STRUCTURAL SCIENCE CRYSTAL ENGINEERING MATERIALS

ISSN 2052-5206

# Polymorph sampling with coupling to extended variables: enhanced sampling of polymorph energy landscapes and free energy perturbation of polymorph ensembles

Eric J. Chan[a,b]* and Mark E. Tuckerman[b,c,d]

[a]Chemistry Department, Curtin University, Bentley, WA, 6102, Australia, [b]Department of Chemistry, New York University, New York City, NY, 10003, USA, [c]Courant Institute of Mathematical Science, New York University, New York City, NY, 10003, USA, and [d]New York University-East China Normal University Center for Computational Chemistry at NYU Shanghai, 3663 Zhongshan Road North, Shanghai 200062, China. *Correspondence e-mail: eric.j.chan@curtin.edu.au

Received 26 October 2023
Accepted 9 February 2024

Edited by A. Nangia, CSIR–National Chemical Laboratory, India

This article is part of a collection of articles covering the seventh crystal structure prediction blind test.

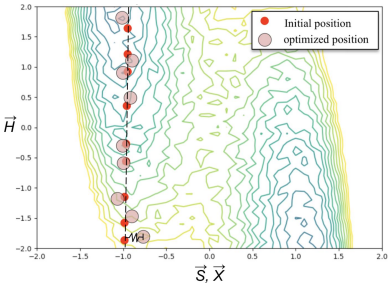
**Keywords:** crystal structure prediction (CSP); polymorphism; enhanced sampling; lattice ensemble free energies.

**Supporting information:** this article has supporting information at journals.iucr.org/b

A novel approach to computationally enhance the sampling of molecular crystal structures is proposed and tested. This method is based on the use of extended variables coupled to a Monte Carlo based crystal polymorph generator. Inspired by the established technique of quasi-random sampling of polymorphs using the rigid molecule constraint, this approach represents molecular clusters as extended variables within a thermal reservoir. Polymorph unit-cell variables are generated using pseudo-random sampling. Within this framework, a harmonic coupling between the extended variables and polymorph configurations is established. The extended variables remain fixed during the inner loop dedicated to polymorph sampling, enforcing a stepwise propagation of the extended variables to maintain system exploration. The final processing step results in a polymorph energy landscape, where the raw structures sampled to create the extended variable trajectory are re-optimized without the thermal coupling term. The foundational principles of this approach are described and its effectiveness using both a Metropolis Monte Carlo type algorithm and modifications that incorporate replica exchange is demonstrated. A comparison is provided with pseudo-random sampling of polymorphs for the molecule coumarin. The choice to test a design of this algorithm as relevant for enhanced sampling of crystal structures was due to the obvious relation between molecular structure variables and corresponding crystal polymorphs as representative of the inherent vapor to crystal transitions that exist in nature. Additionally, it is shown that the trajectories of extended variables can be harnessed to extract fluctuation properties that can lead to valuable insights. A novel thermodynamic variable is introduced: the free energy difference between ensembles of $Z' = 1$ and $Z' = 2$ crystal polymorphs.

## 1. Introduction

*In-silico* crystal structure prediction (CSP) has gained significant interest among material engineers, chemical control specialists, and solid-state organic chemists (Chan *et al.*, 2021; Davey & Garside, 2000; Desiraju, 1989, 2001; Hartman, 1973; Mullin, 2001; Price, 2013, 2004). However, the precise design of molecular building blocks for targeted packing motifs and desired physical properties remains a challenge in crystal engineering (Bernstein, 2008, 2002; Dunitz, 1995; Gavezzotti, 2006; Kitaigorodskiy *et al.*, 1965; Kitaigorosky, 1973). This is due not only to the complex physical laws governing the packing of molecular crystals but also to the role of crystallization kinetics and factors such as nucleation and growth, necessitating practical experimentation as the primary means




OPEN ACCESS
Published under a CC BY 4.0 licence

of design. While new approaches to CSP continue to emerge (Bier *et al.*, 2021; Day *et al.*, 2003; Neumann *et al.*, 2008; Price, 2018; Reilly *et al.*, 2016; Schneider *et al.*, 2016; Yu & Tuckerman, 2011; Zhu *et al.*, 2014), the most successful methods often remain closely guarded industrial secrets (Neumann, 2008; Hunnisett *et al.*, 2024*a*, Hunnisett *et al.*, 2024*b*).

Despite the wealth of crystal data in the Cambridge Structural Database (CSD), many reliable CSP approaches still rely on energy-based techniques and configurational sampling, rather than being data-driven. While machine learning has made significant strides in predicting protein structures (Jumper *et al.*, 2021; Silver *et al.*, 2018, 2016), the application of similar breakthroughs in CSP remains a challenge. CSP primarily relies on computational chemistry and molecular simulation techniques (Allen & Tildesley, 1987; Frenkel & Smit, 2002; Hermann *et al.*, 2017; Parr & Weitao, 1995; Tuckerman, 2010) to bridge the gap between theory and experimental evidence.

Simulation-driven CSP methods focus on two main objectives: polymorph sampling and ranking stability. Polymorph sampling involves configuration sampling algorithms that require a molecular structure as input, which is the primary topic of this report. Ranking stability aims to accurately predict energy differences between pre-generated polymorph configurations and benefits from insights into crystal nucleation and growth (Case *et al.*, 2016; Reilly *et al.*, 2016; Yang & Day, 2021*a*,*b*; Hermann *et al.*, 2017; Hoja *et al.*, 2017; Wengert *et al.*, 2021; Rossi *et al.*, 2016).

Polymorph sampling methods frequently utilize Monte Carlo (MC) sampling, basin hopping (BH), molecular dynamics (MD) or evolutionary algorithms (Neumann *et al.*, 2008; Price, 2004; Yu & Tuckerman, 2011; Rosso *et al.*, 2002; Sobol, 1977; Zhu *et al.*, 2014; Bier *et al.*, 2021). Most MC methods such as BH or Sobol sampling involve a subsequent molecular energy optimization from an higher energy test configuration to identify local minima. A key question pertains to how the probability of generating a structure correlates with a compound's intrinsic ability to crystallize in nature.

Effective and efficient polymorph sampling algorithms must adapt as molecular systems grow in complexity, which may involve an increased number of torsional degrees of freedom or more molecules in the asymmetric unit ($Z'$). *In silico* polymorph screening can often be incomplete (Case *et al.*, 2016; Sobol, 1977). CSP faces the challenge of dimensionality as the number of configuration variables increases, making exhaustive searches less feasible. To address this challenge, low-$Z'$ values are often used, and CSP employs pseudo-random (PR) or quasi-random (QR) sampling methods (van Eijck & Kroon, 1999; Case *et al.*, 2016), or enhanced sampling schemes (Hasenbusch & Schaefer, 2010; Laio & Parrinello, 2002; Liu *et al.*, 2005; Yu & Tuckerman, 2011). These techniques expedite the exploration of the configuration space, ensuring adequate sampling of a wide range of crystal structures and polymorphs.

Polymorphs often exhibit complex and diverse structures, and efficiently sampling them within a practical time frame in computational simulations can be challenging. The systems themselves are not inherently non-ergodic, but achieving ergodicity within a reasonable time frame presents a formidable challenge. This challenge also is primarily addressed with enhanced sampling techniques.

This report introduces an enhanced sampling approach, adapted from QR, BH and temperature-accelerated methods, with similarities to umbrella sampling. The method involves stepwise propagation of molecular coordinates represented by extended variables (EVs) in a heat bath, allowing a broader distribution of possible unit cells to be randomly sampled at each step. EVs are reference variables acting as a tool and, as described later, are not necessarily atomic coordinates. During each step, the EVs are held fixed, which biases a pseudo-random polymorph sampling stage. This method is referred to as 'Extended Variable Coupled to Crystal Polymorph Monte Carlo' sampling (EVCCPMC or EVCCP).

The reasoning to design and test the EVCCP algorithm for the possibility of enhanced sampling of crystal structures was due to the obvious relation between extended variables being representative of a molecular configuration in the gaseous phase and the statistically biased generation of corresponding polymorphs being representative of inherent vapor to crystal transitions that can exist in nature.

This report aims to introduce and demonstrate EVCCP sampling conceptually within the context of CSP. EVCCP was trialed as part of a recent blind test (Hunnisett *et al.*, 2024*b*). The specific focus of this report is to demonstrate a modification that enables replica exchange, known as the EVCCP modified replica exchange (EVCCPMRE) approach. It is worth noting that this modification conceptually allows for the exchange of non-ordinal extended variables, such as those associated with space group symmetry or $Z'$. However, this is outside the scope of this report and the authors intend to conduct a comprehensive study of EVCCPMRE, particularly focusing on the exchange of space group as a variable, separately as part of a future investigation.

Within the EVCCP framework, EVs exhibit ergodic behavior. This enables the calculation of thermodynamic properties pertaining to ensembles of crystal polymorph configurations. Specifically, it facilitates the calculation of Free Energy Differences (FED) between variables, such as stoichiometric ratios or different values of $Z'$. The latter serves as a qualitative measure of a molecule's propensity to form a polymorph with a specific $Z'$, distinct from a comparison of selected minimum energy polymorphs.

The system of coumarin polymorphs (Shtukenberg *et al.*, 2017) was used as the benchmark compound for this investigation and comparisons are made between EVCCP and vanilla pseudo-random polymorph generation (van Eijck & Kroon, 1999). Coumarin was deemed ideal because it is well suited for rigid body approximation and there are experimentally known forms with different $Z'$ that crystallize in the same space group.

## 2. Method description

The EVCCP framework considers two independent atomic systems that describe identical sets of components (*i.e.* the cluster of $Z'$ molecules). One system represents the crystal polymorph and the other is a reference system containing extended variables (EVs) (Abrams & Tuckerman, 2008; Laio & Parrinello, 2002; Maragliano & Vanden-Eijnden, 2006; Ciccotti & Meloni, 2011). To elaborate, if the desire is to sample crystal polymorphs with $Z' = 4$ then the EVs would represent an isolated cluster of four molecules in the gas phase. Both systems contain atomic coordinates ($\mathbf{R}_i$), where $\mathbf{R} \in \mathbb{R}^3$ and the subscript $i$ represents the $i$th atom in either system.

In EVCCP, the atomic positional coordinates ($\mathbf{R}_i$) are mapped onto collective variables (CVs) for molecular centers and orientations (Euler angles). A matrix is used to describe the unit-cell parameters for the crystal polymorph (*i.e.* parallelepiped).

An orthogonal simulation box is used for the reference system that has a volume much greater than the volume occupied by all the atoms. $\mathbf{X} \in \mathbb{R}^d$ is a vector of coordinates with dimensionality ($d$) representing the CVs in the crystal system ($d = Z' \times 6$) and $\mathbf{H}$ is a vector representing the unit-cell parameters. $\mathbf{H} \in \mathbb{R}^9$ are the vector coordinates for a parallelepiped or unit cell. $\mathbf{H} \equiv \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$, where $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are the unit-cell vectors in Å. $\mathbf{S}$ are EVs which correspond with $\mathbf{X}$ (see Fig. 1).

The partition function [$\mathbb{Z}(\beta)$] describing the coupled systems is

$$\mathbb{Z}(\beta) = \iiint \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\,\mathrm{d}\mathbf{S} \exp\left[-\beta\left[U(\mathbf{X},\mathbf{H}) + U(\mathbf{S}) + \frac{k}{2}(\mathbf{X}-\mathbf{S})^2\right]\right]. \tag{1}$$

In equation (1), $U(\mathbf{X}, \mathbf{H})$ is the potential energy surface (PES) of the crystal polymorphs, $U(\mathbf{S})$ is the reference system potential and $\frac{k}{2}(\mathbf{X}-\mathbf{S})^2$ is a harmonic coupling term with a spring constant ($k$). For simplicity, in this study the reference system is chosen to be not self-interacting [*i.e.* $U(\mathbf{S}) = 0$]. However, this is not at all a strict requirement and might be exploited for further investigation. The combined potential of the two systems is initially represented using

$$U(\mathbf{X},\mathbf{H},\mathbf{S}) = U(\mathbf{X},\mathbf{H}) + U(\mathbf{S}) + \frac{k}{2}(\mathbf{X}-\mathbf{S})^2. \tag{2}$$

In EVCCP, the evolution of the polymorph system is adiabatically (quasi-static) decoupled from that of the reference system because each $\mathbf{X}$ and $\mathbf{H}$ are evaluated while $\mathbf{S}$ remains unchanged and vice-versa, *i.e.* the evolution or change of either system is subject to largely separate characteristic timescales. If $k = 0$, the $\mathbf{X}$ and $\mathbf{H}$ are obtained by a minimization of the polymorph structure energy $U(\mathbf{X}, \mathbf{H})$ from some initial configuration ($\mathbf{X}_0$, $\mathbf{H}_0$). This final configuration will also have a generation probability $P(\mathbf{X}, \mathbf{H})$. Instead, $k > 0$ and $\mathbf{X}_0$ are instantiated with some reference coordinate $\mathbf{S}$ and independently $\mathbf{H}_0 \sim \mathcal{U}_{[\mathbf{a},\mathbf{b}]}$, where $\mathbf{a}$ and $\mathbf{b}$ are limits on unit-cell vectors defined to specify a range of polymorph densities. The configurational energy currently undergoing minimization is denoted as $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$. This notation is employed here to represent the potential associated with the joint probability [*i.e.* $P(\mathbf{S})^*P(\mathbf{X}, \mathbf{H}|\mathbf{S}) = P(\mathbf{X}, \mathbf{H}, \mathbf{S})$], with $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ specifically emphasizing the fixed nature of $\mathbf{S}$. Exploiting the conditional probability $P(\mathbf{X}, \mathbf{H}|\mathbf{S})$ is a conceptual underpinning of EVCCP. In the next section estimates for $P(\mathbf{X}, \mathbf{H}|\mathbf{S})$ are made by statistical inference.

$\mathbf{S}$ are stepwise propagated with temperature ($T$)[1] according to the metropolis MC updating scheme (Metropolis *et al.*, 1953). The configuration energy used for these updates (detailed in Appendix *A*) is given by

$$U(\mathbf{X},\mathbf{H}|\mathbf{S}) \equiv U(\mathbf{X}_{\min},\mathbf{H}_{\min}|\mathbf{S}) = U_{\min} + \frac{k}{2}(\mathbf{X}_{\min}-\mathbf{S})^2. \tag{3}$$

Here $U_{\min}$ and $\mathbf{X}_{\min}$ are the respective energy and CV coordinate of the polymorph that was generated from a subset number of polymorphs ($M$) generated each EVCCP step which share the same fixed $\mathbf{S}$ coordinate (essentially argmin [$U(\mathbf{X}, \mathbf{H}, \mathbf{S})$]). To determine $U_{\min}$ the harmonic energy penalty for the $m$th polymorph $\left[\frac{k}{2}(\mathbf{X}_m - \mathbf{S})^2\right]$ must be taken into account. $U_{\min}$ and $\mathbf{X}_{\min}$ are elements of subsets of $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M\}$ and $\{\mathbf{H}_1, \mathbf{H}_2, \ldots \mathbf{H}_M\}$ under the scope of a set value for $\mathbf{S}$. Thus,

$$\begin{aligned} U_{\min} &= U_{\mathrm{adb}}(\mathbf{X}_{\min},\mathbf{H}_{\min}|\mathbf{S}) \\ &= \mathrm{Min}\left[\left\{\begin{array}{c} U_{\mathrm{adb}}(\mathbf{X}_1,\mathbf{H}_1|\mathbf{S}) + \frac{k}{2}(\mathbf{X}_1-\mathbf{S})^2, \\ U_{\mathrm{adb}}(\mathbf{X}_2,\mathbf{H}_2|\mathbf{S}) + \frac{k}{2}(\mathbf{X}_2-\mathbf{S})^2, \\ \ldots, \\ U_{\mathrm{adb}}(\mathbf{X}_M,\mathbf{H}_M|\mathbf{S}) + \frac{k}{2}(\mathbf{X}_M-\mathbf{S})^2 \end{array}\right\}\right] - \frac{k}{2}(\mathbf{X}_{\min}-\mathbf{S})^2, \end{aligned} \tag{4}$$

where $U_{\mathrm{adb}}(\mathbf{X}_m, \mathbf{H}_m|\mathbf{S})$ represents the unbiased energy component of the $m$th polymorph generated adiabatically by using the biased PES $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ and $\mathbf{X}_0 = \mathbf{S}$. This strategy

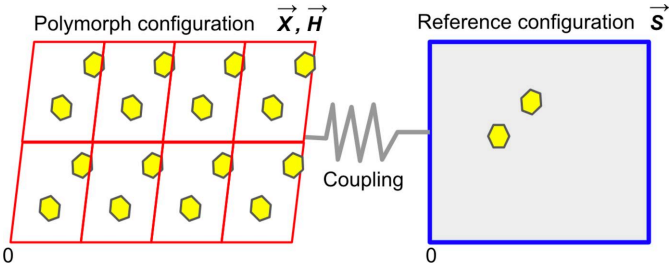



**Figure 1**
Schematic representation of the crystal polymorph and reference systems in the extended variable coupled to crystal polymorph Monte Carlo (EVCCPMC) scheme. In both systems, the components are identical sets of (extended) variables – $\mathbf{X}$ or $\mathbf{S}$ – illustrated using yellow six-membered rings. Red parallelograms are the unit-cell parameters of a crystal ($\mathbf{H}$), in contrast with the reference system (blue orthogonal box). The gray spring connecting the two systems represents a harmonic coupling [see equation (2)]. As shown in this diagram, the state of the (extended) variables between systems need not be the same, but will be similar as a result of coupling.

[1] If $\beta = 0$ then this is the special case of $T = \infty$ condition equivalent with the random search method.

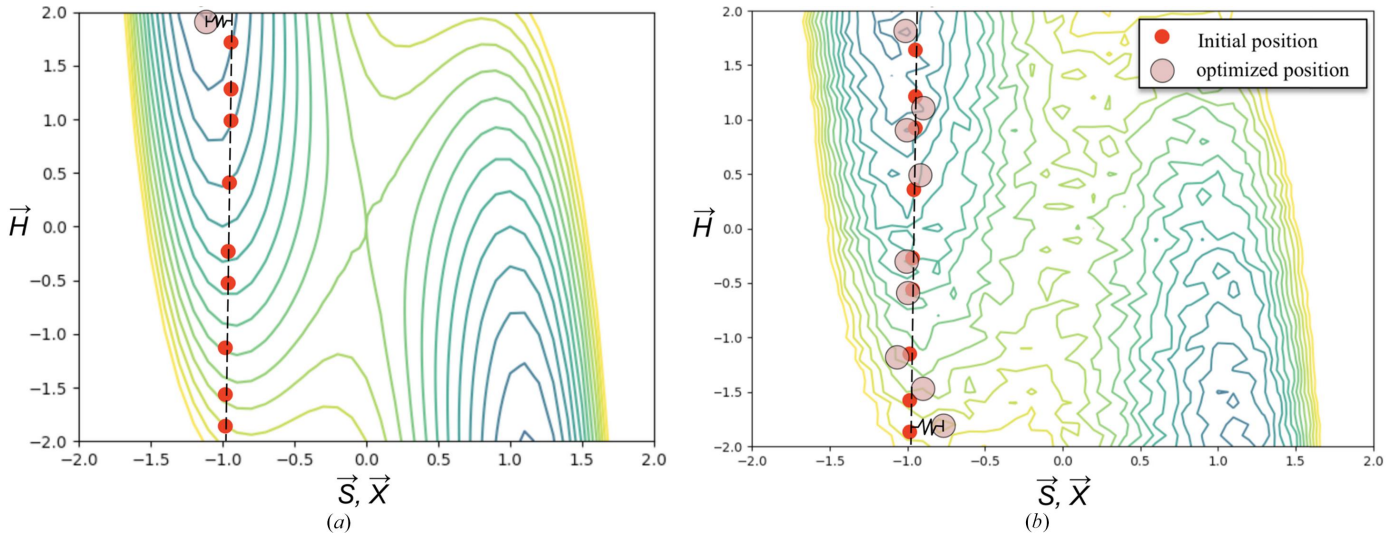



**Figure 2**
Schematic representation of the mini-batch polymorph generation workflow in EVCCP. (*a*) Colored contours represent a smooth potential energy surface (PES) function $U(\mathbf{X}, \mathbf{H})$ and (*b*) is the same surface only roughed with arbitrary noise to create pockets of local minima. In both plots, red circles are the $M$ initial points which have the same reference EV $\mathbf{S}$ indicated by the vertical black dashed line. For the initial positions, $\mathbf{H} \sim \mathcal{U}_{[-2.0,2.0]}$ with $\mathbf{X} = \mathbf{S}$. Pink circles indicate the final coordinates. The difference between $\mathbf{S}$ and $\mathbf{X}$ (highlighted for one point with a black spring) will make a contribution to $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ via a harmonic coupling term [see equation (3)]. When the optimization is performed on the smooth surface all points end at the same minimum. In contrast, on the roughed surface the mini-batch global minimum $U_{\min}$ will be the best approximation for the optimal solution.

involving optimization of multiple polymorphs to obtain $U_{\min}$ as part of each EVCCP step is likened to the particle swarm approach (Kennedy & Eberhart, 1995) and also has similarities with umbrella sampling (Torrie & Valleau, 1977), as demonstrated schematically in Fig. 2.

A flowchart indicating the most relevant steps for coding of the EVCCPMC algorithm is shown as Fig. 3. Initialization of the algorithm requires $\mathbf{S}_0$ (iter = 0, *i.e.* 0th iteration) which is read from a previously known or randomly selected polymorph. This $\mathbf{S}_{\text{iter}} = \mathbf{S}_0$ is input to the structure generator which is used to generate the mini-batch of $M$ energy-minimized structures (see Fig. 2). It is the global minimum energy polymorph from the mini-batch that is used to obtain $U(\mathbf{X}, \mathbf{H}|\mathbf{S}_{\text{iter}})$. The next step is the metropolis update for $\mathbf{S}_{\text{iter+1}}$, *i.e.* generation of a test configuration $\mathbf{S}_{\text{test}} = \mathbf{S}_{\text{iter}} + \Delta\mathbf{S}$. A different batch of $M$ structures are generated using $\mathbf{S}_{\text{test}}$ as the seed EV

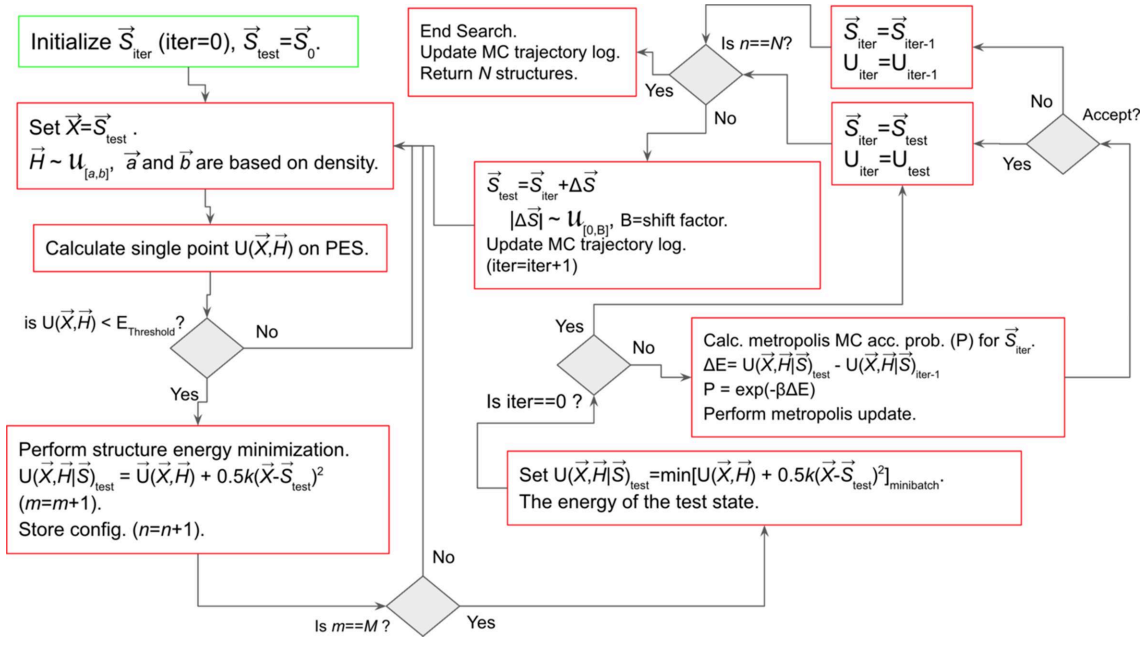



**Figure 3**
Flowchart of the EVCCPMC algorithm. In this schematic, the symbol '=' represents the assignment operation, while '==' is used for conditional evaluation.

resulting in the test configuration energy $U(\mathbf{X}, \mathbf{H}|\mathbf{S}_{\text{test}})$ required for the update. This procedure is iterated (iter = iter +1) to determine each $\mathbf{S}_{\text{iter}}$ until $N$ structures have been sampled. For the algorithm to work effectively, an $E_{\text{threshold}}$ must be declared so that polymorph configurations with $U(\mathbf{X} = \mathbf{S}_{\text{test}}, \mathbf{H}) > E_{\text{threshold}}$ can be rejected prior to the polymorph energy optimization because values for $\Delta\mathbf{S}$ and $\mathbf{H}$ are generated randomly ($|\Delta\mathbf{S}| \sim \mathcal{U}_{[0,B]}$ where $B = k_B T|\Delta\mathbf{S}|_{\max}$) and can lead to unsuitable pre-optimization configurations. Each MC update for $\mathbf{S}_{\text{test}}$ is related to the corresponding estimate for $P(\mathbf{X}, \mathbf{H}|\mathbf{S}_{\text{test}})$ *i.e.* the polymorphs sampled by the polymorph generator for $\mathbf{S}_{\text{test}}$. Each step in the EVCCPMC trajectory contains both $\mathbf{S}_{\text{iter}}$, representing a node in a Markov chain of EV coordinates, as well as the collection of all $M$ local minimum polymorphs.

## 3. Polymorph generation probabilities

When using PR or EVCCP sampling, the respective probabilities for polymorph generation $P(\mathbf{X}, \mathbf{H})$ or $P(\mathbf{X}, \mathbf{H}|\mathbf{S})$ can be estimated using

$$P(\mathbf{X}_0, \mathbf{H}_0) = \frac{N_{\text{hits}}}{N} \quad \text{or} \quad P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S}) = \frac{N_{\text{hits}}}{M}, \tag{5}$$

where $N_{\text{hits}}$ is the total number of hits obtained for a specified polymorph ($\mathbf{X} = \mathbf{X}_0$, $\mathbf{H} = \mathbf{H}_0$), $N$ is the total number of polymorphs generated for a PR search and $M$ (the mini-batch size) is the number of polymorphs generated for a EVCCP step. Some benchmark for the dependence of these ratios on sampling sizes pertaining to this work will now be demonstrated for different $Z'$.

$P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S})$ is further evaluated through histogramming and 2D-projection of $P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S} - \mathbf{S}_0)$. All sampling was performed in space group $P2_12_12_1$ with $Z' = 1$, 2 or 3 corresponding with the known coumarin polymorphs (*i.e.* form V, III and IV).

### 3.1. Sampling of probability distributions

Fig. 4 are the results from PR [Figs. 4(*a*)–4(*c*)] and EVCCP [Figs. 4(*d*)–4(*f*)] sampling of the set of $N, M = \{5, 10, 20, 100, 1000\}$ for $Z' = 3$ (results for $Z' = 1, 2$ are made available in the supporting information). For this part of the study only a gentle coupling ($k = 2$) was applied for EVCCP such that effects are mostly resulting from setting $\mathbf{X} = \mathbf{S}$. Figs. 4(*a*) and 4(*d*) compare the overall post-sampling unbiased optimized energy distributions $P[U(\mathbf{X}, \mathbf{H})]$ which for EVCCP results are denoted as $P[U_{\text{unb}}(\mathbf{X}, \mathbf{H}|\mathbf{S})]$. As expected, the estimate of these distribution functions becomes smoother as sample size ($N$) increases. For pseudo-random searches the $P[U(\mathbf{X}, \mathbf{H})]$ distribution appears more characteristic of the Maxwell Boltzmann distribution, the strong departure from this as shown in EVCCP sampling demonstrates the expected key differences between $P(\mathbf{X}, \mathbf{H})$ and $P(\mathbf{X}, \mathbf{H}|\mathbf{S})$. In general, $P[U_{\text{unb}}(\mathbf{X}, \mathbf{H}|\mathbf{S})]$ is bimodal exhibiting peaks roughly situated at $\langle U(\mathbf{X}, \mathbf{H})\rangle$ as well as $U(\mathbf{X}_0, \mathbf{H}_0)$.

Figs. 4(*b*), 4(*c*) and 4(*e*), 4(*f*) compare the polymorph sampling using a measure of the variance of the molecular center components in $\mathbf{X}$ from those components corre-

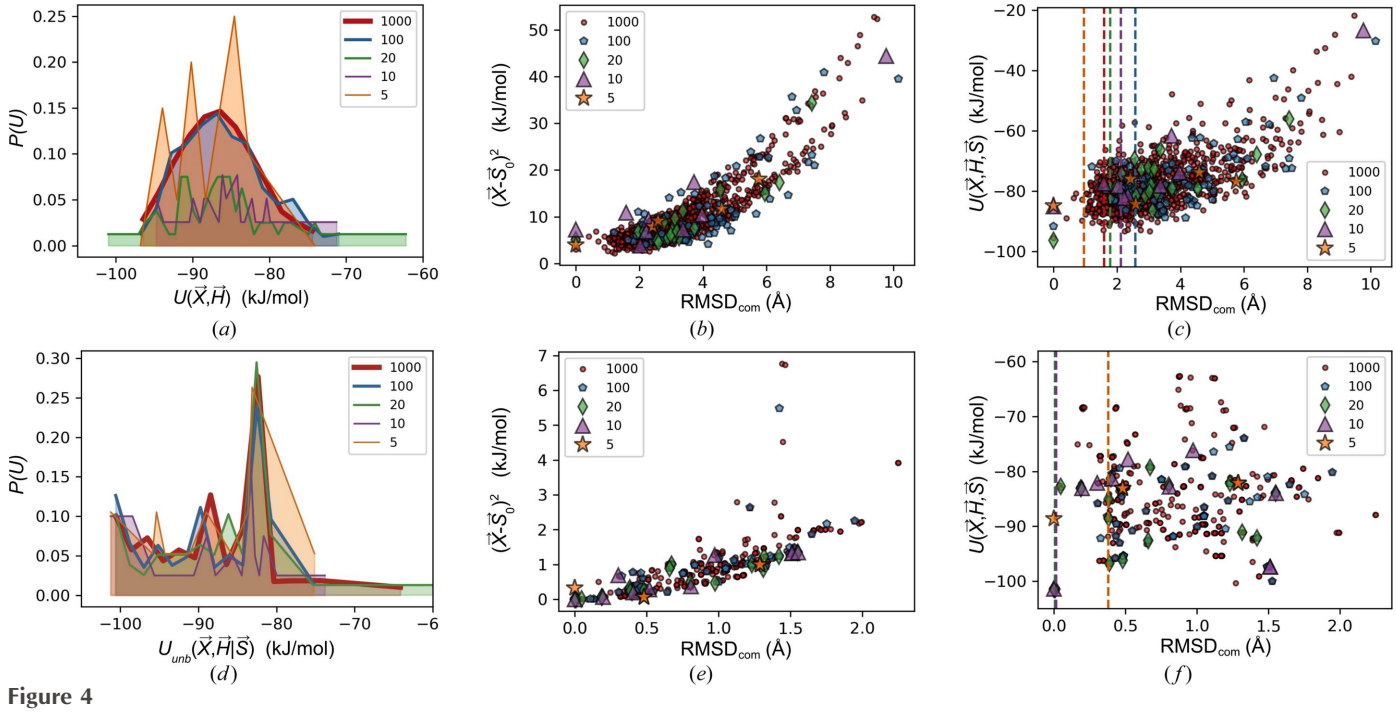



**Figure 4**
Plots for random (*a*)–(*c*) and EVCCP (*d*)–(*f*) coumarin $Z' = 3$ polymorph data generated from searches. A different marker and color was used to differentiate the corresponding sample size ($N$, $M$) as indicated in the legend. The $\mathbf{S}_0$ coordinate is that of form IV. (*a*) and (*d*) compare $P[U(\mathbf{X}, \mathbf{H})]$ and $P[U_{\text{unb}}(\mathbf{X}, \mathbf{H}|\mathbf{S})]$ histograms; (*b*) and (*e*) compare plots of $\text{RMSD}_{\text{com}}$ (*i.e.* a measure of the variance associated with an identified minima sampled) against the biasing penalty of $U(\mathbf{X}, \mathbf{H}|\mathbf{S}_0)$; (*c*) and (*f*) are $\text{RMSD}_{\text{com}}$ versus $U(\mathbf{X}, \mathbf{H}, \mathbf{S})$ [see equation (2) and main text for explicit details].

sponding with the minimum energy polymorph that was identified in that particular sampling distribution [*i.e.* $\mathbf{X}_{\min}$ from $U_{\min}$ as shown in equation (4)]. The variance measure for molecular centers ($\mathrm{RMSD}_{\mathrm{com}}$) is thus defined as,

$$\mathrm{RMSD}_{\mathrm{com}} = \left[(\mathbf{X}_{\mathrm{com}} - \mathbf{X}_{\mathrm{min,com}})^2\right]^{1/2}. \quad (6)$$

That is, polymorphs with $\mathrm{RMSD}_{\mathrm{com}} = 0$ represent the global minimum polymorph identified in a distribution. In Figs. 4(*b*) and 4(*e*) the $\mathrm{RMSD}_{\mathrm{com}}$ is plotted against the harmonic biasing penalty term in equation (2) $\left[\frac{k}{2}(\mathbf{X} - \mathbf{S})^2\right]$. As expected, the value of this bias ($k = 2$) increases with the $\mathrm{RMSD}_{\mathrm{com}}$. Also, the overall magnitudes of $\mathrm{RMSD}_{\mathrm{com}}$ are much lower for EVCCP generated polymorph distributions.

Figs. 4(*c*) and 4(*f*) are plots of $\mathrm{RMSD}_{\mathrm{com}}$ against the polymorph energy $U(\mathbf{X}, \mathbf{H}, \mathbf{S})$ from equation (2) so that each polymorph energy includes the biasing contribution relative to $\mathbf{S}_0$ irrespective of how it was generated. Vertical dashed lines represent when the $\mathbf{X}$ was set to be $\mathbf{S_0}$ for coumarin form IV. Figs. 4(*c*) and 4(*f*) are the proof of the important fact that $\mathrm{RMSD}_{\mathrm{com}} = [(\mathbf{S}_{0,\,\mathrm{com}} - \mathbf{X}_{\mathrm{min,\,com}})^2]^{1/2}$ can be seen to approach zero when sampling from a $P(\mathbf{X}, \mathbf{H}|\mathbf{S})$ distribution as opposed to sampling from $P(\mathbf{X}, \mathbf{H})$.

The plots in Fig. 4 demonstrate the effect of increasing accuracy for $P(\mathbf{X}, \mathbf{H})$ or $P(\mathbf{X}, \mathbf{H}|\mathbf{S})$ as sample sizes $N, M \rightarrow \infty$. The comparison of Figs. 4(*c*) with 4(*f*) suggests that $M$ can be small, in this case $5 < M < 10$, so that $\mathbf{X}_{\min} = \mathbf{S}_0$.

Raw MC generated polymorph data is commonly removed of duplicate configurations prior to stability ranking. The number of distinct or unique structures is denoted as $\mathbb{U}[N]$. Fig. 5 shows bar plots of $\log(\mathbb{U}[N]$ versus $\log(N)$ for the EVCCP searches [Fig. 5(*a*)] with $\log \mathbb{U}[M])$ versus $\log(M)$ for PR searches [Fig. 5(*b*)]. $N, M$ values were ranged between [5, 2000]. Bars are stacked as triplets representing $Z' = 1, 2, 3$ from left to right. The error bars show the ratio $\mathbb{U}[N]/N$ or $\mathbb{U}[M]/M$. Markers on the right of each triplet are the corresponding $P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S}_0)$ or $P(\mathbf{X}_0, \mathbf{H}_0)$ estimate for each set of data [see equation (5)]. The center of each triplet is the $\log(N)$ or $\log(M)$ value for that group. As expected, the results demonstrate that $\mathbb{U}[M]$ and $\mathbb{U}[N]$ increase with $Z'$. Also when $M = N$, $\mathbb{U}[M]$ is lower than $\mathbb{U}[N]$ and $\mathbb{U}[M]/M$ converges to small values with increasing $M$ indicating that EVCCP does generate less unique configurations than a random search (the only exception is the case of $Z' = 1$ PR search). $P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S}_0)$ estimates are higher than $P(\mathbf{X}_0, \mathbf{H}_0)$ and in fact $P(\mathbf{X}_0, \mathbf{H}_0)$ is negligible for $Z' > 1$ which assists to confirm $\mathbf{S}$ will bias sampling for a specific polymorph.

### 3.2. Sampling of polymorph conditional probability distributions and effect of coupling ($k$) magnitude

To demonstrate the $P(\mathbf{X}, \mathbf{H}|\mathbf{S})$ underlying EVCCP, the related probability distribution $P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S}_0 + \Delta\mathbf{S})$ was estimated from sampling [see equation (5)] with a specific deviation ($\Delta\mathbf{S}$ where $|\Delta\mathbf{S}| \sim \mathcal{U}_{[0,B]}$ and $B = |\Delta\mathbf{S}|_{\max}$) from $\mathbf{S}_0$ (*i.e.* $\mathbf{S}_0 + \Delta\mathbf{S} = \mathbf{S}$). It is assumed that $P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S}_0)$ will be highest (*i.e.* $B = 0$), with $P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S}) \rightarrow 0$ as $B \rightarrow \infty$. Differences in the curvature of $P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S} - \mathbf{S}_0)$ distributions for the $Z' = 1, 2, 3$ example forms are mapped as 2D-projections (*i.e.* multi-variate $\rightarrow$ bi-variate) as the $(|\Delta\mathbf{S}_{\mathrm{Eul}}|, |\Delta\mathbf{S}_{\mathrm{com}}|) \in \mathbb{R}^2$ coordinate space (shown in Fig. 6) using the integer values for $[P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S} - \mathbf{S}_0)]^{-1}$ provided in Table 1. The values shown in Table 1 are directly related to the EVCCP parameter ($M$) required to ensure the specific polymorph can be generated. Clearly, the projections becomes narrower as $Z'$ increases. The reasonable fit of Table 1 values using a bi-variate Gaussian function demonstrates the suitability of the Gaussian approximation used for evaluating $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ [see equation (3), and equations (32) and (33) in Appendix *A*].

The expected effect of increasing the spring constant magnitude ($k$) on $P(\mathbf{X}, \mathbf{H}|\mathbf{S})$ distribution is an increased probability at the origin and an overall narrower distribution. This is also shown in the sampling and the comparison between $k = 0$ and $k = 1000$ is shown in Table 1 and Fig. 6 for all $Z'$ examples.

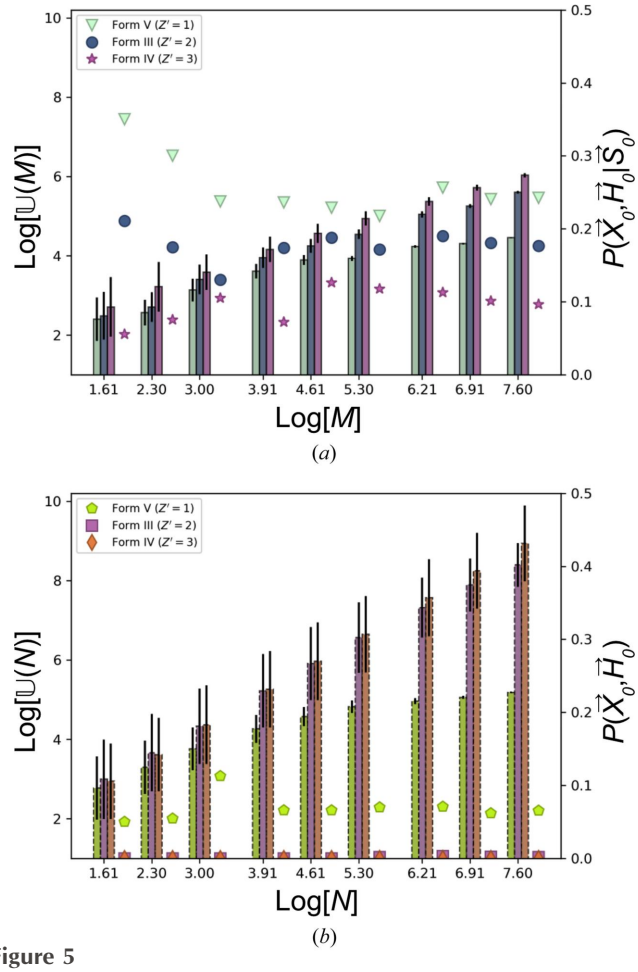



**Figure 5**
Log plots for the number of unique structures $\mathbb{U}$ as a function of the number of polymorphs generated ($M$ or $N$) from the respective EVCCP (*a*) or pseudo-random (*b*) search data. Bars are stacked as triplets for $Z' = 1, 2, 3$ plotting $\log \mathbb{U}$ versus $\log(N)$ with the ratio $\mathbb{U}[N]/N$ as error bars. In addition, different markers on the right of each triplet stack compare the conditional probability $P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S}_0)$ to generate a specific polymorph from EVCCP (*a*) with the $P(\mathbf{X}_0, \mathbf{H}_0)$ from a pseudo-random search (*b*). The scale for $\log \mathbb{U}$ is on the left of each plot with the scale for probabilities on the right.

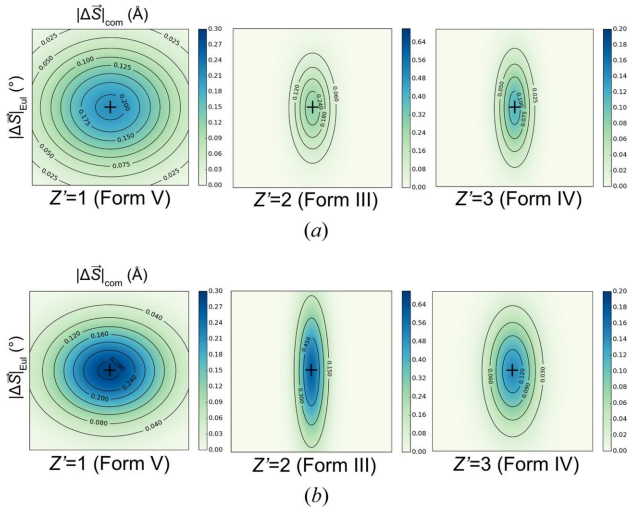



**Figure 6**
Plots of the bi-variate Gaussian fits to Table 1 data depicted as $P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S} - \mathbf{S}_0)$ for coumarin polymorphs with spring constants (*a*) $k = 0$ and (*b*) $k = 1000$. A black cross is positioned at $|\Delta\mathbf{S}| = 0$ for each particular form.

The actual effect of $k$ on polymorph generation each EVCCP step is best appreciated using Fig. 7. In Fig. 7 the reference EV are the same as in Table 1 only that $M = 20$ polymorphs are generated with either $k = 10$ or $k = 1000$. The resulting structures are arranged as a multi-structure projection down b with the origin as the center of the asymmetric unit. The difference between weak versus strong coupling can be clearly seen in that strong coupling results in fewer variants of polymorphs and that displacement from the reference system coordinates (molecules with green outline) between polymorphs is negligible.

## 4. Polymorph sampling with EVCCP

The EVCCP sampling threads were modified to enable replica exchange updates, following established principles in the field (Tuckerman, 2010; Frenkel & Smit, 2002). In this modified replica exchange (EVCCPMRE) approach, only the vector **S** traverses between replicas with different temperatures ($T$). This modification carries a significant implication: the configurational extended variables within **S** can encompass non-ordinal variables that influence structural energy and polymorph generation but remain invariant with respect to other components of **S** during the update moves within each $T$ bath. In other words, these CSP search variables may lack obvious analytical derivatives, such as those linked to space group symmetry or $Z'$. A thorough evaluation of EVCCPMRE involving exchanges of variables such as space group is outside the scope of this investigation. Thus, the PES sampled in this report will be based on the marginal probabilities, given that variables like space group and $Z'$ are held constant in each MC run.

Fig. 8 demonstrates how $\mathbf{S}_{\rm com}$ fluctuate and diffuse about some local minima with a variance relative to $T$.

Tests of polymorph screenings using EVCCPMRE with coupling $k = 1000$ were performed for 13 space groups for both $Z' = 3, 4$ and compared with analogous unbiased PR searches ($N = 8000$). Unless otherwise specified the initial configuration for seeding the MRE $\mathbf{S}_0$ was generated using a smaller preliminary random search ($N = 100$). Bath $T$ were chosen based on the condition that the acceptance rate for exchange moves should be close to 0.5. Baths were set at 263 K, 370 K, 574 K

**Table 1**
Values for inverse conditional probabilities $[P(\mathbf{X}_0, \mathbf{H}_0|\mathbf{S} - \mathbf{S}_0)]^{-1}$ rounded to the nearest integer *i.e.* the average mini-batch size ($M$) needed to generate a specific polymorph with $\mathbf{S}_0$.

1(*a*), 1(*b*) and 1(*c*) are, respectively, $Z' = 1, 2, 3$ for coumarin form V, III and IV ($k = 0$). Rows index the $|\Delta\mathbf{S}|_{\rm max}$ increments for molecular centers $|\Delta\mathbf{S}_{\rm com}|$ (0.1–1.5 Å), with columns being increments for Euler angles $|\Delta\mathbf{S}_{\rm Eul}|$ (5–50°). The effect of $k = 1000$ is shown as 1(*d*), 1(*e*) and 1(*f*).

Table 1(*a*)

| $\lvert\Delta\mathbf{S}\rvert_{\rm max}$ | 5 | 10 | 25 | 30 | 50 |
|---|---|---|---|---|---|
| 0.1 | 4 | 4 | 6 | 6 | 21 |
| 0.2 | 6 | 5 | 8 | 7 | 16 |
| 0.5 | 11 | 10 | 10 | 9 | 17 |
| 1.0 | 9 | 10 | 14 | 14 | 19 |
| 1.5 | 11 | 11 | 11 | 16 | 18 |

Table 1(*b*)

| $\lvert\Delta\mathbf{S}\rvert_{\rm max}$ | 5 | 10 | 25 | 30 | 50 |
|---|---|---|---|---|---|
| 0.1 | 6 | 6 | 8 | 9 | 19 |
| 0.2 | 8 | 10 | 12 | 11 | 31 |
| 0.5 | 57 | 52 | 89 | 67 | 123 |
| 1.0 | 320 | 145 | 267 | 800 | 533 |
| 1.5 | 388 | 227 | 320 | 1600 | 800 |

Table 1(*c*)

| $\lvert\Delta\mathbf{S}\rvert_{\rm max}$ | 5 | 10 | 25 | 30 | 50 |
|---|---|---|---|---|---|
| 0.1 | 11 | 12 | 13 | 18 | 267 |
| 0.2 | 27 | 21 | 39 | 43 | 200 |
| 0.5 | 178 | 89 | 200 | 267 | – |
| 1.0 | 800 | 526 | – | – | – |
| 1.5 | – | – | 1538 | 800 | – |

Table 1(*d*)

| $\lvert\Delta\mathbf{S}\rvert_{\rm max}$ | 5 | 10 | 25 | 30 | 50 |
|---|---|---|---|---|---|
| 0.1 | 2 | 3 | 5 | 7 | 11 |
| 0.2 | 5 | 5 | 6 | 10 | 31 |
| 0.5 | 8 | 8 | 9 | 12 | 48 |
| 1.0 | 10 | 8 | 12 | 14 | 23 |
| 1.5 | 7 | 11 | 13 | 14 | 42 |

Table 1(*e*)

| $\lvert\Delta\mathbf{S}\rvert_{\rm max}$ | 5 | 10 | 25 | 30 | 50 |
|---|---|---|---|---|---|
| 0.1 | 3 | 4 | 3 | 5 | 6 |
| 0.2 | 7 | 5 | 5 | 7 | 21 |
| 0.5 | 5 | 57 | 40 | 22 | – |
| 1.0 | 7 | 145 | 100 | 178 | 200 |
| 1.5 | 61 | 265 | – | 320 | 800 |

Table 1(*f*)

| $\lvert\Delta\mathbf{S}\rvert_{\rm max}$ | 5 | 10 | 25 | 30 | 50 |
|---|---|---|---|---|---|
| 0.1 | 7 | 9 | 8 | 11 | 100 |
| 0.2 | 15 | 9 | 18 | 17 | 267 |
| 0.5 | – | 100 | – | – | – |
| 1.0 | 320 | – | – | – | – |
| 1.5 | 198 | – | 1600 | 784 | – |

and 1142 K (roughly exponentially spaced). Since commonly $P(\mathbf{X}_0,\mathbf{H}_0) < [8000]^{-1}$, each search is not considered exhaustive, thus identifying the same global minima during comparison is not a requirement for enhanced sampling. In principle, the high $T$ bath acts to scan for new global minima using large displacements in EV space whereas the lower $T$ baths can harvest information about polymorphs in surrounding super basins (Yang & Day, 2021*a*).

In Fig. 8(*a*) the molecular center EVs are plotted in projection down the *x*-axis of the reference system with the initial positions for $\mathbf{S}_{\mathrm{com},0}$ indicated using black squares. Fig. 8(*d*) plots energy output $U(\mathbf{X},\mathbf{H}|\mathbf{S})$ stepwise from a test simulation. The trajectories are from simulation with 200 steps of MC (no replica exchange) compared with MRE. The trajectory data shown for Fig. 8 was for $Z' = 3$ searches in space group $P2_12_12_1$ starting with a predetermined global minimum polymorph that has $U(\mathbf{X},\mathbf{H}) = 103.8$ kJ mol$^{-1}$ (isostructural with coumarin form IV). In the 370 K MC run, the final EV position for $\mathbf{S}_{\mathrm{com}}$ regenerates the initial polymorph [shown in Fig. 8(*b*)], whereas in the 370 K MRE bath this is not the case, the resulting structure is shown in Fig. 8(*c*) with overlay of the unit cell and EVs (red circles) in Fig. 8(*a*). The effect of MRE is clear upon inspection of Fig. 8(*a*) with an example indicated using a black arrow. In contrast when exchange is disabled, EVs in comparative reference system do not have same degree of spacial coverage in the same number of MC steps.

### 4.1. EVCCPMRE variant schemes

EVCCPMRE was implemented using the Python interpreter as wrapper code to drive a modified version of the codes for the crystal structure generator *UPACK* (van Eijck & Kroon, 1999). Variations of EVCCPMRE were tested in order to identify if certain modifications of the workflow could further enhance the sampling and subsequent screening results. Despite the utilization of *UPACK* code for this work, these concepts are transferable and can be coded using many other publicly available crystal structure generators. The basic MRE algorithm as previously described is referred to as MRE0 (baths = 4, $k = 1000$, steps = 200, cycles = 1, $M = 10$, $N = 8000$).

A variant algorithm, MRE1, evaluates the addition of a history dependent biasing potential as a Gaussian kernel (Laio & Parrinello, 2002) placed along the EV path in attempts to further enhance sampling. In the MRE1 scheme, the historical biasing potential adds an energetic penalty to $U(\mathbf{X},\mathbf{H}|\mathbf{S})$ if the EV test-state $\mathbf{S}$ had already been previously visited.

The MRE2 scheme incorporates an additional forced update move which is referred to as 'forced-relaxation'. The forced-relaxation move causes the system replicas to reset the EV state of each $T$ bath, thus descend back into a local basin that was previously detected 'on-the-fly'. The update occurs at the start of each MC cycle after a set number of MC steps. MRE2 is based on a notion that the EVs re-visit a low-energy

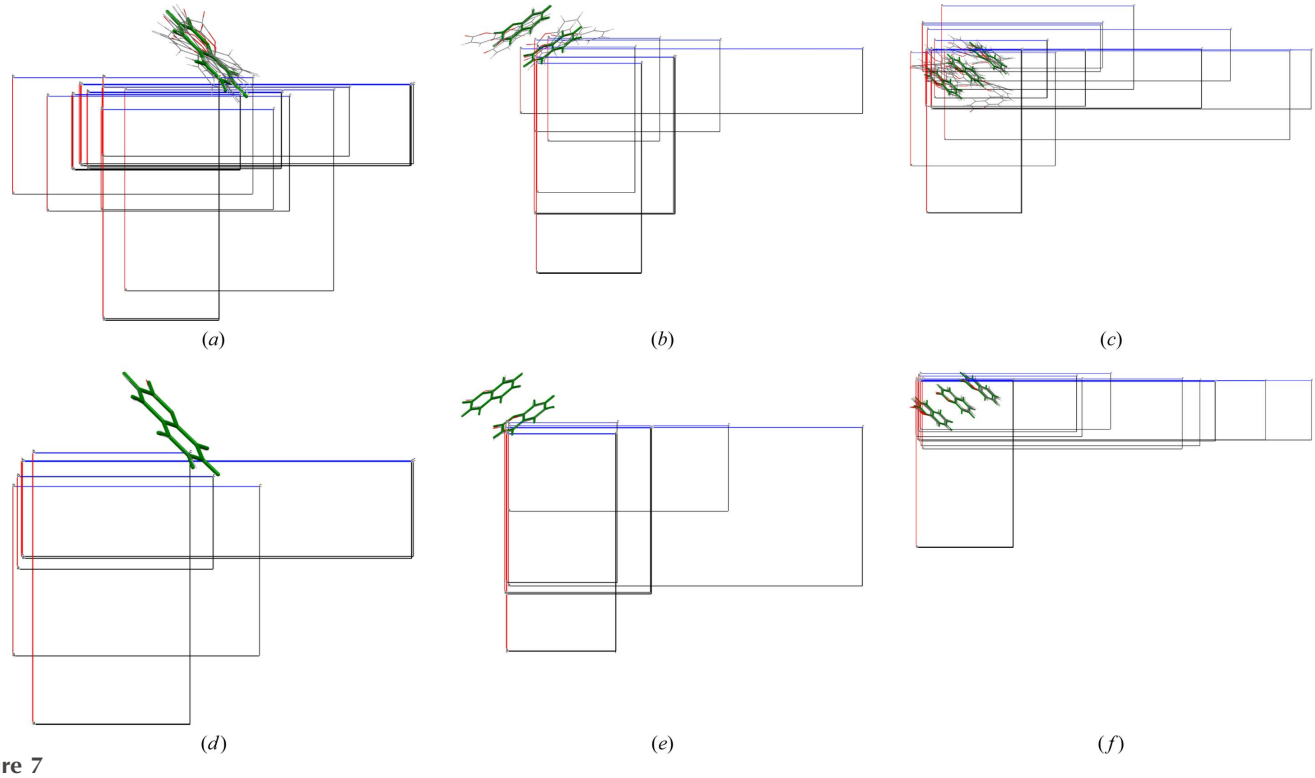



**Figure 7**
Visual representation of the effect of magnitude of harmonic coupling ($k$) during EVCCP sampling using different $Z'$. Displayed are the asymmetric units and unit cells projected down the *b*-axis as a multi-structure overlay between EVs (molecules with green outline) and corresponding 20 polymorphs generated (gray outline) for a single EVCCPMC step with a known coumarin polymorph for the reference EV. (*a*),(*d*) are $Z' = 1$ form V with (*b*),(*e*) $Z' = 2$ form III and (*c*),(*f*) $Z' = 3$ form IV. The effect of weak coupling $k = 10$ [(*a*)–(*c*)] versus strong $k = 1000$ [(*d*)–(*f*)] is remarkable and at the heart of understanding the concept behind EVCCPMC.

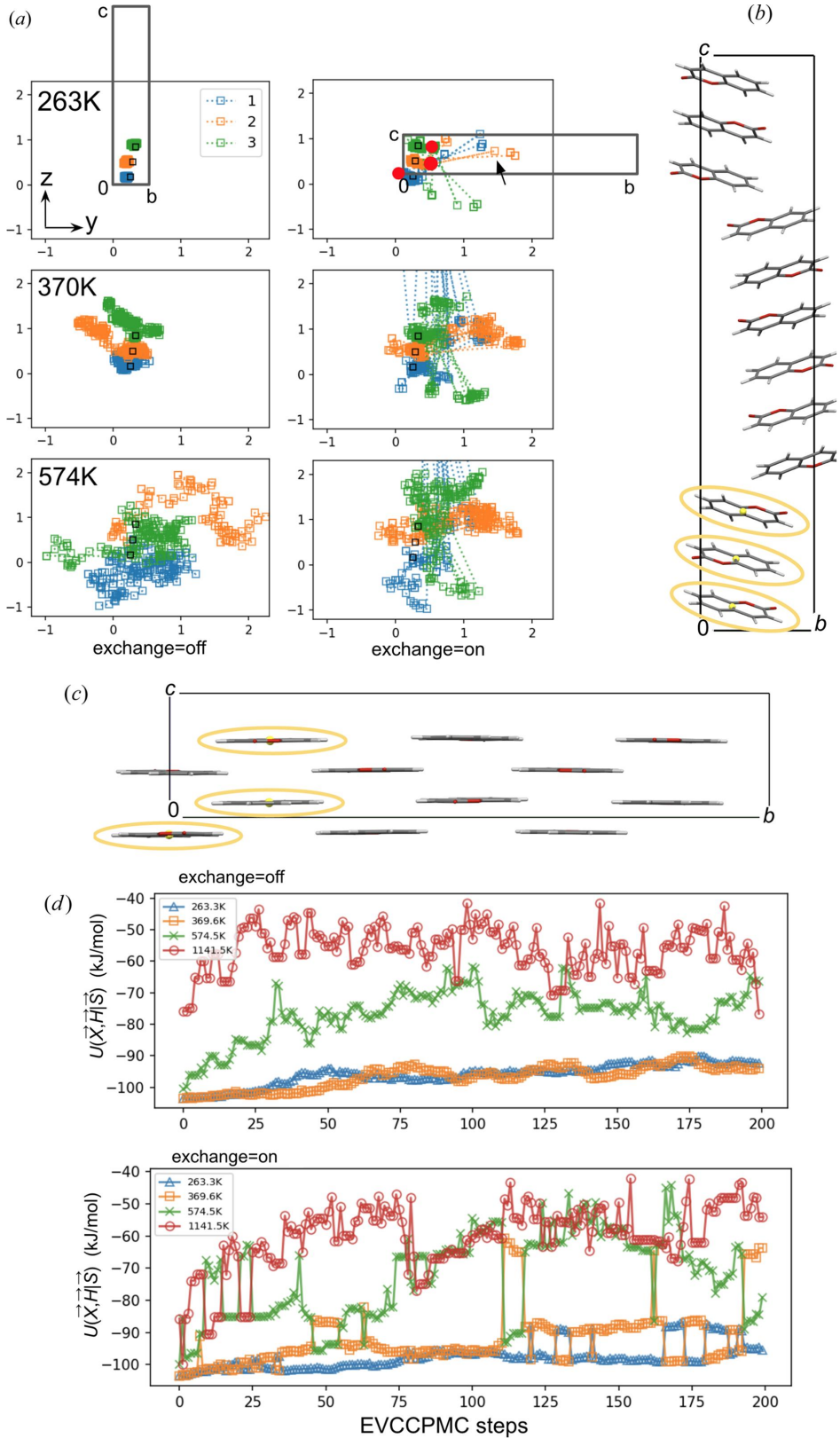



**Figure 8**
Comparison between EVCCPMC(exchange = off) and EVCCPMRE(exchange = on) for 263 K, 370 K, 574 K and 1142 K $T$ bath 200-step trajectories with $Z' = 3$ in space group $P2_12_12_1$. (*a*) Molecular center components of **S** (molecules 1, 2 and 3) projected down reference system $x$-axis. Initial positions are highlighted with black squares. The black arrow indicates a configuration exchange event between baths. The last position in the EVCCPMRE 370 K bath is indicated with red circles. Overlays of pre-optimized unit cells for $U_{min}$ polymorphs that were generated in the 100 K bath final step are shown with corresponding image of post-optimization structure for (*b*) MC and (*c*) MRE with yellow ovals outlining the molecules of the asymmetric unit. (*d*) The stepwise configuration energy $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ is plotted for each bath.

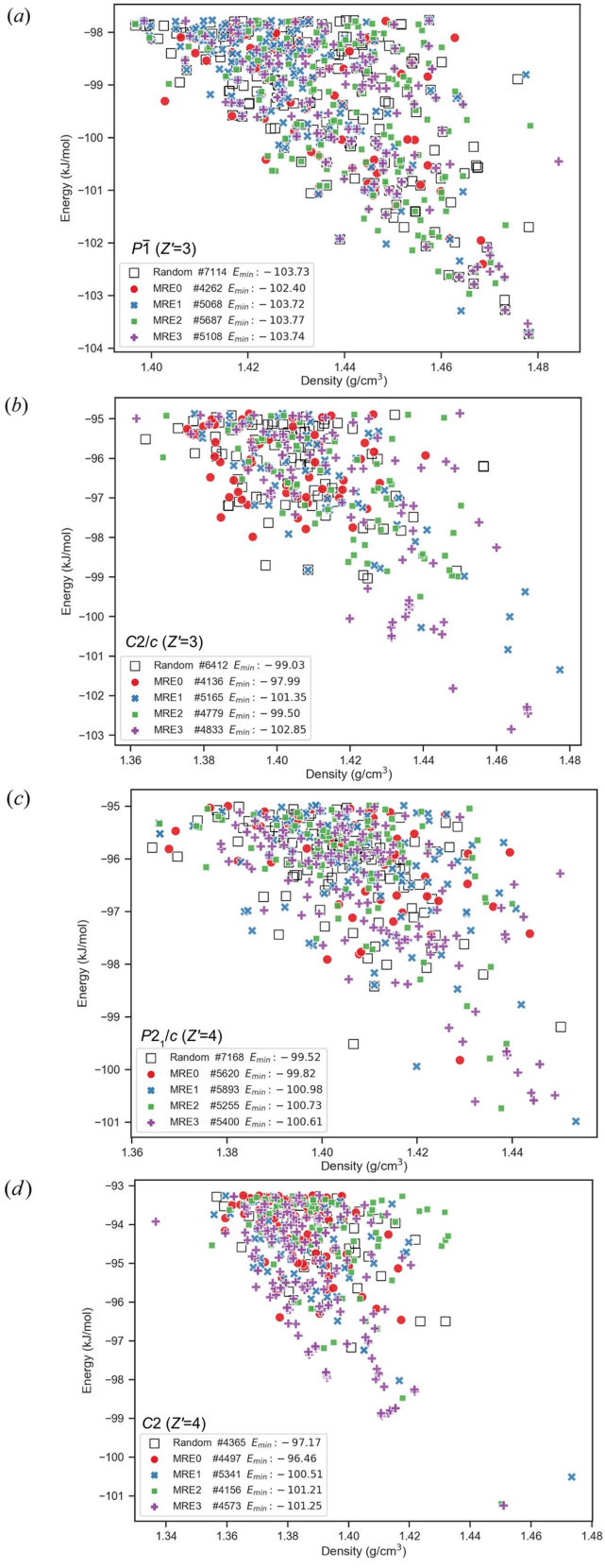



**Figure 9**
Example sets of polymorph energy versus density landscape plots resulting from the $N$ = 8000 searches with EVCCPMRE ($k$ = 1000) variant searches as differently colored markers and pseudo-random search results as the square black outlines. (*a*) and (*b*) are examples from $Z'$ = 3 with (*c*) and (*d*) from $Z'$ = 4.

basin after some predetermined length along the sampling path trajectory. When running MRE2 the mini-batch size was reduced ($M$ = 5) so that twice as many MC steps will be run (baths = 4, steps = 40, cycles = 10, $M$ = 5, $N$ = 8000).

The final modification, MRE3 tests the performance of a global forced-relaxation resetting which occurs after several MC-cycles during the overall search (termed 'relaxation-restart'). The MRE3 is similar to an MRE2 update, yet differs in that all polymorphs over a number of cycles from all replicas are evaluated for the unbiased [$U$(**X**, **H**)] global minimum required for restarting the MRE. This takes longer to run because the basic MRE0 does not perform structure optimizations on-the-fly to absolute full convergence each MC step (full convergence occurs during post-processing), but also differs because EVs for structures based on $U$(**X**, **H**|**S**) will be affected by the harmonic coupling $k$.

### 4.2. EVCCPMRE efficiency and comparison with pseudo-random search method

To facilitate a stepwise comparison of the sampling performance between variant searches, we devised a representative metric for the number density of unique configurations identified within a low-energy window. This was loosely based on earlier work with replica exchange ergodic metrics (Thirumalai *et al.*, 1989; Whitfield *et al.*, 2002). The choice was an average energy for the window of 20 lowest energy ranked structures made relative by using a mean-shifted value (*i.e.* $\langle U\rangle_{T20} - \langle U\rangle$). Monitoring of the relative $\langle U\rangle_{T20} - \langle U\rangle$ parameter was calculated stepwise from each set of search statistics as a function of number of polymorphs generated ($N$, corresponding with the number of MC steps). Evaluation of overall search performances was made using typical energy versus density landscapes from unbiased structure energy optimization of resultant distinct polymorphs. The notation $E = U$(**X**, **H**) is used for the unbiased polymorph energy and $E_{\rm min}$, $\langle E\rangle_{T20}$ and percentiles of $E$ (0.05%, 1.0%, 5.0%) were also evaluated.

Comparative landscape plots for $Z'$ = 3 searches (space group $P\bar{1}$ and $C2/c$) and $Z'$ = 4 searches (space groups $P2_1/c$ and $C2$) are shown as Fig. 9. Corresponding examples for the $\langle U\rangle_{T20} - \langle U\rangle$ metrics plotted as a function of $N$ are shown in Fig. 10. The comparison of the $\langle U\rangle_{T20} - \langle U\rangle$ metrics in Fig. 10 demonstrates that EVCCPMRE sampling is more efficient at sampling low energy configurations. For variant searches the metric was able to reach the same value from the corresponding PS search in a smaller number of steps. For EVCCPMRE the metric appears to always converge to values lower than for PR. This does not imply that at this stage of development EVCCPMRE is more efficient at identifying new global minima ($E_{\rm min}$), since this would also be highly system dependant, rather that the EVCCPMRE sampling algorithm was behaving as intended, and the concept was correctly implemented.

Search metrics were evaluated over all space groups and summarized for larger intervals of $N$ = 2000, 4000, 6000 and

8000 (see Table 2). In Table 2, the measures $R_g$ and $\delta E_g$ are used to summarize comparisons over the many space groups.

$$R_g = \frac{1}{N_{\rm SG}} \sum^{\rm SG} \mathfrak{P} \begin{cases} \mathfrak{P} = & 1 \text{ if } \mathbb{E}[\text{MRE}] < \mathbb{E}[\text{PR}] \\ \mathfrak{P} = & 0 \text{ if } \mathbb{E}[\text{MRE}] \geq \mathbb{E}[\text{PR}] \end{cases} \quad (7)$$

Here $\mathbb{E}[\ldots]$ represents whether a $\langle U\rangle_{T20} - \langle U\rangle$ metric or $E_{\rm min}$ value is from MRE or PR sampling. The $R_g$ value is the ratio between [0,1] for when $\mathbb{E}[\ldots]$ is less for MRE than for random sampling averaged over all the space group searches ($N_{\rm SG}$) for a particular $Z'$. If $R_g > 0.5$ it means that more MRE searches gave the lower $\mathbb{E}[\ldots]$ metric.

$$\delta E_g = \langle \mathbb{E}[\text{MRE}] - \mathbb{E}[\text{PR}] \rangle_{\rm SG} \quad (8)$$

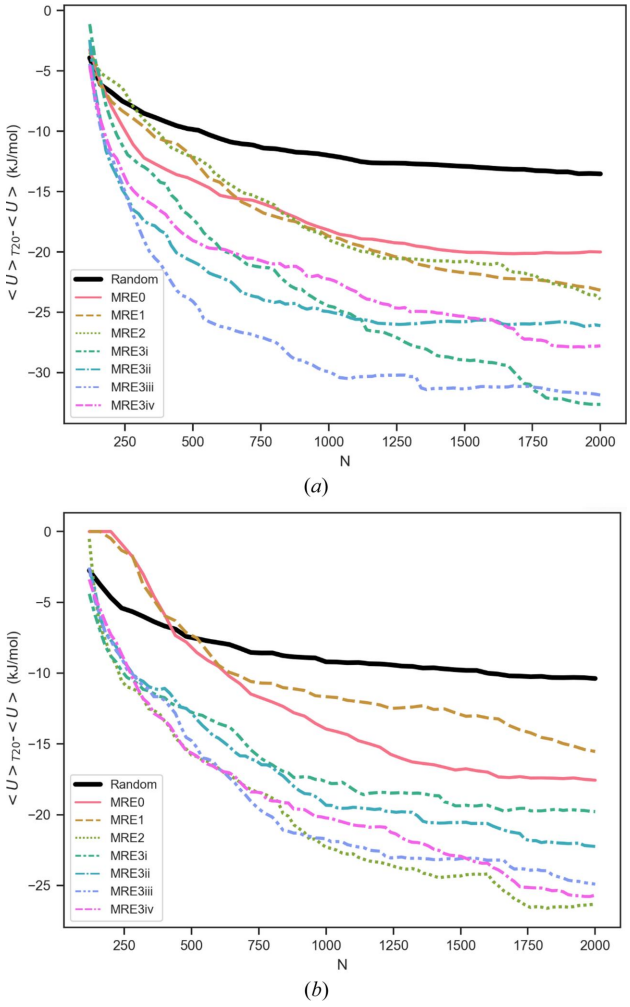



**Figure 10**
Example plots for search performance metric $\langle U\rangle_{T20}$-$\langle U\rangle$ plotted as a functions of the number of structures generated from $N = 0 \rightarrow 2000$. Values from PR search trajectories are shown as black lines. Colored line plots are for EVCCPMRE searches with individual MRE3 runs (labeled i–iv, due to relaxation-restart) also included for comparison. Corresponding with Fig. 9 (*a*) is $Z' = 3$ in space group $C2/c$ with (*b*) $Z' = 4$ using space group $P2_1/c$. The measures from MRE searches obtain lower values in fewer steps and are thus interpreted as achieving a more efficient exploration of lower energy configurations.

represents an average difference, over the space groups tested, between the $\mathbb{E}[\ldots]$ values being compared. The degree to which $E_g < 0$ represents the magnitude by which MRE searches obtained a lower $\mathbb{E}[\ldots]$ metric. From Table 2, the fact that values obtained $\delta E_g$ are slightly lower when running the modified MC updating schemes (algorithms MRE2 and MRE3) suggests that the forced-relaxation or relaxation-restart moves are useful options.

Landscape plots that combine search data for all space groups are provided as Fig. 11, with corresponding $E_{\rm min}$, $\langle E\rangle_{T20}$ values and percentiles provided in Table 3. Interestingly two of the $E_{\rm min}$ structures for $Z' = 4$ were isostructural with the experimental coumarin form I polymorph ($Z' = 1$, $Pca2_1$) and represent the overall global minimum for the searches. It is likely a structure corresponding with experimental form II ($Z' = 2$, $P2_1$) was also generated.

Interestingly, the overall search results suggest any improvements in $E_{\rm min}$, $\langle E\rangle_{T20}$ and percentiles made by

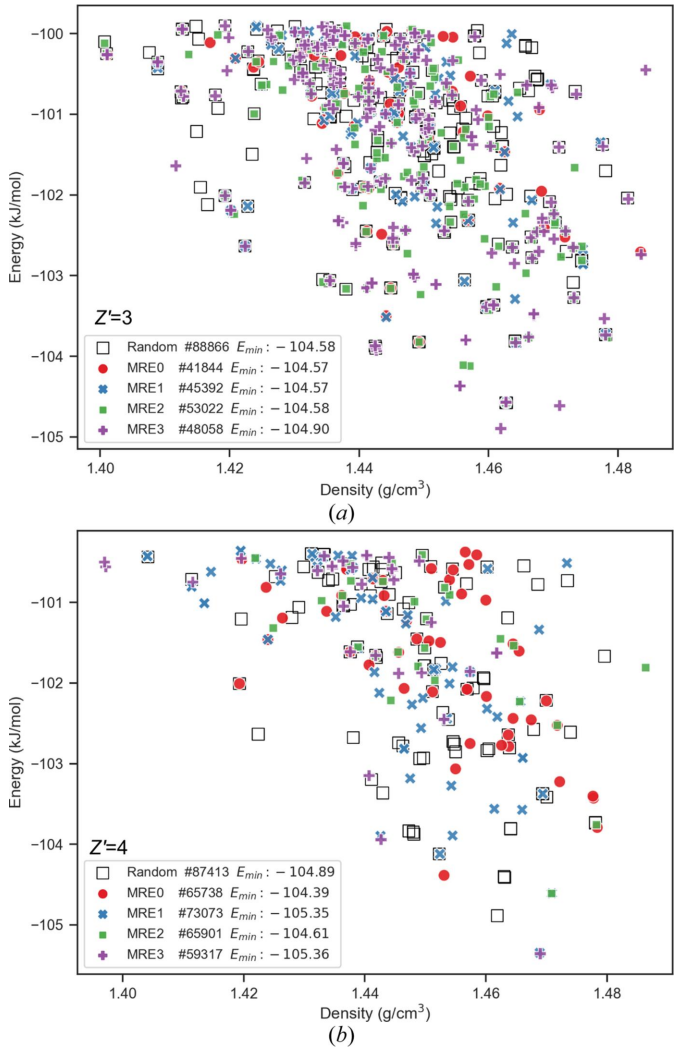



**Figure 11**
Low-energy high-density region from the crystal polymorph landscapes generated for (*a*) $Z' = 3$ and (*b*) $Z' = 4$ across all 13 space groups. Search results from EVCCPMRE variants are colored using different shaded markers with random searches as black square outlines.

**Table 2**
Comparative $Z' = 3, 4$ search summary measures from the different MRE schemes and pseudo-random sampling evaluated for 13 space groups at different $N$ intervals along the search path.

$R_g$ and $\delta E_g$ are measures representing the degree the energy metric differs between MRE and pseudo-random searches. Table 1(*a*) lists $\mathbb{E}[\ldots]$ for $E_{\min}$ and Table 1(*b*) has $\mathbb{E}[\ldots]$ as the $\langle U\rangle_{T20} - \langle U\rangle$ metric.

2(*a*)

| | | | $R_g$ @ $N$ = | | | | $\delta E_g$ @ $N$ = | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $Z'$ | Variant | $M$ | 2000 | 4000 | 6000 | 8000 | 2000 | 4000 | 6000 | 8000 |
| 3 | MRE0 | 10 | 0.23 | 0.15 | 0.23 | 0.23 | 0.98 | 1.27 | 1.17 | 1.23 |
| 3 | MRE1 | 10 | 0.46 | 0.38 | 0.38 | 0.31 | 0.27 | 0.61 | 0.61 | 0.60 |
| 3 | MRE2 | 5 | 0.62 | 0.38 | 0.46 | 0.31 | 0.10 | 0.80 | 0.17 | 0.38 |
| 3 | MRE3 | 5 | 0.46 | 0.46 | 0.38 | 0.46 | 0.27 | 0.37 | 0.41 | 0.13 |
| 4 | MRE0 | 5 | 0.23 | 0.23 | 0.23 | 0.15 | 0.93 | 1.50 | 1.17 | 1.48 |
| 4 | MRE1 | 5 | 0.38 | 0.38 | 0.46 | 0.23 | 0.81 | 1.63 | 1.40 | 1.82 |
| 4 | MRE2 | 5 | 0.31 | 0.31 | 0.23 | 0.23 | 0.12 | 0.76 | 0.94 | 0.93 |
| 4 | MRE3 | 5 | 0.31 | 0.15 | 0.23 | 0.46 | 0.45 | 1.63 | 0.87 | 0.75 |

2(*b*)

| | | | $R_g$ @ $N$ = | | | | $\delta E_g$ @ $N$ = | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $Z'$ | Variant | $M$ | 2000 | 4000 | 6000 | 8000 | 2000 | 4000 | 6000 | 8000 |
| 3 | MRE0 | 10 | 0.92 | 1.00 | 1.00 | 1.00 | −6.37 | −7.57 | −8.47 | −9.14 |
| 3 | MRE1 | 10 | 1.00 | 1.00 | 1.00 | 1.00 | −6.84 | −8.36 | −9.85 | −10.88 |
| 3 | MRE2 | 5 | 1.00 | 1.00 | 1.00 | 1.00 | −12.17 | −14.17 | −15.52 | −16.32 |
| 3 | MRE3 | 5 | 1.00 | 0.92 | 0.92 | 1.00 | −10.91 | −10.48 | −10.65 | −13.27 |
| 4 | MRE0 | 5 | 0.92 | 1.00 | 1.00 | 1.00 | −4.04 | −4.74 | −5.98 | −7.08 |
| 4 | MRE1 | 5 | 1.00 | 1.00 | 1.00 | 1.00 | −4.33 | −6.47 | −7.68 | −8.64 |
| 4 | MRE2 | 5 | 0.92 | 0.92 | 0.92 | 0.92 | −10.44 | −12.93 | −14.11 | −14.83 |
| 4 | MRE3 | 5 | 1.00 | 1.00 | 1.00 | 1.00 | −10.03 | −10.75 | −10.10 | −10.33 |

**Table 3**
Composite unbiased energy metrics [$E = U(\mathbf{X}, \mathbf{H})$ in kJ mol$^{-1}$] evaluated over the multiple search types from the 13 space groups.

The latter three columns are percentiles of $E$.

| Variant | #Hits | $E_{\min}$ | $\langle E\rangle_{T20}$ | $E$@0.05% | $E$@1% | $E$@5% |
|---|---|---|---|---|---|---|
| $Z' = 3$ | | | | | | |
| Random | 88866 | −104.575 | −103.577 | −102.061 | −97.802 | −94.954 |
| MRE0 | 41844 | −104.566 | −102.635 | −101.485 | −97.466 | −94.707 |
| MRE1 | 45392 | −104.574 | −102.760 | −102.005 | −97.712 | −94.858 |
| MRE2 | 53022 | −104.579 | −103.465 | −102.637 | −97.844 | −94.854 |
| MRE3 | 48058 | −104.897 | −103.987 | −103.369 | −98.597 | −95.110 |
| $Z' = 4$ | | | | | | |
| Random | 87413 | −104.886 | −103.598 | −101.706 | −96.486 | −93.387 |
| MRE0 | 65738 | −104.385 | −102.675 | −101.110 | −96.174 | −93.269 |
| MRE1 | 73073 | −105.346 | −103.260 | −101.116 | −96.308 | −93.250 |
| MRE2 | 65901 | −104.611 | −101.783 | −100.188 | −96.004 | −93.309 |
| MRE3 | 59317 | −105.358 | −101.647 | −100.426 | −96.541 | −93.528 |

EVCCP sampling were minimal or ill-defined. There was some expectation that enhancements may not be statistically significant (or highly system dependent) as these tests were restricted to a simple rigid body system and that $N$ is too small (*i.e.* with limited MC steps the simulation EVs remain far from equilibrium behavior). In the limit of identical $N$ in comparison, the overall coverage from EVCCPMRE was not expected to be as good as PR sampling due to the opportunistic exploration of a $k$ biased search space which appears to sacrifice the total number of distinct hits. The results certainly re-demonstrate the robustness and acceptability of the PR and QR sampling strategies for CSP.

As expected the MRE searches do identify different polymorphs in low energy regions especially for the case of MRE2 and MRE3 for $Z' = 3$. It is expected that many more might still be identified for large enough $N$ such that the difference between the total number of distinct hits for MRE search versus PR method is reduced. Interpreting the results shown in Tables 2 and 3 is less useful for initial bench-marking or design of hyper-parameter defaults. It is very likely that EVCCP specific parameters such as $k$ and $\Delta\mathbf{S}$ might need further experimentation or on-the-fly adjustment especially when screening other molecular compounds with varying chemical complexity (*e.g.* molecules with many torsional degrees of freedom).

### 4.3. Benchmark for identification of rare polymorphs

Can EVCCPMRE increase the probability of generating a specific polymorph with a known intrinsically low probability

**Table 4**
Estimates of inverse probabilities for coumarin form IV generation using different $Z' = 3$ searches in space group $P2_12_12_1$.

Search options: $\mathcal{HB}$ = History dependent biasing; $\mathcal{FR}$ = Forced-relaxation updating.

| Search | $\mathcal{HB}$ | $\mathcal{FR}$ | $P(\mathbf{X}_{\rm IV}, \mathbf{H}_{\rm IV})^{-1}$ |
|---|---|---|---|
| Random | No | No | 58000 |
| EVCCPMRE | No | No | 32000 |
| EVCCPMRE | Yes | No | 17778 |
| EVCCPMRE | No | Yes | 11428 |
| EVCCPMRE | Yes | Yes | 8421 |

of occurring from PR sampling? The candidate polymorph used for this part of our study was the experimentally identified form IV ($Z' = 3$) of coumarin, which was previously reported to have a probability $P(\mathbf{X}_{\rm IV}, \mathbf{H}_{\rm IV})$ of 1 in 60000 occurring from PR searches (Shtukenberg *et al.*, 2017).

In order to make this evaluation, a $P(\mathbf{X}_{\rm IV}, \mathbf{H}_{\rm IV})$ estimate was made by running 20 ($Z' = 3$) complete EVCCPMRE searches spawned with different initial CV coordinate ($N$ = 8000) in space group $P2_12_12_1$ and counting the number of times the form IV polymorph was generated. This was recalculated analogously using the PR method where it was found the probability was closer to 1 in 58000. The probabilities from variant EVCCPMRE search options are shown in Table 4. It is remarkable that the combination of both history-dependent biasing potential ($\mathcal{HB}$) and forced-relaxation updates ($\mathcal{FR}$) results in roughly a sixfold increase in the probability of successfully generating form IV. In hindsight, from the analysis of CSP landscapes (as documented in the previous section), any enhanced sampling effect from adding in the $\mathcal{HB}$ was less prominent if not spurious. However, for generating a specified rare polymorph, the utility appears more striking.

### 4.4. Free energy calculation

**4.4.1. Overview.** Typically for studies in molecular simulation of crystal polymorphism, the temperature dependence of the free energy difference (FED) between individual forms is of considerable interest (Day *et al.*, 2003; Hoja *et al.*, 2017; Parrinello & Rahman, 1981; Reilly & Tkatchenko, 2013; Yu & Tuckerman, 2011). This is because a realistic ranking of polymorphs is attained by accurately factoring finite temperature effects. In practice, most methods involve direct calculation of phonon spectra to determine entropic contributions from a vibrational partition function. Strategies also exist which are based entirely on sampling with MD or MC finite temperature simulations. However, all methods are both approximate and computationally expensive (Baroni *et al.*, 2001; Martoňák *et al.*, 2003; Reilly & Tkatchenko, 2015; Nyman & Day, 2015; Frenkel & Ladd, 1984).

In an EVCCPMC simulation, the EV evolution includes contributions from the ensemble of crystal polymorphs to which there is the harmonic tether $k$. The EV are said to be scanning the polymorph probabilities at a finite temperature and it is assumed that thermodynamic ensemble averages such as free energy differences are derivable from a collective of EVCCPMC simulation trajectories. It is believed that the concept can be validated and such a FED estimate might be useful in future as a qualitative measure. To demonstrate, as example of such an ensemble FED estimate was determined, namely the FED between nominal variables $Z' = 1$ and $Z' = 2$. In this study we used the free energy perturbation (FEP) method for the FED calculation (Zwanzig, 1954). FEP from ECVCCP was appealing since it is straightforward to implement, requiring the EVs as input coordinates since a modified potential for $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ could be evaluated (see Appendix *A*).

**4.4.2. Free energy perturbation.** The FED between $Z' = 1$ and $Z' = 2$ can be determined using EVCCPMC as the follows.

Given

$$F(Z') = -\frac{1}{\beta_{\mathbf{S}}}\log\left[\mathbb{Z}(Z', \beta_{\mathbf{S}})\right], \tag{9}$$

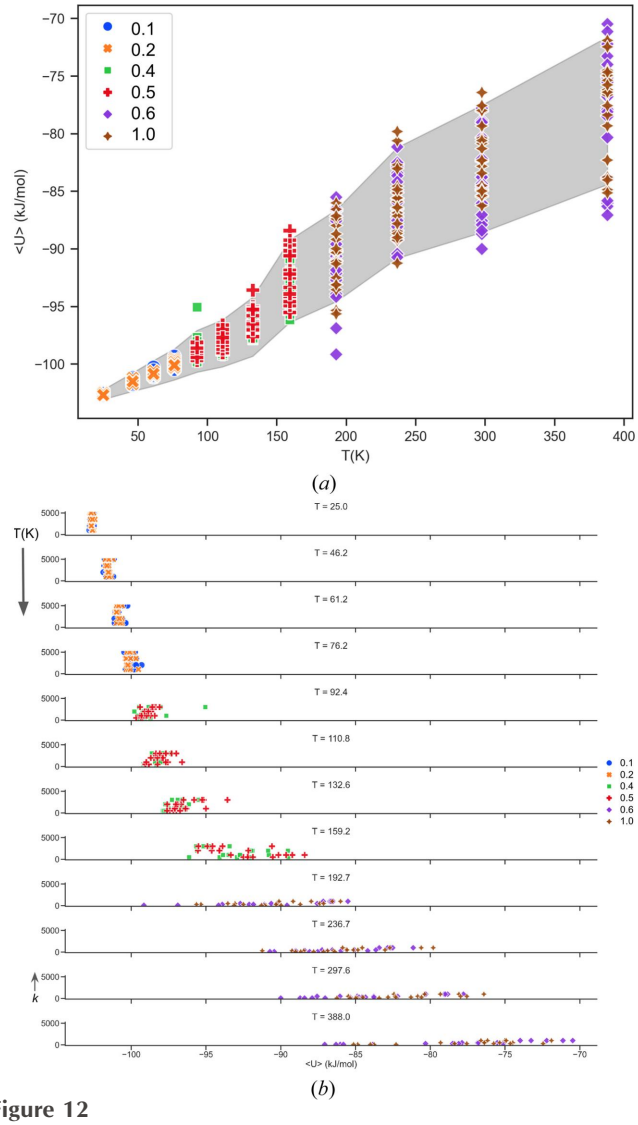



**Figure 12**
The $\langle U \rangle$ values from EVCCPMC simulations at various $T$ are plotted with filled area (gray) representing the RMSD. (*a*) Colored markers for the data points represent the different values for $|\Delta\mathbf{S}_{\rm com}|$ used. (*b*) Row plots of $\langle U \rangle$ are stacked for each $T$. The vertical axis on each subplot corresponds with constant $k$ which takes on values $0 \rightarrow 5000$.

then

$$\Delta F = \left[F\left(\{\mathbf{S}_1\}_{Z'=1}\right) + F\left(\{\mathbf{S}_2\}_{Z'=1}\right)\right] - F\left(\{\mathbf{S}_1, \mathbf{S}_2\}_{Z'=2}\right). \quad (10)$$

To do this a $Z' = 2$ trajectory was generated at a specified $T$ and the resulting EV coordinates are fed back into the polymorph generator with settings for $Z' = 1$. Because a $Z' = 2$ calculation will generate two sets of configurations the energy is re-weighted to account for stoichiometry.

The $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ will be affected both by the harmonic coupling constant ($k$) and the displacement factor ($\Delta\mathbf{S}$) for the shift magnitude which generates MC test positions. For this work, multiple simulation runs were performed with a range of different $\Delta\mathbf{S}$ and $k$ depending on $T$. The deviation of MC acceptance/rejection (AR) ratio from 0.5 was used as a guide (Frenkel & Smit, 2002) to ensure simulation results were reasonable. All $Z' = 2$ ($P2_12_12_1$) runs start from a known predetermined global minimum EV ($\mathbf{S}_{g0}$) corresponding with coumarin form III. Multiple trajectories (steps = 10000) were generated at twelve exponentially spaced $T$ between 25 and 388 K. Spring constants ranged from 100 to 5000 in units of kJ mol$^{-1}$ Å$^{-1}$ for COM displacements or kJ mol$^{-1}$ $\circ^{-1}$ for Euler angles.

**4.4.3. Simulation results.** As expected, it was found that at very low temperatures $U_{\min}$ values do not deviate much since the conditional probability $P(\mathbf{X}_{g0}, \mathbf{H}_{g0}|\mathbf{S}_{g0})$ was reasonably high (up to 0.33 with $k$ = 1000 and $M$ = 20). Also, the EV coordinates $\mathbf{S}$ fluctuate about the point $\mathbf{X}_{g0}$ typical of the behavior for a system tethered to a harmonic spring. As the temperatures are increased the biasing components of the energies were higher and EVs move away from $\mathbf{S}_{g0}$.

The plots of $\langle U\rangle$ are shown in Fig. 12. Each system was considered equilibrated after 2000 MC steps and $\langle U\rangle$ are averaged over the last 8000 steps.

The comparison of the instantaneous $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ for a few different trajectories is depicted in Fig. 13. A plot of the FED values (with $\langle\Delta U\rangle$ representative of the associated error) between $Z' = 2$ and $Z' = 1$ for coumarin crystal polymorph ensembles in space group $P2_12_12_1$ is shown as Fig. 14. To facilitate the FED estimate, recall that a vibrational free energy curve (Nyman & Day, 2015) can be expressed as

$$F_{\text{vib}}(T) = -k_{\text{B}}T\log(\mathbb{Z}_{\text{vib}}) \quad (11)$$

with

$$F_{\text{vib}}(T) = \frac{1}{2}\sum_{i,k}\hbar\omega_{i,k} + k_{\text{B}}T\sum_{i,k}\log\left[1 - \exp\left(\frac{\hbar\omega_{i,k}}{k_{\text{B}}T}\right)\right] = \text{ZPE} + k_{\text{B}}T\Theta, \quad (12)$$

where $\omega_{i,k}$ are phonon frequencies. A $\Delta F_{\text{vib}}(T)$ difference between any two such curves (*e.g.* $\mathcal{A}$ and $\mathcal{B}$) can then be approximated as

$$\Delta_{\mathcal{AB}}F_{\text{vib}}(T) = \left(\text{ZPE}_{\mathcal{B}} - \text{ZPE}_{\mathcal{A}}\right) + k_{\text{B}}T\log\left[\frac{\prod_{i,k}^{(\mathcal{B})}\left(1 - \exp\left(\frac{\hbar\omega_{i,k}^{(\mathcal{B})}}{k_{\text{B}}T}\right)\right)}{\prod_{i,k}^{(\mathcal{A})}\left(1 - \exp\left(\frac{\hbar\omega_{i,k}^{(\mathcal{A})}}{k_{\text{B}}T}\right)\right)}\right] \quad (13)$$

$$\Delta F_{\text{vib}}(T) = C_{\text{ZPE}} + k_{\text{B}}T\mathcal{S}_{\mathcal{V}}.$$

Thus the estimate for the FED between two curves can be grossly simplified as a straight line with a slope ($\mathcal{S}_{\mathcal{V}}$). Despite being unrelated calculations and a high degree of error, there is a correspondence (albeit coincidental) with experimental observations and DFT based phonon calculations (Shtukenberg *et al.*, 2017). For coumarin space group $P2_12_12_1$, experimental form III ($Z' = 2$, $E_{0\,\text{K}} = -103.977$ kJ mol$^{-1}$) and form V ($Z' = 1$, $E_{0\,\text{K}} = -102.366$ kJ mol$^{-1}$) have $E_{0\,\text{K}}$ values close to the global minimum configurations ($E_{\min}$). For fitting the $\mathcal{S}_{\mathcal{V}}$ to

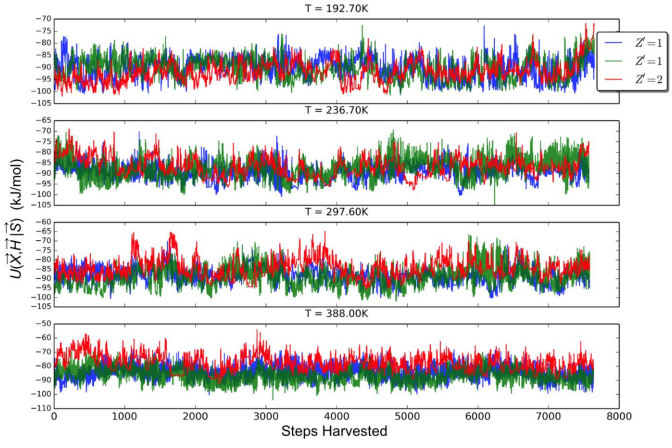



**Figure 13**
Instantaneous $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ taken from EVCCPMC trajectories and used for FEP. The red signal trace is from the $Z' = 2$ simulation. The green and blue traces are a re-sampling of $U$ with $Z' = 1$ using the EVs from the red trace.

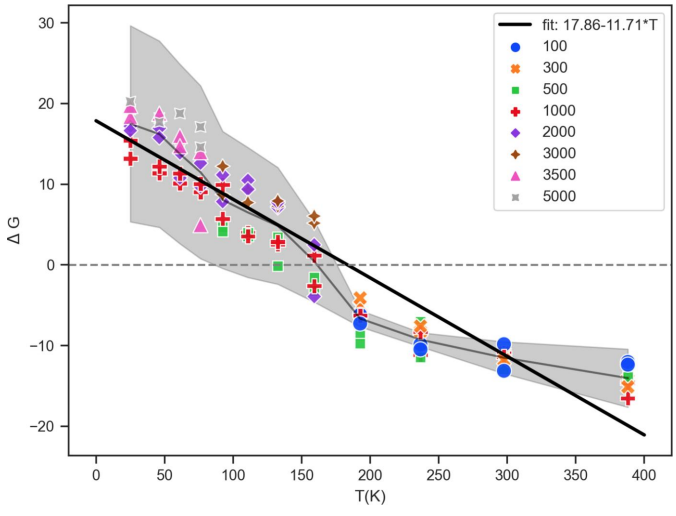



**Figure 14**
The $\Delta G = \Delta F \equiv F^{(Z'=1)} - F^{(Z'=2)}$ versus $T$ linear relation [equation (13)] for coumarin polymorphs in space group $P2_12_12_1$ fitted using data points from EVCCPMC simulations. Different markers represent $k$. The $\langle\Delta G\rangle$ at each $T$ is plotted with a dark gray line. The gray filled area is the magnitude $|\langle\Delta U\rangle|$ centered at $\langle\Delta G\rangle$.

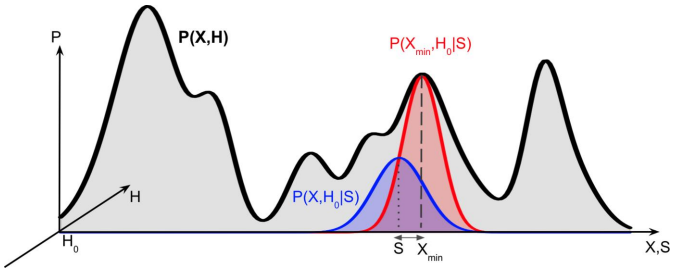



**Figure 15**
The EVCCP approach is based on an ideal relationship between conditional probability densities $P(\mathbf{X}, \mathbf{H}|\mathbf{S})$ which are approximated as Gaussian and the actual probability of sampling any structure $P(\mathbf{X}, \mathbf{H})$. To help illustrate this idea a fictitious 1D plot of the probability distribution for $P(\mathbf{X}, \mathbf{H})$, which is difficult to estimate in practice, is projected down the **H**-axis located at $\mathbf{H}_0$. The red-shaded $P(\mathbf{X}_{\min}, \mathbf{H}_0|\mathbf{S})$ centered on $\mathbf{X}_{\min}$ demonstrates how the probability of obtaining $\mathbf{X}_{\min}$ will be highest when $\mathbf{S} = \mathbf{X}_{\min}$. The blue-shaded $P(\mathbf{X}, \mathbf{H}_0|\mathbf{S})$ is centered on **S** and has its width and height controlled by $k$. The blue distribution shows the actual probability for sampling of $\mathbf{X}_{\min}$.

the data points, the ZPE difference ($C_{\mathrm{ZPE}}$) was not negligible and is fit as an intercept which takes on a positive value indicating that $Z' = 2$ structures are expected to be more favorable at lower temperatures. At higher temperatures a transition point at $\simeq 200$ K occurs. Above this transition temperature $Z' = 1$ polymorphs in $P2_12_12_1$ are predicted to be more favorable due to an entropic stabilization. This is in agreement with results identified using PBE(0)+MDB DFT-based phonon calculations that demonstrated form V should be significantly stabilized when harmonic vibrations and zero-point energies are taken into consideration (Shtukenberg *et al.*, 2017). However, suggesting that the answer to why one polymorph is more likely to crystallize than another may be qualitatively derived by ensemble averaging of many polymorphs is an overly bold statement which still remains invalid.

## 5. Conclusion

A new EV-based approach to CSP has been investigated. The approach relies on harmonic coupling between the reference EV coordinate system and the PR polymorph generator. Results of comparison of EVCCPMRE versus PR based on the coumarin system showed that EVCCP does not always lead to an overall greater yield of polymorphs in the low energy high density region of a landscape. It is believed this might be attributed to the selected $\Delta\mathbf{S}$ and $k$ used in the evaluation but is mostly attributed to the fact that coumarin CSP can be adequately performed within a rigid molecule approximation. It is believed that larger molecular systems with many more degrees of freedom (*i.e.* more rotatable bonds) would benefit from the EVCCPMRE approach in contrast to the PR method.

EVCCPMRE can be modified with history dependent biasing ($\mathcal{HB}$), forced-relaxation ($\mathcal{FR}$) or relaxation-restart ($\mathcal{RR}$) approaches to further enhance sampling. This was evidenced from the sampling statistics for coumarin form IV.

Averaging of $\langle U \rangle$ from EVCCPMC trajectories and performing FEP is one strategy to obtain a FED between ensembles of polymorphs (*i.e.* $Z'$ or space group). The FED tested was for the propensity of either $Z' = 2$ or $Z' = 1$ polymorphs to crystallize and by a sheer coincidence was found to correspond with the attributed entropic stabilization at $T >$ 300 K that lead to the discovery of coumarin form V ($Z' = 1$) from melting and cooling experiments.

## APPENDIX *A*
## Theoretical framework

### A1. Theoretical description for an extended variable reference coupled to crystal polymorph configurations

Generally, a configurational partition function, as applied in molecular simulation, represents the phase space comprising all possible microstates in a system of $N$ molecules (Tuckerman, 2010). The partition function is used to statistically derive thermodynamic properties. We will make a simplification and state that one subset of these microstates encompasses possible crystal polymorphs for a molecule. The configurational variables in this subset are those necessary for describing these polymorphs and can include hidden variable types representing both translational and point symmetry.

Let $\widetilde{\mathbb{Z}}(\beta)$ represent this rudimentary configurational partition function of the crystal polymorph space for a molecular system, defined as follows:

$$\widetilde{\mathbb{Z}}(\beta) = \int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\exp[-\beta U(\mathbf{H}, \mathbf{H})] \tag{14}$$

Here $\mathbf{X} \in \mathbb{R}^d$ has components that are collective variables (CV), which are the positional coordinates for molecular centers (Å) as well as relative molecular orientations in Euler angles (°). In cases where rotations about a bond can occur **X** will also contain these angular coordinates. The number of components in **X** is $d = Z'*6 + n_{\mathrm{tor}}$ where $Z'$ is the number of molecules in the asymmetric unit and $n_{\mathrm{tor}}$ are the number of rotating bonds. $\mathbf{H} \in \mathbb{R}^9$ are the vector coordinates for a parallelepiped or (unit cell). $\mathbf{H} \equiv \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ where **a**, **b** and **c** are the unit-cell vectors in units of Å. It is worth mentioning that in general when $\widetilde{\mathbb{Z}}(N, V, \beta)$ the system volume ($V$) and number of molecules ($N$) are fixed quantities which is not the case here. The states for the system are limited to involving only a certain fixed number of molecules ($Z$) that can occupy a parallelepiped (unit cell) having parameters denoted as **H** (treated as integration variables). Thus, $Z$ is effectively replacing the $N$ in the $\widetilde{\mathbb{Z}}(N, V, \beta)$ formalism. For any such system $Z$ is also related to the space group symmetry (SG) which is considered a hidden constraint variable, held constant, affecting the number of molecules in the asymmetric unit ($Z'$). For example, $\mathbb{Z}(\beta, Z' = 2)$ would represent the configurational partition function for all $Z' = 2$ structures in all possible space groups (there are 230 space groups, however most molecular crystals will commonly manifest in a smaller subset of < 30 of these space groups). $\mathbb{Z}(\beta, Z' = 2)$ can then be further subdivided into subsets of $\mathbb{Z}$ each with a fixed respective space group variable [*e.g.* $\mathbb{Z}(\beta, Z' = 2, \mathrm{SG} = P2_1/c)$]. The 0 K potential energy surface (PES) for the system is

represented with $U(\mathbf{X}, \mathbf{H})$ and is a rough multidimensional surface. $U(\mathbf{X}, \mathbf{H})$ is associated with the crystal energy landscape which is a projection of sampled $U(\mathbf{X}, \mathbf{H})$ for local minimum polymorphs and their respective densities.

Efficient, accurate and thorough sampling of $\mathbb{Z}(\beta)$ is the goal of crystal structure prediction. Historically there are many approaches to do this that go beyond what is necessary to understand this report, however the interested reader can consult the following articles (Reilly *et al.*, 2016; Nyman & Day, 2015; Schneider *et al.*, 2016; Chan *et al.*, 2021; Bier *et al.*, 2021; Yang & Day, 2021*b*; Gavezzotti, 2006; Hoja *et al.*, 2017; van Eijck & Kroon, 1999; Zhu *et al.*, 2014).

In the extended variable coupled to crystal polymorph (EVCCP) scheme, a reference coordinate system with an extended variable (EV) space **S** is introduced and included as an integration variable in the partition function. **S** also represents the same CV coordinates of the molecule as **X**. In this study, the reference system is not self-interacting so $U(\mathbf{S}) = 0$.

$$Z(\beta) = \int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H} \exp[-\beta U(\mathbf{X}, \mathbf{H})] \int \mathrm{d}\mathbf{S} \exp[-\beta U(\mathbf{S})] \tag{15}$$

Energetic coupling between the two systems is introduced which can be written as

$$Z(\beta_{\mathbf{S}}) = \int\int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\,\mathrm{d}\mathbf{S} \exp\big[-\beta_{\mathbf{S}} U(\mathbf{X}, \mathbf{H}, \mathbf{S})\big], \tag{16}$$

where

$$U(\mathbf{X}, \mathbf{H}, \mathbf{S}) = U(\mathbf{X}, \mathbf{H}) + U(\mathbf{S}) + \frac{k}{2}(\mathbf{X} - \mathbf{S})^2. \tag{17}$$

Equation (16) represents the partition function of a system that couples variables **X**, **H**, and **S**. This coupling is achieved by introducing an additional harmonic spring term with coupling strength $k$ that restrains **X** and **S** to take on similar values. This coupling introduces interactions between the different variables, and $\beta_{\mathbf{S}}$ represents the temperature associated with the thermal bath of **S**.

However, it is crucial to emphasize that equation (16) does not imply an adiabatic separation between the systems, and the potential energy surface (PES) $U(\mathbf{X}, \mathbf{H}, \mathbf{S})$ is distinct from $U(\mathbf{X}, \mathbf{H})$. Adiabatic separation would require that the evolution of **X** and **H** is completely decoupled from **S**, meaning that they evolve at different time and frequency regimes.

To handle this, the conditional probability distribution $P(\mathbf{X}, \mathbf{H}|\mathbf{S})$ is introduced. This distribution represents the probability of observing a specific configuration of **X** and **H** given a fixed value of **S**. The existence of this conditional probability distribution indicates that **X** and **H** evolve in a way that is co-dependent with **S**, which allows for the possibility of an adiabatic separation. This conditional distribution relates to the overall joint probability distribution $P(\mathbf{X}, \mathbf{H}, \mathbf{S})$ through the product rule of probability theory, as shown in equation (18). In general this is known as the factorization of a joint probability and equation (16) represents this in terms of a partition function which incorporates an interaction term between co-dependent variables.

$$P(\mathbf{X}, \mathbf{H}, \mathbf{S}) = P(\mathbf{X}, \mathbf{H}|\mathbf{S})P(\mathbf{S}) \tag{18}$$

Also, because

$$\beta U(\mathbf{X}, \mathbf{H}, \mathbf{S}) = -\log[P(\mathbf{X}, \mathbf{H}, \mathbf{S})] + \text{const} \tag{19}$$

implies the relation

$$U(\mathbf{X}, \mathbf{H}) \approx \int \mathrm{d}\mathbf{S}\ \frac{-\log[P(\mathbf{X}, \mathbf{H}|\mathbf{S})]}{\beta} + \text{const} \tag{20}$$

can be defined.

This conditional probability distribution is a subset of the larger overall distribution $P(\mathbf{X}, \mathbf{H})$, from which useful information and properties are extracted. In essence, the introduction of $P(\mathbf{X}, \mathbf{H}|\mathbf{S})$ within the context of equation (16) makes the goal of sampling and extracting these useful ensemble properties more tractable (see Fig. 15). It serves to handle the interactions and dependencies between variables within the system. The notation $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ indicates a modified potential that accounts for the adiabatic decoupling of variables for the relations described in equation (20).

This approach, where variables can now be adiabatically separated and are still co-dependent, shares conceptual similarities with umbrella sampling (Torrie & Valleau, 1977), a technique used to enhance the sampling of specific regions of the configurational space.

It is possible to state the following:

$$\begin{aligned}
P(\mathbf{X}, \mathbf{H}, \mathbf{S}) &= \frac{1}{\mathbb{Z}(\beta_{\mathbf{S}})} \exp[-\beta_{\mathbf{S}} U(\mathbf{X}, \mathbf{H}, \mathbf{S})] \\
P(\mathbf{S}) &= \frac{1}{\int \mathrm{d}\mathbf{S} \exp[-\beta_{\mathbf{S}} U(\mathbf{S})]} \exp[-\beta_{\mathbf{S}} U(\mathbf{S})] \\
P(\mathbf{S}) &= \frac{P(\mathbf{X}, \mathbf{H}, \mathbf{S})}{P(\mathbf{X}, \mathbf{H}|\mathbf{S})} \\
P(\mathbf{S}) &= \frac{1}{\mathbb{Z}(\beta_{\mathbf{S}})} \int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H} \exp[-\beta_{\mathbf{S}} U(\mathbf{X}, \mathbf{H}, \mathbf{S})]
\end{aligned} \tag{21}$$

The $P(\mathbf{S})$ in equation (21) is a familiar sum rule from probability theory. This means that

$$\begin{aligned}
P(\mathbf{X}, \mathbf{H}|\mathbf{S}) &= \frac{P(\mathbf{X}, \mathbf{H}, \mathbf{S})}{P(\mathbf{S})} \\
P(\mathbf{X}, \mathbf{H}|\mathbf{S}) &= \frac{\frac{1}{\mathbb{Z}(\beta_{\mathbf{S}})} \exp[-\beta_{\mathbf{S}} U(\mathbf{X}, \mathbf{H}, \mathbf{S})]}{\frac{1}{\mathbb{Z}(\beta_{\mathbf{S}})} \int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H} \exp[-\beta_{\mathbf{S}} U(\mathbf{X}, \mathbf{H}, \mathbf{S})]} \\
P(\mathbf{X}, \mathbf{H}|\mathbf{S}) &= \frac{\exp[-\beta_{\mathbf{S}} U(\mathbf{X}, \mathbf{H}, \mathbf{S})]}{\int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H} \exp[-\beta_{\mathbf{S}} U(\mathbf{X}, \mathbf{H}, \mathbf{S})]}.
\end{aligned} \tag{22}$$

It is then proposed that the joint probability from equation (21) may be represented as

$$P(\mathbf{X}, \mathbf{H}, \mathbf{S}) = \frac{\exp[-\beta_{\mathbf{S}} U(\mathbf{X}, \mathbf{H}|\mathbf{S})]}{\int\int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\,\mathrm{d}\mathbf{S} \exp[-\beta_{\mathbf{S}} U(\mathbf{X}, \mathbf{H}|\mathbf{S})]}, \tag{23}$$

*i.e.* $U(\mathbf{X}, \mathbf{H}, \mathbf{S})$ is replaced with $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ and **S** is introduced in the denominator of equation (22) as an integration variable. Note that the modified potential $U(\mathbf{X}, \mathbf{H}|\mathbf{S})$ still remains undefined. The denominator represents another partition

function $\mathbb{Z}'(\beta_{\mathbf{S}})$ which now includes the same integration variables as equation (16).

$$\mathbb{Z}'(\beta_{\mathbf{S}}) = \int\int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\,\mathrm{d}\mathbf{S}\exp[-\beta_{\mathbf{S}}U(\mathbf{X},\mathbf{H}|\mathbf{S})] \tag{24}$$

Also the independent construction of

$$\mathbb{Z}(\langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M) = \int\int \mathrm{d}\mathbf{X}\mathrm{d}\mathbf{H}\cdot\delta[U(\mathbf{X},\mathbf{H}) - \langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M] \tag{25}$$

is made. $\mathbb{Z}(\langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M)$ represents a micro-canonical partition function which will be used as a tool at a later stage. The $\langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M$ represents the average over a subset of $M$ microstates, which in practice is estimated from the biased distribution $P(\mathbf{X},\mathbf{H}|\mathbf{S})$, *i.e.* fixed value of $\mathbf{S}$. Equation (25) represents scanning over the entire phase space of $\mathbf{X}$ and $\mathbf{H}$ to identify all configurations where $U(\mathbf{X},\mathbf{H}) = \langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M$. This partition function differs from a conventional micro-canonical definition in that it is representative of a 'Dirac comb' that identifies the subset of configurations which match for $\langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M$. The subscript adb is for 'adiabatic' and is used to differentiate $U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})$ from $U(\mathbf{X},\mathbf{H}|\mathbf{S}) = \langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M + \frac{k}{2}(\langle\mathbf{X}\rangle_M - \mathbf{S})^2$ which will be described later. One can also represent the difference between $\mathbb{Z}(\beta_{\mathbf{S}})$ and $\mathbb{Z}'(\beta_{\mathbf{S}})$ as

$$\langle U(\mathbf{X},\mathbf{H},\mathbf{S})\rangle = -\frac{\partial}{\partial\beta_{\mathbf{S}}}\log[\mathbb{Z}(\beta_{\mathbf{S}})] \tag{26}$$

and

$$\langle U(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle = -\frac{\partial}{\partial\beta_{\mathbf{S}}}\log[\mathbb{Z}'(\beta_{\mathbf{S}})]. \tag{27}$$

The following reformulation of equation (16) into equation (24) by application of equation (25) is made. This is written out as

$$\begin{aligned}\mathbb{Z}'(\beta_{\mathbf{S}}) &= \mathbb{Z}(\beta_{\mathbf{S}})\times\mathbb{Z}(\langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M)\\ \mathbb{Z}'(\beta_{\mathbf{S}}) &= \int\int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\,\mathrm{d}\mathbf{S}\exp[-\beta_{\mathbf{S}}U(\mathbf{X},\mathbf{H},\mathbf{S})]\\ &\quad\times\int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\cdot\delta[U(\mathbf{X},\mathbf{H}) - \langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M],\end{aligned} \tag{28}$$

which is the general form of the partition function used for EVCCP. Equation (28) can be written as

$$\begin{aligned}\mathbb{Z}'(\beta_{\mathbf{S}}) &= \int\int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\,\mathrm{d}\mathbf{S}\exp\left[-\beta_{\mathbf{S}}U(\mathbf{X},\mathbf{H}) + \frac{k}{2}(\mathbf{X}-\mathbf{S})^2\right]\\ &\quad\times\int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\cdot\delta[U(\mathbf{X},\mathbf{H}) - \langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M]\\ \mathbb{Z}'(\beta_{\mathrm{S}}) &= \int\int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\,\mathrm{d}\mathbf{S}\exp\left[-\beta_{\mathrm{s}}\left(\langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M\right.\right.\\ &\quad\left.\left.+\frac{k}{2}(\langle\mathbf{X}\rangle_M - \mathbf{S})^2\right)\right]\end{aligned} \tag{29}$$

where $\langle\mathbf{X}\rangle_M$ denotes the average value of $\mathbf{X}$ that was obtained from configurations where the value of $\mathbf{S}$ was fixed, hence $\mathbf{X}$ and $\mathbf{H}$ are adiabatically decoupled from $\mathbf{S}$ [*i.e.* using fixed $\mathbf{S}$ to sample an estimate for $P(\mathbf{X},\mathbf{H}|\mathbf{S})$ and obtain $\langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\rangle_M$ and $\langle\mathbf{X}\rangle_M$]. Consider now that we can expand the energetic term in equation (29) as follows:

$$\begin{aligned}U(\mathbf{X},\mathbf{H}|\mathbf{S}) &= \left\langle U_{\mathrm{adb}}(\mathbf{X},\mathbf{H}|\mathbf{S})\right\rangle_M + \frac{k}{2}(\langle\mathbf{X}\rangle_M - \mathbf{S})^2\\ &= \frac{1}{M}\sum_{i=1}^{M}\left[U_{\mathrm{adb}}(\mathbf{X}_i,\mathbf{H}_i|\mathbf{S}) + \frac{k}{2}(\mathbf{X}_i-\mathbf{S})^2\right]\\ &= \frac{1}{M}\sum_{i=1}^{M}\left[U_{\mathrm{min}} + \left(U_{\mathrm{adb}}(\mathbf{X}_i,\mathbf{H}_i|\mathbf{S}) - U_{\mathrm{min}}\right)\right.\\ &\quad\left.+\frac{k}{2}(\mathbf{X}_i-\mathbf{S})^2\right]\end{aligned} \tag{30}$$

where $U_{\mathrm{adb}}(\mathbf{X}_i,\mathbf{H}_i|\mathbf{S})$ is the energy of the $i$th polymorph generated from $P(\mathbf{X},\mathbf{H}|\mathbf{S})$ under the influence of the reference $\mathbf{S}$ and

$$\begin{aligned}U_{\mathrm{min}} &= U_{\mathrm{adb}}(\mathbf{X}_{\mathrm{min}},\mathbf{H}_{\mathrm{min}}|\mathbf{S})\\ &= \mathrm{Min}\left[\left\{\begin{array}{c}U_{\mathrm{adb}}(\mathbf{X}_1,\mathbf{H}_1|\mathbf{S}) + \frac{k}{2}(\mathbf{X}_1-\mathbf{S})^2,\\ U_{\mathrm{adb}}(\mathbf{X}_2,\mathbf{H}_2|\mathbf{S}) + \frac{k}{2}(\mathbf{X}_2-\mathbf{S})^2,\\ \ldots,\\ U_{\mathrm{adb}}(\mathbf{X}_M,\mathbf{H}_M|\mathbf{S}) + \frac{k}{2}(\mathbf{X}_M-\mathbf{S})^2\end{array}\right\}\right]\\ &\quad-\frac{k}{2}(\mathbf{X}_{\mathrm{min}}-\mathbf{S})^2\end{aligned} \tag{31}$$

In the last line of equation (30) the value $U_{\mathrm{min}}$ is introduced which is the energy of the lowest energy configuration with the harmonic penalty taken into account. $U_{\mathrm{min}}$ is obtained from sampling the $M$ structures for fixed $\mathbf{S}$. $U_{\mathrm{min}}$ enables further manipulation of equation (30) so that $P(\mathbf{X},\mathbf{H}|\mathbf{S})$ can be approximated with a Gaussian distribution as shown in Fig. 15 [*i.e.* $P(\mathbf{X},\mathbf{H}|\mathbf{S})$ is a Gaussian distribution centered at $\mathbf{X}_{\mathrm{min}}$ and $\mathbf{H}_{\mathrm{min}}$)]. A Gaussian approximation can be made which is based on the square of the deviation of the variables $\mathbf{X}_i$ and $\mathbf{H}_i$ from those of the minimum energy polymorph configuration $\mathbf{X}_{\mathrm{min}}$ and $\mathbf{H}_{\mathrm{min}}$ that was identified in the mini-batch generation. Within the proposed Gaussian approximation equation (30) becomes

$$\begin{aligned}U(\mathbf{X},\mathbf{H}|\mathbf{S}) &= \frac{1}{M}\sum_{i=1}^{M}\left[U_{\mathrm{min}} + \left(\mathbf{X}_i-\mathbf{X}_{\mathrm{min}}\right)^2 + \left(\mathbf{H}_i-\mathbf{H}_{\mathrm{min}}\right)^2\right.\\ &\quad\left.+\frac{k}{2}\left(\mathbf{X}_i-\mathbf{S}\right)^2\right]\end{aligned} \tag{32}$$

A Gaussian approximation implies that in equation (30) the $\langle U_{\mathrm{adb}}(\mathbf{X}_{\mathrm{min}},\mathbf{H}_{\mathrm{min}}|\mathbf{S})\rangle_M = U_{\mathrm{min}}$. The summation is represented as an integral with a $\delta$ function and equation (32) is rewritten as follows:

$$\begin{aligned}U(\mathbf{X},\mathbf{H}|\mathbf{S}) &= \int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\left[U_{\mathrm{min}} + \left(\mathbf{X}-\mathbf{X}_{\mathrm{min}}\right)^2 + \left(\mathbf{H}-\mathbf{H}_{\mathrm{min}}\right)^2\right.\\ &\quad\left.+\frac{K}{2}\left(\mathbf{X}_i-\mathbf{S}\right)^2\right]\times\delta\left[\mathbf{X}-\mathbf{X}_{\mathrm{min}}\right]\cdot\delta\left[\mathbf{H}-\mathbf{H}_{\mathrm{min}}\right]\\ U(\mathbf{X},\mathbf{H}|\mathbf{S}) &\equiv U(\mathbf{X}_{\mathrm{min}},\mathbf{H}_{\mathrm{min}}|\mathbf{S}) = U_{\mathrm{min}} + \frac{k}{2}(\mathbf{X}_{\mathrm{min}}-\mathbf{S})^2\end{aligned} \tag{33}$$

Equations (28) and (29) are then rewritten as follows:

$$\mathbb{Z}'(\beta_{\mathbf{S}}) = \int\int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\,\mathrm{d}\mathbf{S}\exp\left[-\beta_{\mathbf{S}}U(\mathbf{X},\mathbf{H},\mathbf{S})\right] \times \int\int \mathrm{d}\mathbf{X}\,\mathrm{d}\mathbf{H}\cdot\delta\left[U(\mathbf{X},\mathbf{H}) - U_{\min}\right] \tag{34}$$

or

$$\mathbb{Z}'(\beta_{\mathbf{S}}) = \int\int\int \mathrm{d}\mathbf{X}_{\min}\,\mathrm{d}\mathbf{H}_{\min}\,\mathrm{d}\mathbf{S}\exp\left[-\beta_{\mathbf{S}}\left(U_{\min}+\frac{k}{2}(\mathbf{X}_{\min}-\mathbf{S})^2\right)\right] \tag{35}$$

As shown in Fig. 15 the utility of this approach is to approximate the probability distribution $P(\mathbf{X},\mathbf{H})$ with many smaller Gaussian distributions which are centered at **S**. Because **S** can deviate from **X** based on the value of the coupling constant ($k$) this means that the coupling constant also acts to adjust the spread of the Gaussian distribution which affects the resolution of how well the approximation of $P(\mathbf{X},\mathbf{H})$ can be made since this approximation is based on sub-sampling many $P(\mathbf{X},\mathbf{H}|\mathbf{S})$. For very large values of $k$ a reconstruction of $P(\mathbf{X},\mathbf{H})$ from integration of $P(\mathbf{X},\mathbf{H},\mathbf{S})$ would be as close as possible to the true $P(\mathbf{X},\mathbf{H})$. However for smaller $k$ the reconstruction of $P(\mathbf{X},\mathbf{H})$ is smoothed.

### A2. Calculation of free energy differences

The following equations outline the scheme used for calculating the free energy difference between polymorph ensembles $\mathbb{Z}_{\mathcal{A}}(U_{\min},\beta_{\mathbf{S}},Z'=2,\mathrm{SG}=P2_12_12_1)$ and $\mathbb{Z}_{\mathcal{B}}(U_{\min},\beta_{\mathbf{S}},Z'=1,\mathrm{SG}=P2_12_12_1)$ sampled using EVCCPMC. The Zwanzig relation (Zwanzig, 1954) provides a prescription for calculating the free energy difference $\Delta F_{\mathcal{AB}}$ from the ratio between $\mathbb{Z}_{\mathcal{A}}$ and $\mathbb{Z}_{\mathcal{B}}$ such that

$$\frac{\mathbb{Z}_{\mathcal{B}}}{\mathbb{Z}_{\mathcal{A}}} = \left\langle \exp\left[-\beta(U_{\mathcal{B}}-U_{\mathcal{A}})\right]\right\rangle_{\mathbf{S}(\mathcal{A})} \tag{36}$$

where the subscript with angle brackets $\langle\ldots\rangle_{\mathbf{S}(\mathcal{A})}$ denotes averaging taken with respect to the collective variables **S** trajectory generated using $\mathbb{Z}_{\mathcal{A}}$, *i.e.* $Z'=2$.

$$\Delta F_{\mathcal{AB}} = \frac{-1}{\beta_{\mathbf{S}}}\log\left(\frac{\mathbb{Z}_{\mathcal{B}}}{\mathbb{Z}_{\mathcal{A}}}\right) \tag{37}$$

It is appropriate to express such a formulation with respect to the energetic coupling terms [see equation (33)] used for EVCCPMC so that $U_{\mathcal{A}}$ and $U_{\mathcal{B}}$ are representative of

$$\begin{aligned} U_{\mathcal{A},\min} &= U\left(X_1^{(a)},X_2^{(a)},H_{12}^{(a)}|S_1^{(a)},S_2^{(a)}\right)+0.5k\left(X_1^{(a)}-S_1^{(a)}\right)^2 \\ &\quad +0.5k\left(X_2^{(a)}-S_2^{(a)}\right)^2 \\ U_{\mathcal{B},\min} &= 0.5\Big(U\left(X_1^{(b)},H_1^{(b)}|S_1^{(a)}\right)+k\left(X_1^{(b)}-S_1^{(a)}\right)^2 \\ &\quad +U\left(X_2^{(b)},H_2^{(b)}|S_2^{(a)}\right)+k\left(X_2^{(b)}-S_2^{(a)}\right)^2\Big) \end{aligned} \tag{38}$$

where the superscript $(a)$ and $(b)$ denote which ensemble had generated the particular variable, also $U(X_1,X_2,H_{12}|S_1,S_2)$, $U(X_1,H_1|S_1)$ and $U(X_2,H_2|S_2)$ are lattice energies which are scaled relative to $Z'$. Such a workflow requires generating a MC trajectory for $Z'=2$ to obtain $S_1^{(a)}$ and $S_2^{(a)}$ coordinates at a specified $\beta_S$. This is followed by two separate $Z'=1$ recalculations to obtain $U_{\mathcal{B},\min}$. So for $Z'=1$ the input $S$ is fixed and comes directly from the $Z'=2$ calculation. The $S^{(a)}$ will experience a field effect and forces created by $X^{(b)}$ and $H^{(b)}$ as it is coupled with the conditional distribution of polymorphs that can be generated using the fixed $S^{(a)}$, *i.e.* $P(X^{(b)},H^{(b)}|S^{(a)})$. The corresponding energy is evaluated as $U(X^{(b)},H^{(b)}|S^{(a)}) + k(X^{(b)}-S^{(a)})^2$.

## APPENDIX *B*

### Further computational details

The algorithm for pseudo-random (PR) sampling of polymorphs and structure optimization was made available as part of the program *UPACK* (van Eijck & Kroon, 1999). The chosen force field parameters and charge assignments comply with the generalized amber force field (Wang *et al.*, 2004).

An initial polymorph description or 'test structure' as atomic positions and unit-cell parameters (Cartesian coordinates) or as CVs (**X** and **H**) is randomly sampled from a uniform distribution. The initial position in the polymorph landscape is then subjected to a density test so that the unit-cell parameters are sensible and within specified tolerances. Given the density test is satisfied, gradient descent moves are made based on the chosen force field and the structure is energy minimized and becomes a local minimum on the 0 K PES. If at any point within the optimization part workflow the structure energy goes above a certain threshold then the test structure is rejected.

In practice the $P(\mathbf{X},\mathbf{H})$ associated with an identified polymorph at the local minimum depends on the boundaries associated with the gradients surrounding **X** and **H**. This probability is complex and not a uniform distribution. $P(\mathbf{X},\mathbf{H})$ from PR sampling considers polymorphs as local minima on the PES and must have some dependence on the $U(\mathbf{X},\mathbf{H})$. Generally, $P(\mathbf{X},\mathbf{H})$ will take on a shape that is difficult to estimate even for very simple molecular systems and rarely evaluated in practice using statistical methods.

## Acknowledgements

The authors express gratitude to Professor Julian Gale, at Curtin Institute for Computation, for providing access to academic computing and library resources. The authors also extend thanks to the team at PAWSEY supercomputing research centre for ongoing access to the Nimbus facility. Special thanks must go to Julian Gale, Jutta Rogal, Peter Spackman and Leslie Vogt for their valuable discussions on the topic of CSP.

## Funding information

The following funding is acknowledged: Materials Research Science and Engineering Center, NYU (award No. DMR-1420073).

## References


Abrams, J. B. & Tuckerman, M. E. (2008). *J. Phys. Chem. B*, **112**, 15742–15757.

Allen, M. P. & Tildesley, D. J. (1987). *Computer Simulation of Liquids.* Clarendon Press.

Baroni, S., Gironcoli, S. de, Dal Corso, A. & Giannozzi, P. (2001). *Rev. Mod. Phys.* **73**, 515.

Bernstein, J. (2002). *Polymorphism in Molecular Crystals.* Oxford University Press.

Bernstein, J. (2008). *Crystal Polymorphism,* pp. 87–109. Springer Netherlands.

Bier, I., O'Connor, D., Hsieh, Y.-T., Wen, W., Hiszpanski, A. M., Han, T. Y.-J. & Marom, N. (2021). *CrystEngComm*, **23**, 6023–6038.

Case, D. H., Campbell, J. E., Bygrave, P. J. & Day, G. M. (2016). *J. Chem. Theory Comput.* **12**, 910–924.

Chan, E. J., Shtukenberg, A. G., Tuckerman, M. E. & Kahr, B. (2021). *Cryst. Growth Des.* **21**, 5544–5557.

Ciccotti, G. & Meloni, S. (2011). *Phys. Chem. Chem. Phys.* **13**, 5952.

Davey, R. J. & Garside, J. (2000). *From Molecules to Crystallizers - an Introduction to Crystallization*. Oxford Chemistry Primers, Vol. 86. Oxford Science Publications.

Day, G. M., Price, S. L. & Leslie, M. (2003). *J. Phys. Chem. B*, **107**, 10919–10933.

Desiraju, G. R. (1989). *Crystal Engineering: the Design of Organic Solids.* Elsevier.

Desiraju, G. R. (2001). *Nature*, **412**, 397–400.

Dunitz, J. D. (1995). *X-ray Analysis and the Structure of Organic Molecules.* Weinheim, New York: VCH.

Frenkel, D. & Ladd, A. J. C. (1984). *J. Chem. Phys.* **81**, 3188–3193.

Frenkel, D. & Smit, B. (2002). *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed. Academic Press.

Gavezzotti, A. (2006). *Molecular Aggregation: Structure Analysis and Molecular Simulation of Crystals and Liquids.* Oxford University Press.

Hartman, P. (1973). *Crystal Growth: an Introduction*, Vol. 1. Amsterdam: North Holland Publishing Co.

Hasenbusch, M. & Schaefer, S. (2010). *Phys. Rev. E*, **82**, 046707.

Hermann, J., DiStasio, R. A. & Tkatchenko, A. (2017). *Chem. Rev.* **117**, 4714–4758.

Hoja, J., Reilly, A. M. & Tkatchenko, A. (2017). *WIREs Comput. Mol. Sci.* **7**, e1294.

Hunnisett, L. M., Francia, N., Nyman, J., Abraham, N. S., Aitipamula, S., Alkhidir, T., Almehairbi, M., Anelli, A., Anstine, D. M., Anthony, J. E., Arnold, J. E., Bahrami, F., Bellucci, M. A., Beran, G. J. O., Bhardwaj, R. M., Bianco, R., Bis, J. A., Boese, A. D., Bramley, J., Braun, D. E., Butler, P. W. V., Cadden, J., Carino, S., Červinka, C., Chan, E. J., Chang, C., Clarke, S. M., Coles, S. J., Cook, C. J., Cooper, R. I., Darden, T., Day, G. M., Deng, W., Dietrich, H., DiPasquale, A., Dhokale, B., van Eijck, B. P., Elsegood, M. R. J., Firaha, D., Fu, W., Fukuzawa, K., Galanakis, N., Goto, H., Greenwell, C., Guo, R., Harter, J., Helfferich, J., Hoja, J., Hone, J., Hong, R., Hušák, M., Ikabata, Y., Isayev, O., Ishaque, O., Jain, V., Jin, Y., Jing, A., Johnson, E. R., Jones, I., Jose, K. V. J., Kabova, E. A., Keates, A., Kelly, P. F., Klimeš, J., Kostková, V., Li, H., Lin, X., List, A., Liu, C., Liu, Y. M., Liu, Z., Lončarić, I., Lubach, J. W., Ludík, J., Maryewski, A. A., Marom, N., Matsui, H., Mattei, A., Mayo, R. A., Melkumov, J. W., Mladineo, B., Mohamed, S., Momenzadeh Abardeh, Z., Muddana, H. S., Nakayama, N., Nayal, K. S., Neumann, M. A., Nikhar, R., Obata, S., O'Connor, D., Oganov, A. R., Okuwaki, K., Otero-de-la-Roza, A., Parkin, S., Parunov, A., Podeszwa, R., Price, A. J. A., Price, L. S., Price, S. L., Probert, M. R., Pulido, A., Ramteke, G. R., Rehman, A. U., Reutzel-Edens, S. M., Rogal, J., Ross, M. J., Rumson, A. F., Sadiq, G., Saeed, Z. M., Salimi, A., Sasikumar, K., Sekharan, S., Shankland, K., Shi, B., Shi, X., Shinohara, K., Skillman, A. G., Song, H., Strasser, N., van de Streek, J., Sugden, I. J., Sun, G., Szalewicz, K., Tan, L., Tang, K., Tarczynski, F., Taylor, C. R., Tkatchenko, A., Touš, P., Tuckerman, M. E., Unzueta, P. A., Utsumi, Y., Vogt-Maranto, L., Weatherston, J., Wilkinson, L. J., Willacy, R. D., Wojtas, L., Woollam, G. R., Yang, Y., Yang, Z., Yonemochi, E., Yue, X., Zeng, Q., Zhou, T., Zhou, Y., Zubatyuk, R. & Cole, J. C. (2024). *Acta Cryst.* B**80**, https://doi.org/10.1107/S2052520624008679.

Hunnisett, L. M., Nyman, J., Francia, N., Abraham, N. S., Adjiman, C. S., Aitipamula, S., Alkhidir, T., Almehairbi, M., Anelli, A., Anstine, D. M., Anthony, J. E., Arnold, J. E., Bahrami, F., Bellucci, M. A., Bhardwaj, R. M., Bier, I., Bis, J. A., Boese, A. D., Bowskill, D. H., Bramley, J., Brandenburg, J. G., Braun, D. E., Butler, P. W. V., Cadden, J., Carino, S., Chan, E. J., Chang, C., Cheng, B., Clarke, S. M., Coles, S. J., Cooper, R. I., Couch, R., Cuadrado, R., Darden, T., Day, G. M., Dietrich, H., Ding, Y., DiPasquale, A., Dhokale, B., van Eijck, B. P., Elsegood, M. R. J., Firaha, D., Fu, W., Fukuzawa, K., Glover, J., Goto, H., Greenwell, C., Guo, R., Harter, J., Helfferich, J., Hofmann, D. W. M., Hoja, J., Hone, J., Hong, R., Hutchison, G., Ikabata, Y., Isayev, O., Ishaque, O., Jain, V., Jin, Y., Jing, A., Johnson, E. R., Jones, I., Jose, K. V. J., Kabova, E. A., Keates, A., Kelly, P. F., Khakimov, D., Konstantinopoulos, S., Kuleshova, L. N., Li, H., Lin, X., List, A., Liu, C., Liu, Y. M., Liu, Z., Liu, Z.-P., Lubach, J. W., Marom, N., Maryewski, A. A., Matsui, H., Mattei, A., Mayo, R. A., Melkumov, J. W., Mohamed, S., Momenzadeh Abardeh, Z., Muddana, H. S., Nakayama, N., Nayal, K. S., Neumann, M. A., Nikhar, R., Obata, S., O'Connor, D., Oganov, A. R., Okuwaki, K., Otero-de-la-Roza, A., Pantelides, C. C., Parkin, S., Pickard, C. J., Pilia, L., Pivina, T., Podeszwa, R., Price, A. J. A., Price, L. S., Price, S. L., Probert, M. R., Pulido, A., Ramteke, G. R., Rehman, A. U., Reutzel-Edens, S. M., Rogal, J., Ross, M. J., Rumson, A. F., Sadiq, G., Saeed, Z. M., Salimi, A., Salvalaglio, M., Sanders de Almada, L., Sasikumar, K., Sekharan, S., Shang, C., Shankland, K., Shinohara, K., Shi, B., Shi, X., Skillman, A. G., Song, H., Strasser, N., van de Streek, J., Sugden, I. J., Sun, G., Szalewicz, K., Tan, B. I., Tan, L., Tarczynski, F., Taylor, C. R., Tkatchenko, A., Tom, R., Tuckerman, M. E., Utsumi, Y., Vogt-Maranto, L., Weatherston, J., Wilkinson, L. J., Willacy, R. D., Wojtas, L., Woollam, G. R., Yang, Z., Yonemochi, E., Yue, X., Zeng, Q., Zhang, Y., Zhou, T., Zhou, Y., Zubatyuk, R. & Cole, J. C. (2024). *Acta Cryst.* B**80**, https://doi.org/10.1107/S2052520624007492.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). *Nature*, **596**, 583–589.

Kennedy, J. & Eberhart, R. (1995). In *Proceedings of ICNN'95. International Conference on Neural Networks*, Vol. 4, pp. 1942–1948.

Kitaigorodskiy, A. I., Mnyukh, Y. V. & Asadov, Y. G. (1965). *J. Phys. Chem. Solids*, **26**, 463–464.

Kitaigorosky, A. I. (1973). *Molecular Crystals and Molecules.* Academic Press.

Laio, A. & Parrinello, M. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 12562–12566.

Liu, P., Kim, B., Friesner, R. A. & Berne, B. J. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 13749–13754.

Maragliano, L. & Vanden-Eijnden, E. (2006). *Chem. Phys. Lett.* **426**, 168–175.

Martoňák, R., Laio, A. & Parrinello, M. (2003). *Phys. Rev. Lett.* **90**, 075503.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087–1092.

Mullin, J. W. (2001). *Crystallization*, 4th ed. Oxford: Butterworth–Heinemann.

Neumann, M. A. (2008). *J. Phys. Chem. B*, **112**, 9810–9829.

Neumann, M., Leusen, F. & Kendrick, J. (2008). *Angew. Chem. Int. Ed.* **47**, 2427–2430.
Nyman, J. & Day, G. M. (2015). *CrystEngComm*, **17**, 5154–5165.
Parr, R. G. & Weitao, Y. (1995). *Density-Functional Theory of Atoms and Molecules.* Oxford University Press.
Parrinello, M. & Rahman, A. (1981). *J. Appl. Phys.* **52**, 7182–7190.
Price, S. L. (2004). *Adv. Drug Deliv. Rev.* **56**, 301–319.
Price, S. L. (2013). *Acta Cryst.* B**69**, 313–328.
Price, S. L. (2018). *Faraday Discuss.* **211**, 9–30.
Reilly, A. M., Cooper, R. I., Adjiman, C. S., Bhattacharya, S., Boese, A. D., Brandenburg, J. G., Bygrave, P. J., Bylsma, R., Campbell, J. E., Car, R., Case, D. H., Chadha, R., Cole, J. C., Cosburn, K., Cuppen, H. M., Curtis, F., Day, G. M., DiStasio, R. A. Jr, Dzyabchenko, A., van Eijck, B. P., Elking, D. M., van den Ende, J. A., Facelli, J. C., Ferraro, M. B., Fusti-Molnar, L., Gatsiou, C.-A., Gee, T. S., de Gelder, R., Ghiringhelli, L. M., Goto, H., Grimme, S., Guo, R., Hofmann, D. W. M., Hoja, J., Hylton, R. K., Iuzzolino, L., Jankiewicz, W., de Jong, D. T., Kendrick, J., de Klerk, N. J. J., Ko, H.-Y., Kuleshova, L. N., Li, X., Lohani, S., Leusen, F. J. J., Lund, A. M., Lv, J., Ma, Y., Marom, N., Masunov, A. E., McCabe, P., McMahon, D. P., Meekes, H., Metz, M. P., Misquitta, A. J., Mohamed, S., Monserrat, B., Needs, R. J., Neumann, M. A., Nyman, J., Obata, S., Oberhofer, H., Oganov, A. R., Orendt, A. M., Pagola, G. I., Pantelides, C. C., Pickard, C. J., Podeszwa, R., Price, L. S., Price, S. L., Pulido, A., Read, M. G., Reuter, K., Schneider, E., Schober, C., Shields, G. P., Singh, P., Sugden, I. J., Szalewicz, K., Taylor, C. R., Tkatchenko, A., Tuckerman, M. E., Vacarro, F., Vasileiadis, M., Vazquez-Mayagoitia, A., Vogt, L., Wang, Y., Watson, R. E., de Wijs, G. A., Yang, J., Zhu, Q. & Groom, C. R. (2016). *Acta Cryst.* B**72**, 439–459.
Reilly, A. M. & Tkatchenko, A. (2013). *J. Chem. Phys.* **139**, 024705.
Reilly, A. M. & Tkatchenko, A. (2015). *Chem. Sci.* **6**, 3289–3301.
Rossi, M., Gasparotto, P. & Ceriotti, M. (2016). *Phys. Rev. Lett.* **117**, 115702.
Rosso, L., Mináry, P., Zhu, Z. & Tuckerman, M. E. (2002). *J. Chem. Phys.* **116**, 4389–4402.
Schneider, E., Vogt, L. & Tuckerman, M. E. (2016). *Acta Cryst.* B**72**, 542–550.
Shtukenberg, A. G., Zhu, Q., Carter, D. J., Vogt, L., Hoja, J., Schneider, E., Song, H., Pokroy, B., Polishchuk, I., Tkatchenko, A., Oganov, A. R., Rohl, A. L., Tuckerman, M. E. & Kahr, B. (2017). *Chem. Sci.* **8**, 4926–4940.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. & Hassabis, D. (2016). *Nature*, **529**, 484–489.
Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K. & Hassabis, D. (2018). *Science*, **362**, 1140–1144.
Sobol, I. (1977). *USSR Comput. Math. Math. Phys.* **16**, 236–242.
Thirumalai, D., Mountain, R. D. & Kirkpatrick, T. R. (1989). *Phys. Rev. A*, **39**, 3563–3574.
Torrie, G. M. & Valleau, J. P. (1977). *J. Comput. Phys.* **23**, 187–199.
Tuckerman, M. E. (2010). *Statistical Mechanics: Theory and Molecular Simulation.* Oxford University Press.
van Eijck, B. P. & Kroon, J. (1999). *J. Comput. Chem.* **20**, 799–812.
Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. (2004). *J. Comput. Chem.* **25**, 1157–1174.
Wengert, S., Csányi, G., Reuter, K. & Margraf, J. T. (2021). *Chem. Sci.* **12**, 4536–4546.
Whitfield, T. W., Bu, L. & Straub, J. (2002). *Physica A*, **305**, 157–171.
Yang, S. & Day, G. (2021*a*). *ChemRxiv*, 10.26434/chemrxiv-2021-jh6cj.
Yang, S. & Day, G. M. (2021*b*). *J. Chem. Theory Comput.* **17**, 1988–1999.
Yu, T.-Q. & Tuckerman, M. E. (2011). *Phys. Rev. Lett.* **107**, 015701.
Zhu, Q., Sharma, V., Oganov, A. R. & Ramprasad, R. (2014). *J. Chem. Phys.* **141**, 154102.
Zwanzig, R. W. (1954). *J. Chem. Phys.* **22**, 1420–1426.