Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 3270

Crystal structure prediction of organic molecules by machine learning-based lattice sampling and structure relaxation

Takuya Taniguchi *[a] and Ryo Fukasawa [b]

Predicting the crystal structures of organic molecules remains a formidable challenge due to intensive computational cost. To address this issue, we developed a crystal structure prediction (CSP) workflow that combines machine learning-based lattice sampling with structure relaxation *via* a neural network potential. The lattice sampling employs two machine learning models—space group and packing density predictors—that reduce the generation of low-density, less-stable structures. In tests on 20 organic crystals of varying complexity, our approach achieved an 80% success rate—twice that of a random CSP—demonstrating its effectiveness in narrowing the search space and increasing the probability of finding the experimentally observed crystal structure. We also characterized which molecular and crystal parameters influence the success rate of CSP, clarifying the effectiveness and limitation of the current workflow. This study underscores the utility of combining machine learning models with efficient structure relaxations to accelerate organic crystal structure discovery.

Received 11th July 2025
Accepted 30th September 2025

DOI: 10.1039/d5dd00304k

rsc.li/digitaldiscovery

Introduction

Crystal structure prediction (CSP) of organic molecules has far-reaching implications for both pharmaceutical and materials science, offering critical insights into controlling polymorphism in drug development and facilitating the rational design of novel functional materials. Managing crystal structures is paramount, as it directly influences drug solubility and stability.[1] Controlling crystal structures is also important for organic semiconductors because the electronic conductivity of π-electron systems varies with molecular arrangement.[2,3] Reliable and efficient method of CSP would enable the selective production of crystal structures with desirable physicochemical properties, addressing a key challenge in pharmaceutical formulation and material design.[4–6] The ability to accurately predict crystal structures promises to be transformative, driving significant advances across a wide range of scientific and industrial domains.[7]

Predicting the crystal structure of an organic molecule is challenging, due to the weaker atomic interactions unique to organic crystals.[8] Unlike inorganic crystals, which often rely on stronger bonds, organic crystals are stabilized by relatively weak intra- and inter-molecular interactions such as van der Waals forces, hydrogen bonds, and π–π stacking. Even minor variations in these interactions can give rise to entirely different crystal structures, making accurate prediction difficult. In addition, many organic molecules exhibit considerable conformational flexibility because of rotatable bonds, significantly increasing the number of possible configurations.[9] Even for relatively rigid molecules, identifying the global energy minimum is still computationally intensive. Consequently, the interplay between molecular flexibility and subtle intermolecular interactions renders the accurate prediction of organic crystal structures a challenge.

In general, CSP can be divided into two stages: structure generation (or exploration) and structure relaxation. To address the challenges in these stages, many CSP workflows have been proposed, as tackled in the series of CSP blind tests.[10] In the structure generation (exploration), the quasi-random method, genetic algorithms, particle swarm optimization, and Bayesian optimization have been developed.[11–18] The quasi-random approach stochastically arranges the lattice parameters, as well as the positions and orientations of molecules, to cover the search space and explore a broad range of possible crystal structures.[11] However, it yields a large number of candidates, many of which are less dense and less stable. Genetic algorithms and Bayesian optimization aim to identify global minimum through iterative active sampling.[12–17] Genetic algorithms work by modifying or combining local optimized structures, while Bayesian optimization speculates a black-box function by regression and performs active sampling based on an acquisition function. Although both methods are expected to find a global minimum after many rounds or iterations, they also produce numerous less dense and less stable structures in earlier rounds. Recently, researchers have reported a CSP

[a]*Center for Data Science, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan. E-mail: takuya.taniguchi@aoni.waseda.jp*
[b]*Graduate School of Advanced Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan*

approach using a generative adversarial network (GAN) to produce more realistic crystal structures.[19] While this method is innovative, the optimization logic of GANs can be difficult to interpret, and the technique may be limited to specific molecular families or crystal systems that have sufficient training data.

Regarding structure relaxation, conventional approaches typically rely on force fields or density functional theory (DFT) calculations. Force fields enable rapid structural relaxation, but their accuracy may not match that of DFT. In contrast, DFT calculation affords more accurate results depending on calculation level, but is computationally expensive, time-consuming, and requires extensive computational resources. In recent years, neural network potentials (NNPs) trained on DFT data have gained attention for achieving near-DFT-level accuracy at a fraction of the cost.[20–26] For organic crystals, some pre-trained base models such as PFP and ANI have demonstrated efficacy and can, in some instances, surpass quantum chemical methods in accuracy.[27,28] NNPs can also be fine-tuned for specific systems by additional training, making them highly versatile.[29–31] Consequently, NNPs are increasingly used to filter or rank candidate structures within CSP workflows, offering a promising balance between computational efficiency and accuracy.

Although a variety of CSP methods have been proposed as described, there is still a need for approaches that reduce the generation of less dense, less stable structures to improve the efficiency of CSP. Indeed, leveraging predicted density or volume to guide the search is a recognized strategy to enhance efficiency. For example, the recently developed low-energy region explorer (LoreX) predicts an optimal cell volume from fundamental atomic properties to constrain the initial sampling space.[32] This constrain-then-sample approach is highly effective for inorganic systems. Our work builds on this concept but adapts it specifically for organic molecules by employing a different sample-then-filter strategy. We use molecular fingerprint to predict space groups and a target crystal density. The predicted density is then used as a criterion to filter randomly sampled lattice parameters, accepting or rejecting them prior to crystal structure generation. This approach is tailored to capture how the unique functional groups of organic molecules influence crystal packing, and could be combined with genetic algorithms or Bayesian optimization for a synergistic effect. It is also pivotal to investigate the effectiveness of NNP for organic crystals for advancing organic CSP. In this study, we developed a workflow, named SPaDe-CSP, that leverages Space group and Packing Density predictor (SPaDe) to decrease the production of low-density, unstable structures, followed by structure relaxation *via* NNP (Fig. 1). Specifically, we narrowed the search space by predicting space group candidates and crystal density. To clarify which processes were key to CSP success, we evaluated the performance of these machine learning models using a representative molecule. We also examined the generalizability of this workflow on a validation dataset and assessed how SPaDe-CSP improves success rate compared to that of random-CSP, a baseline relying on random structure generation.

Methods

Data curation

The dataset for machine learning was collected from the Cambridge Structure Database (CSD version 5.44). The search conditions are $Z' = 1$, organic, not polymeric, R-factor < 10, no solvent. To ensure the quality and representativeness of our training data, the search result was further filtered based on the statistical distributions of key crystallographic parameters, as shown in Fig. 2. We defined ranges for the lattice lengths ($2 \leq a, b, c \leq 50$ Å) and angles ($60 \leq \alpha, \beta, \gamma \leq 120°$). These criteria were established to encompass the vast majority (>97.9%) of the initial search result, thereby systematically removing extreme outliers and potential erroneous entries from the tails of the distributions. Additional filters for molecular weight (≤1500 g mol^{-1}) and Z value (≤16) were also applied. The data size after these filtering was 170 278. To ensure sufficient data for pattern recognition *via* machine learning, we restricted space groups containing more than 100 entries, resulting in 32 space groups. The data size after this space group filtering was 169 656, and this dataset was used for ML. This ML-dataset covers 99.6% of the filtered search result.

ML training

The curated ML-dataset was split into training and test subsets by 8 : 2. Two machine learning models were constructed for space

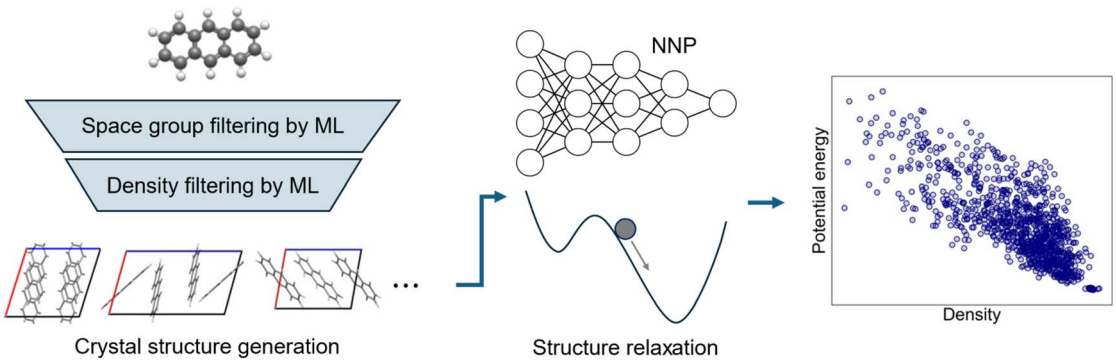

Fig. 1 CSP workflow of this work (SPaDe-CSP). MLs are used for filtering space group candidates and crystal density in crystal structure generation, followed by structure relaxation *via* NNP, affording the energy-density diagram of a molecule.

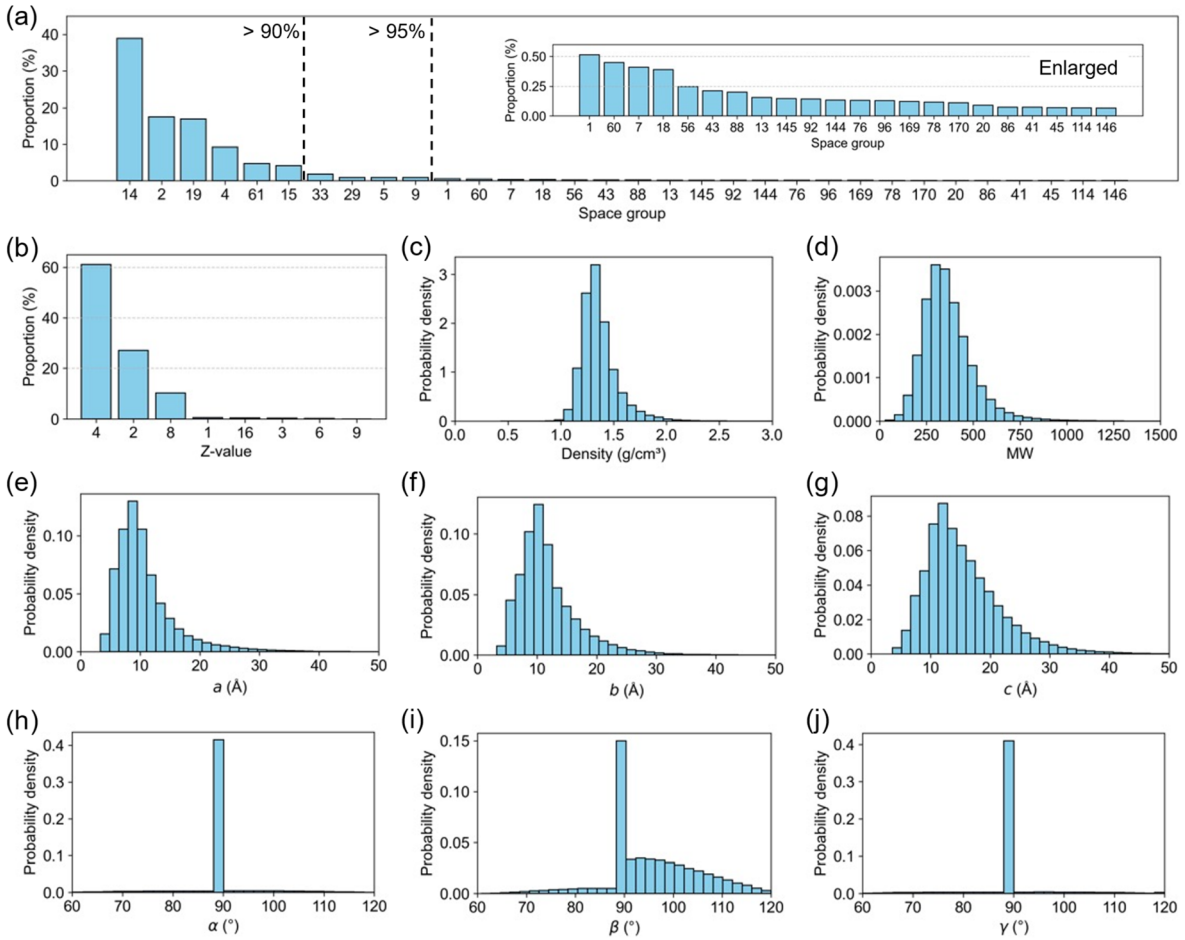

Fig. 2 The distribution of crystallographic information in the ML-dataset. Proportions of (a) space groups and (b) Z values. Probability densities of (c) crystal density, (d) molecular weight, (e) length of a-axis, (f) length of b-axis, (g) length of c-axis, (h) α angle, (i) β angle, and (j) γ angle.

group prediction and density prediction, trained by the training subset and evaluated by the test subset. For both predictions, MACCSKeys was used as molecular fingerprint for the interpretation of the ML result. LightGBM, random forest, and neural network was compared for the ML model. As the loss functions, we used cross-entropy loss for space group prediction and L2 loss for density prediction. Since ML classifier output the probabilities of 32 classes, we set the probability threshold to filter the space group candidates. We evaluated the accuracy and the number of space group candidates in the threshold range of 1×10^{-10} and 0.5 using test subset. For the prediction of crystal density, regression models of LightGBM, random forest, and neural network was compared as well. The prediction ability was evaluated by mean squared error (MAE) and determination coefficient R^2. The molecular fingerprint and the ML training are implemented using rdkit and scikit-learn packages in Python.

CSP

To validate the efficiency of ML-based lattice sampling, we extracted the molecular structures from reference crystal structures. The geometry of an individual molecule was extracted from the corresponding CIF file and then optimized using a pretrained neural network potential PFP[21] version 6.0.0 at MOLECULE mode on Matlantis (**https://matlantis.com/**), software as a service style material discovery tool. We performed the optimization using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method with a residual force threshold of 0.05 eV Å^{-1}.[33]

In the structure generation of random-CSP, when a molecular structure is provided, the PyXtal's function 'from_random' generates crystal structures until 1000 valid structures are generated. The space group is randomly selected from among 32 candidates for each iteration.

In the structure generation of SPaDe-CSP, the SMILES string is converted to a MACCSKeys vector, and the space group candidates and crystal density are predicted by trained LightGBM models. One of the predicted space group candidates is then randomly selected, and lattice parameters are sampled within predetermined ranges of $2 \leq a, b, c \leq 50$ and $60 \leq \alpha, \beta, \gamma \leq 120$. We checked whether the sampled space group and lattice parameters satisfied the density tolerance using molecular weight and Z value, and if they did, we placed the molecules in the lattice. This initial structure generation continues until 1000 crystal structures are produced for each run, and we repeat the run 10 times to evaluate the efficacy and variation of CSP for each compound.

For both CSP approaches, we use the same structure relaxation procedure. The generated structures are optimized with PFP[21] version 6.0.0 at CRYSTAL_U0_PLUS_D3 mode. We employ the limited BFGS (L-BFGS) algorithm, allowing up to 2000 iterations and imposing a residual force threshold of 0.02 eV Å^{-1}. The structural relaxations were performed using the FrechetCellFilter to simultaneously minimize both atomic forces and unit cell stresses, thereby optimizing the atomic positions and lattice parameters. Throughout this process, the FixSymmetry constraint was applied to ensure the initial space group was preserved. To quantify agreement between experimental and calculated structures, we compute the root-mean-square deviation (RMSD) of 30 molecules using the COMPACK algorithm.[34] Both structure generation and optimization are implemented *via* the PyXtal and ASE libraries.[35,36]

Results and discussion

Statistical analysis of curated dataset

We extracted a dataset from the Cambridge Structural Database (CSD) under several filtering criteria for space group classification and density regression (see Method section). In this study, we focus on space groups with more than 100 data entries to guarantee prediction accuracy, resulting in 32 space group candidates with the data size of 169 656 (named ML-dataset). These 32 space groups comprise 99.6% of the search result, reflecting the variety of organic crystals (Fig. 2a). The most frequent space group 14 ($P2_1/c$) occupies nearly 40% of the ML-dataset. The next frequent space groups are group 2 ($P\bar{1}$) and 19 ($P2_12_12_1$), and the top 10 space groups account for 96.0% of the ML-dataset.

Because the Z value is determined by the space group in nearly all cases, the distribution of Z value reflects that of space groups (Fig. 2b). For example, crystals in space groups 14 and 19 always have $Z = 4$, resulting in the most frequent Z value. Space groups 2 and 4 correspond to $Z = 2$, and their combined frequency is the frequency of $Z = 2$. Although there are a few exceptions where Z takes a different value, in general Z value depends on the space group, so we evaluated that there is no need to predict Z value.

Crystal density, molecular weight, and lattice lengths each exhibit a single-peaked smooth distribution. Density mostly falls between 1.0 and 2.0 g cm^{-3}, with a peak around 1.3 g cm^{-3} (Fig. 2c). Molecular weight peaks around 300 g mol^{-1} with a long tail extending to higher values (Fig. 2d). The lengths of a- and b-axes peak around 10 Å, with a longer tail on the right side (Fig. 2e and f). The length of c-axis peaks around 13 Å and shows a broader distribution than a- and b-axes (Fig. 2g). Lattice angles are often constrained to 90° by the space group symmetry, resulting in unique distributions (Fig. 2h–j). Among these, β angle has a different distribution because the triclinic and monoclinic crystal systems, which account for 72.6% of the ML-dataset, has a unique angle that is frequently larger than 90°. As angles move away from 90°, their frequency decreases in all cases.

Prediction of space group and crystal density

To achieve efficient lattice sampling, we employed two machine learning tasks: space group prediction and density prediction. For space group prediction, we used MACCSKeys as the molecular fingerprint and LightGBM as the prediction function based on predictability and interpretability. Among 32 space

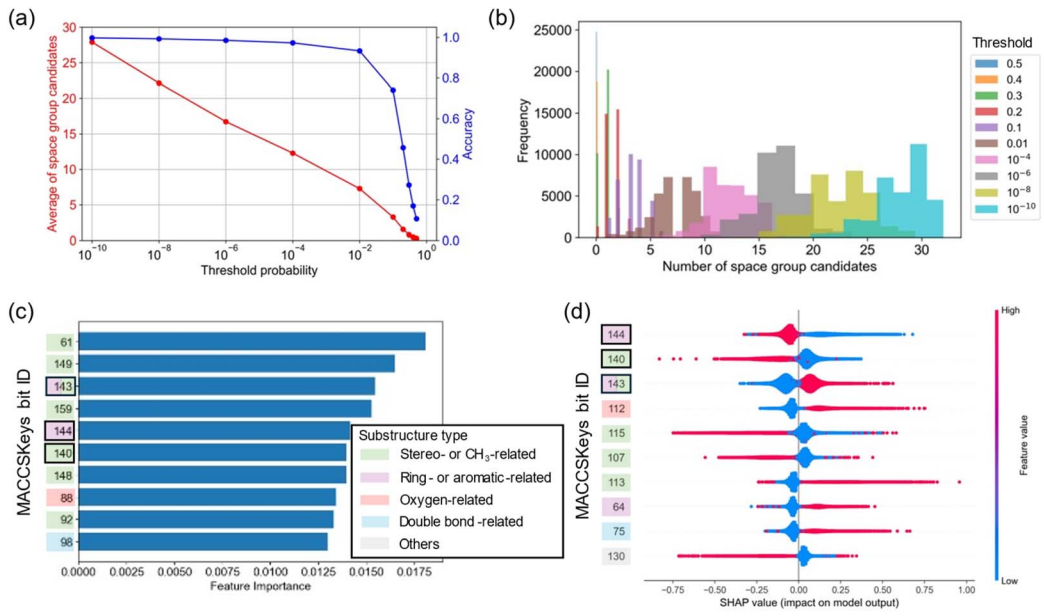

Fig. 3 ML result of space group classification. (a) The dependence of threshold probability on the average of the number of space group candidates and the prediction accuracy. (b) The distribution of the number of space group candidates. (c) Top 10 MACCSKeys bits based on feature importance of the trained model. MACCSKeys bits are categorized into some substructure types as highlighted in colors, based on our knowledge. (d) Top 10 MACCSKeys bits based on SHAP analysis of the trained model. Highlighted bits with black frame line appeared in common with feature importance.

groups, those with predicted probabilities exceeding a given threshold were extracted as candidates, and the accuracy was evaluated whether the true space group was contained in the candidates. When the threshold was below 10^{-2}, the accuracy was above 0.90, but it dropped sharply as the threshold increased (Fig. 3a). This phenomenon occurs because, as the threshold is raised, the number of space group candidates becomes smaller, thereby decreasing the chance that the true space group survives. At a threshold of 10^{-10}, an average of 28 space groups survived, and the number of space group candidates is widely distributed between 20 and 32 depending on molecules (Fig. 3b). As the threshold increased, the distribution narrowed and shifted to the left because the number of candidates becomes smaller. Although the prediction accuracy depends on the threshold, space group classifier works to narrow down the number of candidates if we set suitable threshold such as $<10^{-2}$. For comparison with baselines, if we choose a space group randomly from 32 candidates, the accuracy of the random selection is 3.1%. If we choose a space group randomly based on the frequencies of each class in the training subset, the accuracy of the weighted random selection is 22.2%. ML model achieves higher accuracy than these baselines.

It is important to ensure interpretability of the trained model. An examination of the top-ranking substructures in feature importance of LightGBM indicates that structural characteristics such as stereochemistry and the presence of methyl groups have strong influence (Fig. 3c and SI Fig. 1). This observation suggests that molecular conformation such as the type and number of substituents, and the ring environment serve as major determinants in classifying space groups. Because space groups describe the symmetry of crystals, factors such as substituents, ring structures, and stereochemical substructures are often critically important. Consequently, assigning high importance to these features within the model is justified.

Furthermore, Shapley additive explanations (SHAP) analysis was conducted to interpret whether these features contribute positively or negatively.[37,38] Here, we present the result for the most frequent space group since positive or negative effect can be visualized in each class (Fig. 3d). Consistent with the feature importance findings, the top 10 SHAP features include the presence of multiple six-membered rings, the number of methyl groups, and oxygen-related substructures. SHAP analysis further clarifies. For instance, bit ID 143 represents a substructure in which a bond transitions from "not aromatic" to "aromatic" and then back to "not aromatic," and it exhibits a positive contribution. Because this bit broadly captures configurations where two aromatic rings are connected by a rotatable single bond, it is hypothesized that such compounds can adopt diverse conformations upon crystallization and readily form stable packing arrangements *via* π–π or CH–π interactions between the aromatic rings.

Next, we performed regression of crystal density using the combination of MACCSKeys and LightGBM as well. The metrics for the training ($R^2 = 0.85$, MAE = 0.0044) and test subsets ($R^2 = 0.80$, MAE = 0.049) showed no significant deviation, indicating that overfitting did not occur (Fig. 4a). Since mean model, which assumed no correlation between molecular structure and density, afforded MAE = 0.125 g cm^{-3}, the prediction accuracy was sufficient. However, a negative bias was observed, where the predicted values were higher in low-density regions and lower in high-density regions (Fig. 4b). We reasoned that the source of this bias lies in attempting to predict a crystal property solely from the molecular structure. Because crystal density is

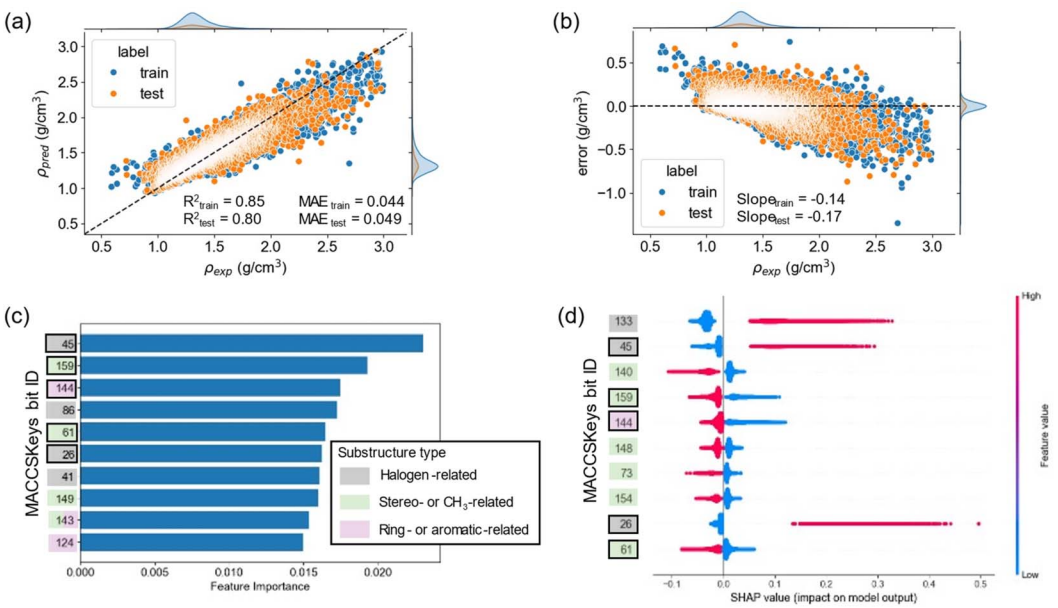

Fig. 4 ML result of density regression. (a) The parity plot and distribution of crystal density. (b) Error plot and distribution of the prediction, where error was defined as predicted value minus experimental value. (c) Top 10 MACCSKeys bits based on feature importance of the trained model. MACCSKeys bits are categorized into some substructure types as highlighted in gray, green, and purple. (d) Top 10 MACCSKeys bits based on SHAP analysis of the trained model. Highlighted bits with black frame line appeared in common with feature importance.

influenced by intermolecular interactions, using only molecular structure to make such predictions has inherent limitations. This negative bias also appeared in other ML models, including graph neural networks in our preliminary validation. Nevertheless, given that most errors are distributed around zero, we consider the regression performance to be acceptable.

We also interpreted the trained model using feature importance and SHAP analysis. Unlike the case of space group, substructures related to halogens had a substantial impact. Among the top 10 bits in the feature importance, four corresponds to halogen-related substructures, with the top-ranking bit indicating whether Br atom was present (Fig. 4c and SI Fig. 2). The SHAP analysis revealed that all top-ranking substructures related to halogens contributed positively, whereas other substructures contributed negatively (Fig. 4d). Halogens are generally incorporated into the molecule by substituting for hydrogen; in the ML-dataset, the average density of crystals without halogens was 1.309 g cm^{-3} (72.6%), whereas that of crystals with halogens was 1.514 g cm^{-3} (27.4%). This difference exerted the strongest influence on the density prediction. Besides halogens, substructures involved in hydrogen bonding—for example, the presence of carbonyl oxygen—also contributed positively, though to a lesser extent than halogens in the density prediction. Thus, since the combination of MACCSKeys and LightGBM effectively captured the relationship between molecular structure, space group, and density, we adopted these ML models for CSP.

CSP of a model molecule

In the SPaDe-CSP workflow, we used these ML models to narrow down the space group candidates and the lattice combinations (Fig. 5a). When a SMILES was given, we begin by predicting the space group and density (step 1), and sampling lattice constants within its defined range (step 2). The sampled lattice is then checked against the predicted density (step 3). If it falls within the acceptable range, a molecule is placed within the cell to generate an initial, unrelaxed structure (step 4). Finally, this structure undergoes relaxation to find a local energy minimum (step 5). Two hyperparameters are introduced in this process: one is the probability threshold for filtering space groups, and the other is the tolerance window (w) for the crystal density. To investigate the dependence on these parameters, we performed CSP using a model molecule (CSD code: NISNAE) (Fig. 5b). This molecule has a simple structure, and the crystal belongs to the space group 4 ($P2_1$) *via* typical intermolecular interactions. Hydrogen bonding chains are formed along the b-axis, and these chains are arranged by van der Waals forces. For structure relaxation, we used the pretrained neural network potential PFP,[21] which was used for structure optimization of organic crystals with sufficient accuracy.[27]

We compared the CSP success rates over ten trials, each of which involved generating 1000 initial structures followed by structure relaxation on PFP. A trial was recognized as successful if at least one among 1000 relaxed structures matched with the reference structure based on root mean square deviation of 30

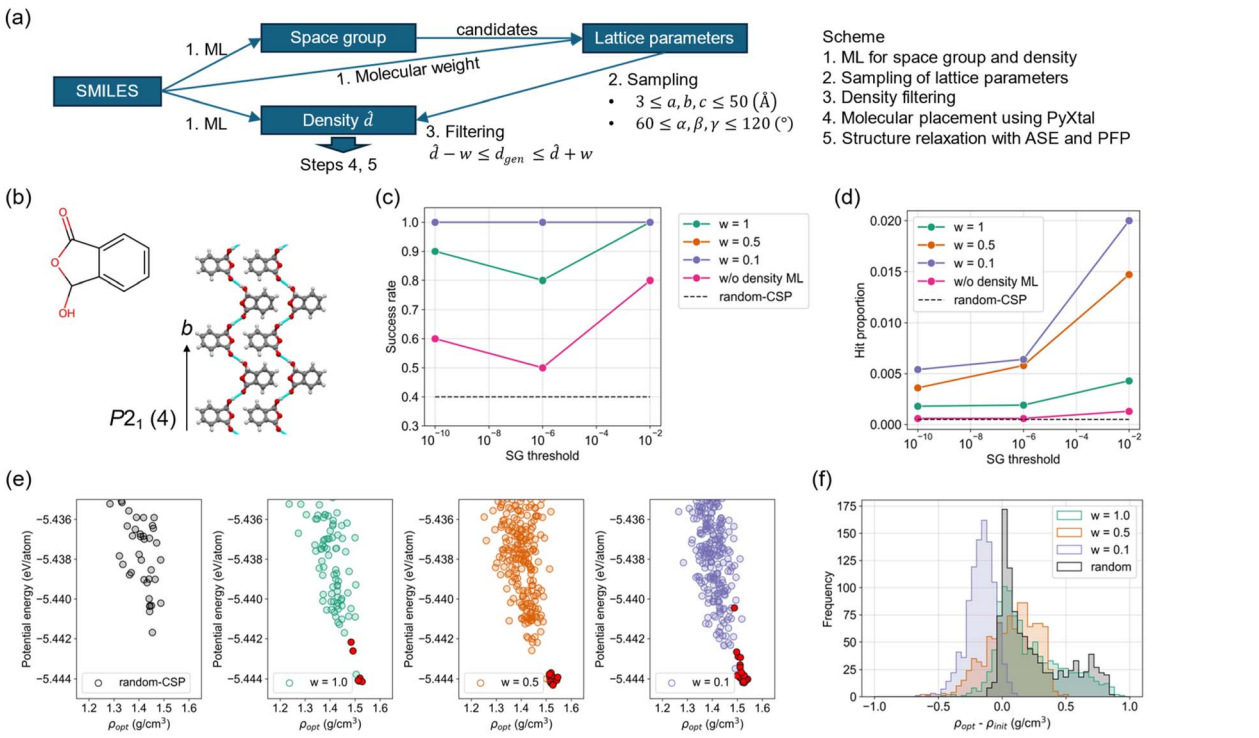

Fig. 5 Dependence of hyperparameters on CSP of a model molecule. (a) Scheme to generate valid crystal structures using ML models. (b) Molecular and crystal structures of the model molecule (CSD code: NISNAE). (c) Success rate of CSP depending on space group threshold and density tolerance window, w. (d) Hit probability of predicted structures matching with the reference structure. (e) Energy-density diagram of the model molecule at the space group threshold of 10^{-2}. Red plots show the structures matching with the reference structure based on $RMSD_{30}$ < 0.8. (f) Distribution of the difference of crystal density before and after structure relaxation.

molecules ($RMSD_{30}$). When density tolerance $w = 0.1$ and 0.5, the success rate was 1.0 regardless of the space group threshold (Fig. 5c). In contrast, when $w = 1.0$, the success rate was not 1.0 at lower thresholds but reached 1.0 at higher threshold. These results exceeded the success rate obtained when structures were generated without any ML (random-CSP). Moreover, when using only the space group prediction (*i.e.*, without ML for density), the success rate was intermediate among those values.

When we calculated the probability of encountering a structure that matches with the reference structure, the hit probability increases as the space group threshold becomes larger and the density tolerance becomes smaller (Fig. 5d). For example, at $w = 0.1$ and threshold of 10^{-10}, 5 out of 1000 generated structures matches with the reference structure on average. At threshold of 10^{-2}, 20 out of 1000 structures matched on average. Since space group prediction narrows down space group candidates and density prediction constraints lattice parameters close to stable structures, their synergistic effect contributes to the increase of success rate and hit probability.

Typical energy-density diagrams at threshold of 10^{-2} verify the effectiveness of ML models and the stability of predicted structures matching with the reference structure. Random-CSP of the first trial among 10 trials did not find a matched structure, while SPaDe-CSP with any density tolerance found several structures matching with the reference (Fig. 5e). The structure with the lowest potential energy matches with the reference structure, and this verifies the adaptability of PFP. The number of structures in the region of high dense and low energy structures increased depending on narrowing the density tolerance, leading to the increase of hit probability.

Here, it is important to understand how density constraint works on the structure relaxation. Comparing the crystal densities of the initial unrelaxed structures with those after structure relaxation, the density difference in each density tolerance at threshold of 10^{-2} showed characteristic distributions (Fig. 5f). The density difference of random-CSP distributed positive region, which means that relaxed structures are more dense than initial unrelaxed structures. This is because initial structures are generated by relatively large lattice parameters and then optimized to more dense structure through structure relaxation. The distribution at $w = 1.0$ showed similar character. In contrast, the density difference at $w = 0.5$ is distributed across both positive and negative values. This indicates that overly dense unrelaxed structures were generated and then became less dense structures through structural relaxation. The density difference at $w = 0.1$ is distributed in more negative values, indicating that too dense unrelaxed structures are generated. It is estimated that an overly dense structure is in a steep region of the potential energy surface due to intermolecular repulsion, so it would require fewer iterations for structure relaxation than a less dense structure. Indeed, setting a smaller density tolerance shortened the optimization time for each structure on average (SI Fig. 3). On the other hand, a smaller density tolerance restricts the acceptance criteria, necessitating more time to generate valid structures. Consequently, striking a balance between these effects led us to select $w = 0.5$ as the most appropriate tolerance value, and we employed this setting in the subsequent validation.

Generalization ability

To assess the generalization performance of SPaDe-CSP, we tested whether it could predict 20 crystal structures including the model molecule (Fig. 6). They were selected to represent a range of space groups and molecular structures. In each compound, molecular geometry was extracted from the reference crystal structure to validate the workflow. We then

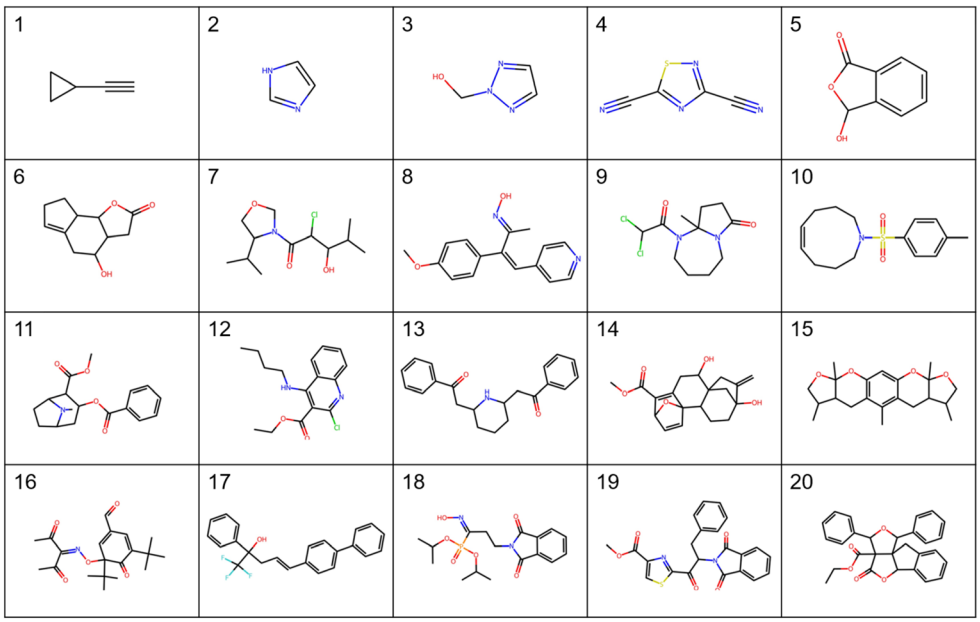

Fig. 6 Molecular structures used for validation in ascending order by molecular weight.

performed random- and SPaDe-CSP to evaluate success rates over ten trials, each of which involved 1000 structure generation and relaxation. Since the limitation of space group threshold differs depending on molecules, the success rate at the maximum threshold among several attempts are used for this comparison (SI Table 1). The proportion of compounds for which at least one predicted structure matched the reference structure was 80.0% (16 out of 20) for SPaDe-CSP (Table 1). This value is twice as high as that achieved by random-CSP (40.0%, 8 out of 20), indicating the effectiveness of ML-based filtering. Moreover, among these compounds, 81.3% (13 out of 16) saw an improved success rate with SPaDe-CSP compared to random-CSP. Based on the results, we grouped these crystals into three categories: (1) those for which random-CSP worked well and there is no room for improvement; (2) those for which the SPaDe-CSP improved the success rate; and (3) those for which SPaDe-CSP did not improve the success rate even though there is room for improvement.

The first category includes entry numbers 4 and 6 (CEBYUD and BAQBUR), which resulted in a success rate of 1.0 using both random- and SPaDe-CSP (Table 1). A key factor in this category is that these crystals either have fewer lattice degrees of freedom or have smaller lattice size. CEBYUD belongs to the space group $P3_2$, giving only two degrees of freedom for the lattice (lengths of the *a*- and *b*-axes). Even with random-CSP, once $P3_2$ is selected from 32 space group candidates, it is more likely to produce an initial structure that converges to the correct local minimum due to low degrees of freedom for the lattice. Random- and SPaDe-CSP showed similar distributions in the energy-density diagram, while SPaDe-CSP yielded slightly more stable high-density structures (Fig. 7a and b). The distribution of initially generated space groups confirms that SPaDe-CSP effectively narrowed the candidates including the experimentally observed one, whereas random-CSP sampled a wider range (Fig. 7c). In addition, SPaDe-CSP preferentially generated structures closer to the experimental density, and the difference between initial and optimized densities was smaller compared to random-CSP (Fig. 7c). These observations highlight that even when the success rate is saturated, the predictors in SPaDe-CSP improve the quality of the initial structural pool, thereby reducing unnecessary optimization steps.

The other crystal BAQBUR belongs to the space group *P*1 and has a small unit cell due to $Z = 1$. A small unit cell reduces the complexity of the search space. This leads to fewer local minima on the potential energy surface, making it easier to converge on the global minimum (SI Fig. 4). Furthermore, a *P*1 crystal have enough flexibility to represent the same crystal in multiple lattice sets due to six degrees of freedom for lattice. Consequently, the high success rate of CSP for BAQBUR can be attributed to both its small unit cell and the characteristics of the *P*1 space group.

The second category includes systems for which SPaDe-CSP achieved higher success rates than random-CSP (Table 1). Although the degree of improvement varies, compounds that had higher success rates of random-CSP tend to show higher success rates with SPaDe-CSP. In addition to the model molecule used for the hyperparameter study, 12 other molecules fell into this category (entry numbers 1, 2, 3, 5, 7, 8, 9, 10, 11, 13, 15, 16, 19). Crystals in this category tend to have limited lattice degrees of freedom or moderate unit cell sizes. In a case, FAMDUS has space group *P*1 and a larger molecular weight than BAQBUR. This demands a larger cell and thus more possible combinations of lattice parameters, decreasing the success rate in random-CSP. With SPaDe-CSP, however, its success rate rose to 1, presumably because space group prediction narrowed down the candidates from 32 to 15, and density prediction preferentially filtered lattice combinations near the stable structure.

Table 1 CSP metrics of validation dataset including NISNAE

Entry	CSD code	M (g mol^{-1})	N_{rot}	SG	Z	N_{DoF}	V (Å^3)	Random-CSP	SPaDe-CSP
1	MEYCIC	66.1	0	*Pbca*	8	3	860.727	0.2	0.5
2	IMAZOL15	68.1	0	$P2_1/c$	4	4	355.508	0.1	0.9
3	MOTLAL	99.1	1	$P2_1/c$	4	4	467.487	0	0.1
4	CEBYUD	136.1	0	$P3_2$	3	2	426.65	1	1
5	NISNAE	150.1	0	$P2_1$	2	4	335.252	0.3	1
6	BAQBUR	194.2	0	*P*1	1	6	246.267	1	1
7	FAMDUS	263.8	0	*P*1	1	6	355.334	0.7	1
8	LOMPUY	268.3	4	$P2_1/c$	4	4	1484.313	0	0.5
9	WURVEM	279.2	1	$P2_1/c$	4	4	1254.9	0	0.3
10	CINYOO01	279.4	4	$P2_1/c$	4	4	1439.328	0	0.1
11	COCAIN10	303.4	2	$P2_1$	2	4	807.479	0.1	0.7
12	HUFXAH	306.8	3	$P\bar{1}$	2	6	813.864	0.2	0.2
13	JOLLUT	321.4	6	$P2_1/c$	4	4	1739.135	0	0.1
14	GEZPIK	330.4	6	$P2_1/c$	4	4	1641.079	0	0
15	XILPAN	344.4	1	$P2_1/c$	4	4	1859.966	0	0.1
16	BESLOE	361.4	0	*Pbca*	8	3	3996.051	0	0.5
17	HETTUZ	368.4	5	$P2_1/c$	4	4	1750.583	0	0
18	SIKFIB	382.3	5	$P2_1/c$	4	4	1982.361	0	0
19	QEVWUJ	420.4	8	$P2_1/c$	4	4	1978.025	0	0.2
20	CIDTAN	440.5	4	$P2_1/c$	4	4	2272.42	0	0

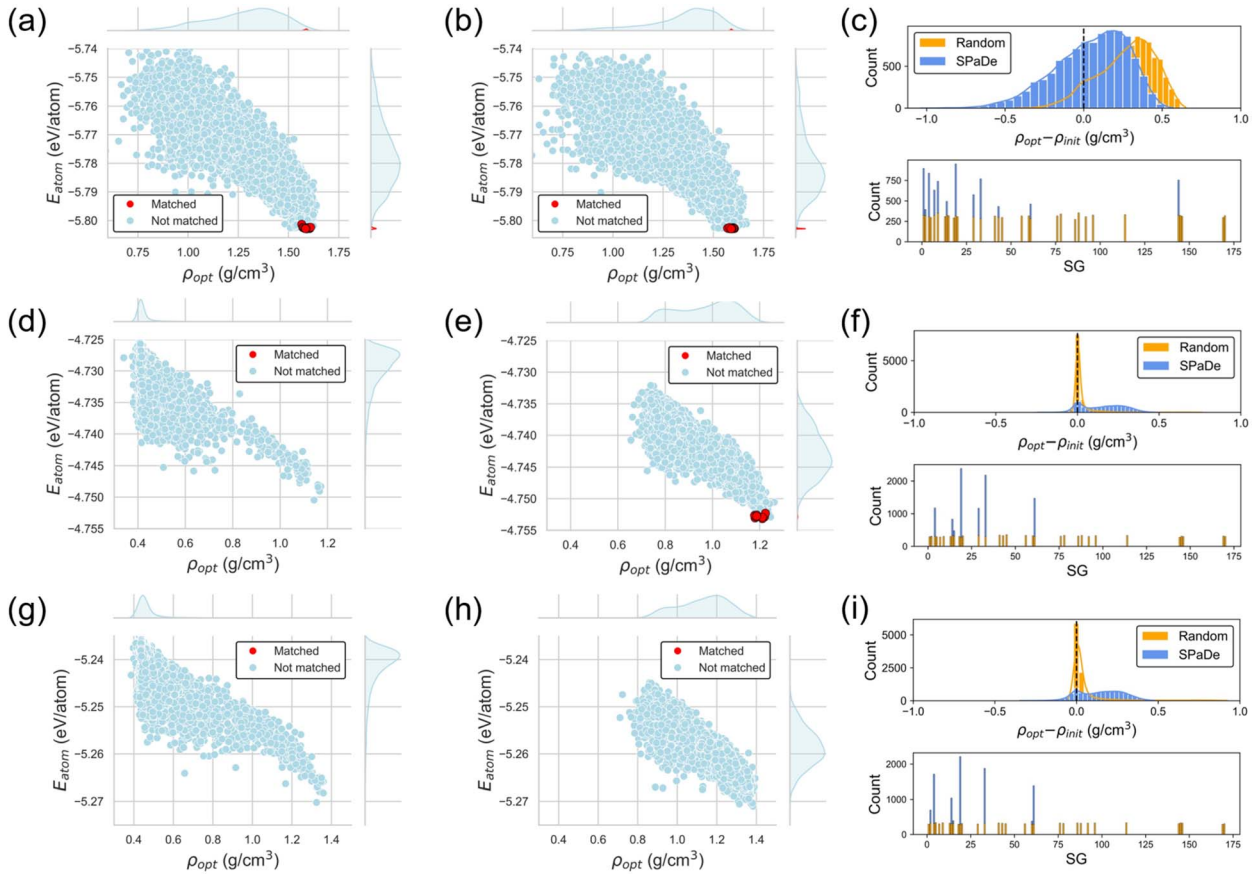

Fig. 7 Energy-density diagrams of some compounds by random- and SPaDe-CSP. CSP result of (a–c) CEBYUD in the first category, (d–f) BESLOE in the second category, and (g and f) HETTUZ in the third category. Left panels (a, d and g) show the results of random-CSP, and middle panels (b, e and h) show those of SPaDe-CSP, and right panels (c, f and i) show the distribution of space groups and the difference in density between the initial and optimized structures. The results of 10 trials are integrated in each graph.

In this category, two crystals NISNAE and COCAIN10 belongs to the space group $P2_1$. In both cases, the success rate of SPaDe-CSP increased more than double compared to random-CSP. The key difference between them lies in their molecular weights: COCAIN10 (303.35 g mol^{-1}) is larger and more structurally complex than NISNAE (150.13 g mol^{-1}), making CSP more challenging. Accordingly, NISNAE, which had a higher success rate in random-CSP, also showed a higher success rate under SPaDe-CSP.

Another noteworthy example with strong SPaDe-CSP effects is BESLOE and MEYCIC, which belong to the space group *Pbca*. This space group is attributed to the orthorhombic crystal system, and has all angles fixed at 90°. This is why only the three lattice lengths can change once the correct space group is selected. Since this space group corresponds to $Z = 8$, the unit cell is relatively large than cells with more frequent $Z = 2$ and 4, resulting in more possible lattice combinations. Random-CSP of BESLOE afforded relatively low success rates for these crystals because the difference between the initial and optimized densities of the random-CSP is distributed near zero, which means that the loose initial structures hardly became dense through structural relaxation (Fig. 7d and f). With SPaDe-CSP, the success rate improved to 0.5 by increasing the probability of generating more dense initial structures (Fig. 7e and f). MEYCIC resulted in higher success rate than BESLOE probably due to smaller molecular weight (SI Fig. 4).

The final category comprises structures for which SPaDe-CSP did not improve the success rate even though there is room for improvement. This category contains 5 structures in total, one belonging to $P\bar{1}$ and four belonging to $P2_1/c$ (entry numbers 12, 14, 17, 18, and 19). HUFXAH which belongs to $P\bar{1}$ has six degrees of freedom for lattice and $Z = 2$, showing a success rate of 0.2 under both random- and SPaDe-CSP. When compared with crystals of space group $P1$ (BAQBUR and FAMDUS) which achieved higher success rates, HUFXAH requires twice the unit cell volume of those other systems due to $Z = 2$. Even though ML narrows down the space group candidates, six lattice degrees of freedom and a larger cell would lead to unchanged success rate with SPaDe-CSP.

The other four structures resulted in success rates of 0 with both random- and SPaDe-CSP. It is sure that SPaDe-CSP improved to generate more stable high-density structures than random-CSP, while the matched structure based on $RMSD_{30}$ was not obtained (Fig. 7g–i and SI Fig. 4). This insists that molecular arrangement was not matched with the reference even when the lattice was sufficiently similar. This is probably

because the inter-sectional effect of degree of freedom for lattice and lattice/molecular size made it difficult to match with the reference structure.

To quantify the factors that determine the success or failure of CSP, we investigated which parameters are correlated to the success rate. Molecular-and crystal-level descriptors included in Table 1 are used for this analysis. Since some descriptors have high correlations with each other, we picked up one to exclude multi-collinearity (SI Fig. 5). For clarity in interpretation, a linear model was adopted, and multiple approaches to incorporate descriptors were tested (SI Table 2). Based on the Bayesian information criteria (BIC), we selected the linear regression model that uses the descriptors calculated according to the following formula.

$$X = rN_{\mathrm{DoF}}{}^{2}Z + (1 - r)M$$

Here, N_{DoF} is the degree of freedom for lattice, Z is Z value in a unit cell, and M is the molecular weight, each of which are regularized by maximum value. The coefficient r is the ratio to consider crystal-level and molecular-level effects. The r was optimized to 0.942 by random-CSP, and the linear regression resulted in adjusted R^2 of 0.635, showing that the success rate was sufficiently explained by the constructed descriptor (Fig. 8a). Since the descriptor considers crystal-level effect $N_{\mathrm{DoF}}{}^{2}Z$ nearly 95%, such tendency largely depends on the first term. The quadratic term $N_{\mathrm{DoF}}{}^{2}$ should capture that nonlinear growth in the search space. When this descriptor was used for the regression of SPaDe-CSP, the linear regression also afforded sufficient adjusted R^2 of 0.631, indicating the robustness of the descriptor (Fig. 8b). The regression lines obtained for random- and SPaDe-CSP have nearly the same slope but different intercepts. Since the difference of intercept should reflect the effect of MLs, it can be inferred that SPaDe-CSP provides an overall improvement of 20–30% in the success rate (Fig. 8b).

Although the present benchmark focused on comparing SPaDe-CSP with random-CSP to isolate and demonstrate the effect of space group and density filtering, it should be emphasized that these predictors can also be incorporated into global optimization frameworks such as genetic algorithms (GA) or Bayesian optimization (BO). In GA, for example, the space group predictor reduces the candidate set from 32 common space groups to 7–8 on average (threshold = 10^{-2}, applicable to ~90% of molecules), while the density predictor further constrains lattice parameters. These predictors therefore provide effective guidance for defining the search domain and generating the initial population, and are expected to synergistically improve the efficiency of GA- or BO-based CSP workflows.

Limitations

The SPaDe-CSP workflow improved the success rate compared to random-CSP, but there are several limitations. The first limitation is the optimization of molecular conformations. In the current study, CSP was performed using molecular conformations extracted from reference crystal structures. While this constraint is useful for evaluating the effects of machine learning in lattice sampling, it cannot be applied when the crystal structure is unknown. In such cases, it is necessary to generate possible molecular conformations and perform CSP to search for the globally stable structure like the case of ROY polymorphs.[39] For molecules with many rotatable bonds, the number of possible conformations increases, leading to larger search space for CSP.

The second limitation is the temperature-effect on the crystal stability. The lattice parameters of organic crystals are known to be more susceptible to temperature changes than inorganic crystals, leading to larger thermal expansion coefficients. In the 20 crystals used for the present verification, those that appeared most stable at 0 K generally matched the reference structures. However, in other cases, the stability of crystal structures at 0 K could differ from that near room temperature. Accounting for temperature effects requires calculating the Gibbs free energy, for which various computational approaches have been proposed.[40–43] Yet, DFT-based calculations of Gibbs free energy are computationally expensive. Because NNPs reduce such computational demands, they are considered relatively easy to

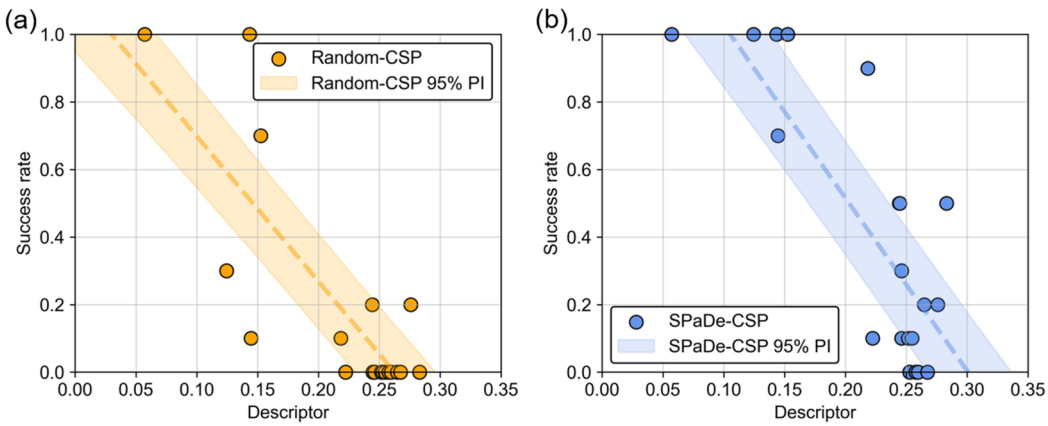

Fig. 8 Relationship between CSP-descriptor and success rate. (a) Random-CSP. (b) SPaDe-CSP. The dashed lines are the linear regression, and the highlighted region are the 95% prediction interval.

introduce into the SPaDe-CSP workflow. We focused on the efficiency of lattice sampling in this work and therefore will incorporate Gibbs free energy calculations in future work.

We would like to stress once more the distinction between SPaDe-CSP and commonly used structure-search techniques like GA and BO. While these techniques parametrize crystal structures and actively search for the global optimum, the current SPaDe-CSP method is a technique that filters space groups and lattice constants. These approaches are not meant to replace each other one-to-one; rather, they can be synergistically combined.

Conclusions

In this work, we presented an ML-based approach to CSP. By predicting space group candidates and crystal density, we narrowed down the possible lattice combinations, and a neural network potential accelerated structure relaxation with sufficient accuracy. For 80% of the compounds tested, the SPaDe-CSP method successfully predicted the experimental structure, achieving a success rate twice that of random-CSP and indicating the effectiveness of its space group and density filtering. Furthermore, among these compounds, 81.3% showed an improved success rate with SPaDe-CSP over random-CSP. Because the success rate decreases as lattice and molecular sizes increase, we quantitatively investigated the relationship between success rate and structural descriptors. We identified one descriptor that correlated linearly with success rate, reflecting both crystal- and molecule-level structural influences. Although SPaDe-CSP has some limitations, this workflow should aid the efficient design and screening of organic crystal structures.

Author contributions

T. T.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing – original draft, and writing – review & editing. R. F.: validation.

Conflicts of interest

The author received a joint research funding from ENEOS Corporation and licensed Matlantis for using PFP model from Matlantis Corporation, a joint venture of Preferred Networks, Inc. and ENEOS Corporation.

Data availability

The Python code used in this study have been uploaded to **https://github.com/takuyhaa/SPaDe-CSP**, and archived at **https://doi.org/10.5281/zenodo.17214315**. The results of crystal structure prediction have been archived at **https://doi.org/10.6084/m9.figshare.29043899.v1**. Please note that both the CSD Python API and PFP require paid licenses for use.

Supplementary information is available. See DOI: **https://doi.org/10.1039/d5dd00304k**.

Acknowledgements

This study was financially supported by JSPS Grant-in-Aid (22K14747 and 24K17748), the Waseda University Grant for Special Research Projects (2022C-313, 2023C-292, 2023R-050, 2024C-297), JST ACT-X (JPMJAX23DD), Sumitomo Foundation (2432210), and ENEOS Corporation. The corresponding author also thanks for Haruki Hotta and Yumi Kumano for the assistance of analysis of CSP results.

References

1 S. L. Price, From crystal structure prediction to polymorph prediction: interpreting the crystal energy landscape, *Phys. Chem. Chem. Phys.*, 2008, **10**, 1996–2009.
2 J. Yang, S. De, J. E. Campbell, S. Li, M. Ceriotti, *et al.*, Large-scale computational screening of molecular organic semiconductors using crystal structure prediction, *Chem. Mater.*, 2018, **30**, 4361–4371.
3 T. Taniguchi, M. Hosokawa and T. Asahi, Graph comparison of molecular crystals in band gap prediction using neural networks, *ACS Omega*, 2023, **8**, 39481–39489.
4 M. O'Shaughnessy, J. Glover, R. Hafizi, M. Barhi, R. Clowes, *et al.*, Porous isoreticular non-metal organic frameworks, *Nature*, 2024, **630**, 102–108.
5 C. Sato, S. Dekura, H. Sato, K. Sambe, T. Takeda, *et al.*, Proton conduction in chiral molecular assemblies of azolium–camphorsulfonate salts, *J. Am. Chem. Soc.*, 2024, **146**, 22699–22710.
6 G. J. Beran, Frontiers of molecular crystal structure prediction for pharmaceuticals and functional organic materials, *Chem. Sci.*, 2023, **14**, 13290–13312.
7 J. Nyman and S. M. Reutzel-Edens, Crystal structure prediction is changing from basic science to applied technology, *Faraday Discuss.*, 2018, **211**, 459–476.
8 J. Kendrick, F. J. Leusen and M. A. Neumann, Empirical van der Waals corrections to solid-state density functional theory: iodine and phosphorous containing molecular crystals, *J. Comput. Chem.*, 2012, **33**, 1615–1622.
9 M. K. Dudek and K. Drużbicki, Along the road to crystal structure prediction (CSP) of pharmaceutical-like molecules, *CrystEngComm*, 2022, **24**, 1665–1678.
10 L. M. Hunnisett, J. Nyman, N. Francia, N. S. Abraham, C. S. Adjiman, *et al.*, The seventh blind test of crystal structure prediction: structure generation methods, *Acta Crystallogr., Sect. B:Struct. Sci., Cryst. Eng. Mater.*, 2024, **80**, 6.
11 D. H. Case, J. E. Campbell, P. J. Bygrave and G. M. Day, Convergence properties of crystal structure prediction by quasi-random sampling, *J. Chem. Theory Comput.*, 2016, **12**, 910–924.
12 A. R. Oganov, A. O. Lyakhov and M. Valle, How evolutionary crystal structure prediction works—and why, *Acc. Chem. Res.*, 2011, **44**, 227–237.
13 F. Curtis, X. Li, T. Rose, A. Vazquez-Mayagoitia, S. Bhattacharya, *et al.*, GAtor: a first-principles genetic algorithm for molecular crystal structure prediction, *J. Chem. Theory Comput.*, 2018, **14**, 2246–2264.

14 A. Kadan, K. Ryczko, A. Wildman, R. Wang, A. Roitberg, *et al.*, Accelerated organic crystal structure prediction with genetic algorithms and machine learning, *J. Chem. Theory Comput.*, 2023, **19**, 9388–9402.
15 Q. Zhu and S. Hattori, Automated high-throughput organic crystal structure prediction via population-based sampling, *Digital Discovery*, 2025, **4**, 120–134.
16 T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, *et al.*, Crystal structure prediction accelerated by Bayesian optimization, *Phys. Rev. Mater.*, 2018, **2**, 013803.
17 T. Yamashita, H. Kino, K. Tsuda, T. Miyake and T. Oguchi, Hybrid algorithm of Bayesian optimization and evolutionary algorithm in crystal structure prediction, *Sci. Technol. Adv. Mater.:Methods*, 2022, **2**, 67–74.
18 Y. Wang, J. Lv, L. Zhu and Y. Ma, Crystal structure prediction via particle-swarm optimization, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2010, **82**, 094116.
19 Z. Ye, N. Wang, J. Zhou and D. Ouyang, Organic crystal structure prediction via coupled generative adversarial networks and graph convolutional networks, *Innovation*, 2024, **5**, 2.
20 K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko and K. R. Müller, SchNet: a deep learning architecture for molecules and materials, *J. Chem. Phys.*, 2018, **148**, 241722.
21 S. Takamoto, C. Shinagawa, D. Motoki, K. Nakago, W. Li, *et al.*, Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements, *Nat. Commun.*, 2022, **13**, 2991.
22 I. Batatia, P. Benner, Y. Chiang, A. M. Elena and D. P. Kovács, *et al.*, A foundation model for atomistic materials chemistry, *arXiv*, 2023, preprint, arXiv:2401.00096 DOI: **10.48550/arXiv.2401.00096**.
23 C. Devereux, J. S. Smith, K. K. Huddleston, K. Barros, R. Zubatyuk, *et al.*, Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens, *J. Chem. Theory Comput.*, 2020, **16**, 4192–4202.
24 B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, *et al.*, CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling, *Nat. Mach. Intell.*, 2023, **5**, 1031–1041.
25 C. Chen and S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, *Nat. Comput. Sci.*, 2022, **2**, 718–728.
26 P. R. Kaundinya, K. Choudhary and S. R. Kalidindi, Prediction of the electron density of states for crystalline compounds with atomistic line graph neural networks (ALIGNN), *J. Miner. Met. Mater. Soc.*, 2022, **74**, 1395–1405.
27 T. Taniguchi, Exploration of elastic moduli of molecular crystals via database screening by pretrained neural network potential, *CrystEngComm*, 2024, **26**, 631–638.
28 Q. Zhu and S. Hattori, Organic crystal structure prediction and its application to materials design, *J. Mater. Res.*, 2023, **38**, 19–36.
29 P. W. Butler, R. Hafizi and G. M. Day, Machine-learned potentials by active learning from organic crystal structure prediction landscapes, *J. Phys. Chem. A*, 2024, **128**, 945–957.
30 T. Taniguchi, Knowledge distillation of neural network potential for molecular crystals, *Faraday Discuss.*, 2025, **256**, 139–155.
31 H. Kaur, F. Della Pia, I. Batatia, X. R. Advincula, B. X. Shi, *et al.*, Data-efficient fine-tuning of foundational models for first-principles quality sublimation enthalpies, *Faraday Discuss.*, 2025, **256**, 120–138.
32 C. N. Li, H. P. Liang, S. Xu, H. Wang, B. Zhao, *et al.*, LoreX: A Low-Energy Region Explorer Boosts Efficient Crystal Structure Prediction, *J. Am. Chem. Soc.*, 2025, **147**, 9544–9555.
33 R. Fletcher. *Practical Methods of Optimization*, Wiley, New York, 1980, vol. 1.
34 J. A. Chisholm and S. Motherwell, COMPACK: a program for identifying crystal structure similarity using distances, *J. Appl. Crystallogr.*, 2005, **38**, 228–231.
35 S. Fredericks, K. Parrish, D. Sayre and Q. Zhu, PyXtal: a python library for crystal structure generation and symmetry analysis, *Comput. Phys. Commun.*, 2021, **261**, 107810.
36 A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, *et al.*, The atomic simulation environment—a python library for working with atoms, *J. Phys.:Condens. Matter*, 2017, **29**, 273002.
37 S. M. Lundberg and S. I. Lee, A unified approach to interpreting model predictions, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 4765–4774.
38 D. Takagi, K. Ishizaki, T. Asahi and T. Taniguchi, Molecular screening for solid–solid phase transitions by machine learning, *Digital Discovery*, 2023, **2**, 1126–1133.
39 G. J. Beran, I. J. Sugden, C. Greenwell, D. H. Bowskill, C. C. Pantelides, *et al.*, How many more polymorphs of ROY remain undiscovered, *Chem. Sci.*, 2022, **13**, 1288–1297.
40 M. Yang, E. Dybeck, G. Sun, C. Peng, B. Samas, *et al.*, Prediction of the relative free energies of drug polymorphs above zero kelvin, *Cryst. Growth Des.*, 2020, **20**, 5211–5224.
41 V. Kapil and E. A. Engel, A complete description of thermodynamic stabilities of molecular crystals, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**, e2111769119.
42 J. Nyman and G. M. Day, Modelling temperature-dependent properties of polymorphic organic molecular crystals, *Phys. Chem. Chem. Phys.*, 2016, **18**, 31132–31143.
43 N. S. Abraham and M. R. Shirts, Statistical mechanical approximations to more efficiently determine polymorph free energy differences for small organic molecules, *J. Chem. Theory Comput.*, 2020, **16**, 6503–6512.