks are combined with 729 negative examples, giving 1,458 Stage 2 pairs. The remaining 25,560 augmented examples are used in Stage 1. Pair-level prediction is evaluated with Balanced Accuracy, macro-F1, positive-class F1, precision, recall, ROC-AUC, and Matthews Correlation Coefficient (MCC). Synthon ranking is evaluated separately on the annotated positive subset so that interaction quality is not conflated with the binary classification task.

5.2 Model Variants and Ablations

We compare four progressively structured variants under this common protocol. V0 contains only the molecular GNN. V1 adds EFGs motif representations. V2 introduces the candidate-synthon channel and the statistical interaction prior, but scores each candidate independently. V3 is the complete Synthon-CGR model and adds contextual candidate reasoning together with two-stage training on HBAT-derived synthon labels. This sequence is intended to show how explicit chemical motifs, a candidate interaction space, and structure-derived supervision change the behavior of the model. Backbone architectures and pooling strategies were studied separately during development and are reported in Appendix E.1.

5.3 Independent Experimental Validation

We perform an additional evaluation on a dataset generated through co-crystallization experiments conducted in our laboratory. These data are independent of the literature and COD records used to build the training corpus. The experimental validation dataset contains 107 molecular pairs with measured co-crystallization outcomes. Seventeen pairs formed co-crystals and 90 did not. All 107 outcomes are used for external pair-level evaluation.

The 17 experimentally confirmed co-crystals, which involve six drug molecules, also provide a separate interaction-level test. For these pairs, we compare the highest-ranked predicted synthons with expert annotations derived from the corresponding resolved crystal structures. We distinguish between exact recovery of the annotated EFGs motif pair and recovery of the correct interacting atoms inside a broader predicted motif. This distinction is important because the granularity of an EFGs motif does not always coincide with the narrower contact description used by an expert.

Table 1: Ablation results on the common pair-disjoint test set. V0–V3 progressively add motif features, candidate synthons, and the full crystal-supervised reasoning model.

Version	BAcc	F1-macro	$F1_+$	Precision	Recall	ROC-AUC	MCC
V0	0.750	0.753	0.799	0.769	0.832	0.852	0.509
V1	0.789	0.791	0.826	0.808	0.845	0.875	0.584
V2	0.709	0.712	0.788	0.717	0.875	0.825	0.449
V3	**0.873**	**0.874**	**0.893**	**0.884**	**0.902**	**0.940**	**0.749**

5.4 Comparison with Existing Models on the Experimental Validation Dataset

We compare Synthon-CGR with DeepCocrystal, MCC-GCN, and CCGNet on the same 107-pair experimental dataset. The comparison therefore tests how models trained on literature and crystallographic data transfer to molecular pairs evaluated independently in the laboratory. Because the validation set is strongly imbalanced, we interpret Balanced Accuracy and MCC together with class-specific behavior rather than relying on overall accuracy alone. The available CCGNet output contains continuous scores without a specified classification threshold, so only threshold-independent metrics are reported for that model.

6 Results

6.1 Co-crystal Formation Prediction and Ablations

Table 1 shows a clear difference between adding chemically meaningful representations and simply adding more model structure. V1 improves over the GNN-only baseline after EFGs motifs are introduced. Balanced Accuracy increases from 0.750 to 0.789 and MCC from 0.509 to 0.584, suggesting that explicit functional-group information is useful even before individual interactions are modeled. V2 then introduces the candidate-synthon channel and the statistical prior while continuing to score candidates independently. Recall increases to 0.875, but the remaining metrics decline. In other words, exposing plausible motif pairs gives the model a richer interaction space, but this alone does not make the pair-level decision more reliable.

The full V3 model changes this pattern. With contextual candidate reasoning and the second stage of crystal-derived supervision, it reaches $F1_+$ = 0.893, Balanced Accuracy = 0.873, ROC-AUC = 0.940, and MCC = 0.749. These results exceed both V1 and V2 on the common test set. The comparison between V2 and V3 should not be read as isolating a single architectural factor because the contextual reasoner and the training regime change together. It does, however, show that the complete interaction-aware formulation can support explicit synthon prediction without sacrificing strong pair-level classification. This is the behavior we seek from the model: interaction reasoning should add an interpretable prediction target while remaining compatible with the broader molecular context needed for formation prediction.

6.2 Synthon Ranking

Synthon ranking is evaluated on 245 held-out positive pairs with non-empty HBAT-derived masks. Table 2 summarizes performance across ranking depths, with metric definitions given in Appendix D.5. At least one annotated synthon is ranked first for 60.8% of pairs, within the top three for 85.7%, and within the top five for 93.5%. Recall reaches 0.952 at $K = 10$, showing that most annotated interactions are concentrated near the top of the ranking.

Across the complete rankings, Synthon-CGR achieves MAP = 0.7141 and MRR = 0.7525. nDCG@K rises from 0.616 at $K = 1$ to 0.802 at $K = 20$, indicating that relevant candidates are preferentially ordered near the top. Precision@K decreases for larger candidate sets because each structure contains only a few annotated interactions relative to the number of chemically plausible candidates. We therefore interpret the ranking metrics jointly rather than using Precision@K as the primary summary.

Table 2: Synthon-ranking performance on 245 held-out co-crystal pairs with HBAT-derived interaction annotations.

Metric	$K = 1$	$K = 3$	$K = 5$	$K = 10$	$K = 20$
HitRate@K	0.608	0.857	0.935	0.976	0.988
Recall@K	0.327	0.700	0.849	0.952	0.990
nDCG@K	0.616	0.687	0.747	0.788	0.802
MAP@K	0.616	0.630	0.672	0.703	0.713
Precision@K	0.616	0.492	0.374	0.226	0.123

Table 3: Performance on the independent experimental validation dataset of 107 laboratory-tested pairs, including 17 confirmed co-crystals and 90 non-forming pairs.

Model	**BAcc**	**F1-macro**	**Precision**	**Recall**	**MCC**	**Specificity**
Synthon-CGR	**0.636**	**0.526**	**0.235**	0.706	0.200	**0.567**
DeepCocrystal	0.632	0.412	0.208	**0.941**	**0.214**	0.322
MCC-GCN	0.389	0.249	0.121	0.588	-0.195	0.189

A representative held-out example shows the same pattern qualitatively. Carboxyl interactions with aromatic-nitrogen or amino motifs receive the highest scores, around 0.94–0.95, whereas a bromine-carbonyl halogen-bond candidate scores 0.038 and non-complementary pairs remain near zero. Because these are explicit motif-pair hypotheses rather than post-hoc saliency values, they can be compared directly with contacts in a crystal structure.

6.3 Synthon Validation on Experimentally Confirmed Co-crystals

The 17 co-crystals obtained in our independent laboratory experiments provide a second way to examine the interaction predictions. For 6 of these structures, the highest-ranked model prediction exactly matches the expert-annotated synthon motif. In the remaining 11 cases, the predicted EFGs motif is broader than the expert annotation but still contains the experimentally interacting atoms. A typical mismatch therefore concerns the boundary of the functional motif rather than the location of the interaction site itself.

This result highlights an important distinction between model error and representation error. Expert annotations may describe a compact two-atom contact, whereas EFGs assigns the same atoms to a larger functional group. Exact motif matching then counts the prediction as different even when both descriptions point to the same physical interaction. We therefore report exact EFGs-pair agreement and atom-level agreement separately. The comparison also suggests a concrete direction for improving the model: finer or interaction-aware motif definitions could make the predicted synthons align more closely with the resolution used in expert structural interpretation.

6.4 Formation Prediction on the Experimental Validation Dataset

Table 3 shows that the models reach similar aggregate performance through different error profiles. DeepCocrystal has the highest positive-class recall at 0.941, but its specificity is 0.322. Synthon-CGR is less sensitive to positives, with recall of 0.706, but rejects non-forming pairs more often and reaches a specificity of 0.567. Its Balanced Accuracy is 0.636 compared with 0.632 for DeepCocrystal, while DeepCocrystal has the slightly higher MCC, 0.214 compared with 0.200. MCC-GCN performs less consistently on this dataset, with Balanced Accuracy of 0.389 and MCC of -0.195. For CCGNet, only threshold-independent scores are available, giving ROC-AUC = 0.454 and Average Precision = 0.205.

With only 17 positive examples, this dataset is sensitive to a small number of classification changes, so we treat it primarily as a test of transfer to independently generated laboratory data. The models show similar balanced performance but different error profiles, especially in the trade-off between

recovering co-crystals and avoiding false positives. This distinction matters when predictions are used to prioritize experiments.

7 DISCUSSION

The results suggest that co-crystal formation and synthon identification are related but distinct learning problems. V1 shows that functional motifs improve pair-level prediction, whereas V2 indicates that introducing candidate interactions alone is not sufficient. The strongest performance is obtained when candidate synthons are interpreted in context and trained against interactions observed in crystal structures. This suggests that synthon supervision is valuable not only because it defines a chemically plausible interaction space, but because it connects molecular compatibility with specific intermolecular contacts.

This distinction also changes the role of interpretability. Synthon-CGR produces interaction hypotheses that can be evaluated directly against resolved crystal structures rather than explaining a binary prediction post hoc. An annotated synthon appears among the top five candidates for 93.5% of evaluated pairs. In the 17 experimentally confirmed co-crystals, the interacting atoms are recovered in every case, while exact EFGs-level agreement is obtained for only 6 structures. The model therefore often localizes the correct interaction even when the predefined motif boundaries do not match the experimentally assigned synthon exactly.

The independent 107-pair validation further shows why realistic experimental evaluation matters. With only 17 positive outcomes, similar aggregate scores can correspond to substantially different recall and specificity. For prospective screening, these differences translate directly into missed viable coformers or unnecessary experiments. Evaluating both formation and interaction prediction therefore provides a more complete picture of model behavior than either task alone.

Limitations. The current candidate space focuses on hydrogen and halogen bonds and does not explicitly represent other packing interactions such as π-π stacking. Synthon supervision is also limited by the structural annotation pipeline, where crystallographic symmetry and alternative molecular representations can complicate the mapping from observed contacts to RDKit atoms and EFGs motifs. The gap between atom-level recovery and exact motif-level agreement further indicates that EFGs granularity is a source of error. In addition, the held-out evaluation is pair-disjoint rather than molecule-disjoint, so individual molecules may occur with different partners in both training and test sets.

Future Work. Future work will move toward prospective co-crystal design, using Synthon-CGR to prioritize coformers and predict the intermolecular motifs expected to stabilize the resulting crystal. Explicit synthon prediction could also provide an intermediate representation linking co-crystal formation models with crystal structure prediction. In the longer term, this may enable iterative model-guided screening in which experimental outcomes are used to refine subsequent predictions.

8 CONCLUSION

Synthon-CGR shows that co-crystal formation can be modeled together with an explicit, testable hypothesis about the intermolecular interactions underlying that outcome. The key result is not only strong pair-level classification on the held-out test set, but the ability to rank experimentally observed synthons near the top of the candidate space, with HitRate@5 = 93.5%, MRR = 0.753, and MAP = 0.714. Independent validation on 107 experimentally tested pairs further supports the practical relevance of the approach and shows that the predicted interactions remain informative beyond the literature data. At the same time, the difference between atom-level recovery and exact motif agreement highlights an important limitation of the current functional-group representation. Overall, these results support crystal-derived interaction supervision as a useful bridge between molecular compatibility prediction and structurally grounded reasoning, opening a path toward models that do not only predict whether a co-crystal will form, but also provide a chemically interpretable basis for why it may form.

AI USE STATEMENT

We used generative AI tools during manuscript preparation to assist with drafting, restructuring, and refining parts of the text, including improvements to clarity, concision, and organization. AI-assisted suggestions were reviewed and edited by the authors, who retained full control over the scientific content and final wording.

REPRODUCIBILITY STATEMENT

Code, data-processing scripts, and materials supporting the experiments will be released at `https://anonymous.4open.science/r/Synthon-CGR-D8E3`. The paper and appendix describe the dataset construction, model architecture, training procedure, and evaluation protocol. Reproducing the complete data pipeline additionally requires access to the original literature and crystallographic data sources used in this work.

REFERENCES

Soroush Ahmadi, Mohammad Amin Ghanavati, and Sohrab Rohani. Machine learning-guided prediction of cocrystals using point cloud-based molecular representation. *Chemistry of Materials*, 36(3):1153–1161, 2024.

Rebecca Birolo, Rıza Özçelik, Andrea Aramini, Roberto Gobetto, Michele R Chierotti, and Francesca Grisoni. Deep supramolecular language processing for co-crystal prediction. *Angewandte Chemie International Edition*, 64(29):e202507835, 2025.

Geetha Bolla, Bipul Sarma, and Ashwini K Nangia. Crystal engineering of pharmaceutical cocrystals in the discovery and development of improved drugs. *Chemical reviews*, 122(13):11514–11603, 2022.

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.

Gonzalo Colmenarejo. Efgs: a complete and accurate implementation of ertl's functional group detection algorithm in rdkit. *Journal of Chemical Information and Modeling*, 65(3):1061–1066, 2025.

Aurora J Cruz-Cabeza, Peter R Spackman, and Amy V Hall. The interplay between hydrogen bonds and stacking/t-type interactions in molecular cocrystals. *Communications Chemistry*, 7(1):284, 2024.

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

László Fábián. Cambridge structural database analysis of molecular complementarity in cocrystals. *Crystal Growth and Design*, 9(3):1436–1443, 2009.

Peter TA Galek, László Fábián, WD Samuel Motherwell, Frank H Allen, and Neil Feeder. Knowledge-based model of hydrogen-bonding propensity in organic crystals. *Structural Science*, 63(5):768–782, 2007.

Peter TA Galek, James A Chisholm, Elna Pidcock, and Peter A Wood. Hydrogen-bond coordination in organic crystal structures: statistics, predictions and applications. *Structural Science*, 70(1): 91–105, 2014.

Mohammad Amin Ghanavati and Sohrab Rohani. Cocrystal formation prediction: Hybrid ginmordred model outperforms dft-based methods. *Crystal Growth & Design*, 25(8):2717–2729, 2025.

Mohammad Amin Ghanavati, Bahareh Khalili, Dino Alberico, and Sohrab Rohani. Machine learning-driven discovery of multicomponent pharmaceutical solid forms via dualnet: Confidence-aware prediction and ranking of salts and cocrystals. *International Journal of Pharmaceutics*, pp. 126117, 2025.

Jiali Guo, Ming Sun, Xueyan Zhao, Chaojie Shi, Haoming Su, Yanzhi Guo, and Xuemei Pu. General graph neural network-based model to accurately predict cocrystal density and insight from data quality and feature representation. *Journal of Chemical Information and Modeling*, 63(4):1143–1156, 2023.

Minshan Guo, Xiaojie Sun, Jiahui Chen, and Ting Cai. Pharmaceutical cocrystals: A review of preparations, physicochemical properties and applications. *Acta Pharmaceutica Sinica B*, 11(8): 2537–2564, 2021.

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.

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.

Maryam Karimi-Jafari, Luis Padrela, Gavin M Walker, and Denise M Croker. Creating cocrystals: A review of pharmaceutical cocrystal preparation routes and applications. *Crystal Growth & Design*, 18(10):6370–6387, 2018.

Paul Kim, In-Seo Lee, Ji-Yoon Kim, Medard E Mswahili, Young-Seob Jeong, Woo-Jin Yoon, Ho-Seop Yun, Min-Jeong Lee, and Guang J Choi. A study to discover novel pharmaceutical cocrystals of pelubiprofen with a machine learning approach compared. *CrystEngComm*, 24(21):3938–3952, 2022.

Xiaoxiao Liang, Shiyuan Liu, Zebin Li, Yuehua Deng, Yanbin Jiang, and Huaiyu 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*, 196:114201, 2024.

Yukun Liu, Yanfei Liu, Ziang Du, Haochun Lan, Hao Zheng, Dongdong Yang, and Yang Gao. Rational high-throughput screening of cocrystals: An explainable deep learning framework based on morgan fingerprints and cross-attention. *ACS Physical Chemistry Au*, 2026a.

Yukun Liu, Yanfei Liu, Ziang Du, Hao Zheng, Dongdong Yang, and Yang Gao. Explainable machine learning for efficient cocrystal prediction: Leveraging morgan fingerprints to decode local structure-cocrystal formation relationships. *The Journal of Physical chemistry. A*, 130(17):3494, 2026b.

Zahrasadat Momenzadeh Abardeh, Alireza Salimi, and Artem R Oganov. Synthon modularity in crystal structure prediction: designing pomalidomide polymorphs and co-crystals. *CrystEngComm*, 27(32):5514–5526, 2025.

Maciej Przybyłek and Piotr Cysewski. Distinguishing cocrystals from simple eutectic mixtures: phenolic acids as potential pharmaceutical coformers. *Crystal Growth & Design*, 18(6):3524–3534, 2018.

Maciej Przybyłek, Tomasz Jelinski, Julia Słabuszewska, Dorota Ziółkowska, Karina Mroczynska, and Piotr Cysewski. Application of multivariate adaptive regression splines (marsplines) methodology for screening of dicarboxylic acid cocrystal using 1d and 2d molecular descriptors. *Crystal Growth & Design*, 19(7):3876–3887, 2019.

Bhupinder Sandhu, Ann McLean, Abhijeet S Sinha, John Desper, and Christer B Aakery. Assessment of computational tools for predicting supramolecular synthons. *Chemistry*, 3(2):612–629, 2021.

Nandini Sarkar, Nina C Gonnella, Mariusz Krawiec, Dongyue Xin, and Christer B Aakeroy. 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.

JARP Sarma and Gautam R Desiraju. The supramolecular synthon approach to crystal structure prediction. *Crystal growth & design*, 2(2):93–100, 2002.

Isaac J Sugden, Doris E Braun, David H Bowskill, Claire S Adjiman, and Constantinos C Pantelides. Efficient screening of coformers for active pharmaceutical ingredient cocrystallization. *Crystal Growth & Design*, 22(7):4513, 2022.

Aikaterini Vriza, Ioana Sovago, Daniel Widdowson, Vitaliy Kurlin, Peter A Wood, and Matthew S Dyer. Molecular set transformer: attending to the co-crystals in the cambridge structural database. *Digital Discovery*, 1(6):834–850, 2022.

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.

Kun Wang, Xiao Zhao, and Weihua Zhu. Application of molecular electrostatic potential surface to predict supramolecular synthons for rdx/solvent cocrystals. *Crystal Research and Technology*, 54 (12):1900171, 2019.

Jerome GP Wicker, Lorraine M Crowley, Oliver Robshaw, Edmund J Little, Stephen P Stokes, Richard I Cooper, and Simon E Lawrence. Will they co-crystallize? *CrystEngComm*, 19(36): 5336–5340, 2017.

Peter A Wood, Tjelvar SG Olsson, Jason C Cole, Simon J Cottrell, Neil Feeder, Peter TA Galek, Colin R Groom, and Elna Pidcock. Evaluation of molecular crystal structures using full interaction maps. *CrystEngComm*, 15(1):65–72, 2012.

Peter A Wood, Neil Feeder, Matthew Furlow, Peter TA Galek, Colin R Groom, and Elna Pidcock. Knowledge-based approaches to co-crystal design. *CrystEngComm*, 16(26):5839–5848, 2014.

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.

Lulu Zheng, Bin Zhu, Zengrui Wu, Mei Guo, Jinyao Chen, Minghuang Hong, Guixia Liu, Weihua Li, Guobin Ren, and Yun Tang. Pharmaceutical cocrystal discovery via 3d-sminbr: A new network recommendation tool augmented by 3d molecular conformations. *Journal of Chemical Information and Modeling*, 63(14):4301–4311, 2023.

A Dataset Construction, Curation, and Splits

This section provides the dataset-level details that are omitted from the main paper for space. We first describe the deterministic curation rules, then show diagnostic molecular-property distributions, summarize additional metadata, and finally report the exact working split sizes used by the current experimental protocol.

A.1 Molecular Component Filtering

Dataset curation begins with component-level filtering. Each molecular representation is parsed and sanitized with RDKit, after which molecules that fall outside the intended scope of the dataset are removed. A component is rejected if the SMILES string cannot be processed by RDKit (*invalid_smiles*), if it contains multiple disconnected fragments (*multiple_fragments*), or if the molecule belongs to an explicit exclusion list (*excluded*). The latter includes water, common crystallization solvents, and simple inorganic ions.

Molecules with a non-zero net formal charge are also excluded (*net_formal_charge*), while charge-separated representations with an overall neutral charge remain allowed. To restrict the dataset to molecular systems of comparable size, compounds containing more than 80 heavy atoms (*too_many_heavy_atoms*) or having a molecular weight above 1000 Da (*molecular_weight_too_high*) are removed.

The solvent exclusion list contains methanol, ethanol, 1-propanol, 2-propanol, 1-butanol, 2-butanol, *tert*-butanol, acetone, acetonitrile, ethyl acetate, diethyl ether, tetrahydrofuran, 1,4-dioxane, dichloromethane, chloroform, carbon tetrachloride, DMF, and DMSO. The explicitly excluded simple ions are F^-, Cl^-, Br^-, I^-, Li^+, Na^+, K^+, NH_4^+, Mg^{2+}, and Ca^{2+}. Additional exclusions for small inorganic or volatile guests and elemental halogens are disabled by default, since these species may participate in valid molecular co-crystals depending on the chemical system.

A.2 Pair Canonicalization and Label Consistency

After component-level filtering, molecular pairs are standardized before dataset splitting. Pairs containing two identical components ($A = B$) are removed. The order of the two components is then canonicalized by sorting their canonical SMILES, so that (A, B) and (B, A) are represented by the same pair key.

This normalization also allows repeated observations to be handled consistently across data sources. If the same molecular pair is associated with both positive and negative co-crystallization labels, all records for that pair are removed rather than resolving the disagreement heuristically. Repeated records with the same label are collapsed into a single observation. These operations are completed before constructing the held-out test split.

A.3 Data Sources

The curated dataset was assembled from publicly available crystallographic databases and data extracted from the literature. Table 4 summarises the number of molecular pairs contributed by each source.

Table 4: Data sources used to construct the curated co-crystal dataset.

Source	Access	Molecular pairs
Literature data	(see Table 5)	14,714
COD	`https://crystallography.net/cod/`	1,369
Total		**16,083**

The 14,714 molecular pairs contributed by literature sources were extracted from 11 prior studies. Table 5 lists each source with its DOI and the number of pairs contributed before filtering (80,403

in total). After removing duplicates, non-organic entries, and records that failed component- and pair-level validation, 14,714 pairs were retained.

Table 5: List of literature sources contributing molecular pairs to the curated dataset. The "Total after filters" row indicates the number of pairs retained after deduplication and validation.

#	Source	Molecular pairs
1	(Jiang et al., 2021)	8006
2	(Birolo et al., 2025)	39746
3	(Vriza et al., 2022)	9920
4	(Wang et al., 2020)	373
5	(Zheng et al., 2023)	2627
6	(Chen et al., 2024)	7473
7	(Ghanavati et al., 2025)	10980
8	(Liang et al., 2024)	302
9	(Przybyłek & Cysewski, 2018)	226
10	(Kim et al., 2022)	38
11	(Przybyłek et al., 2019)	712
Total		**80,403**
Total after filters		**14,714**

A.4 Negative-Pair Augmentation

Because experimentally reported co-crystallization data are strongly enriched in positive outcomes, additional negative pairs are generated for the training set. Augmentation is applied only after the held-out test set has been reserved, preventing synthetic training pairs from affecting test-set construction.

For each positive source pair (A, B), one component is selected as the fixed molecule A, while a candidate partner C is sampled from a different positive pair. A proposed pair (A, C) is rejected if $C = A$, if the pair is already present in the dataset or has already been generated, or if C is structurally similar to a known coformer of A. Structural similarity is measured using the Tanimoto coefficient computed from ECFP4 fingerprints. A candidate is rejected when its similarity to any known coformer of A is at least 0.6.

At most 100 candidate partners are sampled for each source pair. The procedure uses a fixed random seed of 42 for reproducibility. The similarity threshold is kept fixed throughout augmentation and is not relaxed when the target number of negative examples cannot be reached.

A.5 Molecular Property Distributions

Figures 3–5 summarize the molecular property distributions in the aggregated dataset before final component-level filtering. The dashed vertical lines mark the empirical 95th and 99th percentiles reported in each legend and are included only as descriptive reference points. They do not define the filtering criteria. The deterministic thresholds used during molecular component filtering are specified in Appendix A.1.

A.6 Frequent Molecular Components and Stoichiometric Metadata

Table 6 reports the most frequently occurring individual molecular components in the aggregated source table. The purpose of this summary is to make repeated coformers visible and to show that the corpus is not organized around a single dominant molecule. These exploratory frequency counts are separate from the final train/test protocol reported in Appendix A.7.

Stoichiometric ratios were recovered when they were explicitly available in source records. Refcode-based metadata reported in prior work (Guo et al., 2023) were matched to the molecular pairs where possible, and additional records were recovered from a second published dataset (Zheng et al., 2023). This yielded 144 pairs with an AB ratio and 32 pairs with an A_2B ratio. Stoichiometry is retained as

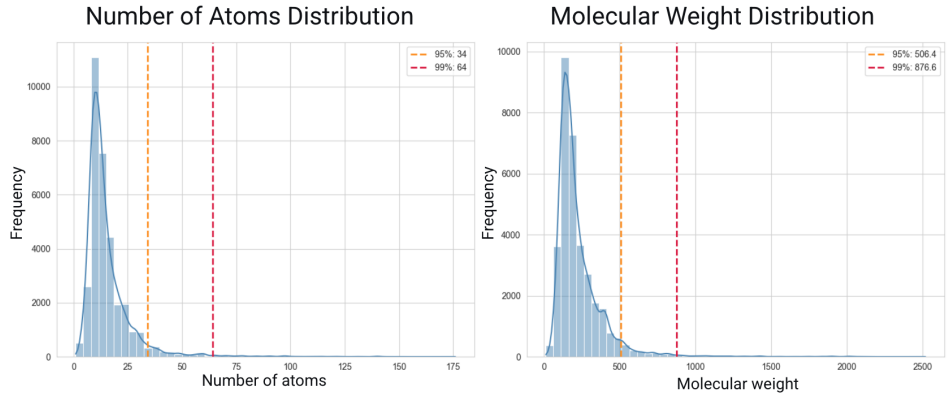

Figure 3: Pre-filtering distributions of molecular size. Left: total atom count per molecule. Right: molecular weight in daltons. Dashed vertical lines mark the empirical 95th and 99th percentiles and are shown for descriptive reference only. The final filtering thresholds are specified in Appendix A.1.

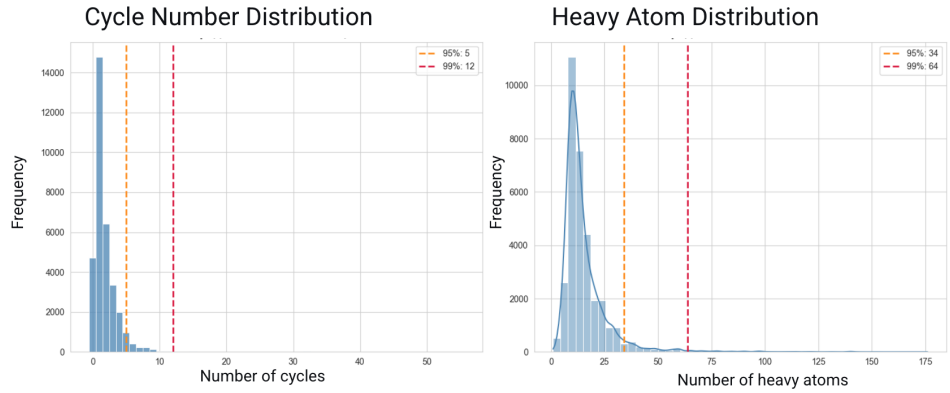

Figure 4: Pre-filtering distributions of molecular structural complexity. Left: number of cycles per molecule. Right: number of heavy atoms per molecule. Percentile markers illustrate the long-tailed character of the source data and do not define the filtering criteria. The final heavy-atom threshold is given in Appendix A.1.

auxiliary metadata; unavailable ratios are left missing rather than inferred, and the current Synthon-CGR input representation does not use the ratio as a model feature.

A.7 Dataset Splits and Augmentation

The current experiments use a common held-out test set for all ablation variants. Exact molecular pairs are disjoint between training and test, while an individual molecule may appear in different splits when paired with different partners. Importantly, negative-pair augmentation is performed only after the test set has been reserved, so synthetic negatives are not used to construct the held-out evaluation set.

The common test set contains 245 positive pairs with a non-empty synthon mask, 105 positive pairs without a usable mask, and 350 original negative pairs. The Stage 2 subset contains 729 positive pairs with non-empty HBAT-derived masks, 75 original negatives, and 654 synthetic negatives. The remaining 25,560 augmented training examples are used for Stage 1 pretraining.

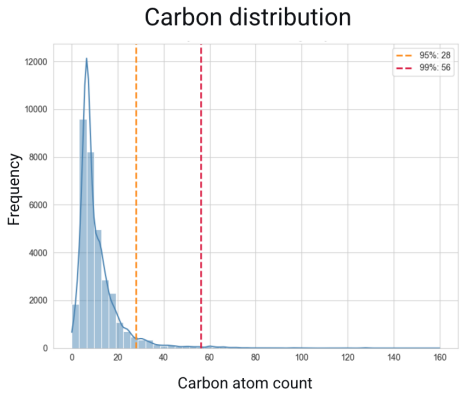

Figure 5: Pre-filtering distribution of carbon-atom count per molecule. This statistic is reported as a dataset diagnostic and is not used as an independent filtering rule. Dashed lines mark the empirical 95th and 99th percentiles.

Table 6: Most frequent individual molecular components in the aggregated source table used for dataset inspection. Count denotes the number of pair records containing the component; Share is the corresponding percentage of records in that source table.

Name	Count	Share (%)
4-pyridin-4-ylpyridine	431	2.6
2,4,6-Trinitrophenol	388	2.3
1,2,4,5-Tetrafluoro-3,6-diiodobenzene	360	2.1
But-2-enedioic acid	340	2.0
1,2-Di(pyridin-4-yl)ethene	263	1.6
Butanedioic acid	240	1.4
2-Hydroxybenzoic acid	237	1.4
Benzoic acid	233	1.4
Oxalic acid	229	1.4
Pyridine-4-carboxamide	207	1.2

B Crystal-Derived Synthon Annotation

The synthon supervision used by Synthon-CGR is constructed by combining functional-group detection on the input molecules with intermolecular contacts observed in experimental crystal structures. This section details the mapping from CIF/HBAT records to motif-pair labels and illustrates the resulting annotations.

B.1 Functional-Group Detection with EFGs

We use the EFGs implementation of Ertl's functional-group detection algorithm (Colmenarejo, 2025). For each molecule, EFGs returns a set of structural motifs, each represented by a pseudo-SMILES type and the indices of the atoms belonging to that motif. These atom-index sets provide the bridge between the molecular GNN representation and crystal-derived interaction annotations: an HBAT contact can be mapped to the EFGs motif containing the corresponding atom.

Figure 6 shows an example of the resulting functional-group decomposition. The visualization is intended to illustrate which local regions of the molecules become candidate interaction motifs; the model receives motif identities and atom membership rather than the rendered colors.

Table 7: Working dataset sizes used by the current experimental protocol. The common test set is reserved before negative-pair augmentation; the synthon-supervised fine-tuning subset is drawn from the augmented training pool.

Stage	Positive	Negative	Total
Original curated dataset	14,335	1,748	16,083
After component/pair filtering	13,859	1,747	15,606
Common test set	350	350	700
Training pool	13,509	1,397	14,906
After negative-pair augmentation	13,509	13,509	27,018
Fine-tuning subset (synthons)	729	729	1,458

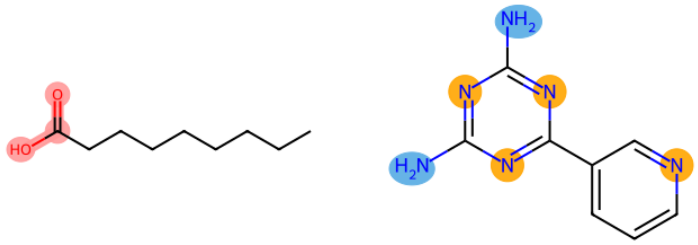

Figure 6: Example of EFGs functional-group detection on a molecular pair. Colored atom regions denote automatically identified functional motifs that are represented by a motif type and a set of atom indices. These motifs are subsequently paired across the two molecules to construct candidate synthons.

B.2 HBAT Interaction Extraction and Atom Mapping

For each available CIF file, the crystal structure is parsed with `gemmi`. Non-hydrogen atoms are extracted, molecular connectivity is reconstructed from interatomic distances in the 0.8–2.2 Å range, and connected components are used to identify molecular copies in the unit cell. The two input SMILES are then matched to corresponding crystal components by comparing elemental compositions.

HBAT is applied to the processed structure to identify intermolecular hydrogen and halogen bonds. For each retained contact, the pipeline records the interaction type, donor and acceptor labels, distance, and angle. We keep only contacts connecting the central molecule and its coformer. When multiple symmetry-related molecular copies are available, the mapping routine selects a copy containing HBAT-annotated atoms when possible; otherwise it falls back to the first matched copy.

To support atom-role matching, atoms are additionally assigned donor or acceptor roles using the SMARTS library stored in `ATOM_ROLE_RULES`, including patterns for carboxylic acids, amides, amines, azo/aromatic nitrogens, sulfonamides, and halogen-bond donors. HBAT atom labels are then mapped to RDKit atom indices and subsequently to EFGs motifs via their stored `atom_ids`.

B.3 Synthon-Mask Construction and Mapping Reliability

For a molecular pair with n_A EFGs motifs in molecule A and n_B motifs in molecule B, the crystal-derived supervision is stored as a binary matrix $m \in \{0,1\}^{n_A \times n_B}$. An entry $m_{ij} = 1$ indicates that motif i of molecule A and motif j of molecule B contain atoms participating in at least one retained HBAT interaction. The same motif pair can therefore receive a positive label even when more than one atom-level contact supports the interaction.

The mapping is intentionally conservative. A non-empty mask is used as positive synthon supervision only when the crystal component, HBAT atom labels, RDKit atom indices, and EFGs mo-

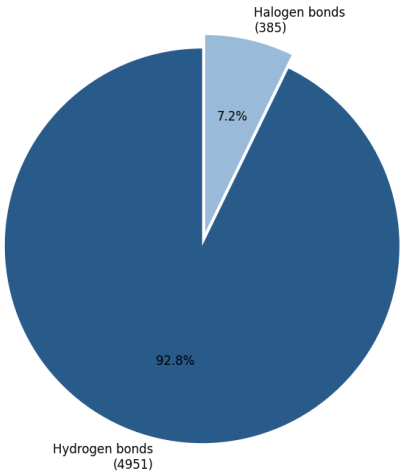

Figure 7: Composition of the 5,336 HBAT-detected intermolecular contacts retained from 1,970 processed crystal structures. Hydrogen bonds account for 4,951 contacts (92.8%) and halogen bonds for 385 contacts (7.2%).

Table 8: Representative HBAT interaction records. The table includes two hydrogen bonds from COD 1505190 and one halogen bond from COD 7057167. Distance and angle are reported in Å and degrees, respectively.

COD ID	Type	Bond ID	Donor	Acceptor	Interaction	Distance (Å)	Angle (°)
1505190	H-bond	#1	CTR:O1	M01:N11	O1–H$\cdots$N11	1.64	165.7
1505190	H-bond	#4	M03:C11	CTR:O1	C11–H$\cdots$O1	2.45	154.0
7057167	X-bond	#1	CTR:C1–I1	M06:N3	C1–I1$\cdots$N3	2.80	175.9

tif membership can be connected consistently. Mapping can fail or become ambiguous because CIF structures contain symmetry-related copies and atom orderings that need not coincide with the canonical molecular representation. Pairs without a reliable non-empty mask remain usable for pair-level formation training but are not treated as positive examples in the synthon-ranking evaluation.

Masks are used only during Stage 2 training. At inference time, Synthon-CGR receives the two molecular structures and candidate motifs but no CIF, HBAT output, or ground-truth mask.

B.4 Annotation Statistics and Representative Examples

Across 1,970 processed crystal structures, HBAT identifies 5,336 retained intermolecular interactions. Hydrogen bonds dominate the resulting supervision: 4,951 contacts (92.8%) are hydrogen bonds and 385 (7.2%) are halogen bonds (Figure 7). The imbalance motivates reporting interaction type explicitly because performance on the rarer halogen-bond cases may be less stable than aggregate synthon-ranking metrics suggest.

Table 8 gives representative raw HBAT annotations before they are mapped to EFGs motif pairs. The donor and acceptor identifiers refer to atoms in the processed crystal structure; distance and angle are the geometric descriptors returned for the detected contact.

To inspect the mapping qualitatively, we also render the two molecular components together with the retained HBAT contacts and the EFGs motifs that contain the interacting atoms. Figure 8 shows one such example. The dashed lines correspond to two annotated hydrogen-bond contacts; the highlighted atom regions indicate the local motifs that are converted into positive synthon entries in the mask.

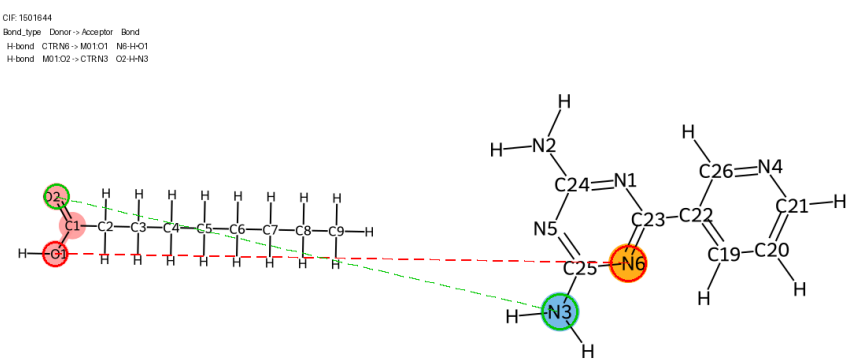

Figure 8: Example of crystal-derived synthon annotation for COD 1501644. The two molecular components are rendered in 2D after mapping from the CIF structure. Dashed lines connect donor and acceptor atoms participating in HBAT-detected hydrogen bonds, while colored atom highlights indicate the corresponding local EFGs motifs used to construct the synthon mask.

C Synthon-CGR Architecture Details

This section gives the implementation-level definition of the two components that are specific to explicit synthon reasoning: the candidate encoder and the SynthonReasoner. Molecular graph encoding and the two prediction channels are summarized in Section 4 of the main paper.

C.1 Candidate Encoder

For a candidate synthon $c = (i, j)$ formed by motif i of molecule A and motif j of molecule B, the candidate embedding is computed as

$$h_c = \mathrm{MLP}([\, h_i \,;\, h_j \,;\, |h_i - h_j| \,;\, h_i \odot h_j \,;\, s_c \,]) \,, \tag{3}$$

where $h_i, h_j \in \mathbb{R}^d$ are the motif embeddings ($d = 128$), $\odot$ denotes the element-wise product, and $s_c \in \mathbb{R}^8$ contains scalar interaction features:

$$s_c = \left[\, p_c,\ \mathbb{1}_{\text{halogen}},\ d_i,\ a_i,\ d_j,\ a_j,\ \tfrac{\min(|i|,20)}{20},\ \tfrac{\min(|j|,20)}{20} \,\right]. \tag{4}$$

Here p_c is the interaction prior, $\mathbb{1}_{\text{halogen}}$ is a halogen-bond flag, $d_\bullet$ and $a_\bullet$ are donor and acceptor scores for the corresponding motif, and $|i|, |j|$ denote motif sizes. Combining ordered motif embeddings with symmetric pair operations ($|h_i - h_j|$ and $h_i \odot h_j$) gives the evaluator access to both motif identity and their pairwise compatibility.

The projection MLP is

$$\mathrm{MLP}(x) = W_2\, \mathrm{ReLU}(W_1 x + b_1) + b_2, \tag{5}$$

with $W_1 \in \mathbb{R}^{d \times (4d+8)}$ and $W_2 \in \mathbb{R}^{d \times d}$.

C.2 SynthonReasoner

The SynthonReasoner refines each candidate in the context of the other candidate interactions available for the same molecular pair. Let $\{h_c\}_{c \in \mathcal{C}_b}$ denote the candidate embeddings for pair b. The update is

$$\tilde{h}_c = h_c + \lambda \cdot \mathrm{MHA}(h_c, \{h_{c'}\}_{c' \in \mathcal{C}_b}, \{h_{c'}\}_{c' \in \mathcal{C}_b}) \,, \tag{6}$$

$$h_c^{(0)} = \tilde{h}_c, \tag{7}$$

$$h_c^{(\ell)} = h_c^{(\ell-1)} + \mathrm{MLP}_\ell\Big(\mathrm{LN}(h_c^{(\ell-1)})\Big) \,, \quad \ell = 1, \ldots, L, \tag{8}$$

$$h_c^{\text{out}} = \mathrm{LN}\Big(h_c^{(L)}\Big) \,. \tag{9}$$

Here MHA is multi-head self-attention with four heads, $\lambda = 0.3$ is the residual scaling factor, $L = 4$ is the number of residual MLP blocks, and LN denotes LayerNorm. Each residual MLP block uses $\text{Linear} \to \text{GELU} \to \text{Dropout}$ with a dropout rate of 0.3.

Attention is restricted to candidates belonging to the same molecular pair, so candidate sets do not interact across examples in a minibatch. When a pair has fewer than two candidates, the attention update is skipped and the residual MLP blocks are still applied. This preserves a defined candidate representation for single-candidate cases without introducing artificial cross-pair context.

D Training and Evaluation Details

D.1 Loss Functions

Stage 1: pair-level pretraining. Stage 1 uses only the binary co-crystal formation objective:

$$\mathcal{L}_{\text{Stage1}} = \text{BCE}(\hat{y}, y), \tag{10}$$

where $\hat{y}$ is the pair-level logit and $y \in \{0, 1\}$ is the formation label. No crystal-derived synthon mask is required at this stage.

Stage 2: synthon-supervised fine-tuning. For examples with mask supervision, the objective combines pair-level classification with candidate-level and ranking-oriented terms:

$$\mathcal{L}_{\text{Stage2}} = \text{BCE}(\hat{y}, y) + \lambda_{\text{syn}}\mathcal{L}_{\text{syn}} + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}} + \lambda_{\text{ctr}}\mathcal{L}_{\text{ctr}}, \tag{11}$$

with $\lambda_{\text{syn}} = 0.3$, $\lambda_{\text{sp}} = 0.001$, and $\lambda_{\text{ctr}} = 3.0$.

Synthon loss. For molecular pair b with candidate mask $m_b \in \{0, 1\}^{n_b}$,

$$\mathcal{L}_{\text{syn}}^{(b)} = \text{BCE}(\hat{s}_b, m_b; w_b)\,, \tag{12}$$

where $\hat{s}_b$ are the candidate logits. To compensate for the fact that a pair generally contains many more negative than positive candidate motif pairs, the positive weight is clipped as

$$w_b = \text{clip}\left(\frac{n_{\text{neg}}}{n_{\text{pos}}}, 2, 10\right). \tag{13}$$

If $n_{\text{pos}} = 0$, the term is reduced to $0.1\,\text{BCE}(\hat{s}_b, m_b)$ to limit the contribution of all-negative candidate sets.

Contrastive loss. When both positive and negative candidates are present, the contrastive term encourages positive candidates to score above negatives:

$$\mathcal{L}_{\text{ctr}}^{(b)} = \text{ReLU}(\bar{s}_{\text{neg}} - \bar{s}_{\text{pos}} + \delta) + 2\,\text{ReLU}(0.7 - \bar{s}_{\text{pos}})\,, \tag{14}$$

where $\bar{s}_{\text{pos}}$ and $\bar{s}_{\text{neg}}$ are the mean sigmoid scores of positive and negative candidates, respectively, and $\delta = 0.5$.

Sparsity term. For batches containing at least one pair with positive candidates,

$$\mathcal{L}_{\text{sp}} = -\frac{1}{|\mathcal{C}|}\sum_{c \in \mathcal{C}} s_c^{(K)}, \tag{15}$$

where $\mathcal{C}$ is the set of candidates in the batch and $s_c^{(K)}$ denotes the K-th largest candidate score with $K = 3$.

D.2 Training Protocol and Optimization

Stage 1. The model is pretrained on 25,560 pairs without synthon masks. This stage uses standard binary formation classification and learns the molecular encoder, global channel, and initial candidate representations from the larger pair-level dataset.

Table 9: Principal architecture, optimization, and loss hyperparameters for Synthon-CGR (V3).

Parameter	Value
Hidden dimension d	128
Synthon dimension	128
GNN layers	2
GATv2 attention heads	4
Dropout (GNN, prediction heads)	0.2
Dropout (candidate evaluator/reasoner)	0.3
Batch size	64
Base learning rate	5×10^{-4}
Evaluator learning-rate multiplier	5.0
Weight decay	10^{-3}
Maximum candidate synthons	100
Number of motif types	1000
SynthonReasoner attention heads	4
SynthonReasoner residual MLP layers	4
Residual attention scale λ	0.3
λ_{syn}	0.3
λ_{sp}	0.001
λ_{ctr}	3.0

Stage 2. The model is fine-tuned on 1,458 pairs in the synthon-supervised subset. The SynthonReasoner and candidate-scoring components receive direct supervision via $\mathcal{L}_{\mathrm{syn}}$ and $\mathcal{L}_{\mathrm{ctr}}$. Evaluator parameters use a $5\times$ larger learning rate than the rest of the model to adapt the candidate-ranking components to the smaller masked dataset.

Optimization. We use Adam with decoupled weight decay of 10^{-3}, gradient clipping at $\|\nabla\|_2 \leq 1.0$, and a ReduceLROnPlateau scheduler that tracks validation F1 (mode=max, factor=0.5, patience=5). Early stopping uses a patience of 20 epochs. The model checkpoint selected by the validation criterion is then used for held-out evaluation; test metrics are not used for optimization.

D.3 Hyperparameters

Table 9 lists the principal architecture and optimization settings of the full V3 model. The same hidden dimension is used for motif and candidate representations so that residual reasoning blocks do not require additional projection layers.

D.4 Training Dynamics

Figure 9 provides a representative optimization trace. The training loss decreases throughout optimization, while validation accuracy and F1 rise rapidly during the early epochs and then stabilize. These curves are shown as training diagnostics only; all reported test-set results in the main paper are computed from the held-out common test set rather than from the validation trajectory.

D.5 Synthon Ranking Metrics

Synthon-ranking metrics are computed only for positive pairs ($y = 1$) with a non-empty HBAT-derived mask. Let $\mathcal{P}$ denote this evaluation subset, $n_{\mathrm{pos}}^{(b)}$ the number of positive candidates for pair b, and $\pi_b(1), \pi_b(2), \ldots$ the candidates of pair b ordered by predicted score from highest to lowest. The metrics capture complementary properties: HitRate@K asks whether any annotated interaction is recovered near the top; Recall@K and Precision@K measure coverage and concentration; nDCG@K and MAP@K reward correct ordering; and MRR focuses on the rank of the first recovered positive candidate.

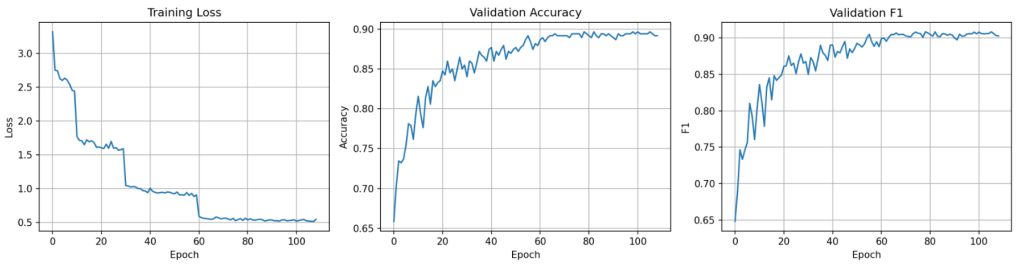

Figure 9: Representative Synthon-CGR training history. Left: training loss. Center: validation accuracy. Right: validation F1. The curves summarize optimization behavior and model selection on validation data; they do not contain held-out test metrics.

HitRate@K.

$$\mathrm{HitRate@K} = \frac{1}{|\mathcal{P}|} \sum_{b \in \mathcal{P}} \mathbb{1}\left[\sum_{k=1}^{K} m_{\pi_b(k)} > 0 \right]. \tag{16}$$

Recall@K.

$$\mathrm{Recall@K} = \frac{1}{|\mathcal{P}|} \sum_{b \in \mathcal{P}} \frac{1}{n_{\mathrm{pos}}^{(b)}} \sum_{k=1}^{K} m_{\pi_b(k)}. \tag{17}$$

Precision@K.

$$\mathrm{Precision@K} = \frac{1}{|\mathcal{P}|} \sum_{b \in \mathcal{P}} \frac{1}{K} \sum_{k=1}^{K} m_{\pi_b(k)}. \tag{18}$$

nDCG@K.

$$\mathrm{nDCG@K} = \frac{1}{|\mathcal{P}|} \sum_{b \in \mathcal{P}} \frac{\mathrm{DCG@K}(b)}{\mathrm{IDCG@K}(b)}, \tag{19}$$

with

$$\mathrm{DCG@K}(b) = \sum_{k=1}^{K} \frac{m_{\pi_b(k)}}{\log_2(k+1)}, \qquad \mathrm{IDCG@K}(b) = \sum_{k=1}^{\min(n_{\mathrm{pos}}^{(b)}, K)} \frac{1}{\log_2(k+1)}. \tag{20}$$

MAP@K.

$$\mathrm{MAP@K} = \frac{1}{|\mathcal{P}|} \sum_{b \in \mathcal{P}} \frac{1}{\min(n_{\mathrm{pos}}^{(b)}, K)} \sum_{k=1}^{K} \mathrm{Prec@k}(b)\, m_{\pi_b(k)}, \tag{21}$$

where $\mathrm{Prec@k}(b) = \frac{1}{k} \sum_{j=1}^{k} m_{\pi_b(j)}$.

MRR.

$$\mathrm{MRR} = \frac{1}{|\mathcal{P}|} \sum_{b \in \mathcal{P}} \frac{1}{\mathrm{rank}_b}, \tag{22}$$

where rank_b is the position of the first positive candidate in the ranking for pair b.

MAP over the full ranking.

$$\mathrm{MAP} = \frac{1}{|\mathcal{P}|} \sum_{b \in \mathcal{P}} \mathrm{AP}(b), \tag{23}$$

where $\mathrm{AP}(b)$ is Average Precision over the complete candidate ranking for pair b, computed with `average_precision_score` from scikit-learn.

Table 10: Preliminary molecular-encoder comparison used to select the GNN backbone for subsequent Synthon-CGR experiments. These results correspond to the encoder-selection configuration and are not the final common-test ablation reported in Table 1.

Model	Accuracy	F1	Precision	Recall	ROC-AUC
GCN	0.9071	0.9364	0.9879	0.8899	0.9632
GAT	0.8949	0.9390	0.9829	0.8988	0.9640
GATv2	0.9115	0.9615	0.9823	0.9416	0.9711
GIN	0.9074	0.9344	0.9886	0.8859	0.9699
GINE	0.9074	0.9505	0.9840	0.9192	0.9666

E Additional Model Diagnostics

E.1 Backbone GNN Selection

Before introducing the full synthon-reasoning architecture, we compared several message-passing backbones for the molecular encoder. Table 10 reports this preliminary encoder-selection experiment. GATv2 gives the highest F1 and ROC-AUC among the evaluated backbones in this configuration and was therefore selected for subsequent Synthon-CGR experiments. This table is an architecture-selection diagnostic from an earlier experimental configuration; its absolute values should not be compared directly with the final common-test ablation in Table 1.

The purpose of this comparison is backbone selection rather than a claim that a particular GNN family is universally preferable for co-crystal prediction. The result only establishes that GATv2 was the strongest of the tested encoders under this development configuration.

E.2 Candidate-Set Statistics and Edge Cases

Under the current candidate-generation settings, a molecular pair produces 7.5 candidate synthons on average (median 6; maximum 171), while 6.3% of pairs produce no candidate under the available EFGs and interaction rules. Candidate set size matters computationally because the SynthonReasoner performs self-attention within each pair. The configured candidate cap therefore bounds the amount of contextual reasoning performed for unusually motif-rich pairs.

Pairs with no generated candidate remain valid inputs for pair-level co-crystal formation prediction through the global molecular channel, but there is no candidate interaction to rank for such examples. They are consequently outside the synthon-ranking evaluation defined in Appendix D.5.