EDGE ARTICLE

Check for updates

All publication charges for this article have been paid for by the Royal Society of Chemistry

Navigating chemical space: multi-level Bayesian optimization with hierarchical coarse-graining

Luis J. Walter [ID] [a] and Tristan Bereau [ID] *[ab]

Molecular discovery within the vast chemical space remains a significant challenge due to the immense number of possible molecules and limited scalability of conventional screening methods. To approach chemical space exploration more effectively, we have developed an active learning-based method that uses transferable coarse-grained models to compress chemical space into varying levels of resolution. By using multiple representations of chemical space with different coarse-graining resolutions, we balance combinatorial complexity and chemical detail. To identify target compounds, we first transform the discrete molecular spaces into smooth latent representations. We then perform Bayesian optimization within these latent spaces, using molecular dynamics simulations to calculate target free energies of the coarse-grained compounds. This multi-level approach effectively balances exploration and exploitation at lower and higher resolutions, respectively. We demonstrate the effectiveness of our method by optimizing molecules to enhance phase separation in phospholipid bilayers. Our funnel-like strategy not only suggests optimal compounds but also provides insight into relevant neighborhoods in chemical space. We show how this neighborhood information from lower resolutions can guide the optimization at higher resolutions, thereby providing an efficient way to navigate large chemical spaces for free energy-based molecular optimization.

1 Introduction

All molecules consist of a limited set of atoms, but their diverse properties arise from the intricate arrangements of these atoms. The vast combinatorial possibilities of such arrangements define the so-called chemical space (CS).[1] Exploring this space to discover new molecules with desired properties is challenging due to its immense size and complexity.[2,3] Traditionally, experimental high-throughput screening is conducted on a small subset of molecular structures to identify candidates with the desired properties. However, this approach is costly and limited by the size of the molecular library.[4,5]

To address these challenges, computational methods have been employed to replace expensive experiments.[6] In particular, molecular dynamics (MD) simulations can be utilized to predict the behavior of molecules based on their structure and empirical force fields.[7–9] Combined with automated, high-throughput setups, they enable the screening of large numbers of molecules.[10] While such simulations can reduce the cost of evaluating molecules for their target properties, they do not inherently facilitate navigation of the vast chemical search space.

Active learning methods—particularly Bayesian optimization (BO)—offer an efficient way to identify promising molecules from the extensive candidate pool. These methods optimize functions where gradient-based approaches are inapplicable.[11,12] As molecular structure–property relationships generally lack gradient information, BO offers a more efficient alternative to uniform or random sampling of molecular space.[13–15] Since BO relies on a covariance function over the input space, a numerical representation of discrete CS is typically used to quantify molecular similarity. For example, autoencoder models can encode molecules into latent representations.[16–18] In contrast to fingerprint methods,[19–22] they do not require a manual feature selection. Although BO helps select promising candidates, it does not reduce the complexity of CS.

Coarse-graining—grouping atoms into pseudo-particles or beads—addresses this complexity by effectively compressing CS. While traditionally employed to accelerate MD simulations, mapping atoms to beads reduces information and results in many-to-one relationships between atomistic and coarse-grained (CG) structures.[9,23] The collective properties of the underlying chemical fragments determine the interactions between the CG beads. Discretizing these interactions enables the use of a transferable CG force field, *i.e.*, a fixed set of interaction or bead types that can be reused across the entire CS.[24] The interaction resolution of such transferable force fields, determined by the number of available CG bead types,

*a*Institute for Theoretical Physics, Heidelberg University, Philosophenweg 19, 69120 Heidelberg, Germany. E-mail: bereau@uni-heidelberg.de

*b*Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany

directly impacts the many-to-one relationship between atomistic and CG structures and therefore the combinatorial complexity of CG CS.[25] Lower-resolution CS representations with fewer available bead types are easier to explore, but the resulting molecular structures lack detailed information.[26] Higher resolutions provide more detailed results, but their CS representations are more challenging to explore. This raises the question of how to combine different coarse-graining resolutions to efficiently explore CS while obtaining detailed molecular results.

In this work, we propose a multi-level BO framework for an efficient exploration of small molecule CS across multiple CG force-field resolutions. Our method combines the reduced complexity of CS exploration at lower resolutions with a detailed optimization at higher resolutions. The Bayesian approach provides an intuitive way to combine information from different resolutions into the optimization. Our method builds upon the work of Mohr *et al.*, who applied BO in a single, relatively low-resolution CG representation of CS to derive molecular design rules.[26] They also conducted optimization in a learned representation of an enumerated CG CS. We build on their approach by integrating multiple CG resolutions into a unified optimization framework.

Our multi-level BO is related to previous multi-fidelity BO efforts,[27–30] which rely on different evaluation costs and accuracies for each fidelity. In contrast, we assume a constant evaluation cost at all levels and instead utilize the varying complexity of our different CG resolutions.

Compared to recently popular generative methods for inverse molecular design,[31] our multi-level BO framework is data efficient and requires no prior training data for the optimization target.

As a demonstration of our method, we optimize a small molecule to promote phase separation in a ternary lipid bilayer. Previous studies[32,33] have shown that molecules embedded within lipid bilayers can modulate their phase behavior. We quantify this phase separation behavior as a free-energy difference, which serves as the objective function for our molecular optimization. We demonstrate that our multi-level BO algorithm effectively identifies relevant chemical neighborhoods and outperforms standard BO applied at a single resolution level. Our proposed approach is versatile and applicable to a broad range of small-molecule optimization tasks where the target property can be expressed as a free-energy difference.

2 Methods

2.1 Overview

We begin by providing an overview of our computational screening methodology. First, we defined multiple CG models with varying resolutions, all using the same atom-to-bead mapping but differing in the assignment of transferable bead types. Higher-resolution models featured more bead types, capturing finer chemical details while still reducing the combinatorial complexity of CS compared to the atomistic level (Fig. 1a). This reduction allowed us to enumerate all possible CG molecules corresponding to a specific region of CS at each resolution. Due to the hierarchical model design, higher-resolution molecules could be systematically mapped to lower resolutions (Fig. 1b).

For the next step of our molecule optimization, we embedded the CG structures into a continuous latent space using a graph neural network (GNN)-based autoencoder, with each resolution encoded separately. This encoding step provided a smooth representation of CS, ensuring a meaningful similarity measure necessary for the subsequent BO.

Finally, a multi-level Bayesian optimization was performed based on all previously encoded CS resolutions. The ground truth values, *i.e.*, the optimization targets, were obtained from MD simulation-based free-energy calculations (Fig. 1c). In our example application, such a free-energy estimate characterized the phase separation behavior of a molecule inserted into a ternary lipid bilayer. The following sections describe each of the molecular discovery steps in detail.

2.2 Multi-resolution coarse-graining of CS

Coarse-graining of molecules generally consists of two steps. First, groups of atoms are mapped to pseudo-particles or beads. Second, the interactions between these beads are defined based on their underlying atomistic fragments. For both steps, the resolution of the coarse-graining can be varied. Assigning larger groups of atoms to single beads results in a lower CG resolution for the mapping step. Interactions between beads can be defined for each bead pair[34,35] or discretized into a limited number of transferable bead types. The number of available bead types then defines the interaction resolution. Various CG models with different approaches to the mapping, discretization, and assignment of bead types exist.[36,37]

Since coarse-graining reduces information, a single CG molecule corresponds to multiple atomistic conformations or chemical compositions. The CG resolution determines how many atomistic structures correspond to a single CG molecule. Representing CS at a lower CG resolution results in fewer combinatorial possibilities for molecules and therefore a smaller CS.[25]

We started the molecule discovery process by directly defining small molecule CS at the high-resolution CG level. To do this, we specified the set of available CG bead types based on the relevant elements and chemical fragments from atomistic CS (Fig. 1a). We used three CG resolution levels for our application. They shared the same mapping of atoms to beads, but differed in the number of available bead types. Our high-resolution model corresponded to the Martini3 model,[24] a versatile CG force field with demonstrated relevance to materials design.[26,38,39] For our model, we ignored Martini3 bead labels, *e.g.*, for hydrogen bonding or polarizability. Further excluding water and divalent ions resulted in a model with 32 bead types per bead size, or 96 bead types in total. The relationship between bead types at different resolutions was hierarchical, meaning that higher-resolution bead types could be uniquely mapped to lower resolutions. In practice, lower-resolution bead types were obtained by averaging the interactions of higher-resolution bead types. For the medium- and low-
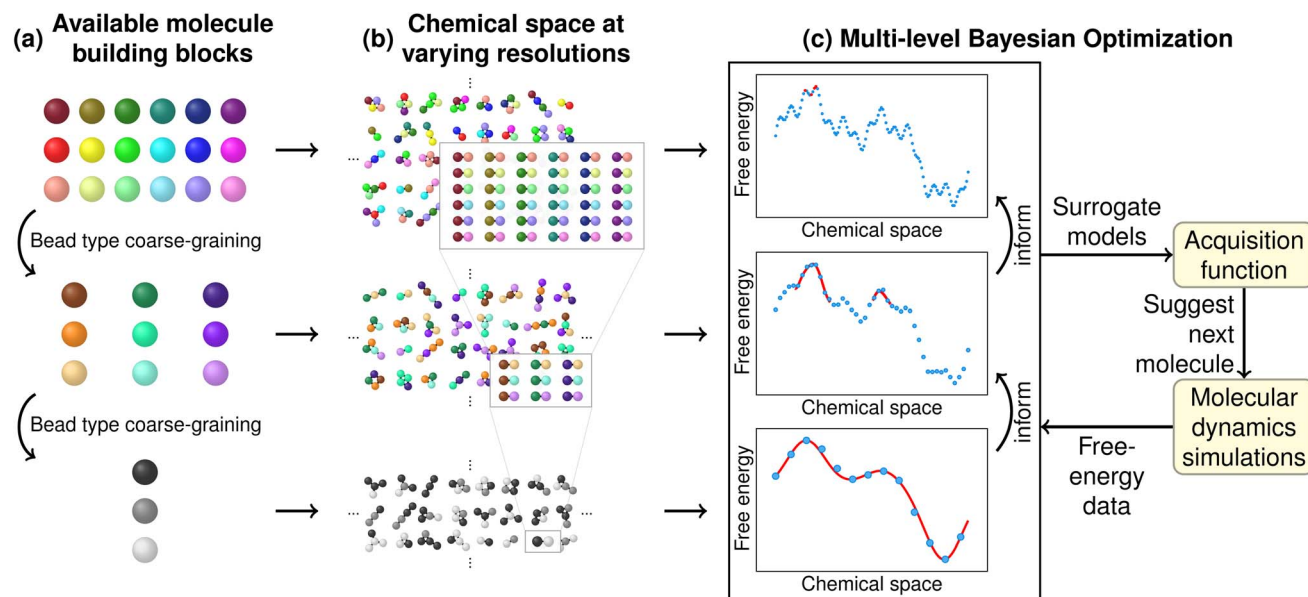
Fig. 1 Overview of our multi-resolution coarse-graining molecule optimization workflow. (a) Definition of multiple coarse-grained (CG) models at varying resolutions. These models share the same atom-to-bead mapping but differ in bead-type assignments, with higher resolutions featuring more bead types to capture finer chemical details. (b) Enumeration of chemical space (CS) at different resolution levels. Higher-resolution molecules can be hierarchically mapped to lower resolutions. (c) Multi-level Bayesian optimization integrating information from all CS resolutions. Molecules are iteratively suggested by an acquisition function and evaluated through molecular dynamics (MD) simulations. The optimization progressively shifts toward higher-resolution evaluations. Optimization at higher-resolution levels is guided by surrogate models at lower resolutions, improving efficiency and accelerating the search for optimal candidates.

resolution models, we derived 45 and 15 bead types, respectively. Section S1.1 of the SI provides further details on the derivation of lower-resolution models.

For all resolutions, we enumerated all possible CG molecules based on the available bead types and the defined molecule size limit of up to four CG beads (Fig. 1b). By directly generating molecules at the CG level, the atomistic resolution was bypassed. Since we assumed bead size-dependent but constant bond lengths and no angle or dihedral interactions, the enumeration of molecules is equivalent to the enumeration of graphs. The small molecule size justified the neglected angle and dihedral interactions. For the three levels of resolution, we obtained chemical spaces of approximately 90 000, 6.7 million, and 137 million molecules, respectively. Section S1.2 of the SI elaborates details on the graph enumeration.

2.3 Chemical space encoding

From the enumeration step, we obtained large sets of molecular graphs. While direct optimization in graph space is possible (*e.g.*, *via* evolutionary algorithms[40–42]), a numerical representation facilitates exploration of CS by enabling distance-based similarity measures. Molecular fingerprints are often used for this purpose[19–22] but require manual feature selection. Instead, we used a learned projection of CS into a low-dimensional, smooth numerical representation.

For the learned encoding, we used a regularized autoencoder (RAE),[43] which offers deterministic behavior compared to the more common variational autoencoder (VAE) architecture.[16,44]

As we only aimed for a smooth embedding, the stochasticity of a VAE was not needed. The built-in regularization of the RAE ensured a well-structured latent space.[43] We used a GNN for the node-permutation invariant encoder,[45,46] which mapped molecular graphs to the five-dimensional latent space. A decoder, composed of fully connected layers, was used to reconstruct node features and the adjacency matrix. Although the decoder was not invariant to node permutations, the reconstruction loss ensured an invariant training of the RAE.

Input and reconstruction node features included bead-type class, size, charge, and octanol–water partition coefficient. The latter was added as a continuous feature to improve latent space structure.

We trained separate RAEs for each CG resolution using the complete set of enumerated molecules. The separated training resulted in lower reconstruction losses and better adaptation to the reduced resolution at lower levels. The loss combined cross-entropy terms for categorical features, a binary cross-entropy for the adjacency matrix, and a mean squared error term for the octanol–water partition coefficient. After training, we retained only the encoder for embedding molecules. The RAE was implemented using the PyTorch and PyTorch Geometric libraries,[47,48] following the architecture of Mohr *et al.*[26] Further details on the RAE architecture, the training, and an analysis of the learned latent space are provided in Section S1.3 and S2.1 of the SI. In the following steps, we performed BO in these learned latent spaces.

2.4 Single-level Bayesian optimization

Before introducing our multi-level BO approach, we first provide an overview of standard BO and our notation (see, *e.g.*, Frazier[12] for a more detailed introduction). We then describe how we extend this approach to combine multiple resolution levels into a single optimization process. BO aims to optimize a black-box function $f : \mathcal{X} \to \mathbb{R}$ that is expensive to evaluate and has no analytical form or gradient information available. The objective is to find the global optimum $x^* = \mathrm{argmin}_{x \in \mathcal{X}} f(x)$ or $x^* = \mathrm{argmax}_{x \in \mathcal{X}} f(x)$ with as few function evaluations as possible. Typically, a Gaussian process (GP) is used as a probabilistic model for $f(x)$, *i.e.*, $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$, defining a multivariate normal distribution with mean function $m(x)$ and a covariance function $k(x, x')$. This covariance kernel quantifies correlations over $\mathcal{X}$. Although various kernel functions exist, a common choice is the radial basis function (RBF) kernel, defined as

$$k\left(x, x'\right) = \exp\left(-\frac{1}{2\xi^2}\|x - x'\|^2\right), \tag{1}$$

where ξ is the length scale parameter. Given training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ with inputs $X = \{x_1, \dots, x_n\}$ and observations $Y = \{y_1, \dots, y_n\}$, the posterior GP provides a predictive mean $\mu(x)$ and variance $\sigma^2(x)$ for any $x \in \mathcal{X}$. The mean and variance are given by

$$\mu(x) = m(x) + k(x, X)K^{-1}(Y - m(X)), \tag{2}$$

$$\sigma^2(x) = k(x, x) - k(x, X)K^{-1}k(X, x), \tag{3}$$

where $K = k(X, X) + \sigma_n\mathbb{I}$ is the covariance matrix of X with an added noise term σ_n.

In BO, the GP model is iteratively updated with new evaluations of the target function. First, the function is evaluated at a set of initialization points. Subsequent evaluations are selected based on the predictive mean and variance of the GP, guided by an acquisition function that balances exploration and exploitation. A common choice for the acquisition function is the expected improvement (EI),[49] which for minimization is defined as

$$\mathrm{EI}(x) = \mathbb{E}_{z \sim \mathcal{N}(\mu(x), \sigma^2(x))}[\max(y^* - z, 0)] \tag{4}$$

with $y^* = \min_{y \in Y} y$. The next evaluation point is determined by $x_{n+1} = \mathrm{argmax}_{x \in \mathcal{X}} \mathrm{EI}(x)$. This process repeats until the evaluation budget is exhausted or a sufficiently good solution is found.

2.5 Multi-level Bayesian optimization

For our multi-level BO approach, we considered $d = 3$ CG resolution levels of CS. At each level $l \in \{1, \dots, d\}$, we defined the mapping of chemical space $\mathcal{X}_l$ to the target free-energy difference y as an unknown function $f_l(x)$. Our goal was to identify molecules at the highest resolution d that are near the optimum, *i.e.*, $x^* = \mathrm{argmin}_{x \in \mathcal{X}_d} f_d(x)$, while leveraging information from the lower-resolution models ($l < d$). Similar to the work of Huang *et al.*, we assumed that each function $f_l(x)$ can be modeled as a correction to the lower resolution

$$f_l(x) = f_{l-1}(x) + \delta_l(x), \tag{5}$$

where $\delta_l(x)$ represents the correction term.[27] The hierarchical bead-type resolutions justified this delta learning approach. We modeled each $\delta_l(x)$ as a GP, *i.e.*,

$$\delta_l(x) \sim \mathcal{GP}\left(0, k_l\left(x, x'\right)\right). \tag{6}$$

with a mean function equal to zero for all x. For all levels, we used an RBF kernel function (eqn (1)) with level-specific length scale parameters ξ_l. By definition of the GP (see eqn (2) and (3)), this delta learning approach corresponds to a GP with a mean prior $m(x)$ equal to the next-lower resolution function $f_{l-1}(x)$. Thus, we can rewrite the GP for $f_l(x)$ as

$$f_l(x) \sim \mathcal{GP}\left(f_{l-1}(x), k_l\left(x, x'\right)\right). \tag{7}$$

At the lowest resolution $l = 1$, no lower-level prior was available. Instead of using a zero prior for $f_1(x)$, we applied a simple model $f_0(x)$ that approximates the free-energy difference of a molecule as the sum of the individual bead free energies.

Until now, we assumed the latent spaces of the different resolutions to be compatible. However, since they were obtained from separate autoencoder trainings, we could not directly use a lower level function $f_l(x)$ as the prior for the GP on level l. Instead, a function $\mathcal{M}_l(x)$ was required that maps points in latent space $\mathcal{X}_l$ to points in latent space $\mathcal{X}_{l-1}$. We determined this mapping from one resolution to a lower one from the known relationships between molecules at different resolutions. Effectively, we had a many-to-one mapping from $\mathcal{X}_l$ to $\mathcal{X}_{l-1}$, which made the mapping $\mathcal{M}_l(x)$ an unambiguous function. Applying this mapping to eqn (7), we get

$$f_l(x) \sim \mathcal{GP}\left(f_{l-1}(\mathcal{M}_l(x)), k_l\left(x, x'\right)\right). \tag{8}$$

as the final probabilistic model for resolution l.

The optimization procedure started at the lowest-resolution level $l = 1$, with initialization molecules selected through weighted k-medoid clustering of the encoded CS. The clustering weights were based on the prior of the lowest resolution and calculated as $w_i = \exp(-f_0(x_i))$.

The length scale parameters ξ_l of the RBF kernels were optimized for each level using the GP marginal likelihood. The kernel noise term σ_n from K in eqn (2) and (3) was fixed to the standard deviation of the calculated free-energy differences. This standard deviation was determined by multiple repeated evaluations of the same molecule (see Section S2.3 of the SI). The multi-level BO implementation used the GPyTorch library.[50]

Although BO is also possible with a batched evaluation of multiple points,[26] we only evaluated one point, *i.e.*, one molecule, at a time. Since each evaluation involved multiple MD simulations, we parallelized over these simulations. We used the EI as the acquisition function on each level. For higher levels $l > 1$, the EI was computed and maximized only over CS regions with expected significant negative free-energy differences. These regions were defined as the neighborhoods of points with promising prior information from the lower level. Restricting the EI calculation to these neighborhoods focused the optimization on the most relevant CS regions and accelerated the EI maximization process. Details regarding the

mapping of points between latent spaces and the calculation of neighborhoods are provided in Section S1.4 of the SI.

Our multi-level BO algorithm transitions to a higher resolution when the prediction error of the GP remains below a predefined threshold for multiple consecutive evaluations. This prediction error serves as a measure of the GP model's convergence. For our application, we empirically set the prediction error threshold to 0.12 kcal mol^{-1} and required three consecutive evaluations below this threshold to trigger the switch. These hyperparameters control the trade-off between exploration at lower resolutions and faster exploitation of promising regions at higher resolutions. Lowering the threshold and increasing the number of required evaluations enhances exploration at lower resolutions, but increases the total number of molecule evaluations needed.

In addition to increasing the resolution level, the algorithm can switch back to the previous lower resolution. Since we want to effectively leverage lower-resolution models, we are only interested in high-resolution evaluations in regions where a reliable prior is available. If the candidate with the maximal EI is too far away from regions with a reliable prior from lower levels, we switch back to the previous resolution level. Specifically, the criterion for switching to resolution level $l - 1$ is defined as $\|x^* - x'\| > 2\xi_l, \ \forall x' \in \{x \in \mathcal{X}_l | x \in X_l \vee \mathcal{M}_l(x) \in X_{l-1}\}$, where X_l denotes the set of already evaluated points at level l.

2.6 Estimating the membrane demixing behavior

For our application, we optimized small molecules to enhance phase separation in a ternary lipid bilayer consisting of 1,2-dipalmitoyl-*sn*-glycero-3-phosphocholine (DPPC), 1,2-dilinoleoyl-*sn*-glycero-3-phosphocholine (DLiPC), and cholesterol (Fig. 2). DPPC and DLiPC differ only in their acyl chains, with DPPC having two saturated 16-carbon chains and DLiPC having two doubly unsaturated 18-carbon chains. The phase separation can be quantified by the DPPC–DLiPC contact fraction.[32] However, directly observing the effect of a molecule on lipid mixing requires long simulations with large bilayer leaflets, which is impractical for high-throughput screening. Alternatively, potential of mean force (PMF) profiles along the axis

perpendicular to the bilayer plane can be compared for pure DPPC, DLiPC, and ternary bilayers.[33] Since PMF calculations (*e.g.*, *via* umbrella sampling[51]) are still computationally expensive, we employed thermodynamic integration (TI)[52,53] calculations at a few key positions in the bilayers as a proxy. Centi *et al.*[33] showed that molecules that influence the demixing or mixing of a DPPC–DLiPC bilayer localize near the bilayer center because the two phospholipids differ only in their carbon tails. To determine a molecule's preferred localization, we performed TI computations at the center ($z = 0$ nm) of the ternary bilayer, at the interface ($z = 1.5$ nm) and in bulk water (Fig. 3), obtaining the free-energies ΔG_{center}, $\Delta G_{interface}$, and ΔG_{water}, respectively. We initially used ΔG_{center} and ΔG_{water} to identify non-inserting molecules, allowing us to skip further free-energy evaluations for these cases. Centi *et al.*[33] showed that molecules that enhance the phospholipid demixing localize near the DLiPC phase. Therefore, we performed a fourth TI calculation at the center of a pure DLiPC bilayer, yielding ΔG_{DLiPC}. Unlike the direct observation of DPPC–DLiPC contacts, ΔG-based scoring was easily parallelized, thereby further reducing the wall time per evaluated molecule. The main optimization target was the free-energy difference, $\Delta\Delta G = \Delta G_{center} - \Delta G_{DLiPC}$, assuming the molecule localizes near the ternary bilayer center. To ensure robust optimization even when molecules localize at the interface or in the water, we combined $\Delta\Delta G$ with a score S defined as a conditional weighted sum of $\Delta G_{water} - \Delta G_{interface}$ and $\Delta G_{interface} - \Delta G_{center}$. In contrast to using a constant $\Delta\Delta G$ for interface- or water-localizing molecules, this score provided a more nuanced direction for optimization, steering it toward relevant regions of CS. Negative $\Delta\Delta G$ values indicated a preference for the DLiPC phase, corresponding to a demixing behavior. Overall, the molecule optimization corresponded to a minimization of $\min(\Delta\Delta G, 0) + S$. Section S1.5 of the SI provides further information on the calculation of the score S.
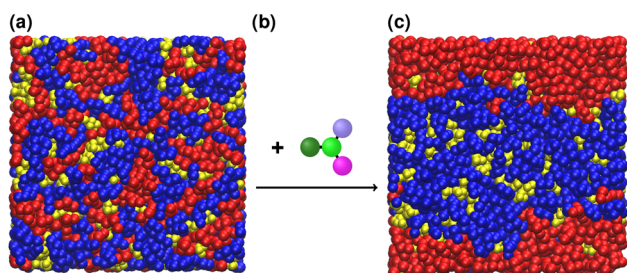
Fig. 2 Influencing phase separation in a lipid bilayer by inserting small molecules. Shown is a top view of a CG ternary lipid bilayer composed of DPPC (blue), DLiPC (red), and cholesterol (yellow). (a) In the mixed state, the bilayer contains small, dispersed lipid patches. (b) Upon inserting specific small molecules, (c) the bilayer transitions to a demixed state with pronounced phase separation between the two phospholipids.
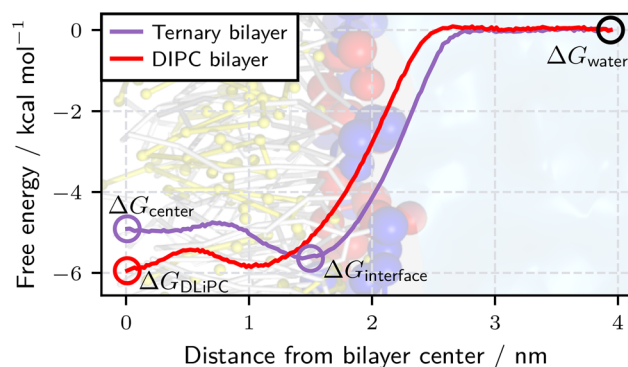
Fig. 3 Estimating the demixing behavior of molecules *via* free-energy calculations at four bilayer depths (circles) as an alternative to potential of mean force (PMF) computations (solid lines). For the molecule optimization, we aim to minimize $\Delta\Delta G = \Delta G_{center} - \Delta G_{DLiPC}$ under the conditions that $\Delta G_{center} < \Delta G_{interface}$ and $\Delta G_{center} < \Delta G_{water}$. The background illustrates the hydrophobic tails (grey), the charged headgroups of DPPC (blue) and DLiPC (red), as well as cholesterol (yellow). The plotted PMFs correspond to a molecule with $\Delta G_{center} > \Delta G_{interface}$, indicating that it localizes at the bilayer interface and therefore does not significantly influence lipid mixing.

2.7 Molecular dynamics simulations

We used MD simulations in a high-throughput manner[54,55] to perform the TI calculations of the free-energy differences. All MD simulations were performed using GROMACS 2024.2.[56,57] Martini3 and Martini3-derived (see Section 2.2) force fields were used for the CG simulations.[24,58] The derived lower-resolution bead types are compatible with the standard Martini3 bead types and can therefore be evaluated within unmodified Martini3 environments.

Our lipid bilayer simulation setup was based on the protocol by Ozturk et al.[59] We used a leap-frog stochastic dynamics integrator with an integration time step of 20 fs (in reduced CG units). All simulations were performed in the NPT ensemble at a temperature of 305 K and pressure of 1 bar,[33] controlled by a semi-isotropic C-rescale barostat.[60] For the TI, we used 26 linearly-spaced λ steps for the decoupling of Lennard-Jones interactions and 10 additional linear λ steps for the decoupling of Coulomb interactions in the case of charged molecules. Since each molecule evaluation required up to four TI calculations, each with up to 36 λ steps, evaluating a single molecule could require up to 144 individual simulations. Further simulation parameters are provided in Section S1.6 of the SI. The package MBAR[61,62] was used to calculate free-energy differences from the MD simulation data.

Membrane systems were generated using the program insane.[63] Following the approach of Centi et al., we used a lipid composition of DPPC : DLiPC : cholesterol in a 7 : 4.7 : 5 ratio.[33] For a bilayer area of 6×6 nm^2, used for the free-energy evaluations, this corresponded to 26 DPPC, 18 DLiPC, and 19 cholesterol molecules per bilayer leaflet. We used the colvars module[64] in GROMACS to calculate or restrain the phospholipid contact fraction. Specifically, the collective variable was defined as the coordination number between the first C1 beads of DLiPC and DPPC with a cutoff distance of 1.1 nm.[33] During the TI simulations, the coordination number was restrained to 65 contacts per leaflet, yielding an average of 2.5 DLiPC molecules within the cutoff per DPPC. This slightly exceeds the 2.15 contacts expected from random lipid placement by insane.[63]

3 Results and discussion

3.1 Multi-level Bayesian optimization

We applied our multi-level BO workflow to identify small molecules that enhance the phase separation of a ternary lipid bilayer, demonstrating its effectiveness in navigating chemical space. We restricted the search to small molecules with up to 16 heavy atoms, corresponding to a maximum of four beads in our CG model. We imposed no additional constraints, such as the presence of specific functional groups, to rigorously test our method. Our multi-level molecule optimization utilized three coarse-graining resolutions, incorporating 15, 45, and 96 distinct bead types. While all three levels use the same spatial coarse-graining, complexity increased with the combinatorial diversity of bead types, spanning approximately 90 000, 6.7 million, and 137 million possible CG molecules. To identify phase separation-enhancing molecules at the highest

resolution, we used lower-resolution models only to guide the search, thereby reducing the complexity of the optimization compared to direct high-resolution exploration. At all levels, a molecule's effect on phase separation was quantified by an MD simulation-derived free-energy difference, $\Delta\Delta G$ (see Section 2.6).

The optimization was conducted within RAE-learned latent embedding spaces, generated from the CG models at each resolution. As a first step, we computed the $\Delta\Delta G$ values for all 15 low-resolution bead types. These results enabled us to construct a cost-effective prior for the low-resolution model, based on an additivity assumption over individual bead values (see Section S2.1 of the SI for a detailed evaluation of this assumption). Using this prior, we initialized the multi-level active learning with 50 low-resolution molecules. Subsequent molecules and their resolution levels were determined iteratively by our multi-level BO algorithm. We evaluated 327 molecules in total: 106 molecules (15 + 50 + 41) at the low resolution, 148 at the medium resolution, and 73 at the high resolution. In each iteration, a single molecule was selected for evaluation using MD simulations. The resulting $\Delta\Delta G$ value was then used to update the BO model, which informed the selection of the following molecule.

Our multi-level BO approach progressively narrows the search space through the three resolution levels. The optimization begins with a broad exploration of low-resolution CS, identifying coarse regions likely to contain molecules with favorable $\Delta\Delta G$ values. Insights from this stage inform the medium-resolution search, allowing the algorithm to focus on more promising sub-regions. This process is further refined at the high-resolution level to pinpoint localized areas within CS that are most likely to yield effective candidates. By leveraging information from the preceding levels, the algorithm bypasses large areas of the CS landscape that are unlikely to yield relevant molecules. Therefore, the number of required evaluations and the overall computational cost are reduced. Fig. 4 presents 2D projections of the encoded CS (black) together with the evaluated molecules. Because each resolution is encoded independently, the representations differ and prevent a direct transfer of points. However, molecules can be readily mapped across latent spaces by leveraging the known mapping between bead types. The figure illustrates the funnel-like optimization: as resolution increases, the search becomes more focused, eventually concentrating on localized sub-regions of chemical space. Many low-resolution candidates display unfavorable $\Delta\Delta G$ values or negligible effects on phase separation (yellow). In contrast, searches at medium and high resolutions increasingly yield molecules with lower $\Delta\Delta G$ values corresponding to a more substantial impact on lipid demixing (orange to red). Fig. 5 further illustrates this trend, showing the distribution of evaluated $\Delta\Delta G$ values across the three resolution levels, including the initialization points at resolution $l = 1$. Candidates from the low-resolution optimization already show lower $\Delta\Delta G$ values relative to the initialization set. However, higher-resolution candidates generally exhibited even stronger phase-separation effects, with medium resolution peaking around -1 kcal mol^{-1} and high resolution around -1.2 kcal mol^{-1}. The
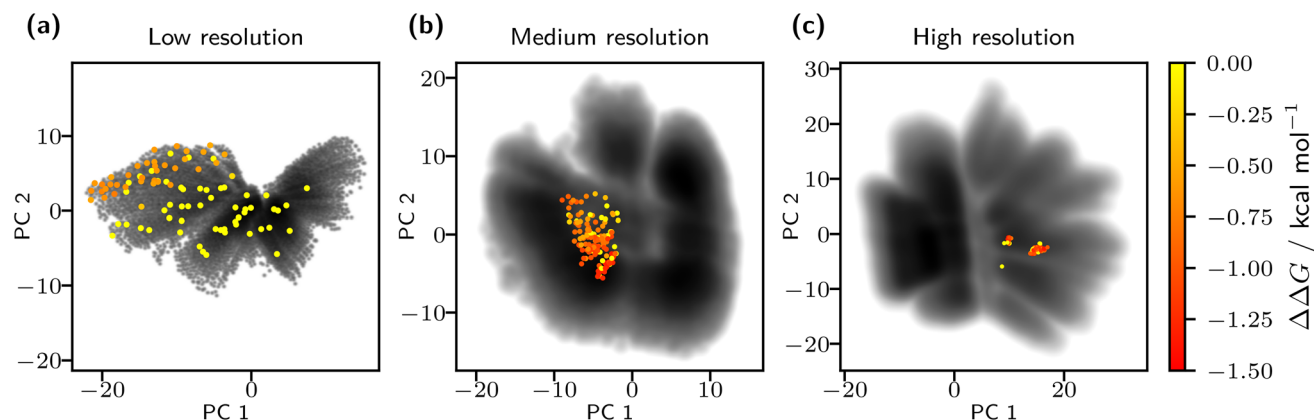
Fig. 4 Encoded chemical spaces and evaluated points for the three levels of resolution. The full chemical spaces are shown as kernel-density estimations of latent space principal component analysis (PCA) projections (black). Evaluated molecules across the three resolutions are overlaid as colored points (yellow to red), where lower $\Delta\Delta G$ values indicate stronger lipid bilayer demixing. Due to separate encodings at each resolution, latent space points are not directly transferable. (a) Optimization proceeds from broad, low-resolution exploration to (b) progressively focused searches in medium and (c) high resolutions.
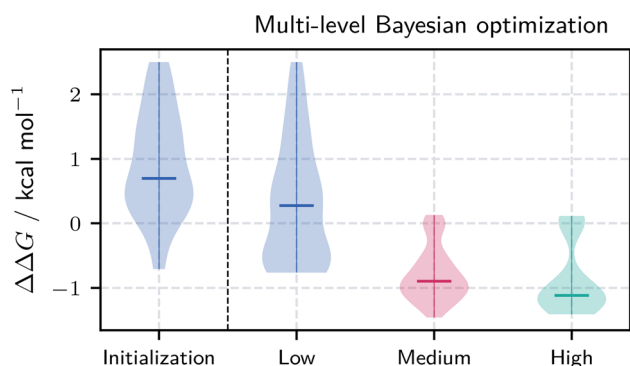
Fig. 5 Distribution of $\Delta\Delta G$ values for all evaluated candidates at different stages of the multi-level Bayesian optimization process. Violin plots show the distributions for the initialization set and candidates evaluated at low, medium, and high-resolution levels. As the optimization progresses to higher resolutions, the distribution of $\Delta\Delta G$ values progressively shifts toward lower (more favorable) values. Horizontal bars indicate the median of each distribution.

promising candidates with pronounced effects on lipid phase separation despite limited evaluations.

3.2 Evaluation of optimized molecules

Following the overall optimization process analysis, we now focus on the top candidate molecules with the lowest $\Delta\Delta G$ values. As the Martini3 CG model (without bead labels)[24] corresponds to our high-resolution model, the optimized molecules do not provide atomistic details but reveal valuable insights into the chemical moieties driving the phospholipid phase separation. The top eight CG molecules, shown in Fig. 6, all display $\Delta\Delta G$ values below -1.3 kcal mol^{-1}, with the best candidate at -1.4 kcal mol^{-1} (top left of the figure). These results indicate a strong effect on the phase separation. A consistent feature across all eight CG molecules is the exclusive
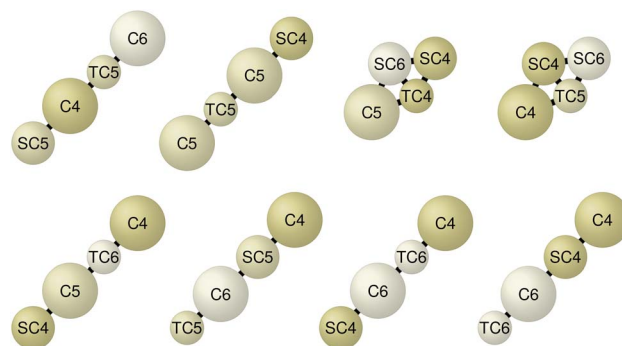
differences between the low- and medium-resolution minima support our hypothesis about the varying smoothness of the free-energy landscape across resolutions.

The computational cost per simulation is the same across all three resolutions. Consequently, the overall computational load at each level is primarily determined by the number of evaluated molecules. For non-inserting molecules, two of the four TI calculations can be omitted (see Section 2.6). As the lowest resolution filtered out most non-inserting molecules, its average computational load per evaluation was slightly lower than at higher resolutions.

We terminated the optimization after 73 high-resolution evaluations, as no further improvement in $\Delta\Delta G$ was observed. The 327 evaluated molecules correspond to less than 3×10^{-4}% of the total high-resolution molecule space. While global optimality is not guaranteed, the workflow identified multiple

Fig. 6 CG structures of the best eight high-resolution molecule candidates identified in the optimization process. The molecules exhibit low free-energy values ($\Delta\Delta G$) below -1.3 kcal mol^{-1}, indicating a strong influence on phospholipid bilayer phase separation. All molecules are exclusively composed of hydrophobic C4, C5, and C6 beads (in different sizes, indicated by prefixes S/T), corresponding to Martini3 types for alkenes, aromatic rings, and sulfide groups, respectively. Six of the eight molecules have an extended/chain-like topology.

presence of hydrophobic C4, C5, and C6 beads in varying bead sizes. These Martini3 beads correspond to alkenes, aromatic rings, and thiol/sulfide groups, respectively.[24] This observation aligns with Barnoud *et al.*, who showed that aromatic groups promote demixing, while aliphatic groups (C1, C2, and C3 beads) favor phospholipid mixing.[32] The two distinct topologies shown in Fig. 6 correspond to the two prominent point clusters in the 2D projection of Fig. 4c. While each cluster contains molecules with a variety of topologies, the highest-scoring molecules within them are predominantly of the two topologies in Fig. 6.

The highest-performing molecules at both low and medium resolution (see Section S2.4 of the SI) exhibit more diverse topologies but share similar trends in bead-type composition. While the low-resolution results already provide preliminary chemical insights, more detailed information—such as the unfavorable contribution of C1, C2, and C3 beads—only becomes evident through the inclusion of higher-resolution models.

Directly measuring bilayer phase separation requires significant simulation time and is therefore computationally expensive. Instead, we estimated demixing effects from free-energy differences. To validate this approach and confirm that the identified candidates indeed promote phase separation, we perform 1600 ns MD simulations (in reduced CG units) of the best candidate (top left in Fig. 6) in a ternary lipid bilayer system. Using this method to evaluate the demixing effect required one to two orders of magnitude more wall time than the free energy-based scoring used for the optimization. As a reference, we employ benzene, previously identified by Barnoud *et al.*[32] as a potent driver of lipid bilayer phase separation. Following their protocol, we use a solute/lipid mass ratio of 4.8% (see Section S2.5 of the SI for composition details). Phase separation was quantified by tracking DLiPC and DPPC contacts over the simulation trajectory. Fig. 7 presents the evolution of these contacts throughout the simulation, with dashed lines indicating average values. Each trajectory's initial 400 ns were discarded to ensure equilibration. Additionally, a control simulation without any added solute was conducted. Compared to this bilayer without solutes, our best candidate substantially reduced DLiPC–DPPC contacts, indicating a pronounced effect on bilayer demixing. Our best candidate also outperforms benzene, producing a greater reduction in the number of contacts, suggesting a stronger influence on phospholipid phase separation.

To identify relevant molecular features and design rules from the set of optimized molecules, we applied LASSO regression analogous to Mohr *et al.*[26] Derived rules could subsequently inform the design of atomistic structures. We analyzed single-bead and bead-pair features across all molecules with $\Delta\Delta G < 0$, yielding 85 features. Higher-order features were not included due to the size of the dataset. Feature extraction and LASSO regression details are provided in Mohr *et al.*[26] The top ten most relevant molecular features, along with their regression coefficients, bootstrapped uncertainties, and frequencies of occurrence, are shown in Fig. 8. Consistent with our earlier analysis of the top eight molecules, the most influential features involve hydrophobic C4, C5, and C6 beads. Moreover, combinations of a regular-sized and tiny or small-sized bead (indicated by T or S) appear relevant. These derived features provide interpretable insights into the physical interaction mechanisms that drive bilayer phase separation. They can be used to design atomistic molecular structures that exhibit the same phase separation behavior.

3.3 Comparison with standard BO

Is multi-level BO computationally advantageous compared to BO using only the high-resolution model? To address this, we performed standard BO with the same number of initial points and total evaluations as in the multi-level case. While BO is typically benchmarked by averaging the cumulative best result across multiple runs to reduce initialization bias, this is computationally infeasible for our bilayer demixing system.
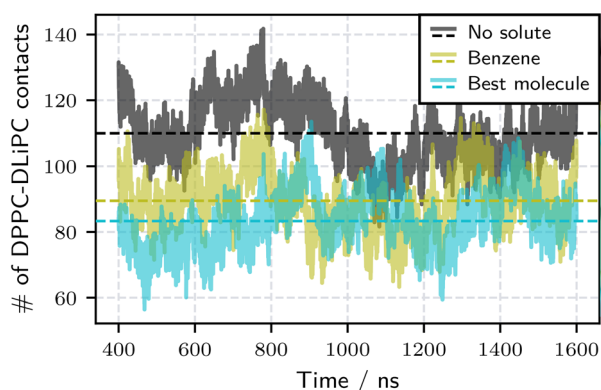
Fig. 7 Time evolution of DPPC−DLiPC lipid contacts in ternary bilayers over 1200 ns CG MD simulations (excluding 400 ns for equilibration). Three conditions are compared: a bilayer without solutes (black), a bilayer containing benzene as a known demixing agent (olive green), and one with the top-performing optimized molecule from Fig. 6 (cyan), each at a solute/lipid mass ratio of 4.8%. Dashed horizontal lines indicate mean contact numbers. The optimized molecule reduces DPPC−DLiPC contacts more than benzene, demonstrating a stronger phase-separation effect.
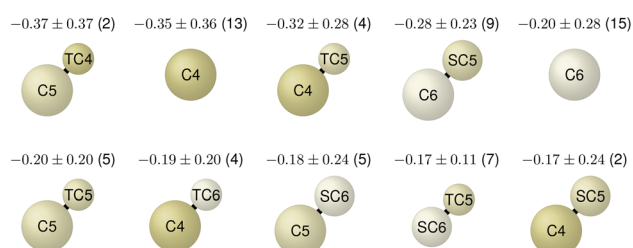
Fig. 8 Top ten most influential molecular features contributing to lipid bilayer phase separation, identified *via* LASSO regression of the optimized CG molecules with $\Delta\Delta G < 0$. Features were limited to single beads and bead pairs. Each panel displays a feature's structure along with its corresponding regression coefficient, bootstrapped uncertainty, and frequency of occurrence within the dataset (number in parentheses). Features only involve hydrophobic C4, C5, and C6 beads and pairs of differing bead sizes.

Instead, we compare the distributions of obtained $\Delta\Delta G$ values and the cumulative best result within single runs. We provide a toy model comparison of results averaged over multiple runs in Section S2.7 of the SI. Fig. 9 presents the progression of the best $\Delta\Delta G$ values for both optimization approaches. The diagram excludes the 50 initialization points and accounts for the 15 additional evaluations required to construct the low-resolution prior for the multi-level approach. The multi-level BO consistently outperforms the standard BO, achieving superior cumulative best values (solid lines) across all resolution levels. Additionally, the distribution (based on the best 50 molecules) and scatter plots in orange and green highlight that multi-level BO not only finds a better overall candidate, but multiple candidates with significantly lower $\Delta\Delta G$ values than the standard BO optimization. The peak of the multi-level BO distribution is shifted toward lower $\Delta\Delta G$ values compared to the standard BO optimization.

3.4 Chemical neighborhood sizes across resolutions

Our multi-level BO algorithm relies on the assumption that the free-energy landscape over the learned chemical representations is smoother at lower resolutions. To test this, we introduce the concept of chemical neighborhoods and analyze their sizes across different resolution levels. We define a chemical neighborhood as a region in chemical space containing similar molecules. Similarity implies that known properties about one molecule help predict properties of its neighbors. Here, neighborhood size is determined by the lengthscale ξ_l of an RBF kernel fitted in a GP regression. This length scale quantifies correlations between points in the latent space and is thus an intrinsic measure of chemical neighborhood size. To obtain the ξ_l, we fit independent GP models to the evaluated molecules at each resolution level, excluding lower-resolution priors to prevent bias. Neighborhood size is then calculated as the
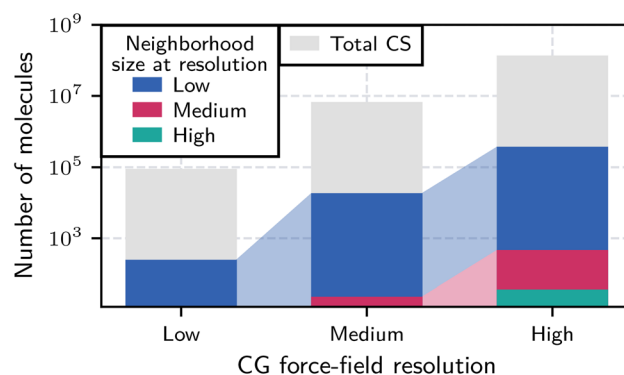

Fig. 10 Chemical neighborhood sizes across different CG resolutions of CS. The chart shows the number of molecules within a chemical neighborhood at low (blue), medium (pink), and high (teal) resolution, derived by fitting the lengthscale of a GP RBF kernel to the evaluated molecule data. The total number of molecules is shown in gray. Lower-resolution neighborhoods are mapped to higher resolutions by considering average molecule densities. These different neighborhood sizes reflect the varying smoothness of the free-energy landscape across different CS resolutions.

average number of neighbors within a distance d, where $d = \alpha\xi_l$ and $\alpha = 0.5$ determines the required similarity for a chemical neighborhood.

Fig. 10 shows relationships between the obtained neighborhood sizes, the total number of molecules in the chemical space, and neighborhoods from lower resolutions mapped to higher resolutions (exact numbers in Section S2.6 of the SI). Considering the logarithmic scale of the y-axis, we observe that neighborhood sizes span several orders of magnitude across the three resolutions. When mapped to medium or high resolution, low-resolution neighborhoods with about 249 molecules expand to about 18 600 and 378 000 molecules. Similarly, a medium-resolution neighborhood with about 23 molecules maps to a neighborhood of 468 molecules at high resolution. This exponential scaling suggests that prior information for many high-resolution molecules can be inferred from relatively few low-resolution evaluations. Section S2.8 of the SI further illustrates this by showing the coverage of the higher-resolution latent spaces by mapping evaluated molecules from the lower resolutions. These results support our assumption of a smoother free-energy landscape at lower resolutions.

4 Conclusions

This work introduces a multi-level Bayesian optimization (BO) framework for efficient exploration of chemical space (CS). Our method employs multiple levels of coarse-graining to exploit the varying smoothness of free-energy landscapes across different model resolutions. By informing the optimization process at higher resolutions with prior knowledge from lower resolutions, we accelerate the search for optimal molecules. Our BO-based algorithm combines information from multiple resolutions in a Bayesian manner, enabling a funnel-like optimization process through CS. This approach allows us to bypass irrelevant regions of CS at higher-resolution representations,
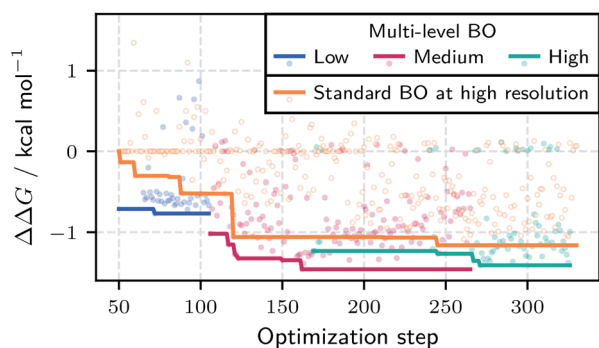

Fig. 9 Progression of the $\Delta\Delta G$ values during multi-level and standard BO runs. Multi-level BO uses evaluations at low (blue), medium (magenta), and high (green) resolutions, while the standard BO (orange) operates only at the high resolution. Solid lines show the current best value during the optimization. Initialization points are excluded. The multi-level case accounts for the 15 extra evaluations for the low-resolution prior. Kernel density estimates (right edge) reflect the distribution of best 50 high resolution candidates. Multi-level BO consistently achieves lower $\Delta\Delta G$ values, as indicated by the shifted distribution.

substantially reducing the number of required molecule evaluations and the overall computational cost. We demonstrate the effectiveness of our method by identifying small coarse-grained (CG) molecules that enhance phase separation in a ternary lipid bilayer. Despite evaluating only approximately 3×10^{-4}% of the total number of high-resolution molecules and assuming no prior knowledge of relevant CS regions, we successfully identified several candidates with a significant impact on lipid bilayer phase separation. Our multi-level approach outperforms standard BO, achieving a better overall best result and obtaining a significantly shifted distribution of evaluated molecules toward stronger effects on phase separation. The optimized CG molecules enable us to extract relevant molecular features and design rules. Our analysis of chemical neighborhood sizes at different resolutions confirms the assumption of smoother free-energy landscapes at lower resolutions. Notably, obtained neighborhood sizes vary by several orders of magnitude, allowing us to get prior information for many molecules at high resolution from a small number of evaluations at low resolution.

In this study, we limited our funnel optimization to the CG level and thus did not derive atomistic structures for the identified candidates. Similar to Mohr *et al.*, atomistic structures could be reconstructed based on the extracted molecular features.[26] Notably, these features provide an intuitive and interpretable summary of the key chemical factors, providing valuable insight into the underlying physical interaction mechanisms. Moreover, the atomistic resolution could be integrated directly into our multi-level optimization framework. Since each CG bead maps to 10^2 to 10^4 atomistic fragments,[54] the atomistic chemical space is vastly larger. Combined with evaluation costs two to three orders of magnitude higher,[65,66] this poses challenges. Nevertheless, these cost differences enable approaches like multi-fidelity BO,[27,30] and high-resolution CG results generally provide an efficient starting point that reduces the number of required atomistic evaluations.

A limitation of our multi-level BO method is its reliance on a hierarchical relationship between resolutions, with higher resolutions required to exhibit sufficient complexity. Although multi-level BO improves efficiency over standard BO for complex optimization landscapes, it may underperform on simpler problems. In our application, the target function—mapping the learned latent representation of CS to free energy—is sufficiently complex and non-smooth to benefit from the multi-level BO strategy. Further work is needed to identify optimal complexity hierarchies and resolution-level differences, which could further enhance efficiency. Another limitation is the increased complexity in implementation and hyperparameter tuning. Multi-level BO requires setting hyperparameters for each resolution, as well as additional parameters for resolution switching. Nevertheless, these hyperparameters are primarily related to the chemical space and can thus be transferred across different molecular optimization tasks.

Beyond its demonstrated application in lipid bilayer phase separation, our multi-level BO framework can solve other optimization problems characterized by free-energy differences. We expect our method to be particularly advantageous in applications with little prior knowledge or training data. Furthermore, integrating our method with a FAIR[67] data infrastructure and automated simulation workflows, such as Martignac,[68] will enhance data management, reproducibility, and end-to-end automation, thereby making the multi-level BO approach more systematic and streamlined.

Our work provides a versatile and efficient molecular design and optimization framework, offering a promising direction for tackling complex chemical search problems.

Author contributions

Luis J. Walter: conceptualization, methodology, software, validation, formal analysis, investigation, writing (original draft, writing – review & editing), visualization, Tristan Bereau: conceptualization, resources, writing (review & editing), supervision, project administration, funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Data availability

The code for the multi-level Bayesian optimization workflow, the simulation setup, the analysis, and the autoencoder training, as well as the autoencoder models and free-energy results, can be found at **https://github.com/BereauLab/Multi-Level-BO-w-Hierarchical-CG**. A representative subset of the simulation data is available on NOMAD at **https://doi.org/10.17172/NOMAD/2025.05.27-1**. We also provide a tutorial showcasing the main concepts of this paper through a simple two-bead molecule optimization: **https://github.com/BereauLab/Molecule-Optimization-w-Hierarchical-CG-Tutorial**.

The supplementary information provides extended methodological details and additional results that support the main findings. See DOI: **https://doi.org/10.1039/d5sc03855c**.

Acknowledgements

Notes and references

1 P. Kirkpatrick and C. Ellis, *Nature*, 2004, **432**, 823.
2 J.-L. Reymond, *Acc. Chem. Res.*, 2015, **48**, 722–730.
3 P. G. Polishchuk, T. I. Madzhidov and A. Varnek, *J. Comput.-Aided Mol. Des.*, 2013, **27**, 675–679.

4 K. Mishra, L. Ganju, M. Sairam, P. Banerjee and R. Sawhney, *Biomed. Pharmacother.*, 2008, **62**, 94–98.

5 R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer and G. S. Sittampalam, *Nat. Rev. Drug Discovery*, 2011, **10**, 188–195.

6 D. C. Fara, T. I. Oprea, E. R. Prossnitz, C. G. Bologa, B. S. Edwards and L. A. Sklar, *Drug Discovery Today: Technol.*, 2006, **3**, 377–385.

7 M. Karplus and G. A. Petsko, *Nature*, 1990, **347**, 631–639.

8 T. Hansson, C. Oostenbrink and W. van Gunsteren, *Curr. Opin. Struct. Biol.*, 2002, **12**, 190–196.

9 T. Bereau, *Modell. Simul. Mater. Sci. Eng.*, 2021, **29**, 023001.

10 N. Stanley and G. De Fabritiis, *In Silico Pharmacol.*, 2015, DOI: 10.1186/s40203-015-0007-0.

11 H. J. Kushner, *J. Basic Eng.*, 1964, **86**, 97–106.

12 P. I. Frazier, *arXiv*, 2018, preprint, arXiv:1807.02811, DOI: 10.48550/arXiv.1807.02811.

13 G. Agarwal, H. A. Doan, L. A. Robertson, L. Zhang and R. S. Assary, *Chem. Mater.*, 2021, **33**, 8133–8144.

14 J. Thompson, W. P. Walters, J. A. Feng, N. A. Pabon, H. Xu, M. Maser, B. B. Goldman, D. Moustakas, M. Schmidt and F. York, *Artif. Intell. Life Sci.*, 2022, **2**, 100050.

15 J. E. Crivelli-Decker, Z. Beckwith, G. Tom, L. Le, S. Khuttan, R. Salomon-Ferrer, J. Beall, R. Gómez-Bombarelli and A. Bortolato, *J. Chem. Theory Comput.*, 2024, **20**(16), 7188–7198.

16 T. N. Kipf and M. Welling, *arXiv*, 2016, preprint, arXiv:1611.07308, DOI: 10.48550/arXiv.1611.07308.

17 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.

18 P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer and P. Friederich, *Commun. Mater.*, 2022, **3**, 93.

19 H. L. Morgan, *J. Chem. Doc.*, 1965, **5**, 107–113.

20 M. K. Warmuth, J. Liao, G. Rätsch, M. Mathieson, S. Putta and C. Lemmen, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 667–673.

21 I. Muegge and P. Mukherjee, *Expert Opin. Drug Discovery*, 2015, **11**, 137–148.

22 R. Gorantla, A. Kubincová, B. Suutari, B. P. Cossins and A. S. J. S. Mey, *J. Chem. Inf. Model.*, 2024, **64**, 1955–1965.

23 W. G. Noid, *J. Chem. Phys.*, 2013, **139**, 090901.

24 P. C. T. Souza, R. Alessandri, J. Barnoud, S. Thallmair, I. Faustino, F. Grünewald, I. Patmanidis, H. Abdizadeh, B. M. H. Bruininks, T. A. Wassenaar, P. C. Kroon, J. Melcr, V. Nieto, V. Corradi, H. M. Khan, J. Domański, M. Javanainen, H. Martinez-Seara, N. Reuter, R. B. Best, I. Vattulainen, L. Monticelli, X. Periole, D. P. Tieleman, A. H. de Vries and S. J. Marrink, *Nat. Methods*, 2021, **18**, 382–388.

25 K. H. Kanekal and T. Bereau, *J. Chem. Phys.*, 2019, **151**, 164106.

26 B. Mohr, K. Shmilovich, I. S. Kleinwächter, D. Schneider, A. L. Ferguson and T. Bereau, *Chem. Sci.*, 2022, **13**, 4498–4511.

27 D. Huang, T. T. Allen, W. I. Notz and R. A. Miller, *Struct. Multidiscip. Optim.*, 2006, **32**, 369–382.

28 C. Fare, P. Fenner, M. Benatan, A. Varsi and E. O. Pyzer-Knapp, *npj Comput. Mater.*, 2022, **8**, 257.

29 P. Mikkola, J. Martinelli, L. Filstroff and S. Kaski, *The 26th International Conference on Artificial Intelligence and Statistics*, Valencia, Spain, 2023, pp. 7425–7454.

30 N. Gantzler, A. Deshwal, J. R. Doppa and C. M. Simon, *Digital Discovery*, 2023, **2**, 1937–1956.

31 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.

32 J. Barnoud, G. Rossi, S. J. Marrink and L. Monticelli, *PLoS Comput. Biol.*, 2014, **10**, e1003873.

33 A. Centi, A. Dutta, S. H. Parekh and T. Bereau, *Biophys. J.*, 2020, **118**, 1321–1332.

34 S. Izvekov and G. A. Voth, *J. Phys. Chem. B*, 2005, **109**, 2469–2473.

35 T. C. Moore, C. R. Iacovella and C. McCabe, *J. Chem. Phys.*, 2014, **140**, 224104.

36 W. Pulawski, M. Jamroz, M. Kolinski, A. Kolinski and S. Kmiecik, *J. Chem. Inf. Model.*, 2016, **56**, 2207–2215.

37 K. Zieba, M. Ślusarz, R. Ślusarz, A. Liwo, C. Czaplewski and A. K. Sieradzan, *J. Phys. Chem. B*, 2019, **123**, 7829–7839.

38 R. Alessandri, F. Grünewald and S. J. Marrink, *Adv. Mater.*, 2021, **33**(24), 202008635.

39 L. R. Kjølbye, G. P. Pereira, A. Bartocci, M. Pannuzzo, S. Albani, A. Marchetto, B. Jiménez-García, J. Martin, G. Rossetti, M. Cecchini, S. Wu, L. Monticelli and P. C. T. Souza, *QRB Discov.*, 2022, **3**, e19.

40 N. van Hilten, J. Methorst, N. Verwei and H. J. Risselada, *Sci. Adv.*, 2023, **9**(11), DOI: 10.1126/sciadv.ade8839.

41 J. Methorst, N. van Hilten, A. Hoti, K. S. Stroh and H. J. Risselada, *J. Chem. Theory Comput.*, 2024, **20**, 1763–1776.

42 S. Lütge, M. Krebs and H. J. Risselada, *J. Phys. Chem. B*, 2025, **129**(9), 2482–2492.

43 P. Ghosh, M. S. M. Sajjadi, A. Vergari, M. Black and B. Scholkopf, *International Conference on Learning Representations*, 2020.

44 D. P. Kingma and M. Welling, *arXiv*, 2013, preprint, arXiv:1312.6114, DOI: 10.48550/arXiv.1312.6114.

45 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, *arXiv*, 2017, preprint, arXiv:1704.01212, DOI: 10.48550/arXiv.1704.01212.

46 W. L. Hamilton, R. Ying and J. Leskovec, *arXiv*, 2017, preprint, arXiv:1706.02216, DOI: 10.48550/arXiv.1706.02216.

47 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2019.

48 M. Fey and J. E. Lenssen, *ICLR Workshop on Representation Learning on Graphs and Manifolds*, New Orleans, USA, 2019.

49 D. R. Jones, M. Schonlau and W. J. Welch, *J. Global Optim.*, 1998, **13**, 455–492.

50 J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger and A. G. Wilson, *Advances in Neural Information Processing Systems*, Montréal, Canada, 2018.

51 G. Torrie and J. Valleau, *J. Comput. Phys.*, 1977, **23**, 187–199.

52 *Free Energy Calculations: Theory and Applications in Chemistry and Biology*, ed. C. Chipot and A. Pohorille, Springer Berlin Heidelberg, 2007.

53 A. S. Mey, B. K. Allen, H. E. Bruce Macdonald, J. D. Chodera, D. F. Hahn, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, A. Rizzi, J. Scheen, M. R. Shirts, G. Tresadern and H. Xu, *Living J. Comp. Mol. Sci.*, 2020, **2**(1), 18378.

54 R. Menichetti, K. H. Kanekal and T. Bereau, *ACS Cent. Sci.*, 2019, **5**, 290–298.

55 C. Hoffmann, A. Centi, R. Menichetti and T. Bereau, *Sci. Data*, 2020, **7**, 51.

56 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1–2**, 19–25.

57 S. Páll, A. Zhmurov, P. Bauer, M. Abraham, M. Lundborg, A. Gray, B. Hess and E. Lindahl, *J. Chem. Phys.*, 2020, **153**, 134110.

58 L. Borges-Araújo, A. C. Borges-Araújo, T. N. Ozturk, D. P. Ramirez-Echemendia, B. Fábián, T. S. Carpenter, S. Thallmair, J. Barnoud, H. I. Ingólfsson, G. Hummer, D. P. Tieleman, S. J. Marrink, P. C. T. Souza and M. N. Melo, *J. Chem. Theory Comput.*, 2023, **19**, 7387–7404.

59 T. N. Ozturk, M. König, T. S. Carpenter, K. B. Pedersen, T. A. Wassenaar, H. I. Ingólfsson and S. J. Marrink, in *Biophysical Approaches for the Study of Membrane Structure—Part B: Theory and Simulations*, ed. M. Deserno and T. Baumgart, Elsevier, 2024, ch. 7, pp. 237–285.

60 M. Bernetti and G. Bussi, *J. Chem. Phys.*, 2020, **153**, 114107.

61 M. R. Shirts and J. D. Chodera, *J. Chem. Phys.*, 2008, **129**, 124105.

62 Z. Wu, D. L. Dotson, I. Alibay, B. K. Allen, M. S. Barhaghi, J. Hénin, T. T. Joseph, I. M. Kenney, H. Lee, H. Li, V. Lim, S. Liu, D. Marson, P. T. Merz, A. Schlaich, D. Mobley, M. R. Shirts and O. Beckstein, *J. Open Source Softw.*, 2024, **9**, 6934.

63 T. A. Wassenaar, H. I. Ingólfsson, R. A. Böckmann, D. P. Tieleman and S. J. Marrink, *J. Chem. Theory Comput.*, 2015, **11**, 2144–2155.

64 G. Fiorin, M. L. Klein and J. Hénin, *Mol. Phys.*, 2013, **111**, 3345–3362.

65 A. May, R. Pool, E. van Dijk, J. Bijlard, S. Abeln, J. Heringa and K. A. Feenstra, *Bioinformatics*, 2013, **30**, 326–334.

66 R. Alessandri, S. Thallmair, C. G. Herrero, R. Mera-Adasme, S. J. Marrink and P. C. T. Souza, in *A Practical Introduction to Martini 3 and its Application to Protein-Ligand Binding Simulations*, AIP Publishing LLCMelville, New York, 2023, pp. 1–34.

67 M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, *Sci. Data*, 2016, **3**, 160018.

68 T. Bereau, L. J. Walter and J. F. Rudzinski, *J. Chem. Inf. Model.*, 2024, **64**, 9413–9423.