Folding Coarse-Grained Oligomer Models with PyRosetta

Theodore L. Fobe, Christopher C. Walker, Garrett A. Meek, and Michael R. Shirts*

Department of Chemical and Biological Engineering

University of Colorado Boulder, Boulder, CO, USA, 80309

E-mail: michael.shirts@colorado.edu

Abstract

Non-biological foldamers are a promising class of macromolecules that share similarities to classical biopolymers such as proteins and nucleic acids. Currently, designing novel foldamers is a non-trivial process, often involving many iterations of trial synthesis and characterization until folded structures are observed. In this work, we aim to tackle these foldamer design challenges using computational modeling techniques. We developed CG PyRosetta, an extension to the popular protein folding python package, PyRosetta, which introduces coarse-grained (CG) residues into PyRosetta, enabling the folding of toy CG foldamer models. Although these models are simplified, they can help explore overarching physical hypotheses about how oligomers can form. Through systematic variation of CG parameters in these models, we can investigate various folding hypotheses at the CG scale to inform the design process of new foldamer chemistries. In this study, we demonstrate CG PyRosetta's ability to identify minimum energy structures with a diverse structural search over a range of simple models, as well as two hypothesis-driven parameter scans investigating the effects of side-chain size and internal backbone angle on secondary structure. We are able to identify several types of secondary structures from single- and double-helices to sheet-like and knot-like structures. We show how side-chain size and backbone bond angle both play an important role in the structure and energetics of these toy models. Optimal side-chain sizes promote favorable packing of side-chains, while specific backbone bond-angles influence the specific helix type found in folded structures.

1 Introduction

The astounding diversity of structure and function found in proteins and nucleic acids demonstrate Nature's ability to design sophisticated macromolecules with ordered 3D structure for specific chemical functions. These examples in biology have inspired the discovery and design of novel non-biological foldamers, oligomeric molecules which can self-assemble into well-structured secondary and tertiary structures.[1,2] Researchers studying foldamers thus aim to design new classes of macromolecules that share similar folding properties as proteins but are constructed from a more diverse set of chemistries. These new non-biological oligomers show promise for a diverse range of applications ranging from therapeutics,[3] antimicrobials,[4,5] catalysts,[6,7] molecular sensors[8] to nanostructured materials.[9,10]

Considerable experimental research over the last twenty years has identified many new types of foldamers and characterized folding patterns formed by these macromolecules.[1] Researchers have explored natural extensions to canonical biopolymers, in addition to foldamers with novel backbone and side chain chemistries. For example, a number of groups have investigated biomimetic chemistries such as β-, γ- and δ-peptides,[11,12] peptoids[13,14] and covalently constrained variants of these peptides.[15–17] Other groups have explored more exotic stabilizing interactions in foldamers, such as π-π interactions and metal coordination.[18,19] The success of this research into novel folding chemistries has suggested a range of new applications these chemistries enable. For example, Porter et al. have shown that β-peptides exhibit similar antibacterial properties as their α-peptide counterparts, without triggering the same immune response and with significantly less proteolytic susceptibility.[20] Additionally, groups have explored the use of artificial foldamers as molecular containers; Ferrand and Huc show aromatic oligomer helices can be designed to recognize a range of simple guest molecules.,[21] and Collie et al. achieved molecular recognition of small alcohols and diols using helical assemblies of oligourea foldamers.[22]

Chemically-novel foldamers have a wealth of future applications if we can figure out how to rationally design them. Currently, formulations for novel foldamers are a tedious process, limited by low-yield synthesis methods and difficulties in characterization for all but the most stable molecules. Most foldamers that resemble existing biopolymers are synthesized using solid-phase synthesis for their respective linker bond type.[23] These solid-phase synthesis methods suffer from poor protective reaction conditions resulting in low polymer length and low final product yield.[24,25] Furthermore, methods for foldamer characterization have additional challenges compared to more well-established biopolymer characterization due to unknown spectroscopic signatures of new foldamer molecules.[1]

In addition, the design of novel backbone chemistry that folds in a desired solvent is somewhat of an art. Gellman describes three challenges when designing novel foldamer molecules.[2]

- First, a novel polymeric backbone must be proposed that has adequate folding propensities and can form structure at the polymer level.
- Second, the foldamer must have an "interesting" chemical function added by design or evolutionary techniques.
- Last, for production, the foldamer must have an efficient synthesis method.[2]

Unfortunately, current foldamer design efforts heavily rely on chemical intuition and in practice results in the need for trial-and-error synthesis to confirm if a foldamer properly folds under particular conditions of interest, including solvent, temperature ranges, polymer lengths, solution pH, and many more.

Current success in foldamer design has stemmed from several biosimilar foldamer backbones that borrow many design choices from traditional biopolymers.[1,2] Using peptide or phosphodiester linkers allows researchers to start with a working design that has been evolutionarily selected to fold under certain conditions. These biopolymers include interactions such as hydrogen bonds in protein backbones, or π-stacking and base-pair hydrogen bonds in nucleic acids.[1,2] Biosimilar foldamers also take advantage of preexisting synthesis techniques allowing comparatively quick synthesis compared to a novel backbone synthesis.[23,26,27] Despite their similarity to existing biopolymers, these biosimilar foldamers are still considerably difficult to design.[1] In order to facilitate the foldamer design process, new ways to propose potentially stable foldamer chemistries are needed. Better tools to formulate novel foldamers would help discover new backbone chemistries that are not biologically inspired. New computational tools to validate new foldamer chemistries would also help incentivize the use of more difficult synthetic techniques.

Existing protein structure prediction software, like the popular protein folding package, Rosetta,[28] leverages tools like fragment assembly,[29,30] rotamer libraries,[31] homology modeling,[32,33] knowledge-based potentials[34] and machine learning,[35,36] to take advantage of the abundance known structures in online databases, such as the Protein Databank,[37] to quickly sample structures that are similar to existing proteins structures.[38] Unfortunately for foldamer structure prediction and design, there is a very limited set of solved structures to guide modeling approaches.[39] Any computational or theoretical approach to modeling foldamers, for now, must be primarily based on physics-based modeling.

Computational methods such as molecular dynamics (MD) simulations are another tool in researchers' arsenal to predict foldamer structures and identify molecular mechanisms driving their folding, and have been used for aromatic helices,[40,41] peptoids,[42–44] and β and δ peptides.[45,46] Enhanced sampling methods, such as temperature replica exchange molecular dynamics, and others can be used to overcome sampling problems of longer foldamer oligomers.[42,44,47] Often MD simulations are used in conjunction with NMR and X-ray crystallography to better illustrate foldamer dynamic properties.[22,48] While MD simulations can be a useful tool in foldamer design, their results can be difficult to interpret and validate without proper physical validation from experimental structures and observables and require substantial effort to develop new parameterizations for each new monomer class.

In this work, we turn to generic or "toy" coarse-grained (CG) models due to the limited foldamer structures available in the literature,[39] in order to interrogate overarching physical hypotheses and principles. Unlike traditional CG models, which are typically derived from an underlying atomistic system,[49–51] generic CG models are often chemically-nonspecific and aim to capture physical phenomena with relatively few model parameters.[52,53] These simple models can capture qualitative insight into the mechanisms that govern the phenomena of interest. Toy models are a common practice in capturing phenomenology with examples in ferromagnetism,[54,55] phase equilibria,[56] amphiphile assembly[53] and protein folding.[57–59]

In this paper, we use generic CG models to explore macromolecule folding using simplified models. Non-specific CG models with relatively few parameters can capture general features of macromolecular folding[57] and give insight into folding principles that are generally applicable to all macromolecular systems, not solely oligopeptides and proteins. Despite not modeling a specific chemical system, these simple models can capture general folding principles of macromolecules, which can inform the selection or design of novel foldamer monomers chemistries. We choose to implement these simple models in Rosetta, an existing protein prediction software, to take advantage of many of the physics-based methods used in traditional protein folding algorithms. For example, Rosetta offers fast sampling of configuration spaces of heteropolymer models using Monte Carlo (MC) minimization methods.[60]

We introduce `cg_pyrosetta` as an extension to Rosetta which adds new CG functionality. In this package, we add new CG models to Rosetta's library allowing users to fold a wide range of CG representations of foldamers in the Rosetta workflow. Using Rosetta's Python-wrapped C++ library[61] we are able to build a set of protocols that identify low-energy structures available to CG foldamer models. Using `cg_pyrosetta`, we can explore the parameter space of CG foldamer models and investigate the underlying physical driving forces of secondary structure formation.

In this paper, we demonstrate cg_pyrosetta's ability to determine CG model minimum energy ensembles with several *ab initio* folding simulations of a variety of CG models. Through parameter scans, we explore simple foldamer hypotheses and present several instances of emergent secondary structure in several CG models. We first discuss the model parameters available in cg_pyrosetta, details of cg_pyrosetta's implementation, the MC minimization algorithm used in this work, and the analysis workflow we developed to identify folded structures. Following this, we share results of several folded structures from a variety of CG models and two preliminary parameter scans based on hypotheses about foldamer stability. The first parameter scan varies the side chain R^{min} and bond-length parameters to explore the effects of side chain size on foldamer secondary structure and the second scan varies backbone bond angles (θ_B) to explore the effects of local monomer geometry on foldamer secondary structure. We conclude with remarks on how cg_pyrosetta can be used in further investigations of foldamers.

2 Methods

2.1 Coarse-grained Model

The CG model implemented in cg_pyrosetta consists of a series of connected CG beads, which can represent whole or parts of the side chain and backbone moieties. This representation was chosen as a starting point as it preserves the notion of individual residues and distinguishes between backbone and side chain interactions. This construction also allows us to build more complex models by adding more interaction sites or interaction potentials to either the backbone or the side chain sites. For example, many commonly-used protein CG models reduce the full-atom representation of a protein to a similar representation while maintaining sufficient information to study protein folding, protein docking and large-scale protein dynamics.[51,62,63]

Our CG model implemented in `cg_pyrosetta` uses the standard Rosetta score functions to implement Lennard-Jones pairwise potentials and the Rosetta molecular mechanics score functions to implement harmonic bond angle potentials, periodic dihedral angle potentials.[34] Electrostatic potentials were not used in this work, but can be explored as needed, using either new or existing Rosetta score terms. Potentials used in `cg_pyrosetta` and their corresponding score terms can be found in Table 1. In this work, we assume bond lengths are rigid at their equilibrium value. This assumption was made in part to reduce the dimensionality of the configuration space we are sampling and in part due to the minimal contribution bond stretching contributes to overall protein motion.[64]

Table 1: Energetic terms, relevant parameters, and potential equations and their corresponding score terms in Rosetta. Pairwise mixing in Rosetta is handled such that, $\epsilon_{i,j}$ is the geometric mean of ϵ_i and ϵ_j and $R_{i,j}^{min}$ is the sum of R_i^{min} and R_j^{min}.

Energy Term	Rosetta Score Term	Parameters	Equation
Lennard-Jones	`fa_atr`, `fa_rep`	$R_{i,j}^{min}$, $\epsilon_{i,j}$	$E_{vdw} = \epsilon_{i,j}\left[\left(\frac{R_{i,j}^{min}}{d_{i,j}}\right)^{12} - 2\left(\frac{R_{i,j}^{min}}{d_{i,j}}\right)^{6}\right]$
Bond Angles	`mm_bend`	$k_{\theta,i}$, $\theta_{0,i}$	$E = k_{\theta,i}\left(\theta - \theta_{0,i}\right)^2$
Dihedrals	`mm_twist`	$k_{\phi,i}$, n_i, $\phi_{0,i}$	$E = k_{\phi,i}\left(1 + \cos(n_i\phi - \phi_{0,i})\right)$

For the majority of the analysis done in this paper, we explore the 1-backbone/1-side chain (*1b1s*) model for a variety of hypotheses. Figure 1 depicts a short segment of a *1b1s* model with relevant topology labeled. We label the torsions, bond angles, and bond lengths consisting of all backbone atoms as ϕ_B, θ_B, and d_{BB}, respectively. Torsions, bond angles, and bond lengths including side chain beads are labeled as ϕ_S, θ_S and d_{BS}, respectively. Since these models are of arbitrary scale, all units in this study are presented in reduced units of length and energy. We define the backbone interaction bead distance, R_B^{min}, and the backbone energy ϵ_B as our base units.

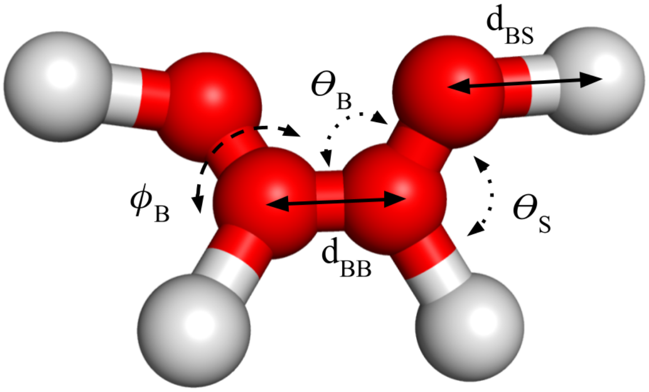


Figure 1: Illustration of the topology and bonded parameters of a 1-backbone/1-side chain model (*1b1s*) containing 4 residues.

We note here that while these foldamer models use traditional molecular dynamics force-fields terms, these models do not explicitly model specific underlying chemistry. These toy models represent a simplification of foldamer systems with the goal of understanding how general macromolecules fold in relation to their model parameters. Using CG models we can quickly prototype several folding hypotheses through systematic variation of CG model parameters.

In the *1b1s* model and other simple models, when all residues are achiral, any type of macromolecular folding would have both left- and right-handed versions of folded structures. These left- and right-handed folded structures are energetically degenerate states and should, with proper sampling, populate equally. For example, in the *1b1s* model, when side chain angles are kept symmetric about their corresponding backbone bead ($\theta_S = 180 - \theta_B$) and torsions can freely rotate, the model is achiral and is expected to have degenerate left- and right-handed minimum energy structures.

All molecular visualizations of CG models in this work were generated using PyMOL 2.5.[65]

2.2 Package Implementation

cg_pyrosetta is implemented as a Python package that adds new CG functionality to PyRosetta, and can be found at https //github.com/shirtsgroup/cg_pyrosetta. This functionality is added using a set of CG residues, CG atomtypes, and CG movers. These additions to PyRosetta work natively in PyRosetta/Rosetta objects and allow for the design and folding of CG models. Natively, PyRosetta creates immutable library objects once data files are read, therefore external data files must be loaded when PyRosetta initializes. This requires new instances of PyRosetta for different types of parameters. cg_pyrosetta initializes PyRosetta using the flags shown in table 2 to create new instances of PyRosetta with different parameter files. cg_pyrosetta can be installed using PyRosetta/Rosetta v2021.45dev61799 or versions after this.

Table 2: Flag descriptions used to add new CG functionality to CG PyRosetta

Flag	Description
-add_atom_types	Adds new atom types to an AtomTypeSet from an external file.
-extra_res_fa	Adds new residue types to a ResidueTypeSet from an external file.
-add_mm_atom_type_set_parameters	Adds new molecular mechanics atom types to an MMAtomTypeSet using an external file.
-extra_mm_params_dir	Adds new molecular mechanics bond-length, bond-angle and torsion parameters from external files.

cg_pyrosetta wraps around PyRosetta and initializes it with the proper CG parameter files, ensuring that most native PyRosetta functionality is still intact while performing cg_pyrosetta simulations. All CG model files are added on top of existing libraries within PyRosetta and therefore do not interfere with existing PyRosetta data files. The MC simulations used in cg_pyrosetta use the preexisting MC framework present in Rosetta, simply with CG foldamer models instead of proteins or other Rosetta-compatible residue types. Additionally, the energy parameters, residue types and atom types added in cg_pyrosetta all interact natively with existing Rosetta energy evaluation and minimization objects. When adding CG residues to PyRosetta, we turn off the energy parameters read in from the `extra.txt` file, as we do not include these in our score function. The exclusion of the energy terms from `extra.txt` is toggle-able with a utility function in `cg_pyrosetta`.

The CG movers added to `cg_pyrosetta` all inherit from the Rosetta mover base class, but are implemented in Python. Rosetta's default torsion movers (`small` and `shear` movers) are implemented to change protein ϕ and ψ angles and are therefore incompatible with the CG models we added to `cg_pyrosetta`. In the CG model space, we are interested in searching over all available degrees of freedom. We therefore developed a CG torsion mover, a CG bond-angle mover, and CG bond-length movers for sampling all degrees of freedom in these CG models. When used on a CG model these movers find all the available degrees of freedom of their respective type and randomly perturb one of them. Combining these new CG movers and Rosetta's existing MonteCarlo and Minimizer objects we develop a folding algorithm that can quickly explore large regions of configuration space and identify minimum energy structures of our first-generation CG models. Further details on the folding algorithm developed are discussed in section 2.3

We use Signac[66] to distribute and manage folding simulations. Due to Rosetta's immutable loading of parameters at the start of each instance of Rosetta, we would be unable to launch several jobs with different parameters using Rosetta's built-in job distributor. Using an external job distributor, like Signac, allows us to run several new instances of `cg_pyrosetta` with different sets of parameters without having to change the immutable definition of parameters internally. Using Signac, we run all folding simulations on single processes distributed in an embarrassingly parallel fashion. Since no information transfer between jobs is required all folding replicas can be run independently of one another.

2.3 Monte Carlo Minimization

Monte Carlo (MC) sampling with minimization, also called basin-hopping, is a global optimization technique[67,68] used to sample folded configurations in many of the folding algorithms in Rosetta.[60] In a traditional MC simulation, a Markov chain of random moves is performed where moves are accepted with the probability given by

$$P(\boldsymbol{x}_i, \boldsymbol{x}_{i-1}) = \min\left(1, e^{-\frac{U(\boldsymbol{x}_i)-U(\boldsymbol{x}_{i-1})}{k_B T}}\right), \tag{1}$$

where $\boldsymbol{x}_i$ is the newly proposed coordinate matrix and $\boldsymbol{x}_{i-1}$ is the coordinate matrix of the last step, U is the potential energy, and $k_B T$ is the effective temperature of the MC process. MC simulations are used to sample the Boltzmann distribution of a system of interest.

In MC minimization, the addition of a minimization step before the MC evaluation changes the MC process into an optimization. Instead of sampling the entire Boltzmann distribution, MC minimization samples local minima on the search for the global minimum. MC minimization is often effective at finding global minima in multi-dimensional systems and has been used to great effect in molecular conformation optimization.[67–69] Monte Carlo minimization is easily applied to the CG models described above using Rosetta's MC framework. Using the CG movers we developed in conjunction with Rosetta's minimizer object, we are able to construct a MC minimization algorithm that can quickly search for minimum energy structures in the simple CG models described above.

We couple MC minimization with simulated annealing to enhance the global-optimization search. Annealing the simulated temperature, T, changes the tolerance of accepting moves over the course of the MC minimization process. At higher temperatures, perturbations which increase the system energy are selected with higher probability, creating a more exploratory search of the potential energy surface. At low temperatures, moves which increase the system energy are accepted with lower probability, causing the MC minimization to get trapped in local minima and exhaustively sample the local potential surface. Combining exploratory and exhaustive sampling ensures solutions consider a wider range of the solution space compared to traditional gradient-descent optimization.

In this study, the temperatures for our MC minimizations follow the geometric series below:

$$T_n = T_0(r_a)^n \text{ for n = 0, 1, 2,}\ldots, \tag{2}$$

where T_0 is the initial simulated temperature, r_a is the annealing rate and n is the index of temperatures. MC minimization simulations are run for N steps at M different temperatures following the annealing schedule above. A flow diagram of the MC minimization with simulated annealing algorithm used in this work is shown in figure 2.

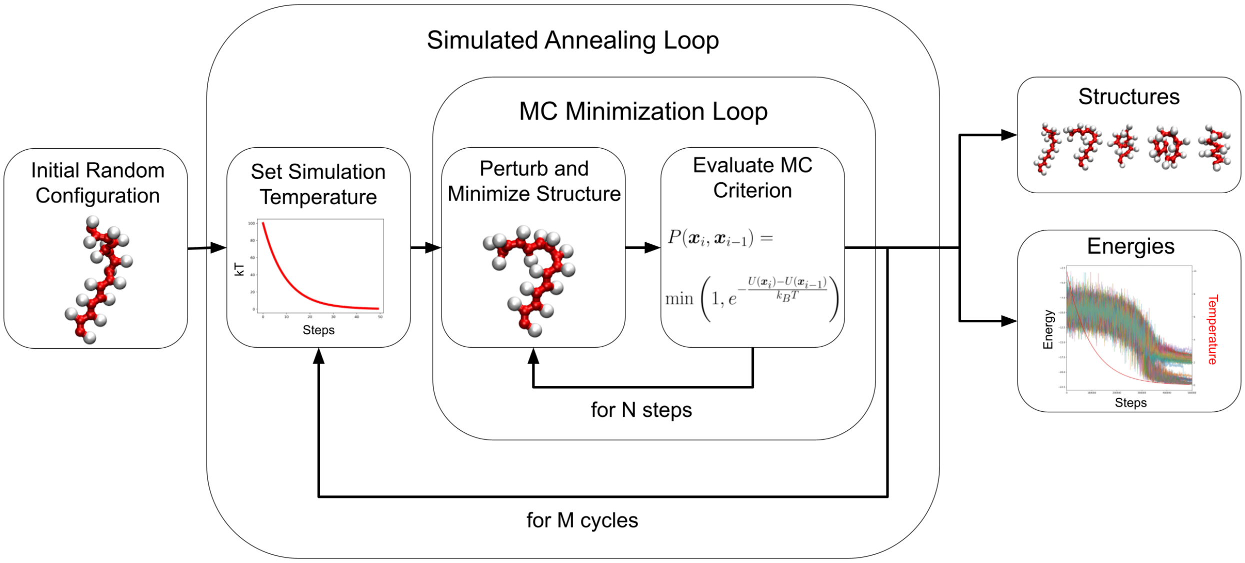

Simulated Annealing Loop
MC Minimization Loop
Initial Random Configuration
Set Simulation Temperature
kT
Steps
Perturb and Minimize Structure
Evaluate MC Criterion
$P(\boldsymbol{x}_i, \boldsymbol{x}_{i-1}) = \min\left(1, e^{-\frac{U(\boldsymbol{x}_i)-U(\boldsymbol{x}_{i-1})}{k_BT}}\right)$
for N steps
for M cycles
Structures
Energies
Energy
Temperature
Steps

Figure 2: Flow diagram depicting how the simulated annealing MC minimization simulations are carried out. Fixed temperature MC simulations with N steps are carried out sequentially at each M different temperature in the annealing schedule. Structures and energies are output during the course of the entire simulation.

Simulation parameters used for the MC minimization with simulated annealing are determined empirically for each folding experiment in this study following a few heuristics. The initial simulation temperature, T_0, starts at the same order of magnitude as the random structure energy distribution, to ensure high energy moves are accepted early on in the simulation. To ensure the annealing process spends sufficient time in both the exploratory and exhaustive regimes, we select an annealing rate, r_a, to slowly transition between the two regimes. The transitions vary from model to model and can give insight into the surface being sampled. We share and examine different transitions in Supporting Information section S1.

The number of simulation steps, N, is selected such that energies are exhaustively sampled at each T_n and the number of annealing cycles, M, is chosen such that the final temperature, T_M is 2–3 orders of magnitude below the initial energy of the foldamer. Figure 3 shows an example of the trajectories of energy over the annealing process. At high temperatures, energies have large fluctuations as the foldamer model explores many configurations. At lower temperatures, the simulation settles into the local minima of the potential energy surface and fluctuates much less. To enable the process to find a consensus minima for a given parameter set, we run 100 replicas of each simulation and verify multiple simulations are populating the minimum structures found.

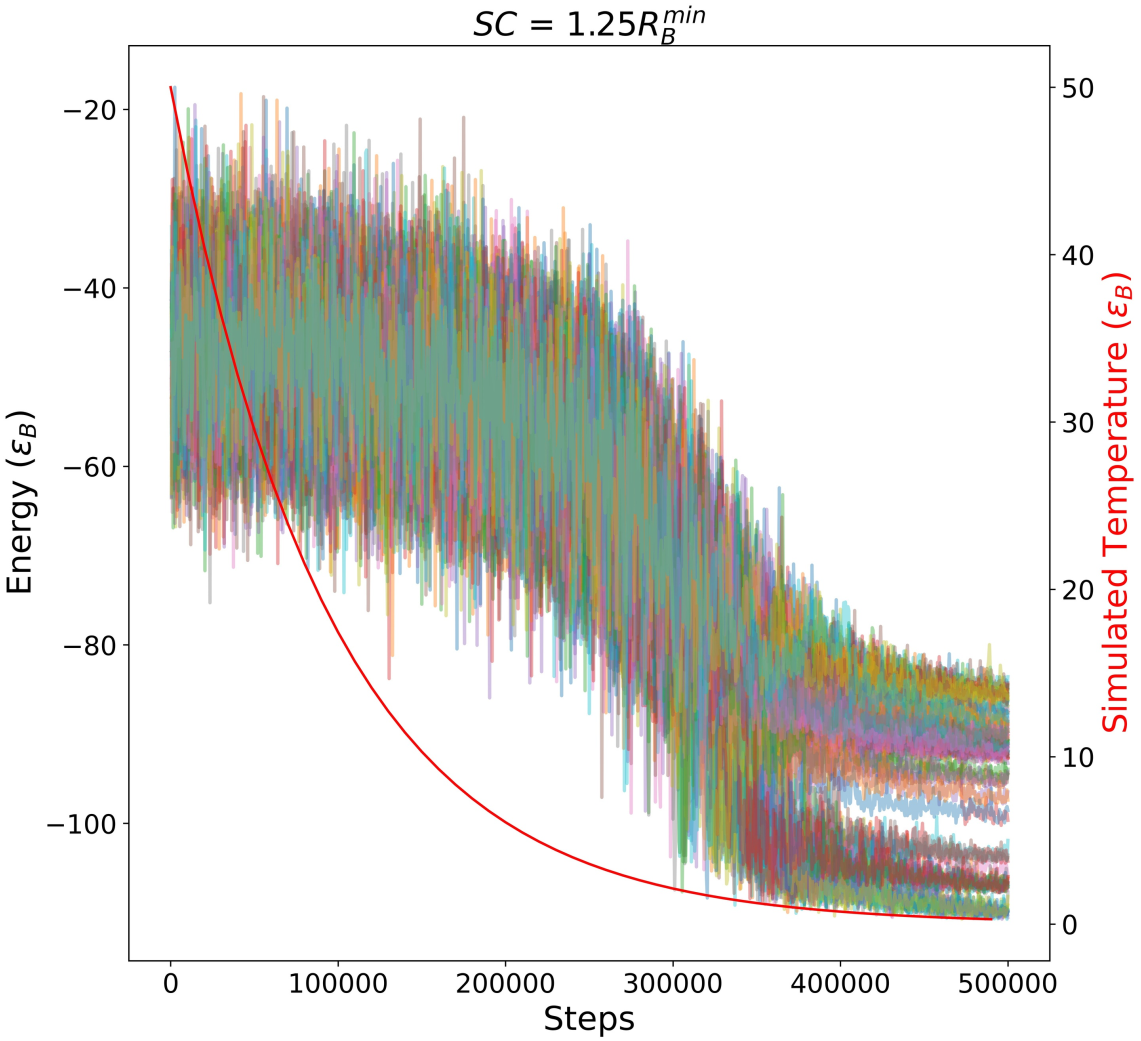


Figure 3: Example energy trajectory of an annealing simulation with 100 replicas. Each replica's energy trajectory is shown in a different color. The annealing temperature of these simulations is drawn as a smooth red line with an axis on the right side. In this particular trajectory, we see a bifurcation of trajectories as the simulation transitions from the exploratory regime to the exhaustive regime, as some go visit the minima close to the global minimum but others are trapped in higher minima. The location and magnitude of the transition region can vary significantly between models.

2.4 Analysis Techniques

2.4.1 RMSD clustering

To identify relevant structures from simulated annealing trajectories we turn to RMSD clustering of structures. Applying clustering to foldamer trajectories enables us to take a large collection of structures and reduce them to a handful of representative structures.[70] In this work, we run clustering on the second half of the folding simulation, to only cluster structures identified in the low-temperature sampling regime.

We compute an RMSD matrix using:

$$RMSD(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\frac{1}{N}\sum_{k=1}^{N} \|x_{ik} - x_{jk}\|^2}, \tag{3}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the coordinate matrices from structure i and j, respectively, x_{ik} is the atom position vector for atom k index within structure i and N is the total number of atoms of each structure. Structures are aligned using MDTraj before calculating RMSDs,[71] such that we are calculating minimum RMSD values between structures. The RMSD matrix was clustered using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method implemented in `scikit-learn`[72] to cluster the RMSD matrix of trajectory frames from our folding simulations. Using the clustering workflow we implemented in `analyze_foldamers`,[73] we can quickly identify representative structures from the MC minimization with simulated annealing process. DBSCAN clustering relies on two hyper-parameters, ϵ and `min_samples`. A data point with `min_samples` other samples within a distance ϵ defines a core point of a cluster. Data points within ϵ of core points without `min_samples` surrounding samples are considered *neighbors* of core points and are included in the cluster. Points outside of ϵ of core points are labeled as noise points.[74,75] DBSCAN is a deterministic density-based clustering approach which has worked well for peptide systems[76] and similar toy models.[77]

To improve RMSD clustering we apply a pre-clustering filter to remove especially noisy trajectory frames, such as those found in the high-temperature regions of the folding simulations. The inclusion of a pre-clustering filter of trajectory structures has made clustering more reliable in several instances of clustering of bio-molecular structures.[77–79] In the pre-clustering filter, the user specifies a filter percentage, and `analyze_foldamers` optimizes a neighbor cutoff radius and neighbor density cutoff using Scipy minimize[80] with the Nelder-Mead method, to reduce the data to that percentage.[73] For this study, we filtered 50% of the original trajectories based on an optimized distance cutoff for each parameter set.

Cluster medoid structures are defined as the structure with the largest similarity score, where similarity scores for each structure within a cluster are calculated using:

$$S_i = \sum_{j}^{N_k} \exp \frac{-RMSD(\boldsymbol{x}_i, \boldsymbol{x}_j)}{d_{scale}}, \tag{4}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent coordinates of structures i and j within cluster k, N_k is the total number of structures in cluster k, $RMSD$ is the operation shown in equation (3) and d_{scale} is a scaling term, often set to the standard deviation of the RMSDs to make the calculation scale invariant.[71] Once a medoid structure for each cluster is selected, we compute the silhouette score of each point within a cluster using:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \tag{5}$$

where $a(i)$ is the average distance between a cluster point and all other points in its own cluster, and $b(i)$ is the distance between a cluster point and the next nearest cluster medoid. Silhouette scores range from 1 to -1, where a value of 1 indicates a structure is well separated from other clusters, a value of 0 indicates a structure is equally distant from its assigned cluster to another cluster, and a value of -1 indicates a structure is closer to another cluster than its assigned cluster.[81] We calculate the silhouette scores for all structures and report the average silhouette score of all non-noise structures to evaluate the quality of clustering. Large silhouette scores imply the cluster was well-defined and structurally distinct from other clusters.

Choosing clustering hyperparameters is a nuanced task. In this work, we aim to identify hyperparameters that yield low numbers of clusters and high average silhouette scores. Finding such hyperparameters represents the simplest way of clustering this data set while still finding structurally distinct clusters. To accomplish this we perform a Pareto-like optimization over a grid search of hyperparameters. Further details of this hyperparameter selection is detailed in Supplemental Information section S2.

Additionally, we identify minimum energy structures of each cluster as the representative structure of that cluster. We choose minimum energy structures of each cluster rather than the identified medoid structure because our sampling method has a minimization component, therefore all identified clusters represent identified local minima, where the structure of interest would be represented by that local minima. Using minimum energy structures of identified clusters ensures good overlap between representative structures from different clusters.

2.4.2 Cluster RMSDs

A necessary condition for well-defined structural minima that are likely to have good folding properties is that there exist well-defined clusters with high structural similarity that are structurally distant from other clusters. We calculate and report the average RMSD *within* all cluster structures to their identified minimum energy structures ($RMSD_{cluster}$), to measure how tightly defined the clusters are. We also calculate RMSD *between* cluster minimum-energy structures ($RMSD_{inter}$) to quantify how structurally different clusters are from one another. With $RMSD_{cluster}$ and $RMSD_{inter}$, we can asses how structurally distinct clusters are from one another and how tightly spread clusters are about their medoids.

2.4.3 Degenerate Cluster Exclusion

For models with achiral residues, we expect to find degenerate left- and right-handed versions of folded structures. Degenerate clusters would have mirror-image structures and also the same energy distribution. To characterize differences between the minimum energy clusters to other clusters we exclude mirror clusters from $RMSD_{inter}$ and energy gap Z-score calculation, such that any minimum cluster $RMSD_{inter}$ and energy gap Z-score is not between degenerate clusters.

We identify mirror clusters by calculating the matrix:

$$A_{ij} = \min\left(RMSD(\boldsymbol{x}_i, \boldsymbol{x}_j), RMSD(\boldsymbol{x}_i, -\boldsymbol{x}_j)\right) \tag{6}$$

where $RMSD$ is the operation shown in equation (3), $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are atom coordinates of cluster minimum energy structures. Structures with smaller mirrored $RMSD_{inter}$ compared to their original $RMSD_{inter}$ are candidates for mirror structures. If the mirror $RMSD_{inter}$ is smaller than the $RMSD_{cluster}$ of the cluster i, cluster j is considered a mirror of cluster i.

2.4.4 Cluster Energy Distributions

Once folding trajectories are clustered, we can evaluate the energetics of each cluster. In order to have sufficient impetus to fold, folded structures must have significantly lower energies compared to competing structures identified in the clustering process. We hypothesize models must have a minimum energy cluster with energies significantly lower than other identified clusters to be considered viable folders. We quantify this value using a z-test for two means between all cluster energy distributions, using:

$$Z_{ij} = \frac{\mu_i - \mu_j}{\sqrt{\frac{(\sigma_i)^2}{n_i} + \frac{(\sigma_j)^2}{n_j}}} \tag{7}$$

where μ_i and μ_j are the means of the energies in the clusters i and j, σ_i and σ_j are the standard deviations of the energies of cluster i and j, and n_i and n_j are the number of samples in each clusters i and j.

Combining cluster energetic information with $RMSD_{cluster}$ illustrates the potential energy landscape of each identified cluster. In figure 4 we can see two minimum energy clusters are identified along with two higher energy clusters. In this example, the two minimum energy cluster corresponds to left- and right-handed helices. The groups of points near 0.08 and 0.12 RMSD away from the minimum energy structure represent structures with frayed helix ends.

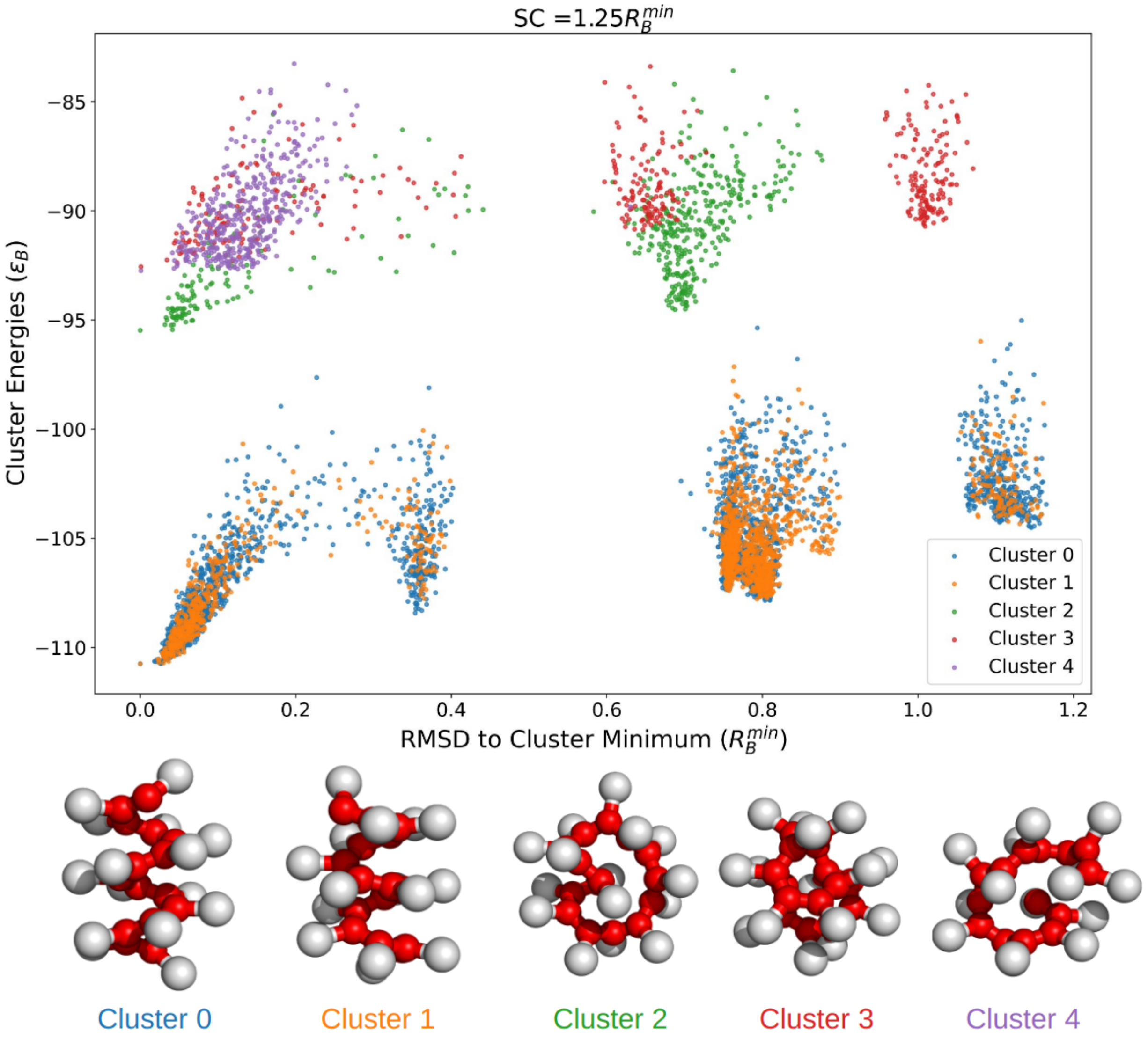


Figure 4: To characterize the potential energy surface of each model, we plot RMSDs from cluster minimum energy structures to individual cluster structures against structure energies. For each cluster, the minimum energy structure will have an RMSD of 0 and will have the lowest identified energy of its respective cluster. Large gaps in the cluster RMSD of each cluster are generally associated with terminal residues misfolding from the minimum energy structure. Minimum energy structures for each cluster are presented below the plot. Clusters 0 and 1 are the degenerate right- and left-handed helical minimum energy clusters identified in this parameter scan. Clusters 2–4 represent higher energy minima found while searching this potential energy landscape. Gaps in the RMSD distribution of each cluster are the result of discrete unfolding events of 1–2 residues. Note carefully that distance along the horizontal axis represents distances from the cluster minima to each structure, *not* the distance between clusters.

In our energetic analysis, we identify the smallest energy Z-score from the minimum energy cluster, which we term the energy gap Z-score of the model. If there are degenerate mirror-image clusters due to the presence of mirror folded states, we exclude the energy gap Z-score between these identified mirror clusters. In the case of figure 4, clusters 0 and 1 are mirror clusters, as they represent both left- and right-hand forms of a helix. We expect a large energetic Z-score value between the minimum energy clusters and other higher energy clusters. For the example shown in figure 4, the energy gap Z-score from cluster 0 or cluster 1 to the next lowest energy cluster (cluster 3) is 3.98. Large energy gap Z-scores represent larger energy gaps between identified clusters. Finally, We note the structural and energetic overlap between degenerate clusters 0 and 1 also indicates good sampling of the configuration space of this model.

2.4.5 Helix Fitting

To classify helical clusters found in these parameter scan we perform a least square fit of the backbone coordinates to the parametric helix equations using methods described in Refs. 82 and 83, which we implemented in `analyze_foldamers`.[73] Further details of this fit are included in Supplemental Information section S3. We find the least-squares fit performed best when fitting to the internal residues of the helical structures, rather than all residues. Excluding the end residues from the least-squares fit removes noise introduced from frayed ends, and more accurately classifies the helix type in the bulk of the structure. From the least-squares fit, we report helix residues per turn for models with helical minimum energy clusters. From the least-square fit, we report the root-mean-squared error of the cylindrical fit, $RMSE_{cyl}$, and the root-mean-squared error of the helix fit, $RMSE_{helix}$, as metrics of the quality of fit. All helix least-squares fits are detailed in the Supplemental Information section S3.

2.5 Identifying Distinct Folded Structures

We are most interested in models with clusters that are structurally compact, are structurally well-separated from other clusters, and are significantly lower in energy than other identified clusters to be considered well-behaved. These are likely candidates for models that would have folded thermodynamic ensembles. In this section, we explore the four metrics for these properties to identify and evaluate well-behaved clusters.

Using metrics defined in previous sections we can identify candidate foldamer models that have structured, low-energy folded states. To reiterate, we evaluate the following metrics for each parameter set:

- **Energy gap Z-score**: The Z-score between the energies of the lowest energy cluster and the 2nd lowest energy cluster, excluding degenerative clusters. This metric evaluates the separation in energy from identified clusters. Larger energy gap Z-scores correspond to energetically distinct minimum energy clusters.

- **Average silhouette score**: The average silhouette score from clustering is used as a metric to evaluate the quality of the clustering. Silhouette scores compare the distance of a point from its own cluster to the next nearest cluster. Larger silhouette scores indicate more distinct clusters.

- **Minimum Inter-minimum RMSD**: The smallest RMSD between the lowest energy cluster's minimum energy and all other clusters' minimum energy structures. Structurally distinct clusters will have large minimum inter-minimum RMSDs indicating that other identified clusters are not structurally similar to the minimum energy cluster.

- **Cluster RMSD**: The average RMSD of all structures in a cluster to its medoid structure. Cluster RMSD is a measure of how structurally spread out a cluster is. Small clusters represent clusters with fewer fluctuations around

We use the following guidelines to differentiate potential foldamers with likely structured folds from those with non-structured folds.

1. Identified minimum energy clusters should have a minimum energy gap Z-score greater than approximately 1.0, indicating that there is little overlap between energy distributions of the minimum energy cluster energy distribution with other clusters.

2. Identified clusters should also have silhouette scores closer to 1.0, indicating a well-separated set of clusters.

3. The minimum energy cluster should also have comparatively large minimum $RMSD_{inter}$ and comparatively small $RMSD_{cluster}$, indicating structurally distinct and well-defined clusters.

4. Additionally, to ensure minima can be found consistently we confirm that identified clusters are found from many replicas of the folding simulation, giving rise to a consensus minimum.

To illustrate these metrics, we plot them on a radar plot for each parameter set. Figure 5 shows an example radar plot for the sample parameter set shown in figure 4. This example parameter set represents a cluster with a well-defined minimum energy cluster, with optimal values for each of the four identified metrics. The axes of these radar plots are chosen such that better performance on each metric results in a larger total area for the model. In particular, $RMSD_{cluster}$ is plotted on an inverted axis so that a larger value indicates a more compact cluster.

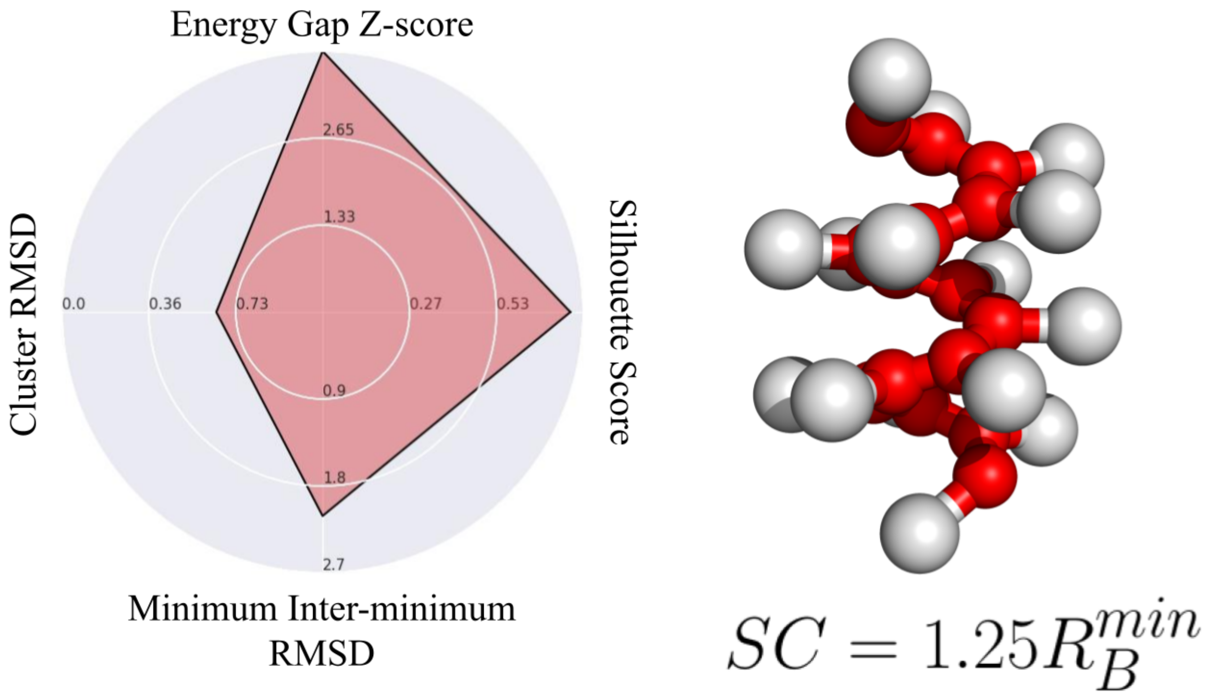


Figure 5: Example radar plot (left) of a well-defined helical cluster (minimum energy structure shown on right). This cluster has a minimum energy gap Z-score of 3.98, an average silhouette score of 0.76, a $RMSD_{cluster}$ of 0.65 R_B^{min} and a minimum $RMSD_{inter}$ of 2.13 R_B^{min}. The radar plots are plotted such that larger colored areas represent better candidate foldamers.

2.6 Simulation Details

2.6.1 Diverse minimum energy structure search

To showcase the diverse toy models available in `cg_pyrosetta`, we performed several MC minimization annealing simulations for the *1b1s* model, the 1-backbone/no-side chain (*1b0s*) model, the 1-backbone/2-side chain (*1b2s*) model, the 1-backbone/3-side chain (*1b3s*) and the 2-backbone/no side chain (*2b0s*) hetero-oligomer model. Findings from this scan are shared in the results and discussion in section 3.1. As several different model parameters and annealing schedules were required for the many parameter scans performed on each of these models, we direct the reader to section S4 for further details on model parameters and annealing parameters.

2.6.2 Side chain size effect on secondary structure

To demonstrate the utility of `cg_pyrosetta` to test structural hypotheses about foldamers, we investigate how side chain size in *1b1s* models changes the model's minimum energy structures, energetics, and ability to form well-folded structures. We perform a series of folding searches to scan over a range of side chain sizes in a *1b1s* model. This parameter scan was inspired by the wide range of side chain sizes observed in natural amino acids. The toy models we investigate in this parameter scan do not directly correspond to amino acid secondary structure, however, the resulting structures can give insight into the effects side chains have on secondary structure propensities.

For this parameter scan we varied the side chain bond length, d_{BS}, and the side chain interaction radius, R_S^{min} in tandem from values of 0.5 to 3.0 R_B^{min} in increments of 0.25 for a foldamer with 15 residues. We use SC to indicate the single value of both R_S^{min} and d_{BS} used for a model for the remainder of the paper. 15 residue homo-polymers were used in this study as we were able to identify single folding helices at this chain length. Longer chain length simulations were tested and are discussed in the Supplemental Information section S5. Bonded modeling parameters were chosen to best replicate a 6 residues-per-turn helix with the fewest model parameters necessary. These model parameters include: backbone bond lengths $d_{BB} = R_B^{min}$, bond angle parameters of $\theta_B = 120°$ and $\theta_S = 180 - \frac{\theta_B}{2}$. Energetic parameters for these models were selected to have practically rigid bond-angles, with a $k_{\theta_B} = k_{\theta_S} = 3750\epsilon_B$ and equivalent side chain and backbone interactions, with $\epsilon_S = \epsilon_B$. No torsion potentials were used for these models, essentially modeling freely rotating torsion angles. Since there are no preferred torsion angles, we expect to find degenerate left- and right-hand helix clusters if a helix cluster is identified.

The annealing schedule used for the side chain size parameter scan used the following annealing parameters: $T_0 = 50\epsilon_B$, $r_a = 0.9$, annealing was run for $M = 50$ cycles, and MC simulations at each temperature are run for $N = 10000$ steps. Each parameter set was repeated with 100 replicas, to identify a consensus minimum energy structure. Individual simulation replicas for the side chain size parameter scan took 50 ± 21 minutes to finish on single core processes, with variations occurring because of differences in minimization times for different models. In this scan, minimum energy clusters sampled from an average of 33 out of 100 simulations, indicating excellent sampling, with the lowest parameter set sampling 4 out of 100 simulations, which indicates that the same minima can be found multiple times. Findings from this experiment are described in the results and discussion in section 3.2.

2.6.3 Internal bond-angle effect on secondary structure

In nature, there are a wide range of helices observed in biopolymer secondary structure. From 3.6 residues-per-turn α-helices in proteins to 10.5 residues-per-turn double helices in B-DNA,[84] biopolymers adopt a wide range of helical secondary structures. Inspired by the diversity of helices observed in nature, we designed a parameter scan in a *1b1s* model to explore the different helix types accessible within these toy models. The θ_B parameter most easily changes the number of residues-per-turn in this model and is the focus of the next parameter scan. For this parameter scan we varied the θ_B from $100°$ to $160°$. Other model parameters include: $d_{BB} = R_B^{min}$, $d_{BS} = R_B^{min}$ and $\theta_S = 180 - \frac{\theta_B}{2}$. Similar energetic parameters to the side chain parameter scan were selected for the bond-angle parameter scan, with $\epsilon_S = \epsilon_B$ and $k_\theta = 3750\epsilon_B$. We borrow the $R_S^{min} = r_{BS} = 1.28R_B^{min}$ side chain parameter from the side chain size parameter scan, $R_S^{min} = r_{BS} = 1.28R_B^{min}$ and $\theta_B = 120°$ had one of the more well-defined helical minimum energy clusters.

The annealing schedule used for the backbone bond angle parameter scan used the following annealing parameters: $T_0 = 50\epsilon_B$, $r_a = 0.925$, annealing was run for $M = 68$ cycles, and MC simulations at each temperature are run for $N = 10000$ steps. Each parameter set was repeated with 100 replicas, to identify a consensus minimum energy structure. Individual simulation replicas for the backbone bond angle, θ_B, parameter scan took 38 ± 9 minutes to finish on single core processes. Minimum energy clusters were sampled on average in 52 out of 100 simulations, indicating excellent sampling, with the lowest parameter set sampling 3 out of 100 simulations, indicating moderate repeatability when using a large number of replicas. Findings from this experiment are described in the results and discussion in section 3.3.

3 Results and Discussion

3.1 Diverse minimum energy structure search

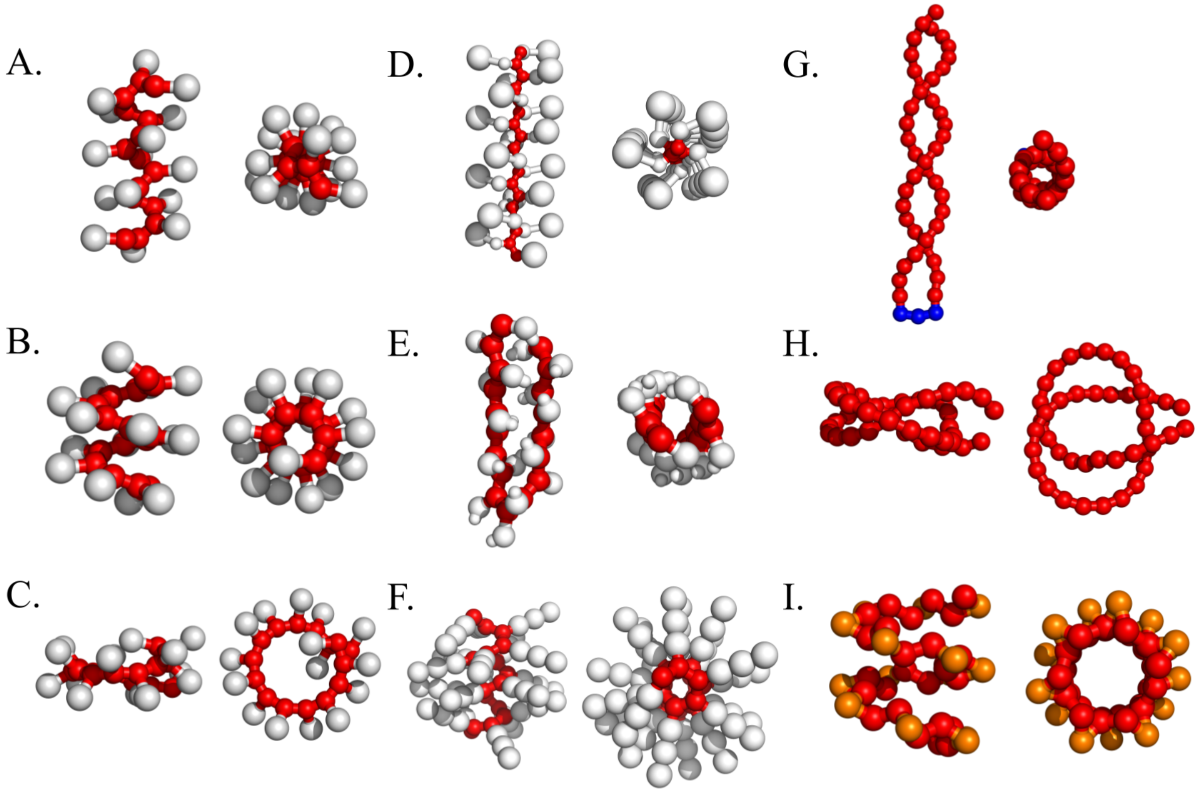


Figure 6: Structures found from a diverse range of folding simulations, with side and top views shown on the left and right, respectively. Structures A, B, D, and F are several types of helices found with homopolymer models, structure C is a loop conformation, structure E is a sheet-like fold, structure G is a double-helix, structure H is a knot-like structure, and structure I is a helix with a heterogeneous backbone. Further detail on each structure can be found in the text of this section. To aid visualization of interaction site placement, especially the backbone, CG spheres are drawn to 0.25 of their original R^{min}. Visualizations with to-scale CG spheres are presented in section S4.

Using `cg_pyrosetta` we can quickly prototype a wide range of CG models to investigate foldamer design principles. Figure 6 showcases a variety of interesting low-energy structures we are able to fold using models generated with `cg_pyrosetta`. A relatively straightforward search for interesting structures identified many types of secondary structure, such as many types of single helices, double helices, sheet-like and knot-like structures. These complex secondary structures were found by modulating parameters in fairly simple toy models. More information on how these folding simulations were set up is described in more detail in section S4.

In figure 6 structure A, B and C are examples of helix and loop configurations found with different parameters using the *1b1s* model. Structure A shows a 3.6 residues-per-turn helix, structure B shows a 6.6 residues-per-turn helix, and structure C shows an 11 residue loop. These structures were found in the θ_B parameter scan using the *1b1s* model presented in section 3.3.

Structure D, E and F shows secondary structures we found using the *1b2s* and *1b3s* models. Structure D is a 1.7 resides-per-turn helix, structure E is a folded sheet-like structure, and structure F is a 5.5 residues-per-turn helix. The longer side chains in the *1b2s* and *1b3s* models adds more conformational flexibility compared to the *1b1s* model, making models folding difficult due to many competing partially folded metastable states. To address this conformational flexibility we added model parameters that aim to reduce the conformational flexibility of these models. For example, in structure D steric clashes between large side chains reduced conformational flexibility. In model F rigid torsion constraints of the side chains also helped reduce conformational flexibility.

Structure G and H are secondary structures found using a *1b0s* chain model. Structure G is a single-stranded foldamer that folds into a double helix where each strand is a 10.9 residues per turn helix. We discovered this structure by placing a rigid hinge residue in the center of two freely rotating strands of backbone beads with $\theta_B = 155°$. Both left- and right-hand versions of these double helix structures are observed in equal frequency. Structure H is a knot-like structure found from a freely rotating strand of backbone beads with a θ_B of 155.

Structure I was generated using a *2b0s* model. Instead of having a uniform backbone geometry, there are 2 different bead types shown as red (R) and orange (O) beads arranged in a ROR repeating unit. The red beads have equilibrium $\theta_B = 120$ and the orange beads have equilibrium $\theta_B = 60$. In this model, we were able to find a 6.3 residues-per-turn helix. For this structure, rigid backbone angle and torsion potentials were used to reduce conformational flexibility and helped form a helix with a heterogeneous backbone. Additionally, this was the first example of a stable helix structure emerging from our toy foldamer model without side chain beads, indicating helix configurations could be stabilized without the need for volume exclusion from SC moieties.

The preceding structures were identified by examining the minimum energy structures of many folding simulations, with specific parameters and simulation details of provided in the supporting information (SI S4). These represent local minimum energy structures on complex potential surfaces, but may or may not represent well-folding models. For a better understanding of how well these structures might fold, one would have to identify all of the minimum energy structural clusters and determine whether they are both energetically and structurally distinct from other clusters of local minima observed in the process. The other two sections in this work perform more rigorous energetic and structural analysis to identify the best folding candidates during two parameter scans performed with `cg_pyrosetta`.

3.2 Side chain size effect on secondary structure

Our analysis workflow is able to identify and characterize a wide range of minimum energy clusters for each foldamer model in the SC parameter scan. Figure 7 shows all minimum cluster structures identified in this parameter scan except for parameter set $SC = 3.0R_B^{min}$. Parameter set $SC = 3.0R_B^{min}$ is excluded as it shares a similar structures and metrics as parameter sets $SC = 2.5R_B^{min}$ and $SC = 2.75R_B^{min}$. Structures are presented with a side-on and top-down view, along with a radar plot with all four metrics previously described. All radar plots share the same axis for comparison between SC values. Summary metrics for the entire SC parameter scan are shown in section S5 and all energy vs cluster RMSD plots are shown in section S8.

A range of ordered and disordered structures were found in this relatively simple parameter scan. Models with parameters $SC = 0.5R_B^{min}$, $SC = 0.75R_B^{min}$ and $SC = 1.0R_B^{min}$ all adopt a disordered globule minimum energy structure. These structures are compact but have no regular structure. Models with parameters $SC = 1.25R_B^{min}$ and $SC = 1.5R_B^{min}$ have helical minimum energy clusters that with 5.5 residues-per-turn and 5.4 residues-per-turn, respectively. Models with even larger side chains with parameters $SC = 1.75R_B^{min}$, $SC = 2.0R_B^{min}$ and $SC = 2.25R_B^{min}$ all have helical minimum energy clusters with 4.7 residues-per-turn, 4.6 residues-per-turn and 4.6 residues-per-turn, respectively. At large side chain sizes, like $SC = 2.5R_B^{min}$ and $SC = 2.75R_B^{min}$, we see minimum energy structures with more extended configuration, with some degree of normal SC packing, but with a fairly disordered backbone. Through modification of the R_S^{min} and d_{BS} parameters, we are able to change the minimum energy conformations of these simple models.

The transition from disordered minimum energy structures in parameters $SC = 0.5$, $SC = 0.75R_B^{min}$ and $SC = 1.0R_B^{min}$ to more ordered helices from parameters $SC = 1.25R_B^{min}$ and higher hints to a minimum necessary side chain volume to stabilize helical conformations. As these models do not have torsion potentials biasing structures into specific equilibrium helix pitches, these emergent helices are primarily driven by the side chain excluded volume and the internal backbone bond-angle of $\theta_B = 120°$. With an internal angle of 120° we expected to find helices with 6 residues-per-turn, but in practice find helices with half-increment helices (~5.5 residues-per-turn and ~4.5 residues-per-turn). These half-increment helices appear more favorable, as they allow hexagonal-like packing of the side chain beads around the helix.

We are able to identify minimum energy clusters for each of these parameter sets. However, not all parameter types yield well-defined minimum energy structures according to the four metrics described earlier. As listed in figure 7, models $SC = 0.5R_B^{min}$, $SC = 0.75R_B^{min}$

and $SC = 1.0R_B^{min}$ all have poor $RMSD_{inter}$ and $RMSD_{cluster}$ scores. Energetically these all share very small energy gap Z-score values, indicating minimum energy clusters have energetic significant overlap with other identified clusters. Parameter $SC = 1.25R_B^{min}$ has a large energy gap Z-score of 4.0, an average silhouette score of 0.76, a minimum cluster $RMSD_{cluster}$ of 0.65 R_B^{min} and a minimum cluster $RMSD_{inter}$ of 2.1 R_B^{min} indicating a very well-behaved minimum energy cluster that is both energetically and structurally distinct from other clusters. Models with parameters $SC = 1.5R_B^{min}$, $SC = 1.75R_B^{min}$ and $SC = 2.0R_B^{min}$ all have reasonably well-behaved silhouette scores and minimum cluster $RMSD_{inter}$, indicating structurally distinct minimum energy clusters. The $SC = 1.5R_B^{min}$ and $SC = 2.0R_B^{min}$ models had the two largest $RMSD_{cluster}$ observed in this parameter scan with both having $RMSD_{cluster}$ values of $0.11R_B^{min}$, indicating a rather large spread of structures for these identified clusters. Energetically, models with parameters $SC = 1.5R_B^{min}$, $SC = 1.75R_B^{min}$ and $SC = 2.0R_B^{min}$ all exhibit small to moderate energy gap Z-scores, indicating that identified minimum energy clusters have some overlap between energetic clusters. Parameter set $SC = 2.25R_B^{min}$ has a large energy gap Z-score of 3.2, but fairly poor silhouette scores, $RMSD_{cluster}$ and $RMSD_{inter}$. Parameters $SC = 2.5$, $SC = 2.75$ and $SC = 3.0$ (not shown), all have good silhouette scores, $RMSD_{cluster}$ values and $RMSD_{inter}$ values, indicating structurally well-defined clusters. However, all three have very small energy gap Z-score values, indicating there are other clusters that are energetically similar to the minimum energy clusters identified in these parameter scans. In section S7 we share several examples of cluster minimum energy structures for clusters of the next lowest energy structures, to give a structural basis for the $RMSD_{inter}$.

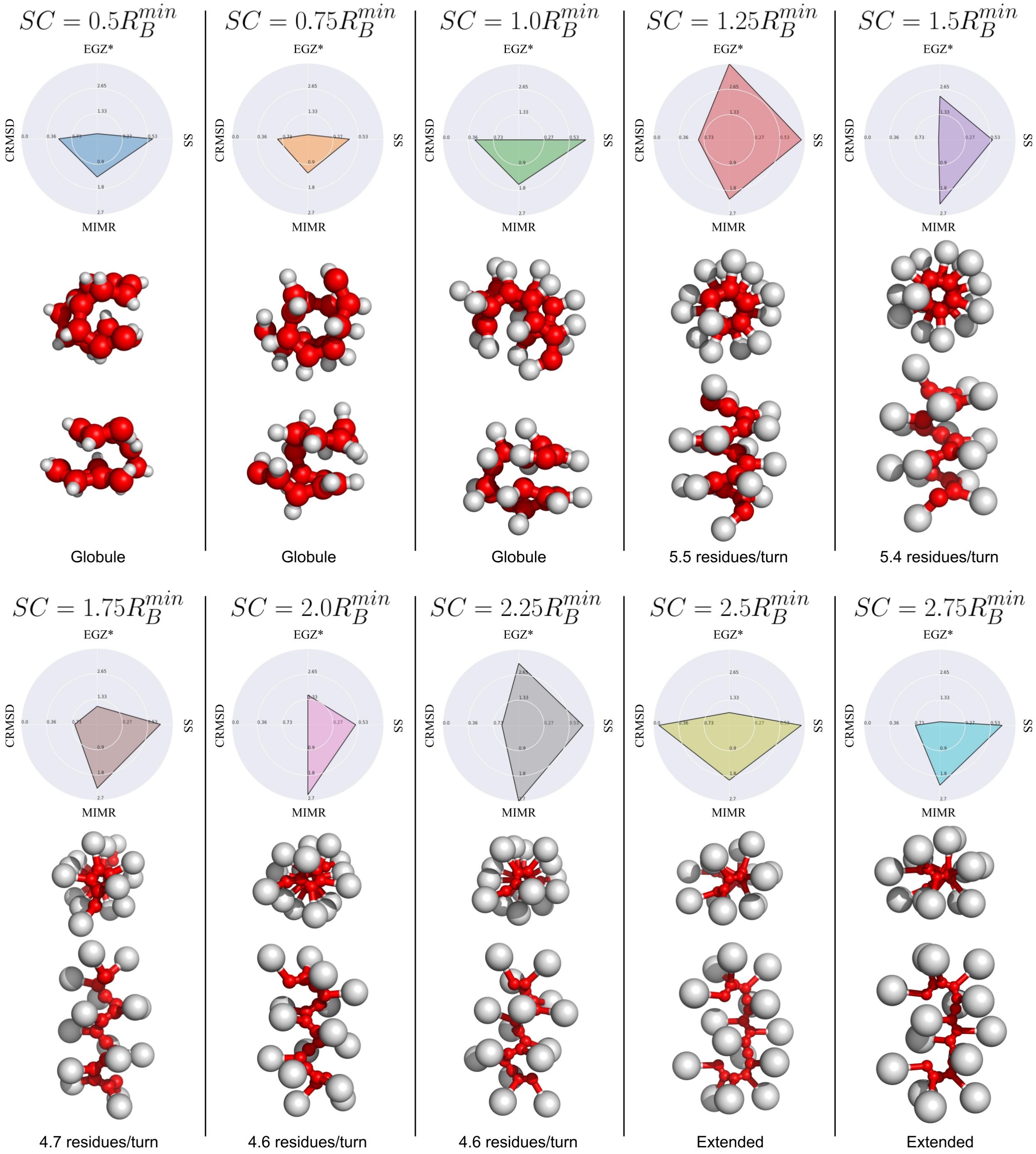


Figure 7: Minimum energy cluster minimum energy structures from the side chain size experiment (extended $SC = 3.0$ not shown), with summary metrics plotted on radar plots. CG spheres are drawn to 0.25 of their original R^{min} to help with the visualization of the backbone structure. Space-filling versions are shown in figure S3. This 1D parameter scan reveals a few types of helical structures, such as the 5.5 and 5.4 residues-per-turn helix found at $SC = 1.25R_B^{min}$ and $SC = 1.5R_B^{min}$, respectively, the 4.7 residues-per-turn helix found at $SC = 1.75R_B^{min}$ and the 4.6 residues-per-turn helix found at $SC = 2.0R_B^{min}$ and $SC = 2.25R_B^{min}$. While helices are possible low-energy structures for many of these parameter sets, not all identified helices have favorable cluster metrics. We observe that no helices were found below $R_S^{min} = 1.25R_B^{min}$.

The SC parameter scan also identifies two distinct modes of helix stabilization. Looking at the average energy of each of the minimum energy clusters over each side chain size, shown in figure 8. At $SC = 1.25R_B^{min}$ the minimum energy cluster has an average energy of around -100 ϵ_B, where at $SC = 2.25R_B^{min}$, we see an average energy of the minimum energy cluster with a value of 0 ϵ_B. Both these parameter sets achieve a large energy gap Z-score, but with different types of stabilization. With larger side chain volumes steric clashes become larger, increasing the overall energy of larger side chain models.

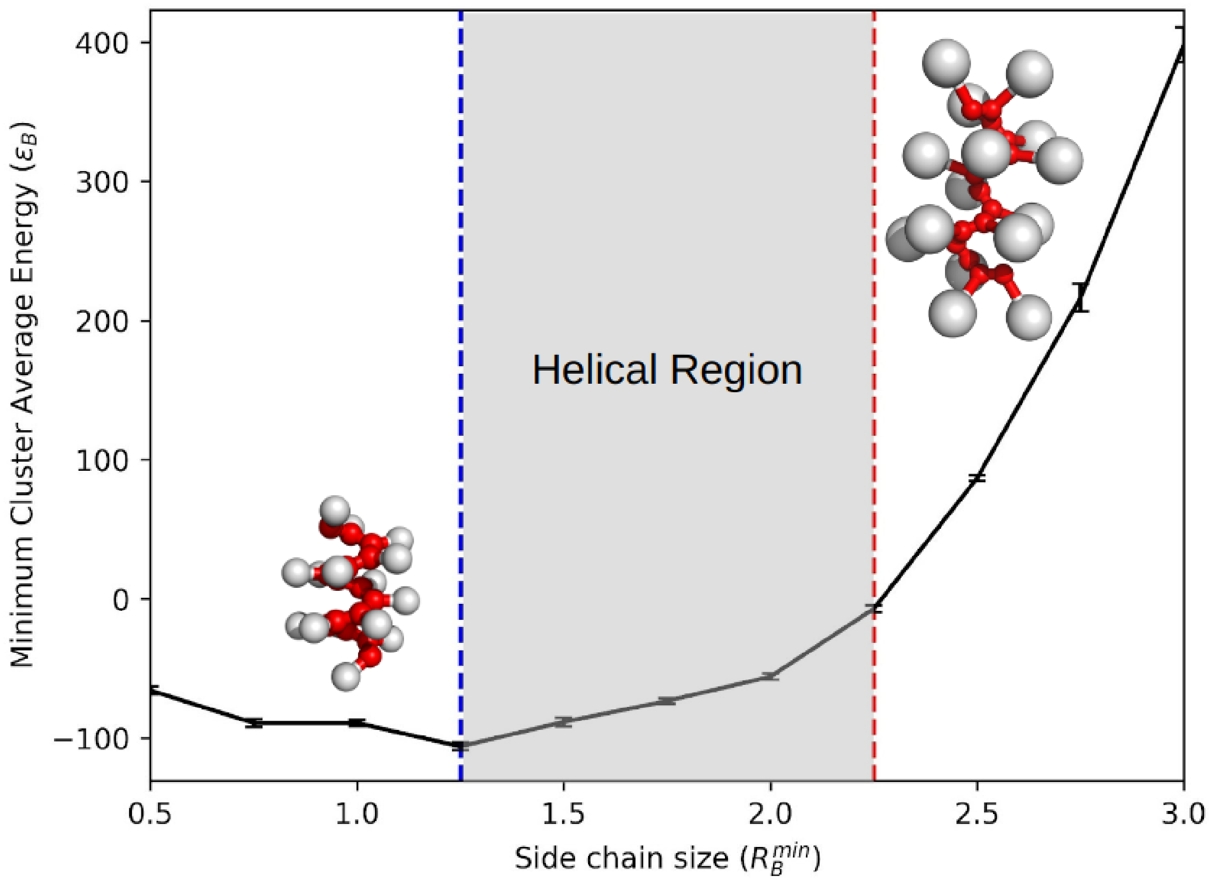


Figure 8: Average energy of all minimum energy clusters in the side chain parameter scan. The blue dashed lined denotes parameter set $SC = 1.25R_B^{min}$ and the red dashed lined denotes parameter set $SC = 2.25R_B^{min}$. These two parameters represent the two minimum energy structures with the most prominent energy gap Z-score. When $SC = 1.25R_B^{min}$, we see a minimum in the average energies, representing a side chain size where favorable interactions are being maximized. At $SC = 2.25R_B^{min}$ and larger side chains, the average energy of the minimum energy clusters are much larger in energy, indicating steric exclusions play a primary role in structure formation. Note spheres are shown with radii 0.25 of their actual value to enable the backbone structure to be seen; space-filling structures are shown in figure S3.

3.3 Internal bond-angle effect on secondary structure

For the bond-angle parameter scan we ran folding simulations for 31 sets of θ_B from $\theta_B = 100°$ to $\theta_B = 160°$ in increments of 2°. Figure 9 shows radar plots of a subset of this parameter scan, with foldamer structures and metrics being shown for increments of 4°. We show structures and metrics up to $\theta_B = 144°$, as all structures beyond this point have similar structures to the minimum energy cluster for $\theta_B = 144°$. All summary metrics for the entire parameter scan are shown in section S6 and all energy vs cluster RMSD plots for the structures shown in this parameter scan are shown in section S9.

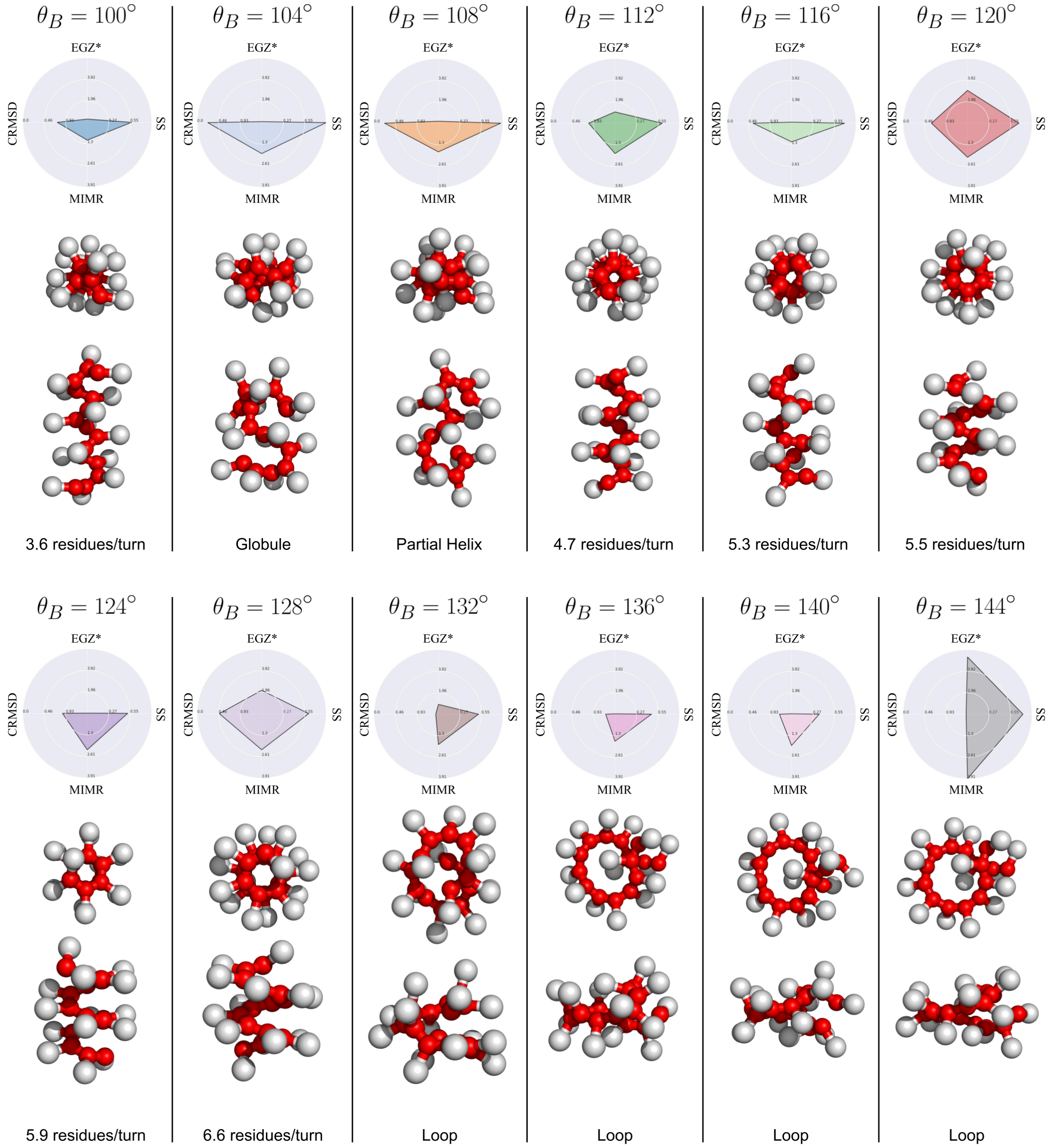


Figure 9: A subset of cluster minimum energy structures with summary metrics plotted on radar plots for the bond-angle parameter scan. CG spheres are drawn to 0.25 of their original R^{min} to help with the visualization of the backbone structure. Scanning over θ_B we identified helices from 4.2 residues-per-turn to 6.6 residues-per-turn. After the partial helix found for $\theta_B = 132^\circ$, larger θ_B models adopt an open loop configuration.

Variation of the internal backbone angle θ_B in the *1b1s* model modulates minimum energy structures through a range of different helices. For the $\theta_B = 100°$ model we identify a 3.6 residues-per-turn helix, for the $\theta_B = 112°$ model we identify a 4.7 residues-per-turn helix, for the $\theta_B = 116°$ model we identify a 5.3 residues-per-turn helix, for the $\theta_B = 120°$ model we identify a 5.5 residues-per-turn helix, for the $\theta_B = 124°$ model we identify a 5.9 residues-per-turn helix and for the $\theta_B = 128°$ model we identify a 6.6 residues-per-turn helix. At lower θ_B values when helices are not achieved we observe partial helical folds ($\theta_B = 108°$) and disordered folds ($\theta_B = 104°$). These misfolded structures arise from bonded geometry restricting the optimal arrangement of non-bonded interactions in a repeating structure. At larger θ_B values helices become less favorable and loop conformations become the dominant structure. For the 15mer foldamer models used here, loop conformations become the primary structure when there are not enough residues to form several helical turns. For models with $\theta_B > 150°$, all identified structures did not fold back onto themselves and resulted in few to no clusters.

Different internal backbone angles θ_B lend themselves to different packing of the side chains. The minimum energy interaction distance for backbone and side chain beads is $2R^{min}$. As we change the θ_B different emergent structures optimize these distances. Helices offer a periodic solution to place backbone and side chain beads near their ideal interaction distance. However, in cases like $\theta_B = 104°$ and $\theta_B = 108°$, the folding simulations are not able to identify a helix that places backbone and side chain beads at this optimal interaction distance for all contacts. In addition, the loop conformations become favorable when more ideal length interactions are satisfied in the loop conformation compared to a helical conformation.

The backbone angle θ_B also governs the energetic characteristics of the minimum energy structures identified in the bond-angle scan. The energy gap Z-score for $\theta_B = 100°$, $\theta_B = 104°$, $\theta_B = 108°$ and $\theta_B = 116°$ all have very small energy gap Z-scores. There are clusters with both helices and disordered structures with similar energies, resulting in poor energy gap Z-scores. This indicates that while helices are low-energy structures, they do not always form well-separated minima. Typical structural changes to helices that result in small energy changes tend to be misfolds of the terminal resides. The more disordered structures with poor Z-scores tend to appear at transitions between helix types. Examples of end misfolds are shown in the section S7. We identified large energy gap Z-scores for the $\theta_B = 112°$ model, $\theta_B = 120°$ model, $\theta_B = 124°$ model, $\theta_B = 128°$ model and $\theta_B = 144°$ model. At larger θ_B few to no clusters were identified, therefore no energetic analysis was performed.

We note for some examples in the θ_B parameter scans we identified cases where sampling of the minimum energy clusters did not adequately capture both left- and right-handed versions of the folded structures. We present energy vs. cluster RMSD plots for all structures presented in the two parameter scans in section S8 and section S9. One instance of poor sampling was identified where only a single helical minimum energy cluster was identified. If true exhaustive sampling was achieved, we would expect to find both left- and right-hand versions of this single helix. Identifying a single helical cluster does not change the calculated energy gap Z-score and $RMSD_{inter}$ from the lowest energy cluster to other clusters; however, it does indicate that in some cases these sampling techniques may not be able to accurately capture the other identified clusters. In particular the $\theta_B = 116°$ model (figure S23) and the $\theta_B = 128°$ (figure S26) model exhibit this poor sampling.

Another example of poor sampling was identified where energetically similar minimum energy clusters had poor overlap between their mirrored minimum energy structures. Here both left- and right-handed clusters were identified, however, due to poor overlap between the mirrored minimum energy structures of each cluster, these clusters were not labeled as mirror clusters, resulting in an extremely small energy gap Z-score where larger values would have been reported. The $\theta_B = 124°$ model (figure S25) exhibited this inadequate sampling.

4 Discussion and Conclusions

In this work, we demonstrate `cg_pyrosetta`'s ability to find both minimum energy structures and putative well-structured ensembles of a variety of toy foldamer models. We have shown that folding CG foldamer models with MC minimization with simulated annealing is a fast and effective way to identify potential global minimum energy structures over a large range of coarse-grained foldamer topologies. We presented a variety of examples of folded structures using a range of toy models. We additionally investigated two folding hypotheses relating model parameters of the *(1b1s)* model to foldamer secondary structure. We identified a rich diversity of secondary structures available to these simple models, including several single helices, double helices, sheet-like structures, and knot-like structures. We found that minimum values of R_S^{min} and d_{BS} are needed to promote helix formation and that certain side chain sizes allow for better packing. Varying θ_B quickly changed the minimum energy structures of these models from 3.6 residues-per-turn helices to more open loop configurations. Using the CG models shown in this work, we can identify structural and energetic changes made to generic macromolecules to help steer the atomistic design of novel foldamer molecules.

With `cg_pyrosetta` we can investigate further folding hypotheses with the goal to inform atomistic foldamer design. While this study was a feasibility study, further folding hypotheses can easily be explored using `cg_pyrosetta`'s framework. In this study, we explored the effects of model changes on an initial 5.5 residue per turn helix identified previously in the study. Along these lines, `cg_pyrosetta` enables exploration of folding effects of model parameters on broad families of secondary structures, giving a clearer picture of the overall effect of these model changes, including increased complexity of monomer sites to better capture molecular structure. Given how computationally inexpensive ($\leq$ 1 hour/simulation running on a single core) these simple MC simulations are, expanding the study to families of models is easily achievable.

In addition, more complex energy functions can be explored with this framework, and such additional terms in the energy functions can bring these models closer to more atomistically realizable models. Using both Lennard-Jones and Coulombic interactions, already implemented in Rosetta, one can investigate the effects of short-range versus long-range interactions on foldamer secondary structure. Directional terms, such as hydrogen bonding terms already implemented in Rosetta allow investigation of anisotropic attraction and its importance in stabilizing foldamer secondary structure. Rosetta has a modular architecture that allows other score function libraries to be easily implemented as well.

cg_pyrosetta can therefore be of use to identify minimally descriptive CG models that can accurately describe the folding of atomistic foldamer systems. These further explorations will continue to expand the number of physical hypotheses that can be explored through simpler coarse-grained, models, which can then inspire the development of atomistic resolution foldamers both computationally and experimentally.

5 Author Contributions

T.L.F., M.R.S., C.C.W., and G.A.M. conceptualized the project and designed methodology; T.L.F. developed the software; all experiments were conducted by T.L.F.; experiments were analyzed by T.L.F; T.L.F. wrote the original manuscript draft; M.R.S., C.C.W., and G.A.M. edited and reviewed the manuscript; M.R.S. supervised the project and obtained resources.

Acknowledgement

T.L.F. was in part supported by the U.S. Department of Education's Graduate Assistance in Areas of National Need (GAANN) fellowship program. This material is based on work supported by the U.S. Department of Energy Office of Science, Office of Basic Energy Sciences, Materials Science and Engineering (MSE) Division, under Award Number DE-SC0018651. This work utilized computational resources from the Colorado Boulder Research Computing Group, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, and the Bridges-2 system, which is supported by NSF ACI-1928147, both at the Pittsburgh Supercomputing Center (PSC). The authors thank Jeffery Gray, Vikram Mulligan, Andrew Watkins, and Douglas Renfrew for insightful discussion on package implementation details. The authors also thank Owen Madin for useful discussion on visualizing results and Isabell Strawn for preliminary work done on this project.

Supporting Information Available

The following files are available free of charge.

- cg_pyrosetta_paper_SI.pdf: Supplemental information detailing energy trajectory regime transitions, clustering hyperparameter selection, helix fitting, folding simulation parameters for section 3.1, foldamer chain length selection, next-lowest energy clusters and all RMSD vs. cluster energy plots for sections 3.2 and 3.3

References

(1) Hill, D. J.; Mio, M. J.; Prince, R. B.; Hughes, T. S.; Moore, J. S. A field guide to foldamers. *Chemical Reviews* **2001**, *101*, 3893–4012.

(2) Gellman, S. H. Foldamers: A Manifesto. *Accounts of Chemical Research* **1998**, *31*, 173–180.

(3) Gopalakrishnan, R.; Frolov, A. I.; Knerr, L.; Drury, W. J.; Valeur, E. Therapeutic Potential of Foldamers: From Chemical Biology Tools To Drug Candidates? *Journal of Medicinal Chemistry* **2016**, *59*, 9599–9621.

(4) Choi, S.; Isaacs, A.; Clements, D.; Liu, D.; Kim, H.; Scott, R. W.; Winkler, J. D.;

DeGrado, W. F. De novo design and in vivo activity of conformationally restrained antimicrobial arylamide foldamers. *Proceedings of the National Academy of Sciences* **2009**, *106*, 6968–6973.

(5) Tew, G. N.; Scott, R. W.; Klein, M. L.; DeGrado, W. F. De Novo Design of Antimicrobial Polymers, Foldamers, and Small Molecules: From Discovery to Practical Applications. *Accounts of Chemical Research* **2010**, *43*, 30–39.

(6) Müller, M. M.; Windsor, M. A.; Pomerantz, W. C.; Gellman, S. H.; Hilvert, D. A Rationally Designed Aldolase Foldamer. *Angewandte Chemie International Edition* **2009**, *48*, 922–925.

(7) Maayan, G.; Ward, M. D.; Kirshenbaum, K. Folded biomimetic oligomers for enantioselective catalysis. *Proceedings of the National Academy of Sciences* **2009**,

(8) Prince, R. B.; Barnes, S. A.; Moore, J. S. Foldamer-Based Molecular Recognition. *Journal of the American Chemical Society* **2000**, *122*, 2758–2762.

(9) Lee, B.-C.; Chu, T. K.; Dill, K. A.; Zuckermann, R. N. Biomimetic Nanostructures: Creating a High-Affinity Zinc-Binding Site in a Folded Nonbiological Polymer. *Journal of the American Chemical Society* **2008**, *130*, 8847–8855.

(10) Nam, K. T.; Shelby, S. A.; Choi, P. H.; Marciel, A. B.; Chen, R.; Tan, L.; Chu, T. K.; Mesch, R. A.; Lee, B.-C.; Connolly, M. D.; Kisielowski, C.; Zuckermann, R. N. Free-floating ultrathin two-dimensional crystals from sequence-specific peptoid polymers. *Nature Materials* **2010**, *9*, 454–460.

(11) Cheng, R. P.; Gellman, S. H.; DeGrado, W. F. β-Peptides: From Structure to Function. *Chemical Reviews* **2001**, *101*, 3219–3232.

(12) Seebach, D.; Hook, D. F.; Glättli, A. Helices and other secondary structures of β- and γ-peptides. *Peptide Science* **2006**, *84*, 23–37.

(13) Gorske, B. C.; Mumford, E. M.; Gerrity, C. G.; Ko, I. A Peptoid Square Helix via Synergistic Control of Backbone Dihedral Angles. *Journal of the American Chemical Society* **2017**, *139*, 8070–8073.

(14) Burkoth, T. S.; Beausoleil, E.; Kaur, S.; Tang, D.; Cohen, F. E.; Zuckermann, R. N. Toward the Synthesis of Artificial Proteins: The Discovery of an Amphiphilic Helical Peptoid Assembly. *Chemistry & Biology* **2002**, *9*, 647–654.

(15) Tošovská, P.; Arora, P. S. Oligooxopiperazines as Nonpeptidic α-Helix Mimetics. *Organic Letters* **2010**, *12*, 1588–1591.

(16) Giuliano, M. W.; Maynard, S. J.; Almeida, A. M.; Guo, L.; Guzei, I. A.; Spencer, L. C.; Gellman, S. H. A γ-Amino Acid That Favors 12/10-Helical Secondary Structure in α/γ-Peptides. *Journal of the American Chemical Society* **2014**, *136*, 15046–15053.

(17) Fisher, B. F.; Gellman, S. H. Impact of γ-Amino Acid Residue Preorganization on α/γ-Peptide Foldamer Helicity in Aqueous Solution. *Journal of the American Chemical Society* **2016**, *138*, 10766–10769.

(18) Horeau, M.; Lautrette, G.; Wicher, B.; Blot, V.; Lebreton, J.; Pipelier, M.; Dubreuil, D.; Ferrand, Y.; Huc, I. Metal-Coordination-Assisted Folding and Guest Binding in Helical Aromatic Oligoamide Molecular Capsules. *Angewandte Chemie International Edition* **2017**, *56*, 6823–6827.

(19) Ziach, K.; Chollet, C.; Parissi, V.; Prabhakaran, P.; Marchivie, M.; Corvaglia, V.; Bose, P. P.; Laxmi-Reddy, K.; Godde, F.; Schmitter, J.-M.; Chaignepain, S.; Pourquier, P.; Huc, I. Single helically folded aromatic oligoamides that mimic the charge surface of double-stranded B-DNA. *Nature Chemistry* **2018**, *10*, 511–518.

(20) Porter, E. A.; Wang, X.; Lee, H.-S.; Weisblum, B.; Gellman, S. H. Non-haemolytic β-amino-acid oligomers. *Nature* **2000**, *404*, 565–565.

(21) Ferrand, Y.; Huc, I. Designing Helical Molecular Capsules Based on Folded Aromatic Amide Oligomers. *Accounts of Chemical Research* **2018**, *51*, 970–977.

(22) Collie, G. W.; Bailly, R.; Pulka-Ziach, K.; Lombardo, C. M.; Mauran, L.; Taib-Maamar, N.; Dessolin, J.; Mackereth, C. D.; Guichard, G. Molecular Recognition within the Cavity of a Foldamer Helix Bundle: Encapsulation of Primary Alcohols in Aqueous Conditions. *Journal of the American Chemical Society* **2017**, *139*, 6128–6137.

(23) Baptiste, B.; Douat-Casassus, C.; Laxmi-Reddy, K.; Godde, F.; Huc, I. Solid Phase Synthesis of Aromatic Oligoamides: Application to Helical Water-Soluble Foldamers. *The Journal of Organic Chemistry* **2010**, *75*, 7175–7185.

(24) Zhang, A.; Ferguson, J. S.; Yamato, K.; Zheng, C.; Gong, B. Improving Foldamer Synthesis through Protecting Group Induced Unfolding of Aromatic Oligoamides. *Organic Letters* **2006**, *8*, 5117–5120.

(25) Francis, A. J.; Resendiz, M. J. E. Protocol for the Solid-phase Synthesis of Oligomers of RNA Containing a 2'-O-thiophenylmethyl Modification and Characterization via Circular Dichroism. *Journal of Visualized Experiments: JoVE* **2017**,

(26) Seo, J.; Lee, B. C.; Zuckermann, R. N. In *Comprehensive Biomaterials II*; Ducheyne, P., Ed.; Elsevier: Oxford, 2017; pp 41–66.

(27) Francis, A. J.; Resendiz, M. J. E. Protocol for the Solid-phase Synthesis of Oligomers of RNA Containing a 2'-O-thiophenylmethyl Modification and Characterization via Circular Dichroism. *Journal of Visualized Experiments : JoVE* **2017**,

(28) Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology* **2011**, *487*, 545–574.

(29) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of protein tertiary

structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology* **1997**, *268*, 209–225.

(30) Abbass, J.; Nebel, J.-C. Enhancing fragment-based protein structure prediction by customising fragment cardinality according to local secondary structure. *BMC Bioinformatics* **2020**, *21*, 170.

(31) Shapovalov, M. V.; Dunbrack, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure (London, England : 1993)* **2011**, *19*, 844–858.

(32) Chivian, D.; Baker, D. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Research* **2006**, *34*, e112.

(33) Raman, S. et al. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **2009**, *77*, 89–99.

(34) Alford, R. F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **2017**, *13*, 3031–3048.

(35) Hiranuma, N.; Park, H.; Baek, M.; Anishchenko, I.; Dauparas, J.; Baker, D. Improved protein structure refinement guided by deep learning based accuracy estimation. *Nature Communications* **2021**, *12*, 1340.

(36) Singh, A. Deep learning 3D structures. *Nature Methods* **2020**, *17*, 249–249.

(37) Berman, H. M. et al. The Protein Data Bank. *Acta Crystallographica. Section D, Biological Crystallography* **2002**, *58*, 899–907.

(38) Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D. *Methods in Enzymology*; Numerical Computer Methods, Part D; Academic Press, 2004; Vol. 383; pp 66–93.

(39) Nizami, B.; Bereczki-Szakál, D.; Varró, N.; el Battioui, K.; Nagaraj, V. U.; Szigyártó, I. C.; Mándity, I.; Beke-Somfai, T. FoldamerDB: a database of peptidic foldamers. *Nucleic Acids Research* **2020**, *48*, D1122–D1128.

(40) Liu, Z.; Abramyan, A. M.; Pophristic, V. Helical arylamide foldamers: structure prediction by molecular dynamics simulations. *New Journal of Chemistry* **2015**, *39*, 3229–3240.

(41) Elmer, S. P.; Park, S.; Pande, V. S. Foldamer dynamics expressed via Markov state models. I. Explicit solvent molecular-dynamics simulations in acetonitrile, chloroform, methanol, and water. *The Journal of Chemical Physics* **2005**, *123*, 114902.

(42) Butterfoss, G. L.; Yoo, B.; Jaworski, J. N.; Chorny, I.; Dill, K. A.; Zuckermann, R. N.; Bonneau, R.; Kirshenbaum, K.; Voelz, V. A. De novo structure prediction and experimental characterization of folded peptoid oligomers. *Proceedings of the National Academy of Sciences* **2012**, *109*, 14320–14325.

(43) Voelz, V. A.; Dill, K. A.; Chorny, I. Peptoid conformational free energy landscapes from implicit-solvent molecular simulations in AMBER. *Biopolymers* **2011**, *96*, 639–650.

(44) Eastwood, J. R. B.; Jiang, L.; Bonneau, R.; Kirshenbaum, K.; Renfrew, P. D. Evaluating the Conformations and Dynamics of Peptoid Macrocycles. *The Journal of Physical Chemistry B* **2022**, *126*, 5161–5174.

(45) Németh, L. J.; Hegedüs, Z.; Martinek, T. A. Predicting Order and Disorder for β-Peptide Foldamers in Water. *Journal of Chemical Information and Modeling* **2014**, *54*, 2776–2783.

(46) Zhao, X.; Jia, M.-X.; Jiang, X.-K.; Wu, L.-Z.; Li, Z.-T.; Chen, G.-J. Zipper-Featured δ-Peptide Foldamers Driven by Donor-Acceptor Interaction. Design, Synthesis, and Characterization. *The Journal of Organic Chemistry* **2004**, *69*, 270–279.

(47) Ahn, S.-H.; Grate, J. W. Foldamer Architectures of Triazine-Based Sequence-Defined Polymers Investigated with Molecular Dynamics Simulations and Enhanced Sampling Methods. *The Journal of Physical Chemistry B* **2019**, *123*, 9364–9377.

(48) Mazzier, D.; De, S.; Wicher, B.; Maurizot, V.; Huc, I. Interplay of secondary and tertiary folding in abiotic foldamers. *Chemical Science* **2019**, *10*, 6984–6991.

(49) Moore, T. C.; Iacovella, C. R.; McCabe, C. Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion. *The Journal of Chemical Physics* **2014**, *140*, 224104.

(50) Davtyan, A.; Dama, J. F.; Voth, G. A.; Andersen, H. C. Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence. *The Journal of Chemical Physics* **2015**, *142*, 154104.

(51) Kolinski, A. Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica* **2004**, *51*, 349–371.

(52) Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *The Journal of Chemical Physics* **2013**, *139*, 090901.

(53) Schmid, F. Toy amphiphiles on the computer: What can we learn from generic models? *Macromolecular Rapid Communications* **2009**, *30*, 741–751.

(54) Ising, E. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik* **1925**, *31*, 253–258.

(55) Onsager, L. Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition. *Physical Review* **1944**, *65*, 117–149.

(56) Dill, K. A.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemsitry, Physics, and Nanoscience*, 2nd ed.; Garland Science, Taylor & Francis Group, LLC, 2010; pp 253–280.

(57) Lau, K. F.; Dill, K. A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* **1989**, *22*, 3986–3997.

(58) Yue, K.; Fiebig, K. M.; Thomas, P. D.; Chan, H. S.; Shakhnovich, E. I.; Dill, K. A. A test of lattice protein folding algorithms. *Proceedings of the National Academy of Sciences* **1995**, *92*, 325–329.

(59) Stillinger, n.; Head-Gordon, n.; Hirshfeld, n. Toy model for protein folding. *Physical Review. E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **1993**, *48*, 1469–1477.

(60) Kaufmann, K. W.; Lemmon, G. H.; DeLuca, S. L.; Sheehan, J. H.; Meiler, J. Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry* **2010**, *49*, 2987–2998.

(61) Chaudhury, S.; Lyskov, S.; Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **2010**, *26*, 689–691.

(62) Levitt, M.; Warshel, A. Computer simulation of protein folding | Nature. *Nature* **1975**, *253*, 694–698.

(63) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *The Journal of Physical Chemistry B* **2007**, *111*, 7812–7824.

(64) Rose, G. D.; Fleming, P. J.; Banavar, J. R.; Maritan, A. A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences* **2006**, *103*, 16623–16633.

(65) Schrödinger, LLC,

(66) Adorf, C. S.; Dodd, P. M.; Ramasubramani, V.; Glotzer, S. C. Simple data and workflow management with the signac framework. *Computational Materials Science* **2018**, *146*, 220–229.

(67) Li, Z.; Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences of the United States of America* **1987**, *84*, 6611–6615.

(68) Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *The Journal of Physical Chemistry A* **1997**, *101*, 5111–5116.

(69) Leary, R. H. Global Optimization on Funneling Landscapes. *Journal of Global Optimization* **2000**, *18*, 367–383.

(70) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *Journal of Chemical Theory and Computation* **2007**, *3*, 2312–2334.

(71) McGibbon, R.; Beauchamp, K.; Harrigan, M.; Klein, C.; Swails, J.; Hernández, C.; Schwantes, C.; Wang, L.-P.; Lane, T.; Pande, V. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **2015**, *109*, 1528–1532.

(72) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON* **2011**, 6.

(73) Walker, C. C.; Meek, G. A.; Fobe, T. L.; Shirts, M. R. analyze_foldamers. 2022; `https://github.com/shirtsgroup/analyze_foldamers`.

(74) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon, 1996; pp 226–231.

(75) Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P.; Xu, X. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems* **2017**, *42*, 1–21.

(76) Lemke, O.; Keller, B. G. Density-based cluster algorithms for the identification of core sets. *The Journal of Chemical Physics* **2016**, *145*, 164104.

(77) Walker, C. C.; Meek, G. A.; Fobe, T. L.; Shirts, M. R. Using a Coarse-Grained Modeling Framework to Identify Oligomeric Motifs with Tunable Secondary Structure. *Journal of Chemical Theory and Computation* **2021**, *17*, 6018–6035.

(78) Wang, K.; Yang, Y.; Chodera, J. D.; Shirts, M. R. Identifying ligand binding sites and poses using GPU-accelerated Hamiltonian replica exchange molecular dynamics. *Journal of computer-aided molecular design* **2013**, *27*, 989–1007.

(79) Zhang, Y.; Skolnick, J. SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry* **2004**, *25*, 865–871.

(80) Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **2020**, *17*, 261–272.

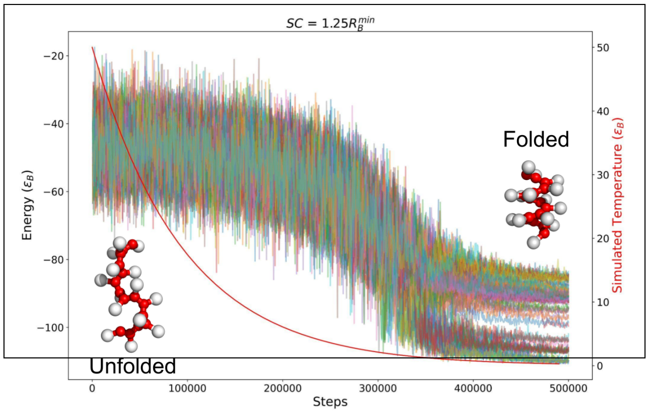
(81) Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **1987**, *20*, 53–65.

(82) Eberly, D. Fitting 3D Data with a Helix. 4.

(83) Eberly, D. Least Squares Fitting of Data by Linear or Quadratic Structures. 55.

(84) Potaman, V. N.; Sinden, R. R. *DNA: Alternative Conformations and Biology*; Landes Bioscience, 2013.

Graphical TOC Entry



Supporting Information for “Folding Coarse-Grained Oligomer Models with PyRosetta”

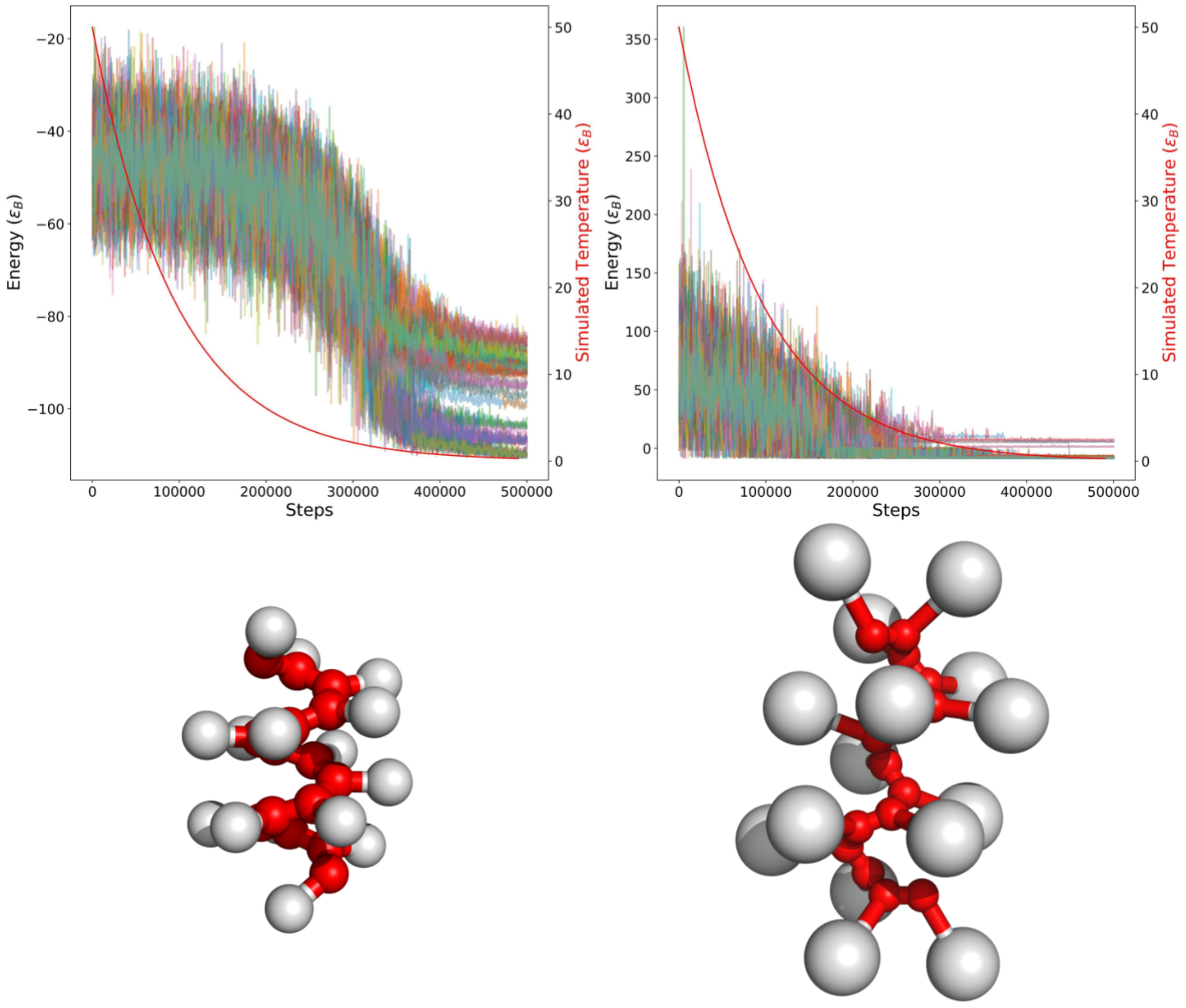
Theodore L. Fobe, Christopher C. Walker, Garrett A. Meek, and Michael R. Shirts*

Department of Chemical and Biological Engineering, University of Colorado Boulder, Boulder, CO

E-mail: michael.shirts@colorado.edu

S1 Energy Sampling Regime Transitions

The transition between the exploratory and exhaustive regimes is both model-dependent and annealing-parameter-dependent. From the side chain parameter scan, we share the energy trajectory of models $SC = 1.25R_B^{min}$ and $SC = 2.25R_B^{min}$ in figure S1. Both models share the same annealing parameters but have characteristically different energy trajectories.



$SC = 1.25R_B^{min}$ $SC = 2.25R_B^{min}$

Figure S1: Energy trajectories and visualisations of the identified minimum energy clusters from $SC = 1.25R_B^{min}$ and $SC = 2.25R_B^{min}$ models. The space-filling spheres are drawn match to the R_S^{min} of each model. The energy trajectory of the $SC = 1.25R_B^{min}$ model has a more pronounced transition between the exploratory and exhaustive phases compared to the $SC = 2.25R_B^{min}$ model.

In the $SC = 1.25R_B^{min}$ model, the side chain R_S^{min} enables energetically favorable packing of side chains. This is reflected in the energy trajectory with a large decrease in energy as the simulation transitions from the exploratory phase to the exhaustive phase. New energetic minima are found as the simulation exhaustively explores the side chain interactions that give the most favorable energies.

In the $SC = 2.25R_B^{min}$ model, the larger side chain R_S^{min} causes steric clashes between side chains resulting in energetically unfavorable packing of the side chains. This is reflected in the energy trajectory with a nominal decrease in energies from the exploratory phase to the exhaustive phase. Similar energetic states are sampled at the high-temperatures and low-temperature phases, indicating that unfolded structures share similar energies to the folded structures.

S2 Hyperparameter Selection

Selecting hyperparameters for a clustering algorithm takes careful consideration. In scikit-learn's implementation of DBSCAN, the two clustering hyperparameters of interest are ϵ and `minSamples`.[72] The clustering ϵ is the distance cutoff between entries within the RMSD matrix used in clustering and is related to the distances used in our folding simulation. We select ϵ values to reflect distance on a similar length-scale of the foldamer model. For this reason, we ensure that clustering ϵ are never less than half of the shortest bond length in the foldamer. This heuristic helps ensure that clusters remain within 1–2 bond lengths from the minimum energy structure in the cluster. We select a minimum number of samples per cluster (`minSamples`) value such that the resulting clusters are not very noisy. We found values between 0.2% to 2% of the total number of frames input into the clustering gave decent clustering.

To select optimal parameters we performed a grid search of clustering hyperparameters with values of ϵ from 0.5 to 4.0 d_B and values of `minSamples` from 50 to 350. For a given parameter set, we plot the resulting average silhouette score against the number of clusters to visualize trade-offs between these two metrics. An example of this Pareto-like plot is shown in figure S2. We exclude identified clustering that gave 3 or fewer clusters because we want to exclude clustering that resulted in two large degenerate clusters and a noise cluster.

We found that this range of hyperparameters populates clustering values near the Pareto front, represented by points at the bottom left corner of figure S2.

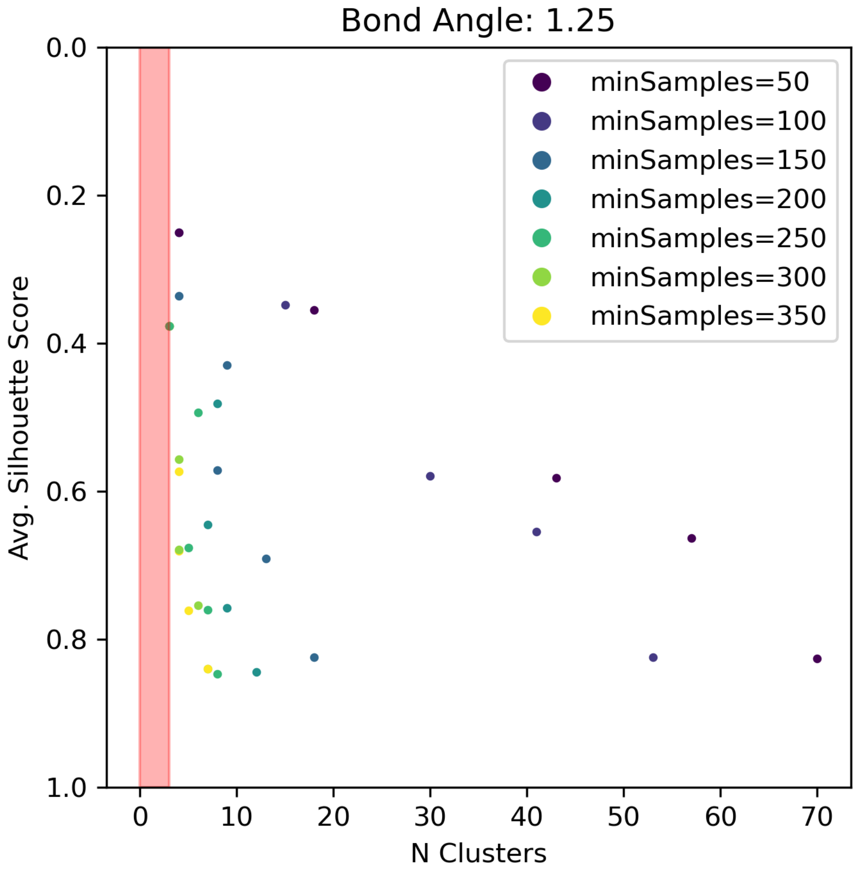


Figure S2: Pareto-like plot used to evaluate optimal clustering hyperparameters. Points that fall in the bottom left corner are optimal clustering solutions that minimize the number of clusters and maximize the average silhouette score. Points that fall within the clear red area have less than 3 identified clusters and are excluded. `minSamples` (labeled) are varied from 50 to 350, while clustering ϵ (not labeled) are varied from 0.5 R_B^{min} to 2.0 R_B^{min}.

With the insight gained from these Pareto-like plots, we define an objective function to select clustering hyperparameters on the Pareto front. We define and select the clustering hyperparameters that minimize the L2 norm of the normalized average silhouette scores and number of clusters, using

$$f(x_1, x_2) = \sqrt{x_1^2 + x_2^2} \tag{1}$$

where x_1 is the normalized $1 - (\text{Avg. Silhouette Score})$ and x_2 is the normalized N Clusters. We invert the axis of x_1 to ensure we are maximizing the average silhouette score. Both sets of metrics are normalized to ensure both metrics are considered equally.

S3 Helix Fitting

To identify and classify helical structures, we perform a least-squares fit of the backbone coordinates of structures of interest to the helix equation. For each structure, we remove the terminal 2 residues from each end to limit the least-squares fit to the internal residues. Additionally, we scale the coordinates by a factor of 100, to help with numerical stability. We first fit the backbone coordinates to an infinite cylinder, defined by a cylinder axis containing point $\boldsymbol{C}$ and having unit-length direction $\boldsymbol{W}$ and radius r. The least-squares error function for defining this cylinder is shown below:

$$E(r^2, \boldsymbol{C}, \boldsymbol{W}) = \sum_{i=1}^{n} \left[(\boldsymbol{X}_i - \boldsymbol{C})^T \left(I - \boldsymbol{W}\boldsymbol{W}^T \right) (\boldsymbol{X}_i - \boldsymbol{C}) - r^2 \right]^2 \tag{2}$$

where n is the number of coordinates and $\boldsymbol{X}_i$ is a coordinate vector.

Once the center and axis of the helix is identified, we rotate the original coordinates to orient the helical axis with the z axis, then we translate the coordinates such that the minimum z-coordinate is placed set to 0. Since the helix equation has periodic symmetry, we expect to find multiple equivalent minima at multiples of the angular frequency. To account for these multiple minima, we add a quadratic angular frequency term to penalize the fit for having angular frequencies that are too large. We fit the coordinates to the helix equation by minimizing the least-square error function shown below:

$$E(\omega, \phi) = \sum_{i=1}^{n} (x_i - r\cos(\omega z_i + \phi))^2 + (y_i - r\sin(\omega z_i + \phi))^2 + \omega^2 \tag{3}$$

where n is the number of coordinates, x_i, and y_i is the x and y coordinate components,

ω is the angular frequency and ϕ is the phase shift of the fitted helix. We use the radius calculated from the cylindrical fitting to reduce the number of parameters needed for this fit.

First, the cylindrical error (equation (2)) is minimized using SciPy's L-BFGS minimizer to find the helical axis. Then we minimize the helix error (equation (3)) using SciPy's basin hopping algorithm with L-BFGS minimization to rigorously search the helix parameter space.[80] Results for these two least-square fittings are reported in root-mean-square errors, $RMSE_{cyl}$ and $RMSE_{helix}$, for the cylindrical and helical fitting respectively.

Below we present the helix RMSE, cylinder RMSE and reported residues per turn for all structures presented in this work. Table 1 shows the helix fitting for the side chain size parameter scan (SC), table 2 shows the helix fitting for the internal bond-angle parameter scan (θ_B) and table 3 shows helix fits for structures presented in the diverse fold section.

Table 1: Helix RMSE, cylinder RMSE and residues per turn from the least-squares fit of minimum energy structures found the in SC size parameter scan (section 3.2). Table rows are colored green for clusters that were visually identified to have helical minimum energy structures. Structures with unfolded terminal residues have larger $RMSE_{helix}$ and $RMSE_{cyl}$ values.

$SC(R_B^{min})$	Helix RMSE (R_B^{min})	Cylinder RMSE (R_B^{min})	Residues/Turn
0.50	9.385	27.664	1.35
0.75	9.274	9.433	0.16
1.00	11.710	15.805	0.41
1.25	0.331	0.119	5.54
1.50	0.634	0.691	5.44
1.75	0.439	0.850	4.71
2.00	0.472	0.261	4.64
2.25	0.523	0.191	4.61
2.50	2.779	16.250	0.58
2.75	2.523	15.698	0.29
3.00	2.547	15.506	0.29

Table 2: Helix RMSE, cylinder RMSE and residues per turn from the least-squares fit of minimum energy structures found in the θ_B parameter scan (section 3.3). Table rows are colored green for clusters that were visually identified to have helical minimum energy structures. Structures with unfolded terminal residues have larger $RMSE_{helix}$ and $RMSE_{cyl}$ values.

θ_B	Helix RMSE (R_B^{min})	Cylinder RMSE (R_B^{min})	Residues/Turn
100	0.028	0.050	3.62
102	0.709	3.035	11.09
104	6.305	31.781	10.75
106	7.312	31.320	0.95
108	5.742	22.286	0.20
110	0.263	0.336	4.55
112	0.214	0.429	4.66
114	6.330	25.106	4.86
116	3.310	3.390	5.28
118	2.359	2.133	5.31
120	0.370	0.377	5.54
122	0.026	0.032	5.76
124	1.023	0.719	5.89
126	2.301	1.788	6.42
128	0.332	0.264	6.55
130	3.379	1.028	7.38
132	13.180	13.804	1.11
134	10.555	14.951	0.51
136	11.290	6.768	1.75
138	10.608	5.131	1.77
140	11.605	5.082	0.29
142	9.719	2.687	14.09
144	5.589	2.643	11.99
146	9.731	7.571	15.78
148	3.940	2.254	13.91
150	14.260	0.916	0.29

Table 3: Helix RMSE, cylinder RMSE and residues per turn from the least-squares fit of minimum energy structures found in the diverse fold section of the main text (section 3.1)

Fold ID	Helix RMSE (R_B^{min})	Cylinder RMSE (R_B^{min})	Residues/Turn
A	0.028	0.050	3.62
B	0.332	0.264	6.55
D	1.07	2.27	1.66
F	3.23	10.10	5.51
G	5.76	4.71	10.86
I	3.43	0.23	6.31

S4 Diverse range of folding simulations simulation parameters

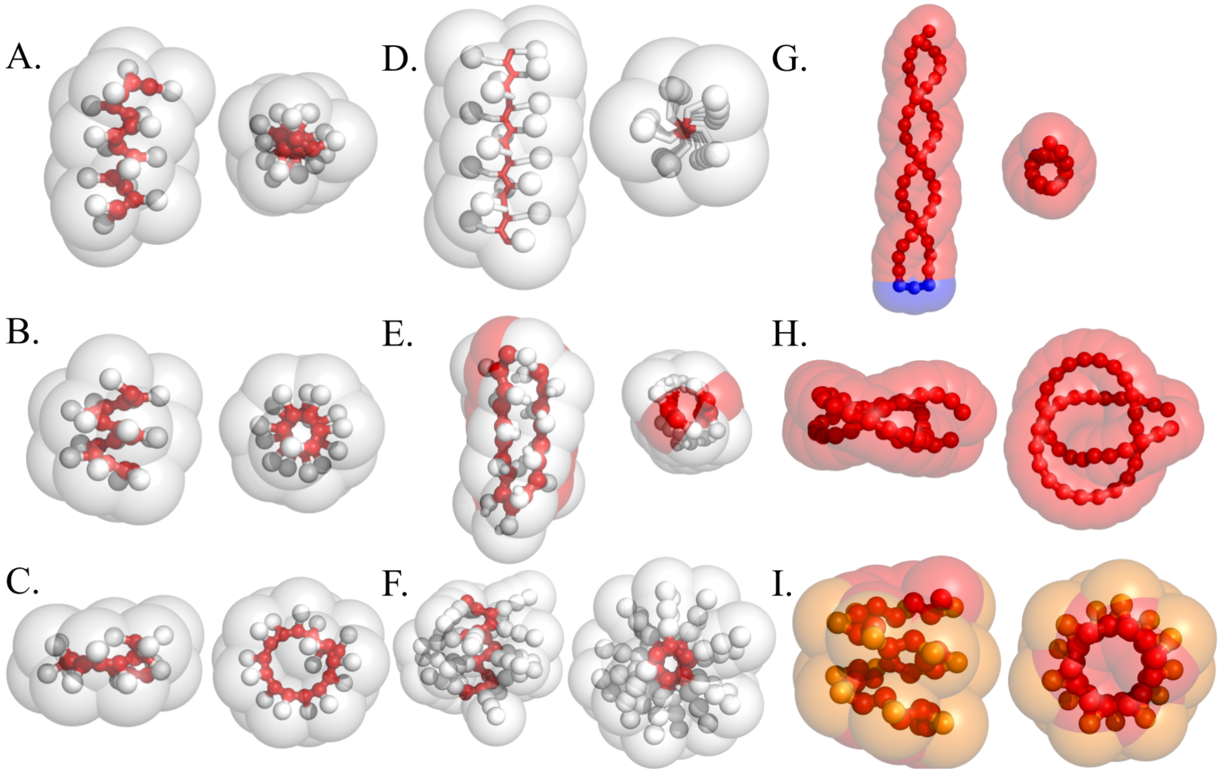


Figure S3: Structures presented in section 3.1 with transparent CG bead radii drawn to scale. Drawing full-scale CG bead radii show the bead packing achieved in these structures. Structures A, B, D, and F are several types of helices found with homopolymer models, structure C is a loop conformation, structure E is a sheet-like fold, structure G is a double-helix, structure H is a knot-like structure, and structure I is a helix with a heterogeneous backbone. Simulation and model parameters of each structure can be found below.

S4.1 1 back-bone / 1 side chain

Structures A, B, and C are helices found while varying parameters in the *1b1s* model. All structures have the same Lennard-Jones parameters where backbone beads have a $R_B^{min} = 1$ a and $\epsilon_B = 1$ and side-chain beads have a $R_S^{min} = 1.27$ and $\epsilon_S = 1$. All bond-lengths are kept rigid at $d_B = R_B^{min}$ and $d_S = 1.27R_B^{min}$. No torsion potentials were applied to these foldamers, allowing a freely rotating backbone. Structure A has a $\theta_B = 100°$ and $\theta_S = 130°$, structure B has a $\theta_B = 120°$ and $\theta_S = 120°$ and structure C has a $\theta_B = 150°$ and $\theta_S = 105°$. All structures have a bond-angle energy parameter of $k_\theta = 7500\epsilon_B$, making all angles effectively rigid. These structures were selected from the two 1D parameter scans as they showcased the range of helices we were able to achieve within a single residue topology.

Annealing parameters for helices A, B, and C are the same as those presented in the bond-angle parameter scan in the main text. The annealing schedule used for the backbone bond angle parameter scan used the following annealing parameters: $T_0 = 50\epsilon_B$, $r_a = 0.925$, annealing was run for $M = 68$ cycles, and MC simulations at each temperature are run for $N = 10000$ steps. Each parameter set was repeated with 100 replicas, to increase the chance we have a consensus minimum energy structures.

S4.2 1 back-bone / 2 side chain

Structure D and E both are examples of folds using the *1b2s* model. S_1 and S_2 denote the parameters involving the first and second side-chain bead respectively. Similar to *1b1b*, model parameters involving back-bone beads are denoted with a B. We use base length units the backbone $R_B^{min} = 1$ and energy units of the backbone $\epsilon_B = 1$. These structures were found in a parameter scan where the size of the terminal side-chain bead for each residue was varied from $R_{S_2}^{min} = r_{S_2} = 0.5R_B^{min}$ to $R_{S_2}^{min} = r_{S_2} = 2.0R_B^{min}$. Other non-bonded parameters (R^{min} and ϵ) and bond-lengths (r) were held constant in this scan at $R_{S_1}^{min} = r_{S_1} = R_B^{min}$, $r_B = R_B^{min}$ and $\epsilon_{S_1} = \epsilon_{S_2} = \epsilon_B$. Bond-angle equilibrium values were set to $\theta_B = \theta_{S1} = \theta_{S2} = 120°$. Bond-angle force constants were the same $k_\theta = 7500\epsilon_B$. A torsion potential is used on the backbone beads to restrict the available conformations the foldamer could adopt with parameters, $k_{\phi_{BBBB}} = 6.0\epsilon_B$, $n_{\phi_{BBBB}} = 1$ and $\phi_{BBBB,o} = 30°$.

The annealing schedule used for the two side chain parameter scan used the following annealing parameters: $T_0 = 50\epsilon_B$, $r_a = 0.9$, annealing was run for $M = 60$ cycles, and MC simulations at each temperature are run for $N = 20000$ steps. Each parameter set was repeated with 100 replicas, to increase the chance we have a consensus minimum energy structures.

S4.3 1 back-bone / 3 side chain

Structure F shows a folded helix using the *1b3s* model. We use S_1, S_2, and S_3 to denote the parameters involving the first, second and third side-chain bead respectively. Similar to the *1b1b* model parameters involving back-bone beads are denoted with a B. This structure was found in a parameter scan varying the side-chain bond angles, θ_{S2} and θ_{S3} in tandem from 90° to 180°. Lennard-Jones parameters for all beads were set to $R_B^{min} = R_{S_1}^{min} = R_{S_2}^{min} = R_{S_3}^{min} = 1$ and $\epsilon_B = \epsilon_{S_1} = \epsilon_{S_2} = \epsilon_{S_3} = 1$, with bond-lengths of $d_B = r_{S_1} = r_{S_2} = r_{S_3} = 1$. The remaining bond-angle parameters were set to $\theta_B = 120°$ and $\theta_{S_1} = 120°$. All bond-angle parameters had a energy constant of $k_\theta = 7500\epsilon_B$, having effectively rigid bond-angles. A torsion potential was applied to the side-chain beads to minimize their configurational flexibility. The side-chain torsion potential had parameters, $k_{\phi_{SSSB}} = 100.0\epsilon_B$, $n_{\phi_{SSSB}} = 1$ and $\phi_{SSSB,o} = 180°$. The strong side-chain torsion potential keeps the side-chain beads rigid during the folding simulation, greatly reducing the available conformations the foldamer can adopt. A less strong torsion was applied to the backbone beads with parameters, $k_{\phi_{BBBB}} = 25.0\epsilon_B$, $n_{\phi_{SSSB}} = 1$ and $\phi_{SSSB,o} = 0°$. This backbone potential was applied to have the foldamer model sample more extended configurations.

The annealing schedule used for the three side chain parameter scan used the following annealing parameters: $T_0 = 50\epsilon_B$, $r_a = 0.95$, annealing was run for $M = 120$ cycles, and MC simulations at each temperature are run for $N = 20000$ steps. Each parameter set was repeated with 100 replicas, to increase the chance we have a consensus minimum energy structures.

S4.4 Double helix

Structure G is a double helix folded with a 1-backbone bead (denoted as B) model with an added rigid hinge residue (denoted as H). These double helix structures were found by scanning through bond-angle parameters of the hinge residue shown in blue. To keep the hinge residue rigid a strong torsion potential with parameters, $k_{\phi_{BHHH}} = 100.0\epsilon_B$, $n_{\phi_{BHHH}} = 1$ and $\phi_{BHHH,o} = 180°$. Bond angle parameters of the hinge parameters were chosen to keep the residues emerging from the hinge parallel with each other. The H-H-H bond-angle parameters were varied from $\theta_{HHH} = 60°$ to $\theta_{HHH} = 160$, while was set to $\theta_{BHH} = 360° - 2\theta_{HHH}$. The backbone and hinge beads have the same Lennard-Jones parameters $R_B^{min} = R_H^{min} = 1$ and $\epsilon_B = \epsilon_H = 1$. All bead-to-bead bond-lengths were set to $d_{BB} = d_{BH} = d_{HH} = 1$. The backbone beads have bond-angle parameters of $\theta_B = 160°$ and $k_{\theta_B} = 4000\epsilon_B$ giving rigid bond-angles. The hinge residue angles have rigid force constants of $k_{\theta_H} = 4000\epsilon_B$.

The annealing schedule used for the two side chain parameter scan used the following annealing parameters: $T_0 = 50\epsilon_B$, $r_a = 0.9$, annealing was run for $M = 50$ cycles, and MC simulations at each temperature are run for $N = 20000$ steps. Each parameter set was repeated with 100 replicas, to increase the chance we have a consensus minimum energy structures.

S4.5 Knot-like structures

Structure H is a knot folded with the 1-backbone bead model, and was found through a bond-angle parameter scan from $\theta_B = 145°$ to $\theta_B = 160°$ with a energy constant $k_{\theta_B} = 4000\epsilon_B$. In this relatively small scan, we found several examples of different knots. The chain beads have Lennard-Jones parameters of $R_B^{min} = 1$ and $\epsilon_B = 1$. All bond-lengths were all set to $d_B = 1$. These models had no torsion parameters and could freely rotate.

The annealing schedule used for the two side chain parameter scan used the following annealing parameters: $T_0 = 50\epsilon_B$, $r_a = 0.9$, annealing was run for $M = 50$ cycles, and

MC simulations at each temperature are run for $N = 20000$ steps. Each parameter set was repeated with 200 replicas, to increase the chance we have a consensus minimum energy structures.

S4.6 Heterogeneous backbone helix

Structure I is a $i \to i + 6.5$ helix folded with the 2-backbone bead model. Here we used CG parameters to try and replicate the torsions and bond-angles found in aromatic oligomers. Beads representing aromatic groups with ortho- and meta-bonding are represented by blue and red beads respectively. Ortho-bonding beads are colored orange and are denoted with O and meta-bonding beads are colored red and denoted by R. The internal angle for ortho-beads is $\theta_{ROR} = 60°$ and the angle for the meta-beads is $\theta_{RRO} = 120°$. These both have force constants of $k_\theta = 4000\epsilon_R$. All bead-to-bead bond-lengths were set to $d_{MM} = r_{RO} = r_{OO} = 1$. All beads share the same pair-wise LJ parameters of $R_R^{min} = R_O^{min} = r_{RR}$ and $\epsilon_R = \epsilon_O = 1$. Torsions were selected to resemble the planar nature of aromatic groups with, $\phi_{RORR} = \phi_{ORRO} = 180°$, with periodicity of $n = 2$. All torsions have a force constant of $k_\phi = 20\epsilon_R$.

The annealing schedule used for the two side chain parameter scan used the following annealing parameters: $T_0 = 50\epsilon_B$, $r_a = 0.9$, annealing was run for $M = 50$ cycles, and MC simulations at each temperature are run for $N = 20000$ steps. Each parameter set was repeated with 100 replicas, to increase the chance we have a consensus minimum energy structures.

S5 Foldamer chain length

The side-chain experiment was carried out at several different chain lengths to assess the optimal chain length for folding these toy foldamer models. *1b1s* homopolymers with lengths of 15 residues, 18 residues, 21 residues, 24 residues, 27 residues and 30 residues were folded with side-chain folding parameters varying from $R_S^{min} = r_S = 0.5$ to $R_S^{min} = r_S = 4.0$. Other model parameters include, Lennard-Jones $\epsilon_B = \epsilon_S = 1$, harmonic bond-angle parameters of $\theta_B = 120$ with a $k_{\theta_B} = 7500$ and $\theta_S = 120$ with a $k_{\theta_B} = 7500$. These models had no torsion potentials and could freely rotate about all bonds.

The number of simulation steps was determined by scaling the 500,000 time steps used in the 15mer simulations timesteps by the number of atoms added in each of the longer models. Resulting in 600,000 time steps for the 18mer simulations, 700,000 time steps for the 21mer simulations, 800,000 time steps for the 24mer simulations, 900,000 time steps for the 27mer simulations, and 1,000,000 time steps for the 30mer simulations.

At each chain length in this experiment, all identified minimum energy clusters were correctly formed helices. We ultimately decided to only use the 15mer model since similar secondary structure had been observed at all chain lengths. At longer chain lengths we observed more identified structures with back-folding of terminal residues, shown in figure S4, often resulting in a larger number of identified clusters. Back-folding was not observed in shorter chain length homopolymers.

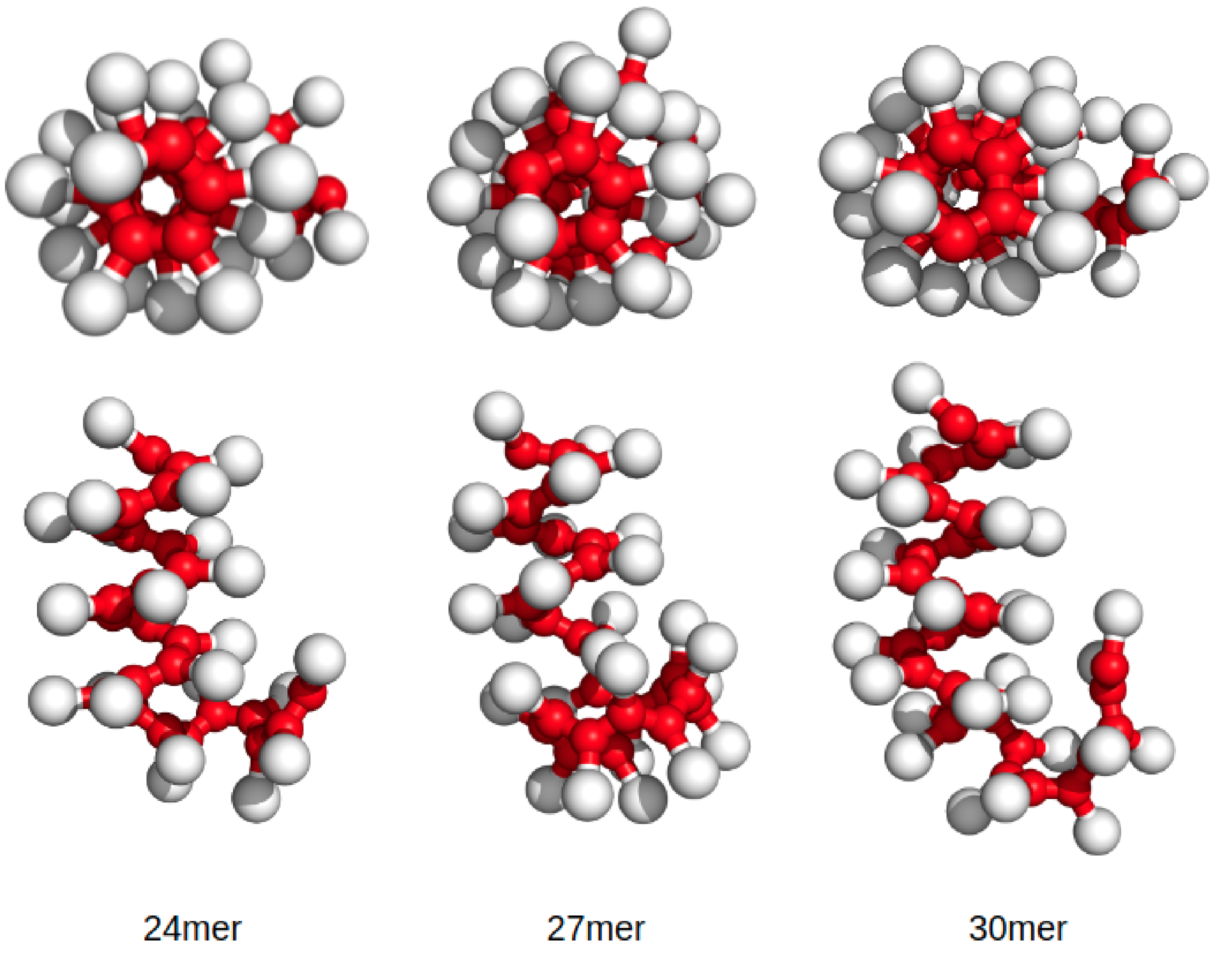


Figure S4: Selected minimum energy structures of parameter $SC = 1.28R_B^{min}$ from the chain length experiment. Misfolded helices become more prevalent in the minimum energy structures of longer chain length models. We note that all longer chain length structures identify the same 5.5 residues per turn helix identified in the main text for $SC = 1.25R_B^{min}$, just with varying degrees of terminal residues misfolding.

S6 Side chain size and bond-angle complete data sets

In the main text, for the sake of brevity, we presented subsets of the side chain size and bond-angle parameter scan in figures 7 and 9. Here we present the entire set of performance metrics (average silhouette score, energy gap Z-score, $RMSD_{cluster}$ and $RMSD_{inter}$) for all parameter sets in both sections. Metrics from both data sets are normalized using equation (4) with metric values across each data set, such that each metric is placed on a scale from 0 to 1.

For the case of $RMSD_{cluster}$ where smaller values indicated better structural clustering, we normalize the negative magnitude of this metric, so that larger bar plots represent smaller $RMSD_{cluster}$.

$$x_{norm} = \frac{x - \min x}{\max x - \min x} \tag{4}$$

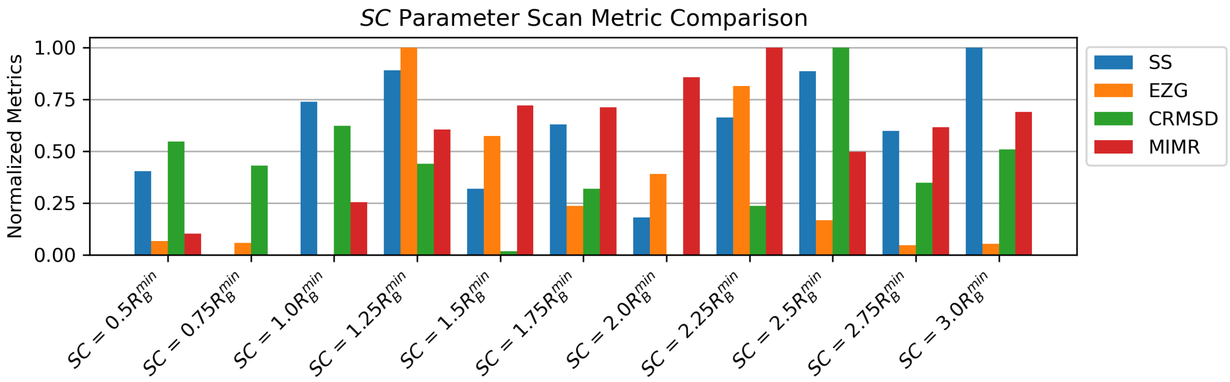


Figure S5: Summary metrics for the side chain size (SC) parameter scan. Metrics are normalized to have values of 0 to 1, where values of 1 indicate better performance. Average silhouette score is abbreviated as SS, energy gap Z-score is abbreviated as EGZ, minimum cluster $RMSD_{cluster}$ is abbreviated as CRMSD, and minimum cluster $RMSD_{inter}$ is abbreviated as MIMR.

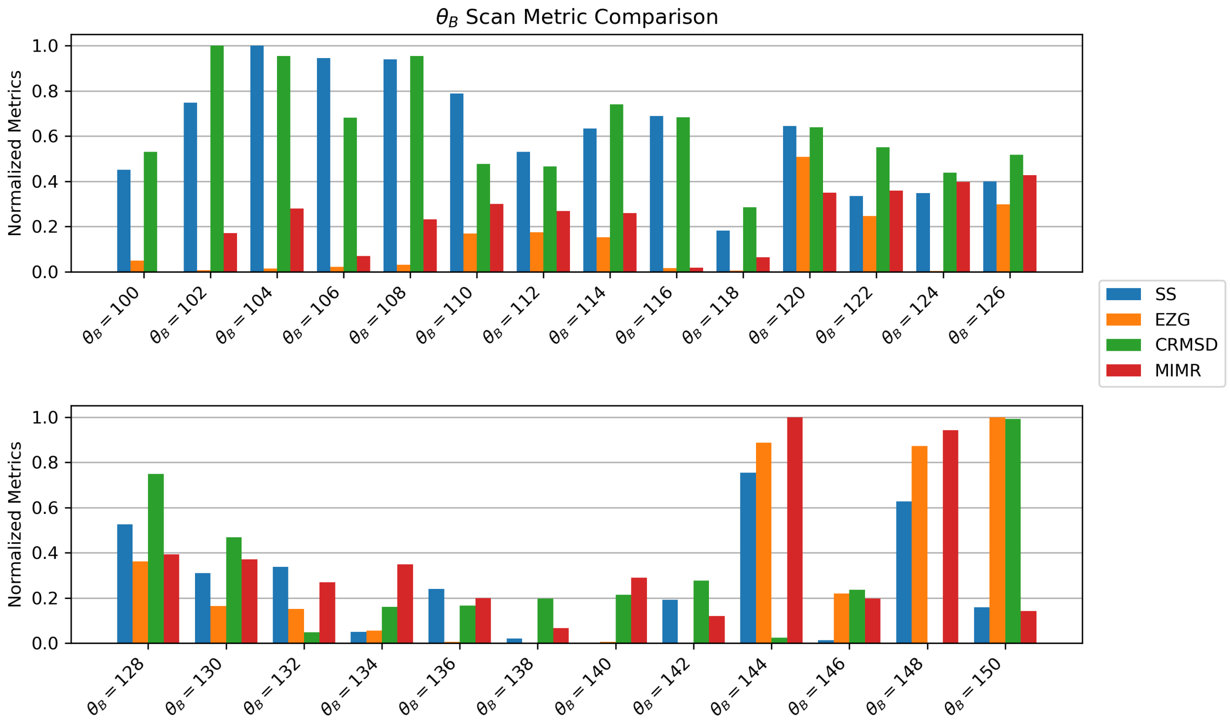


Figure S6: Summary metrics for the internal bond-angle (θ_B) parameter scan. Metrics are normalized to have values of 0 to 1, where values of 1 indicate better performance. Average silhouette score is abbreviated as SS, energy gap Z-score is abbreviated as EGZ, minimum cluster $RMSD_{cluster}$ is abbreviated as CRMSD, and minimum cluster $RMSD_{inter}$ is abbreviated as MIMR.

S7 Next-Lowest Energy Cluster Structures

In the analysis in the main text, we exclusively looked at structures of the minimum energy clusters. Summary metrics like the energy gap Z-score and $RMSD_{inter}$ are meant to give context to the energetic and structural differences between the minimum energy cluster and other identified clusters. While this analysis gives an at-a-glance confirmation of folded minimum energy clusters and their energetic and structural distance from other clusters, it does not delve into the structural changes that give the changes to some of these metrics.

Large energy gap Z-scores come from a variety of different structural changes from the minimum energy cluster. Minimum energy clusters and their next-lowest energy clusters for several models are presented in figure S7. In some cases, like that of the $SC = 1.25R_B^{min}$ model, the structure of the next lowest-energy cluster is an entirely different, non-helical structure. In the $SC = 2.25R_B^{min}$ model, the next-lowest energy cluster is very similar to the minimum energy cluster model with a handful of misfolded residues. In models with large steric clashes, like $SC = 2.25R_B^{min}$ these small structural changes can result in large energetic changes, resulting in the large energy gap Z-score of 8.13 seen in the main text. Some structures, like the $SC = 2.5R_B^{min}$ model have multiple ordered folds. The identified minimum energy cluster is an extended structure with side chains packed in an $i \to i + 2.5$ fashion with a fairly disordered backbone. The next-lowest energy cluster for the $SC = 2.5R_B^{min}$ model is a more ordered $i \to i + 4.5$ helical structure. In this case, both structures have similar energetics, with an energy gap Z-score of 0.67 between them.

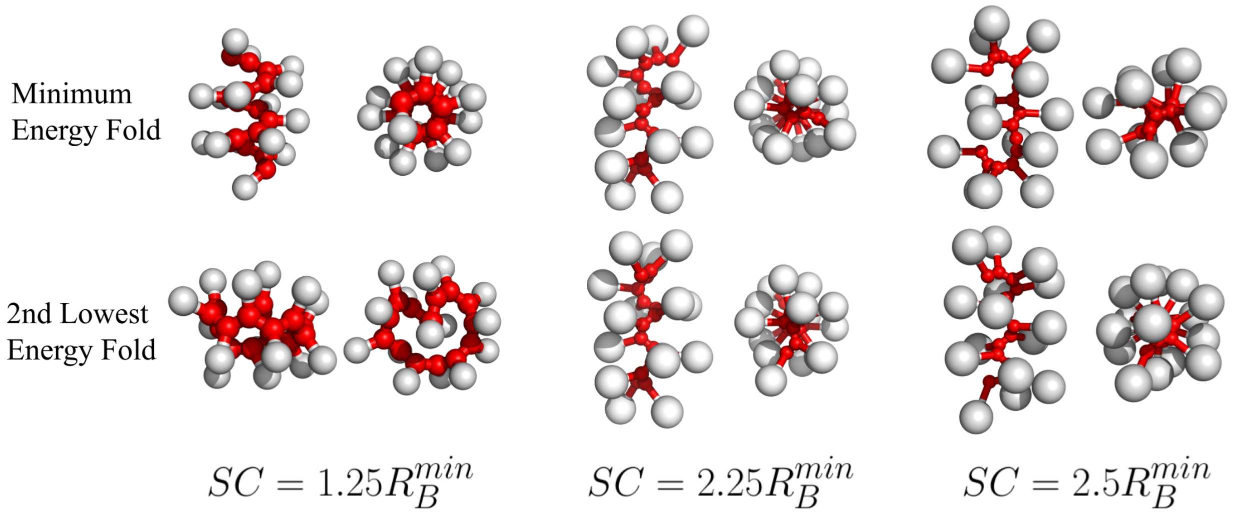


Figure S7: Comparison between the minimum energy cluster structure to the next-lowest energy cluster structure for the $SC = 1.25R_B^{min}$, $SC = 2.25R_B^{min}$ and $SC = 2.5R_B^{min}$ models.The $SC = 1.25R_B^{min}$ and $SC = 2.5R_B^{min}$ models see have large structural changes in the next-lowest energy cluster. The $SC = 2.25R_B^{min}$ model has a small structural change due to two terminal residues misfolding.

S8 RMSD vs. Cluster Energy Plots for the side chain parameter scan

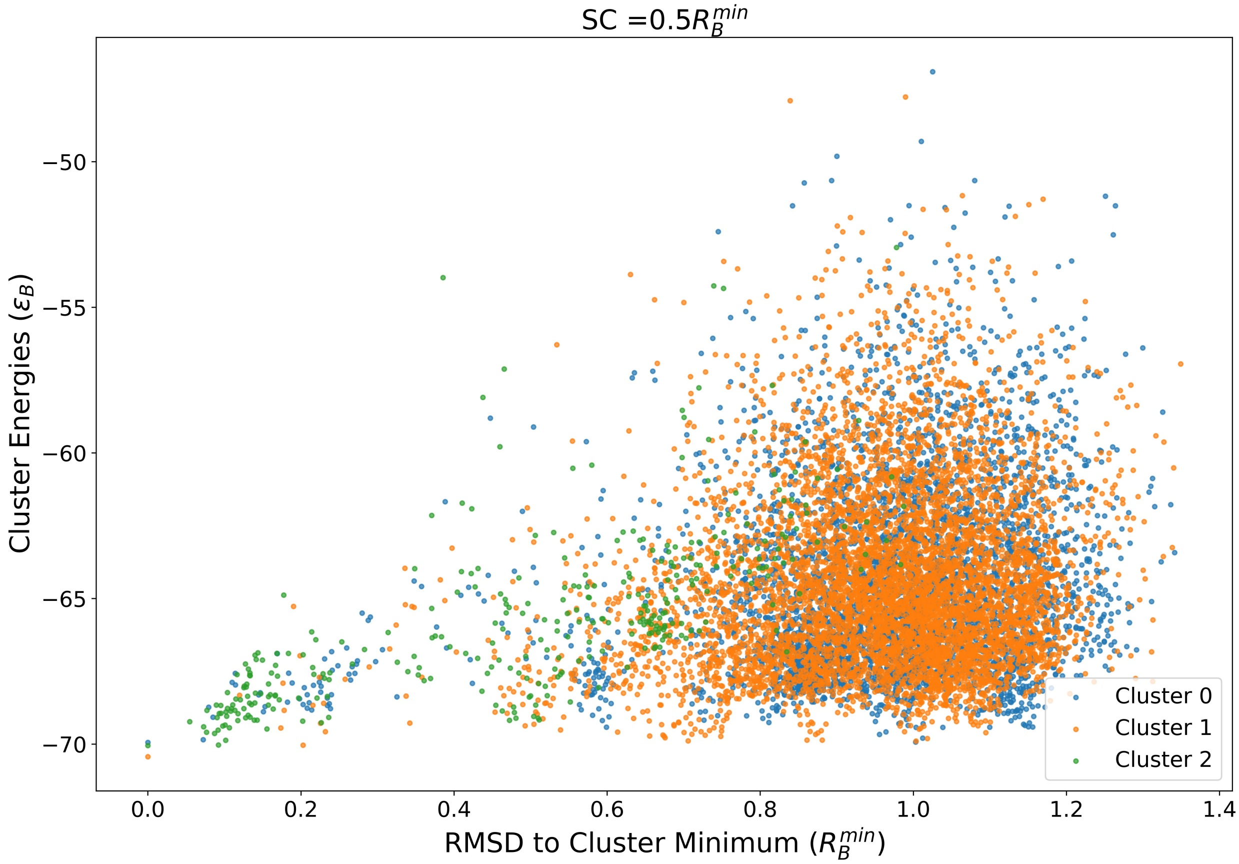


Figure S8: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $SC = 0.5R_B^{min}$ model.

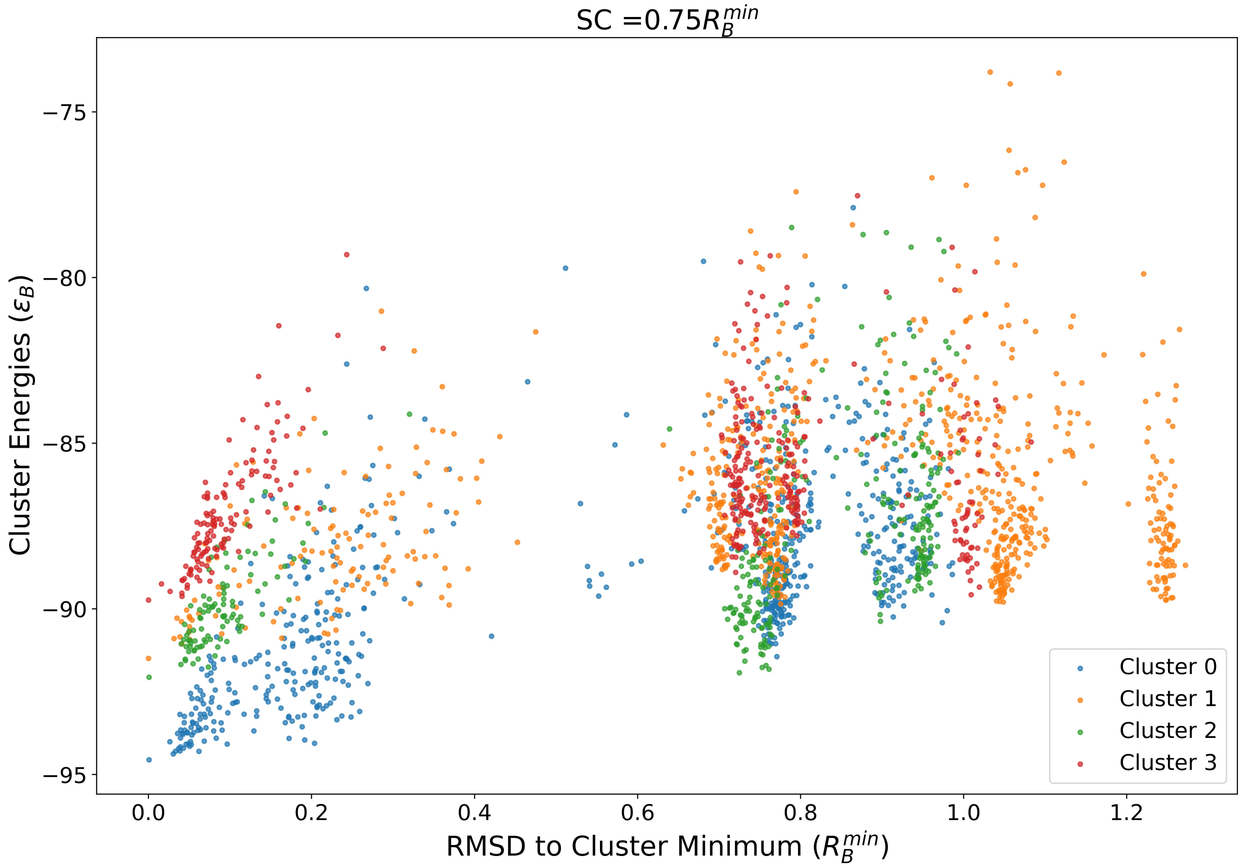


Figure S9: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $SC = 0.75R_B^{min}$ model.

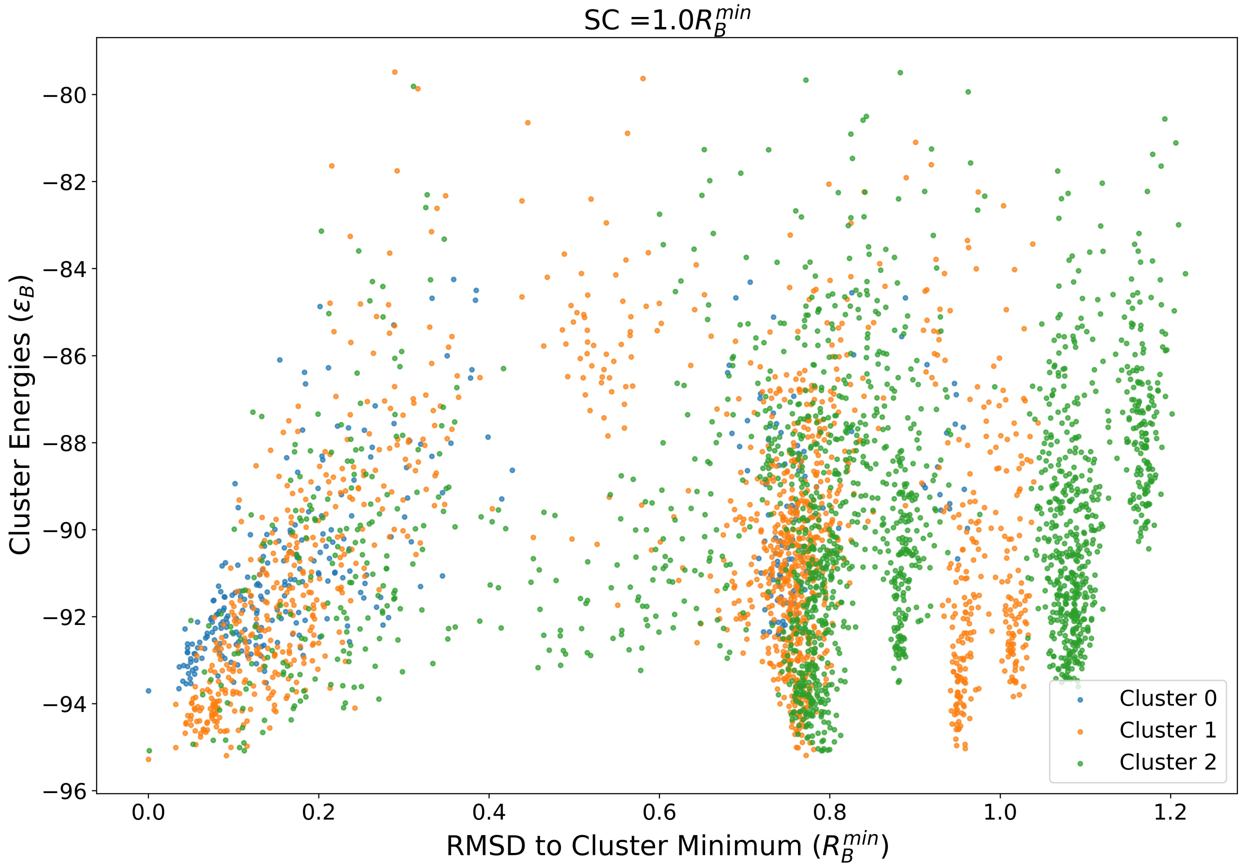


Figure S10: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $SC = 1.0R_B^{min}$ model.

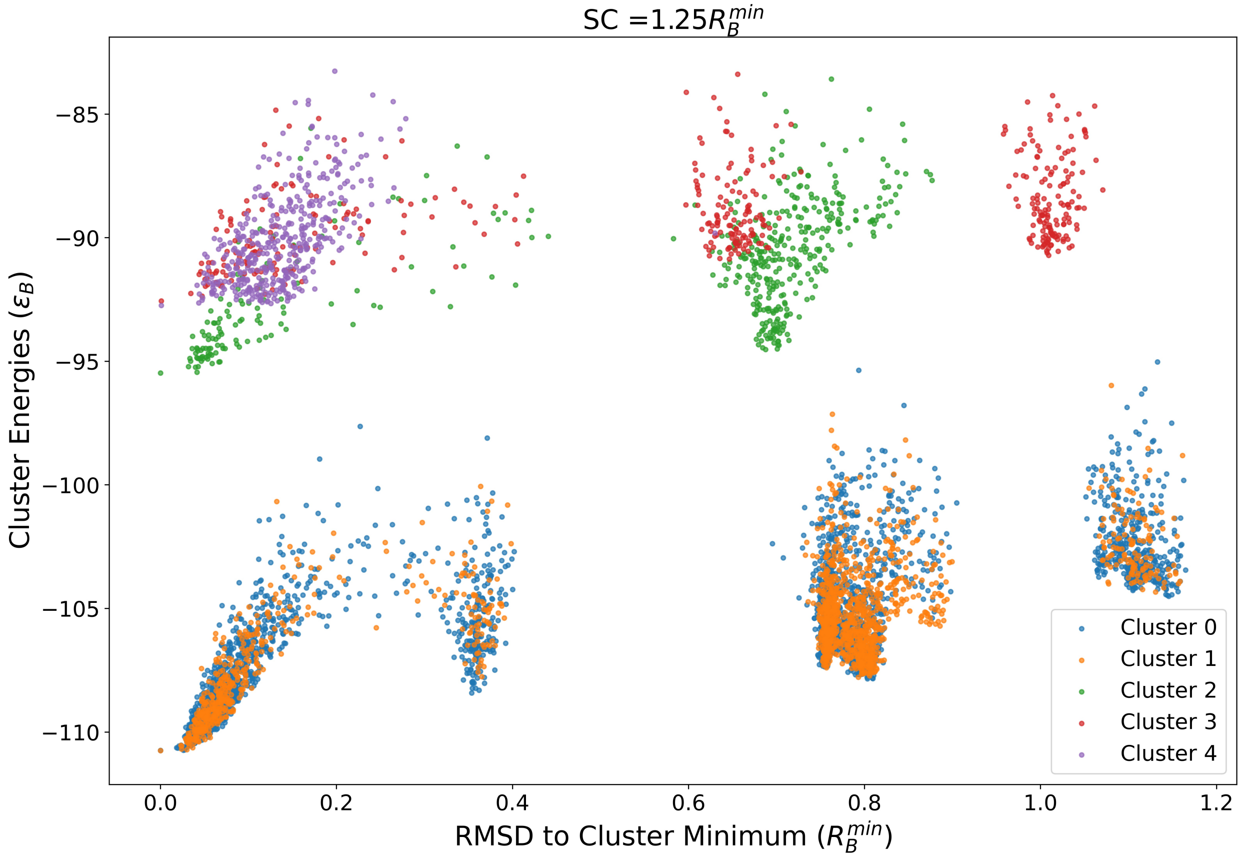


Figure S11: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $SC = 1.25R_B^{min}$ model.

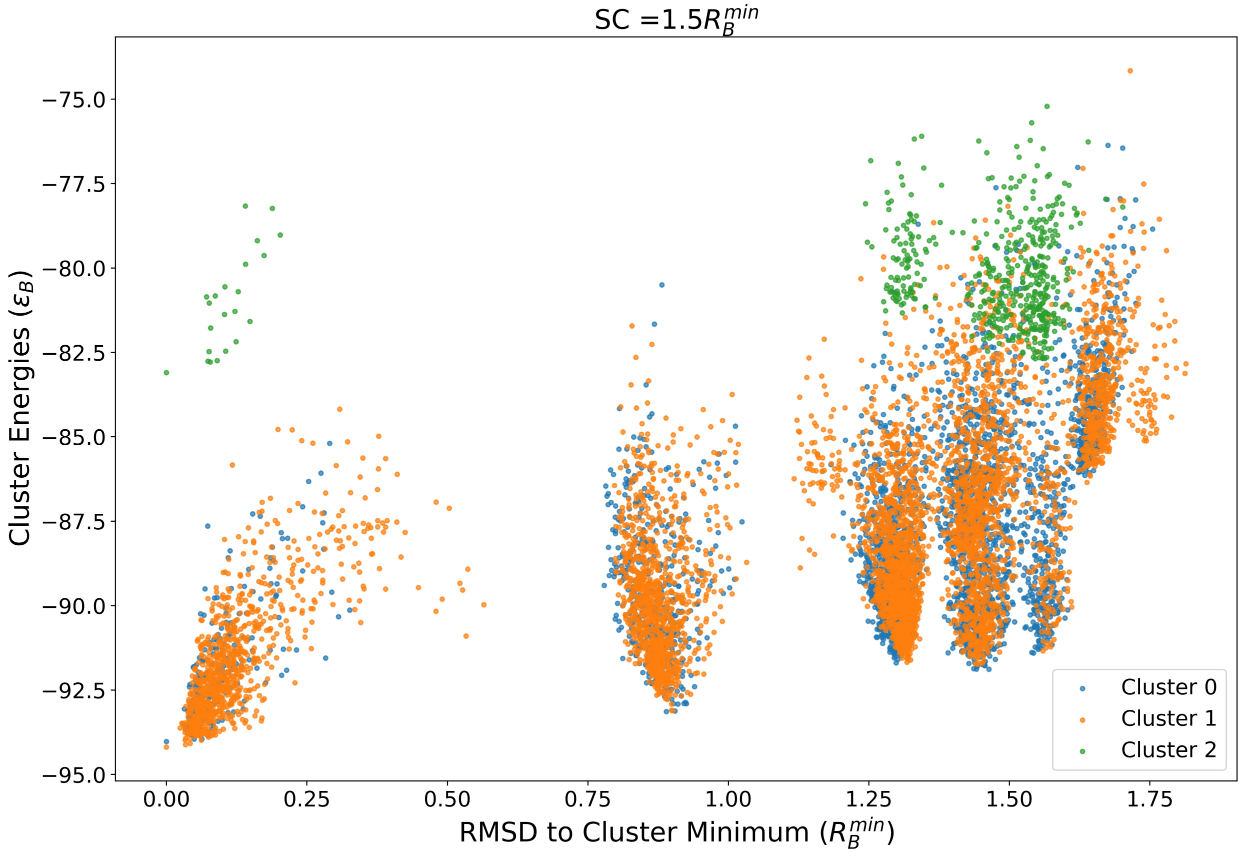


Figure S12: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $SC = 1.5R_B^{min}$ model.

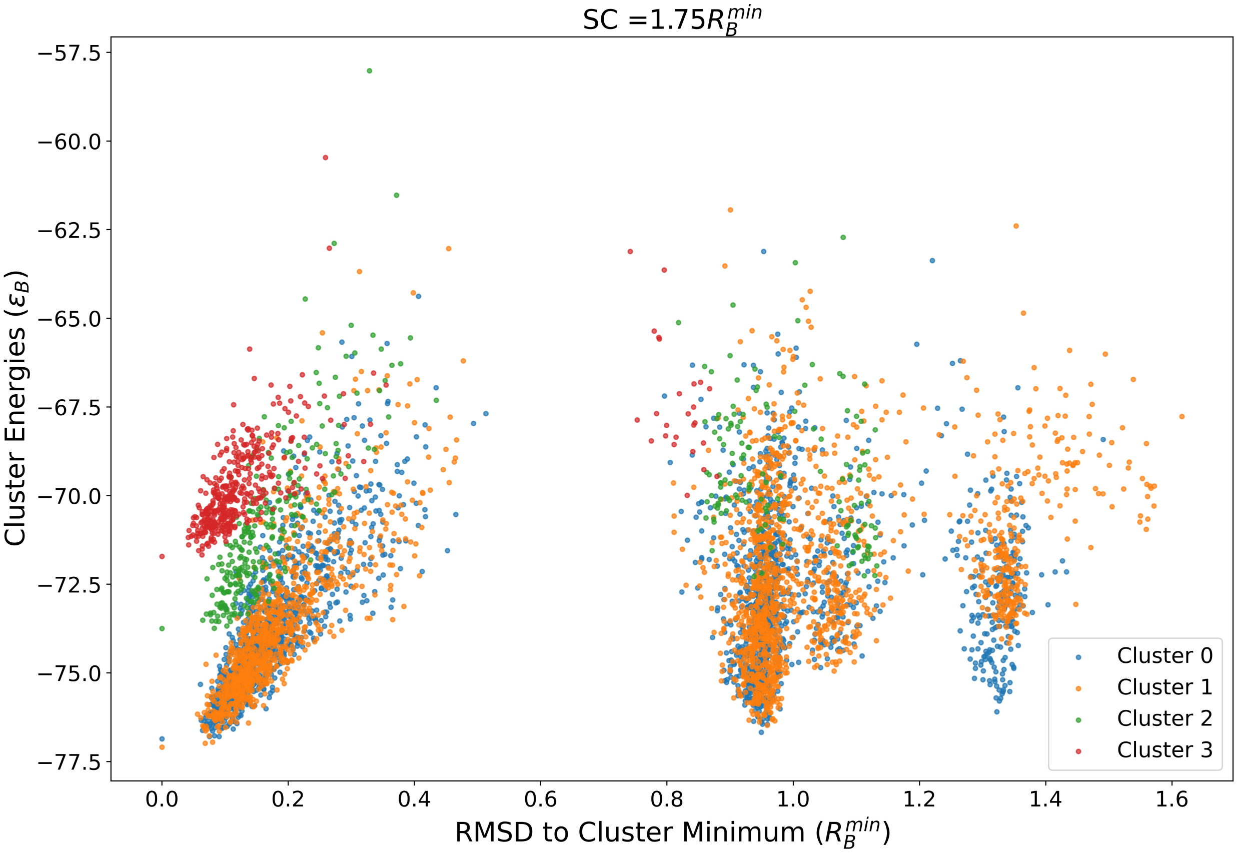


Figure S13: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $SC = 1.75R_B^{min}$ model.

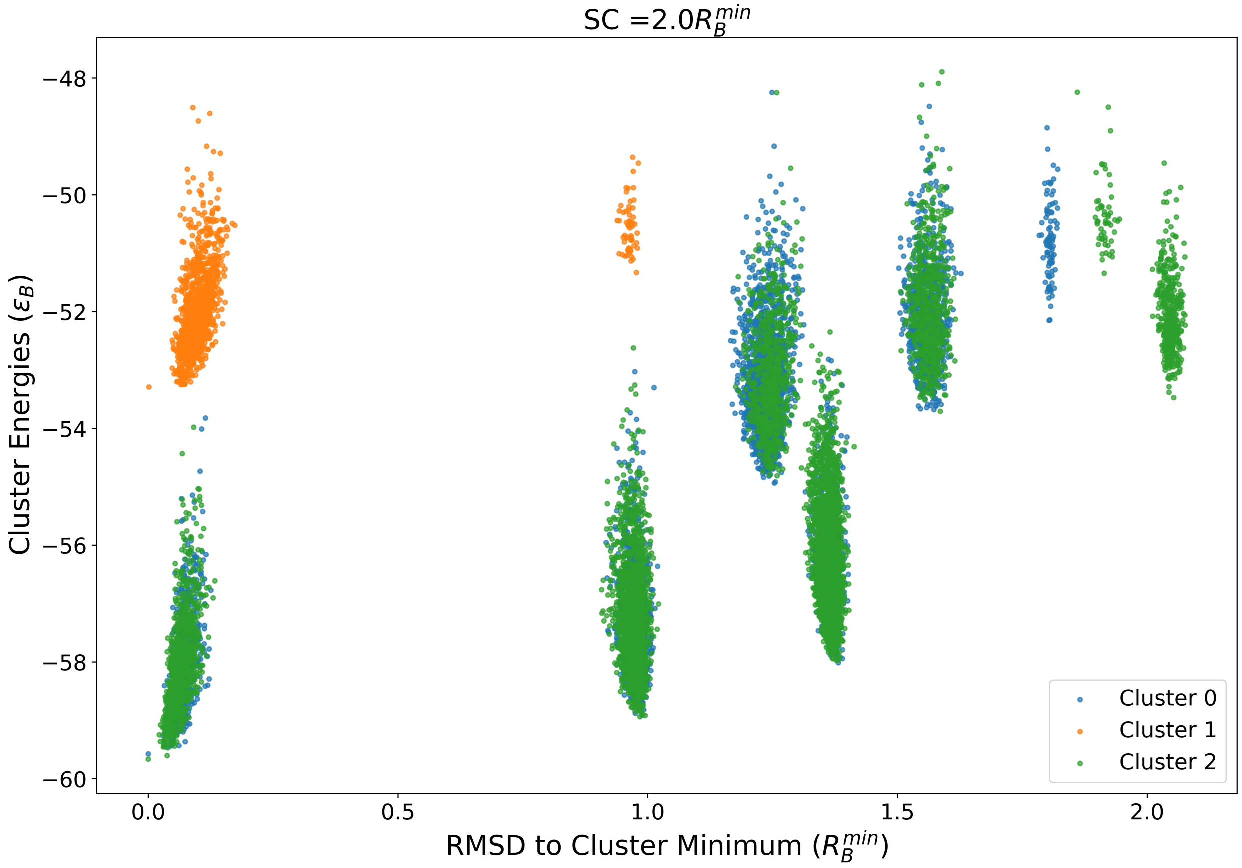


Figure S14: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $SC = 2.0R_B^{min}$ model.

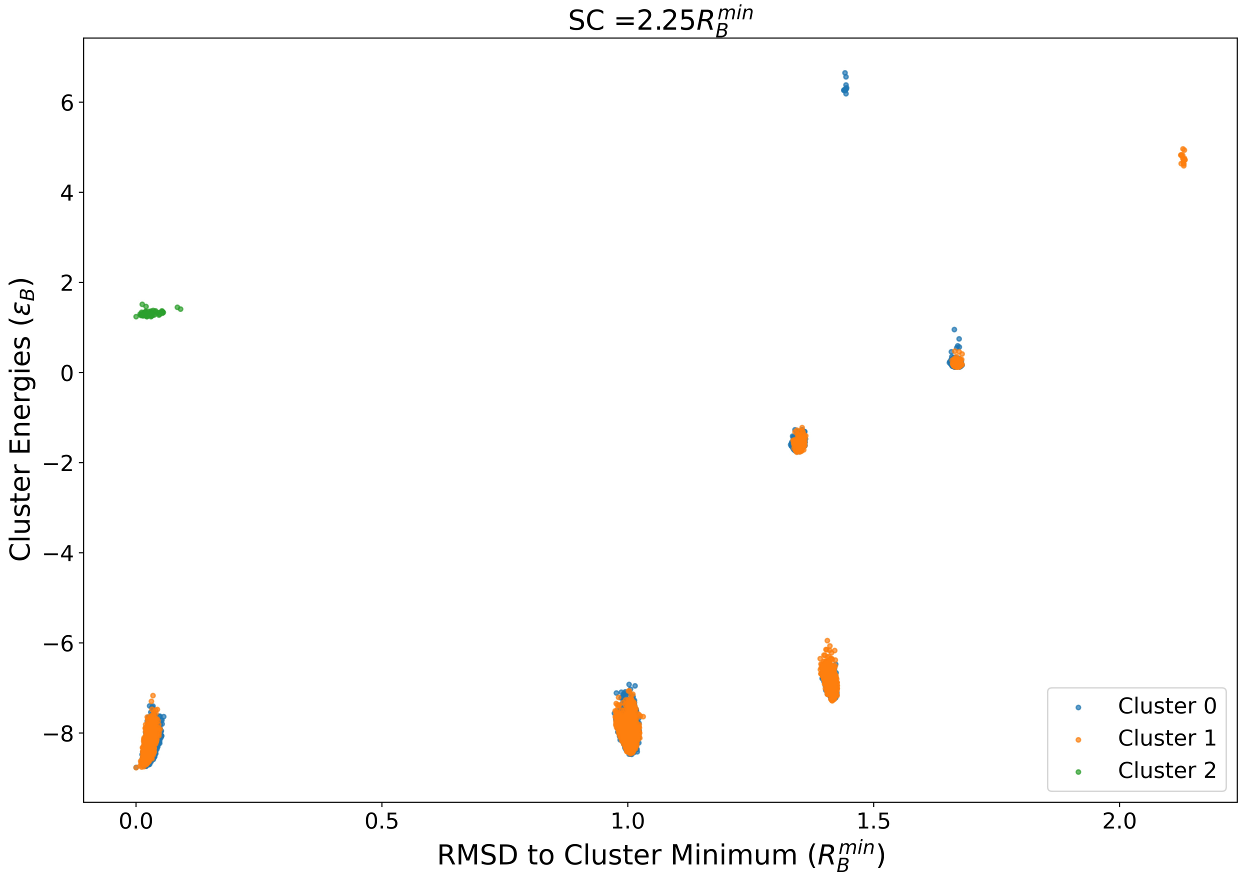


Figure S15: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $SC = 2.25R_B^{min}$ model.

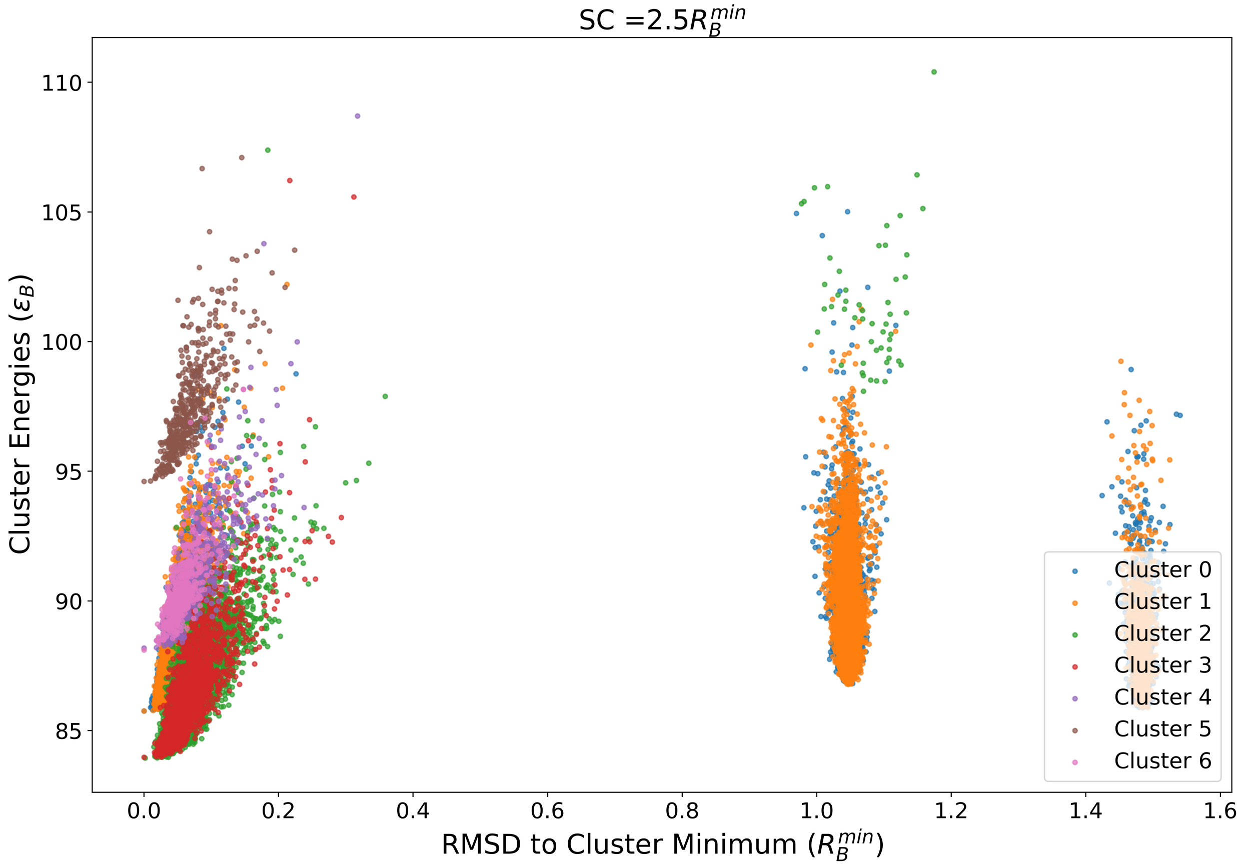


Figure S16: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $SC = 2.5R_B^{min}$ model.

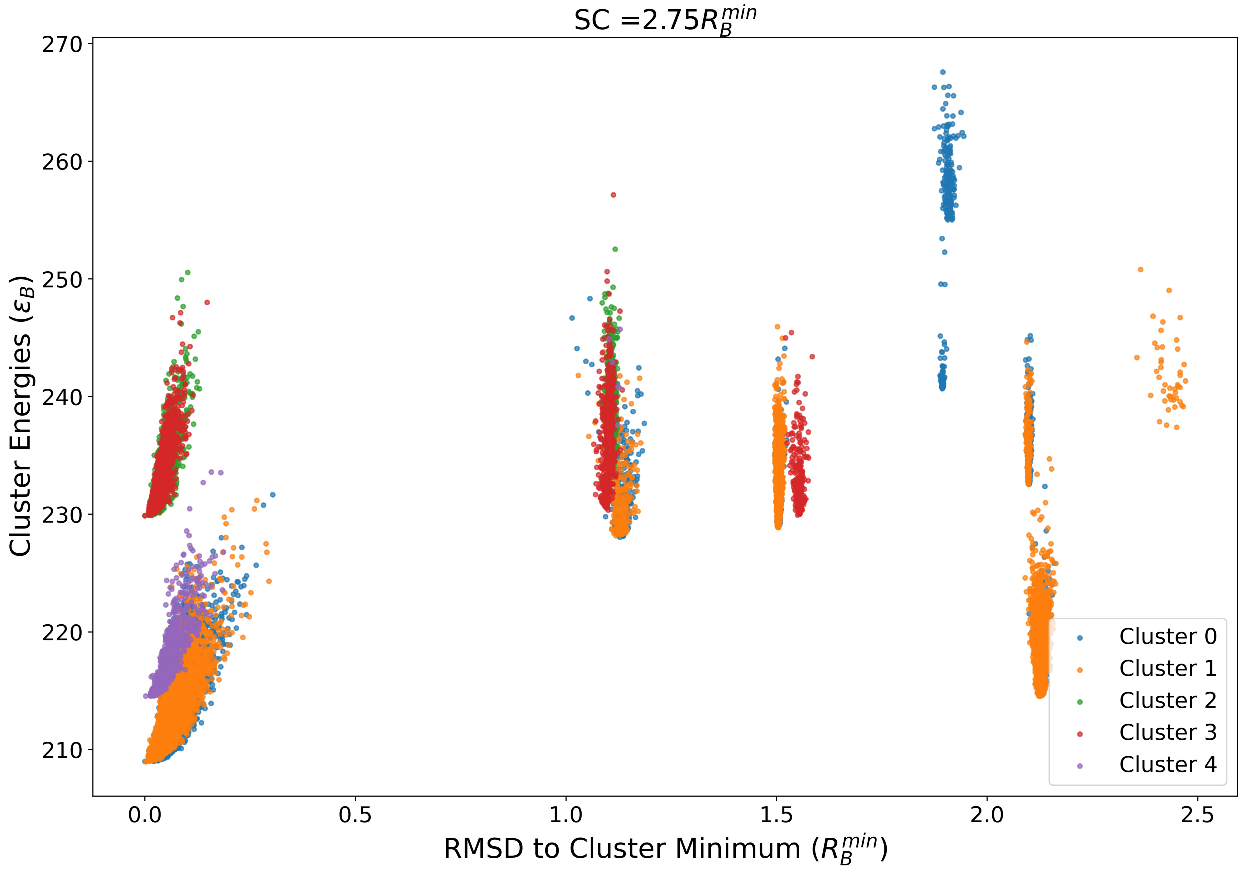


Figure S17: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $SC = 2.75R_B^{min}$ model.

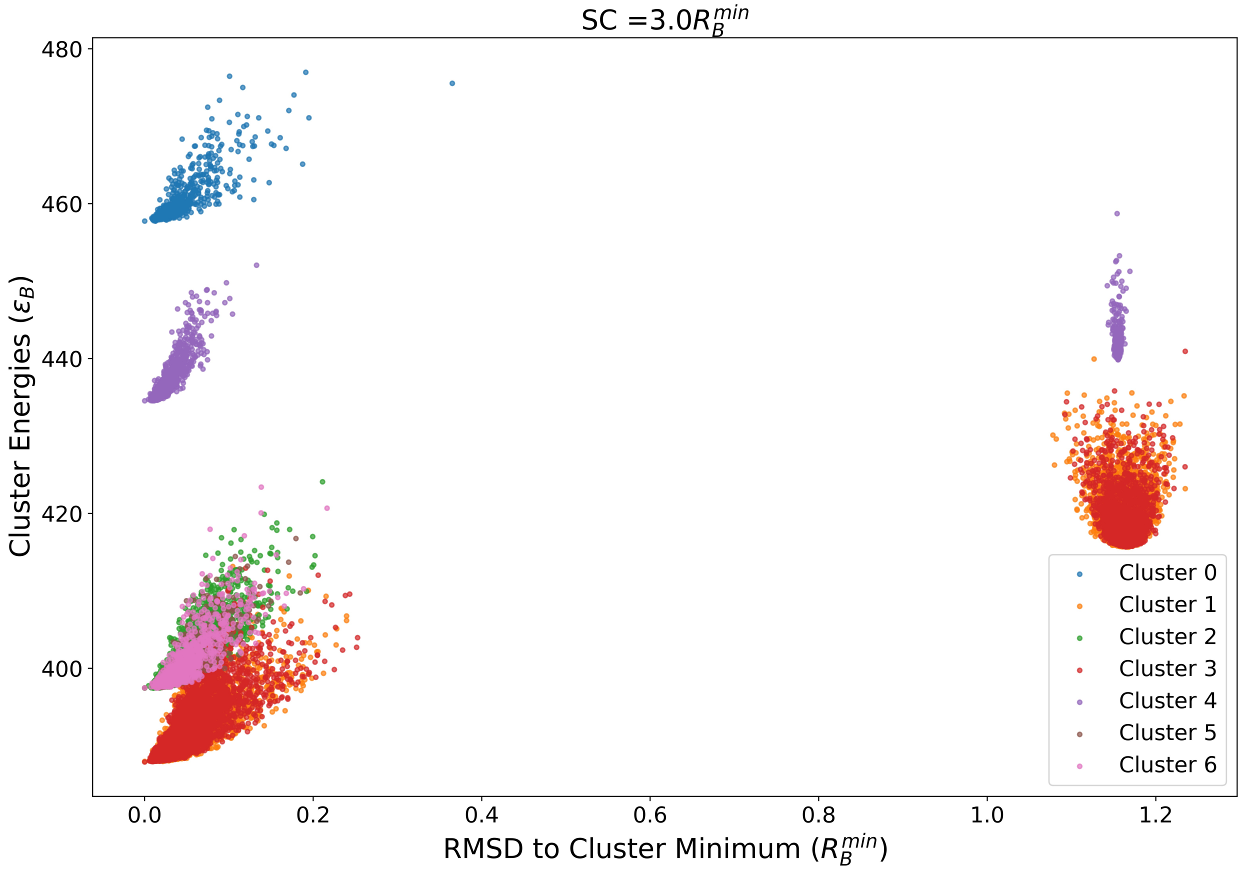


Figure S18: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $SC = 3.0R_B^{min}$ model.

S9 RMSD vs. Cluster Energy Plots for the bond-angle parameter scan

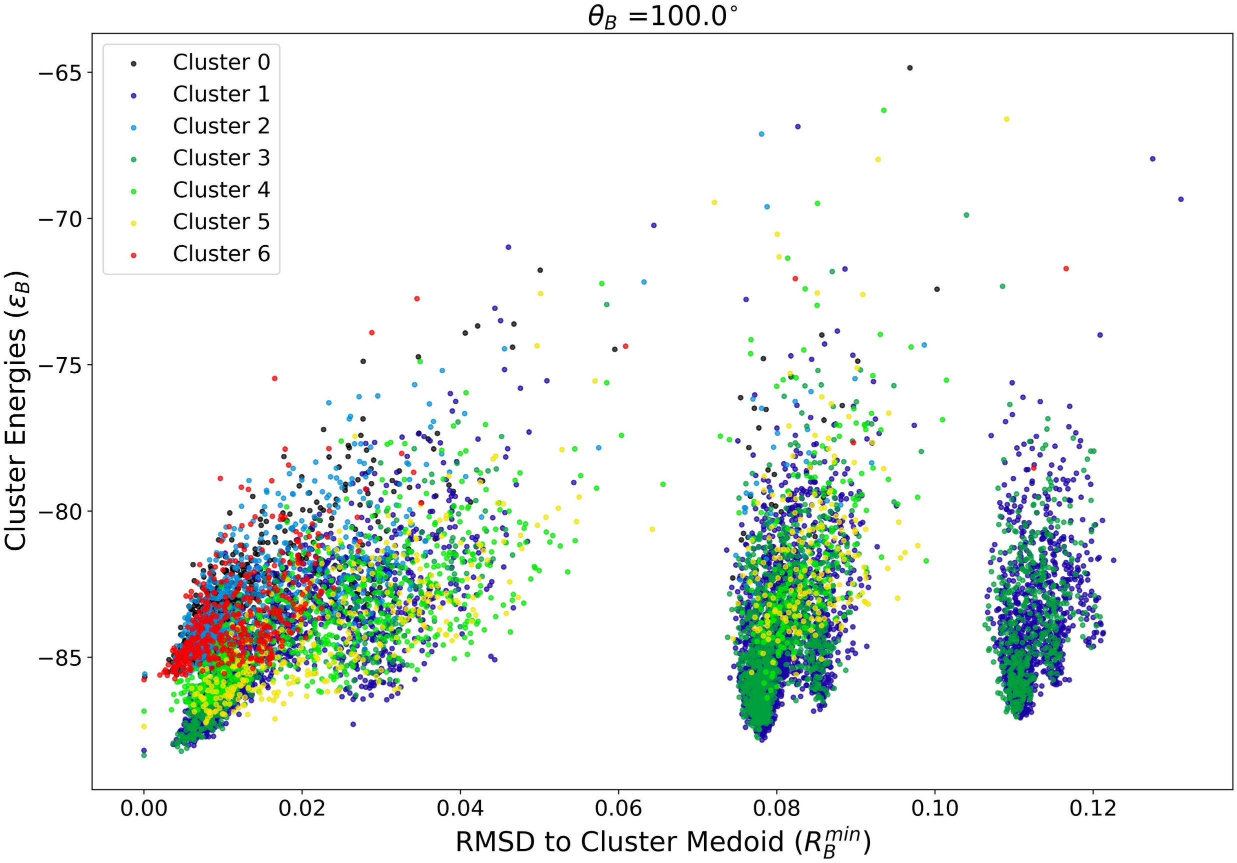


Figure S19: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 100°$ model.

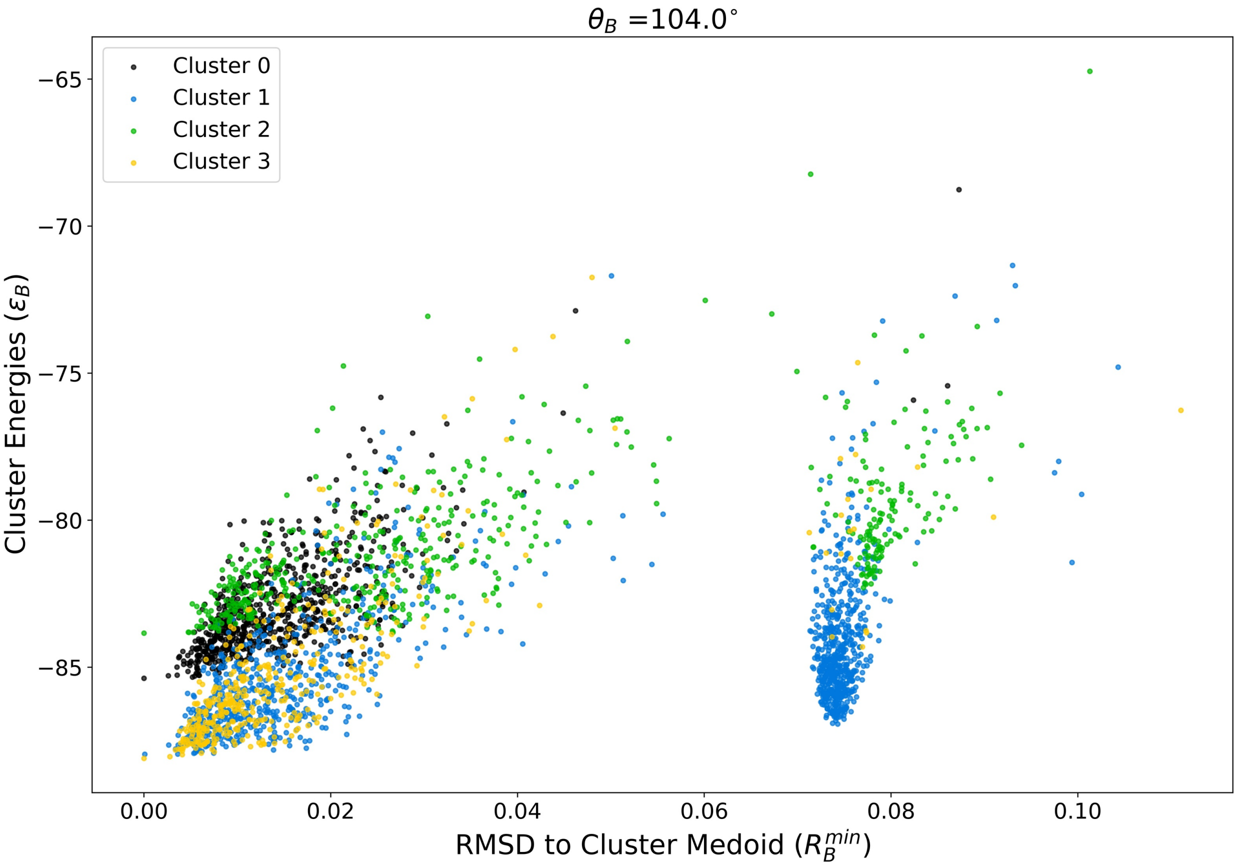


Figure S20: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 104°$ model.

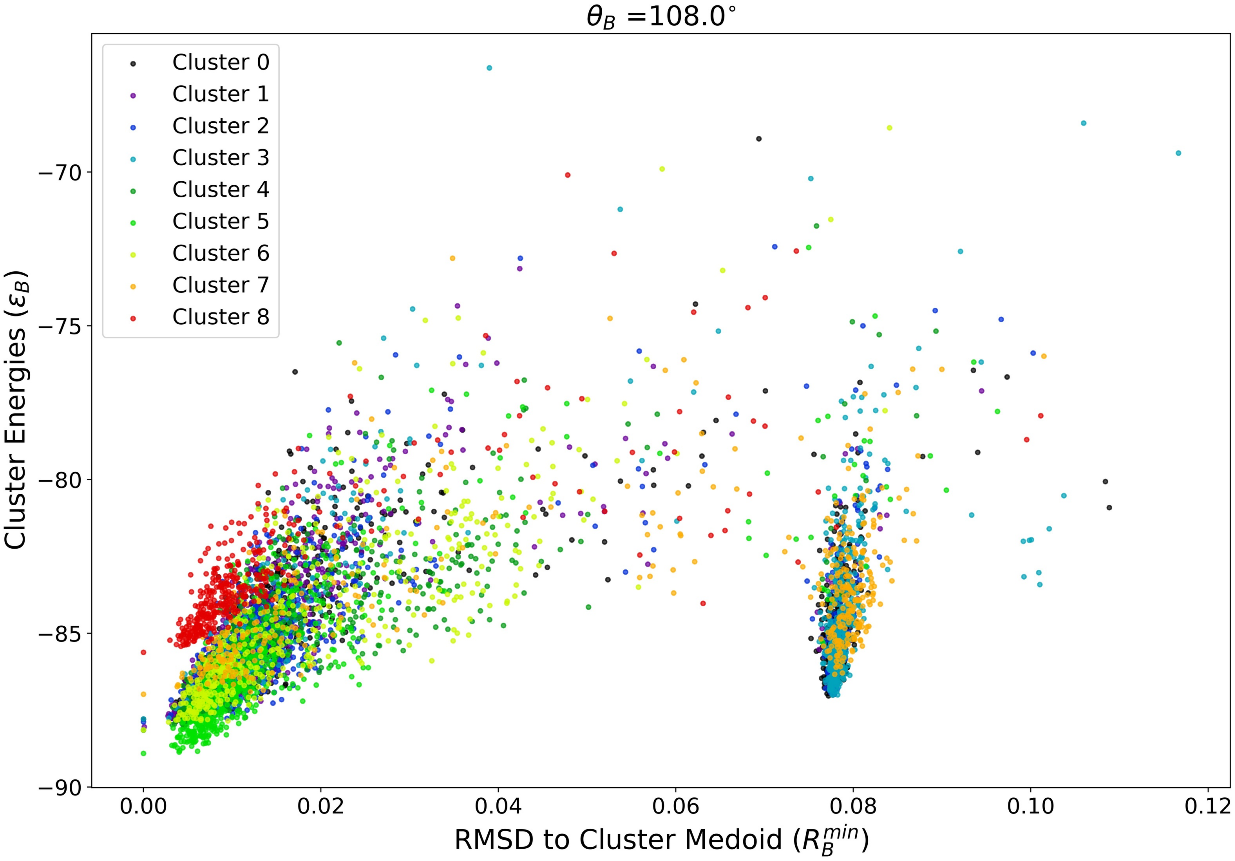


Figure S21: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 108°$ model.

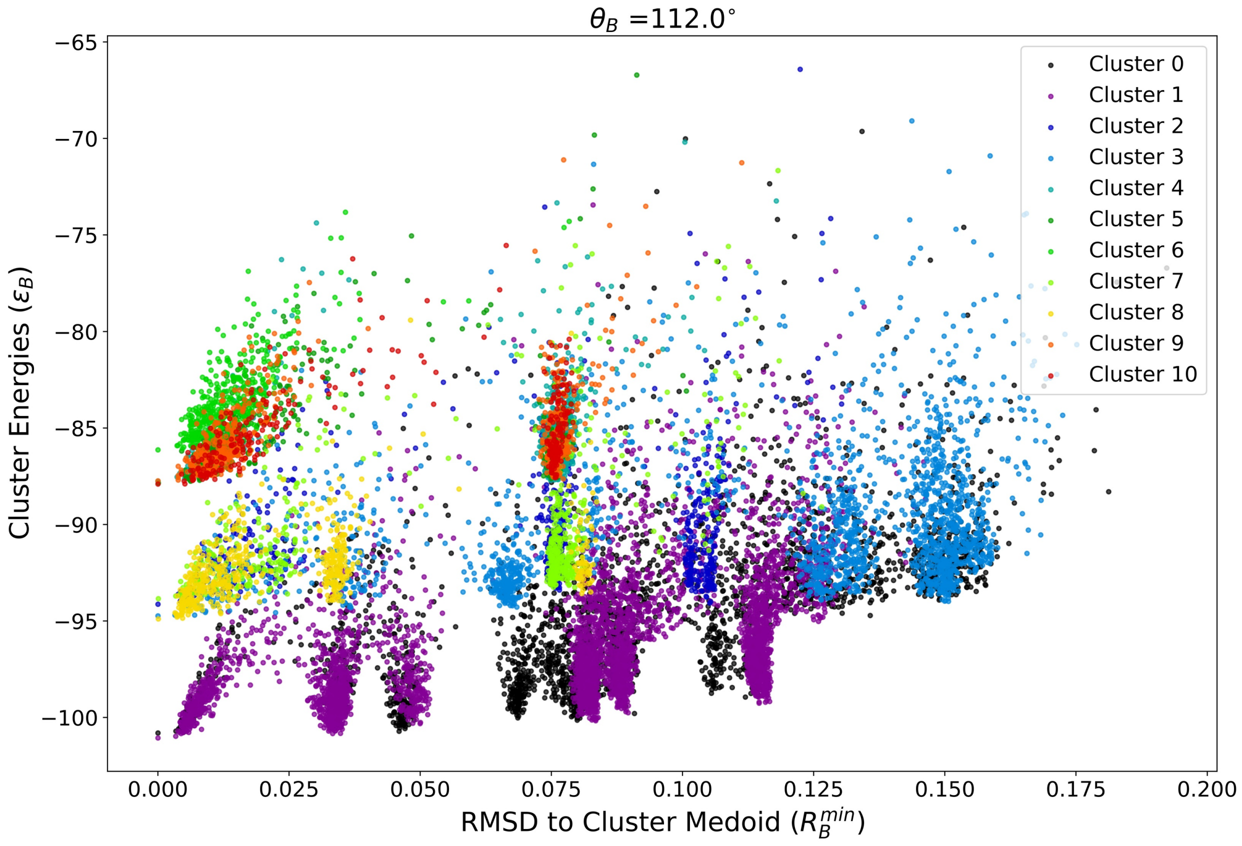


Figure S22: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 112^\circ$ model.

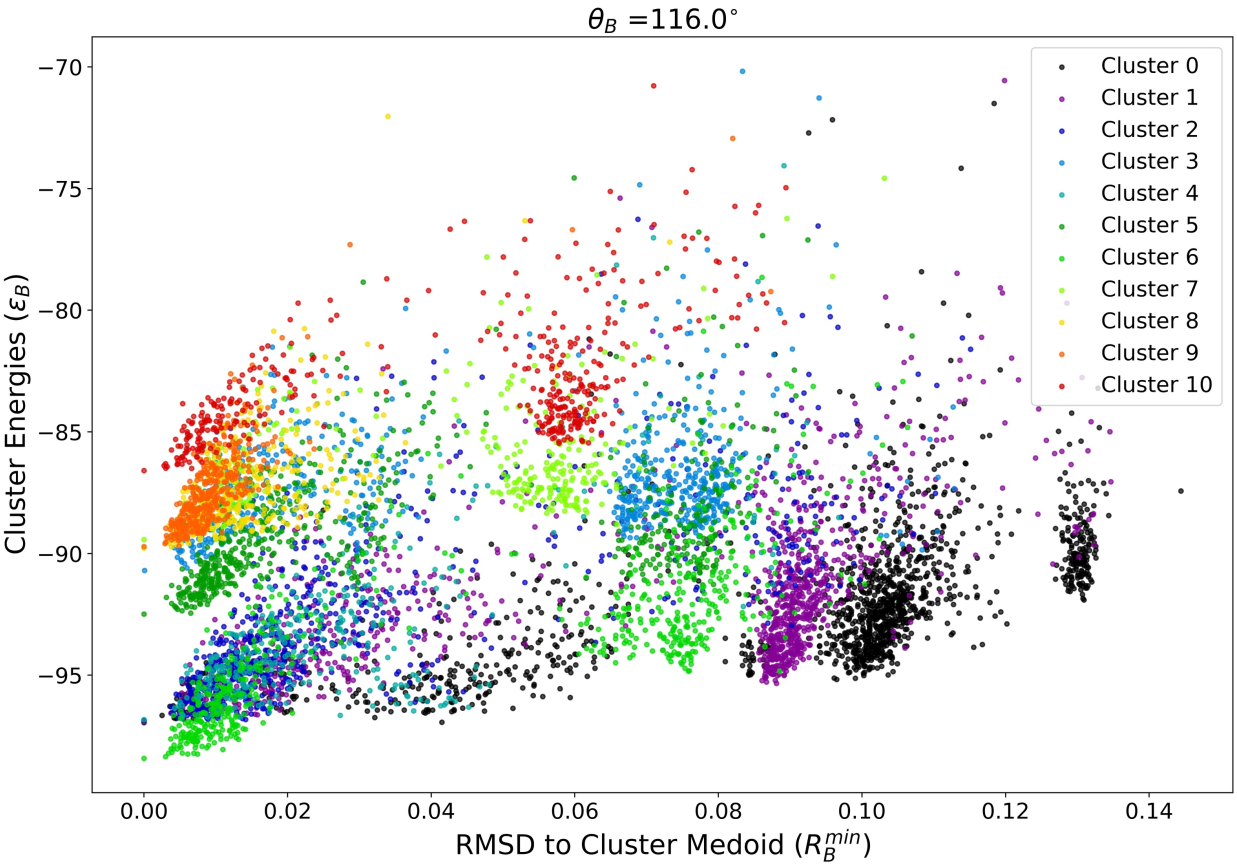


Figure S23: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 116°$ model.

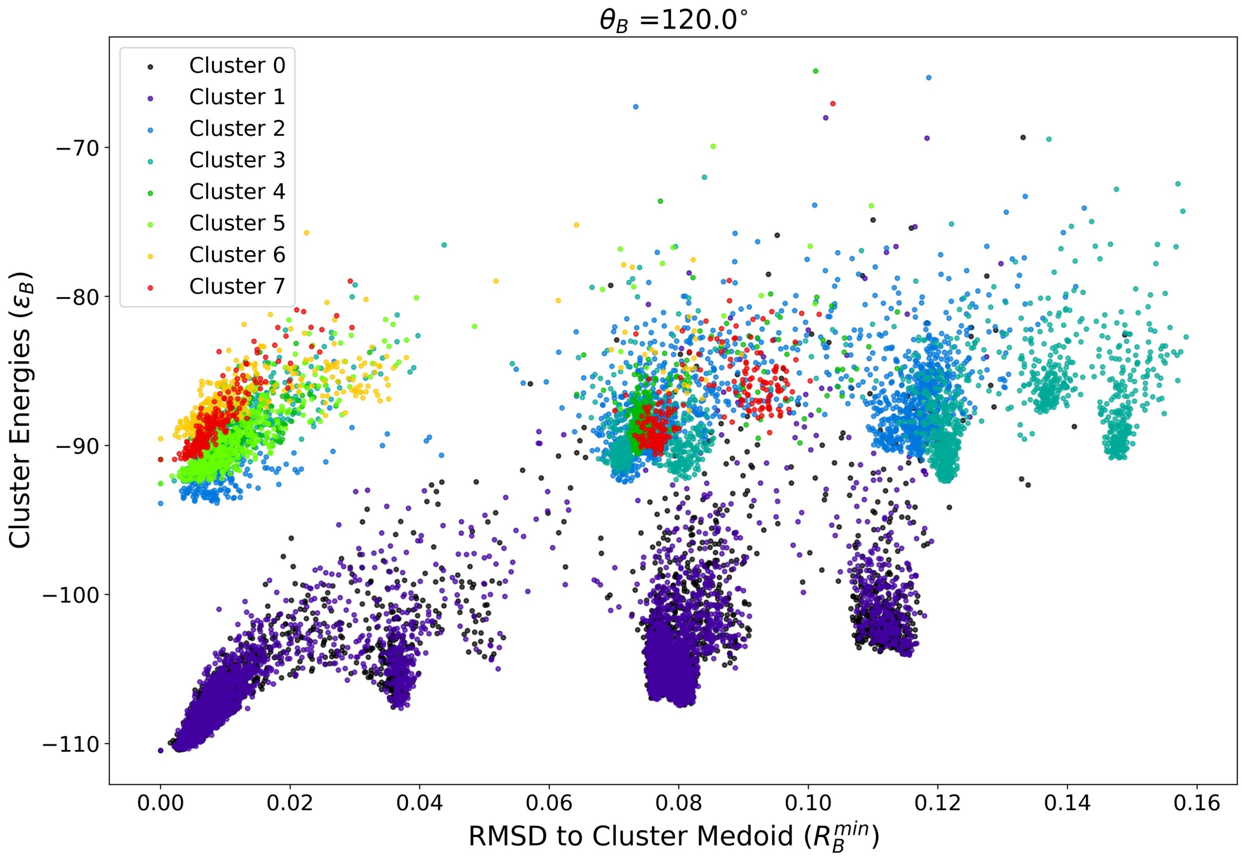


Figure S24: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 120°$ model.

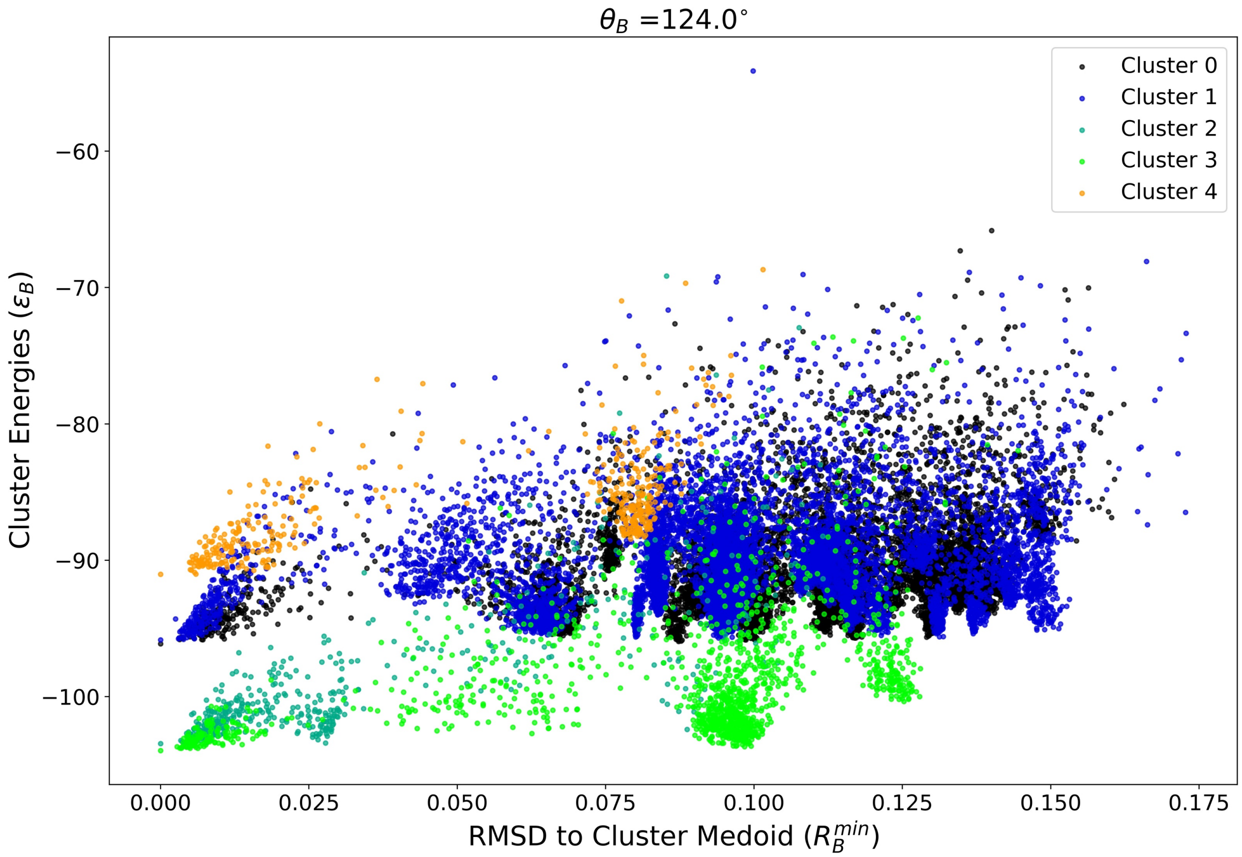


Figure S25: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 124°$ model.

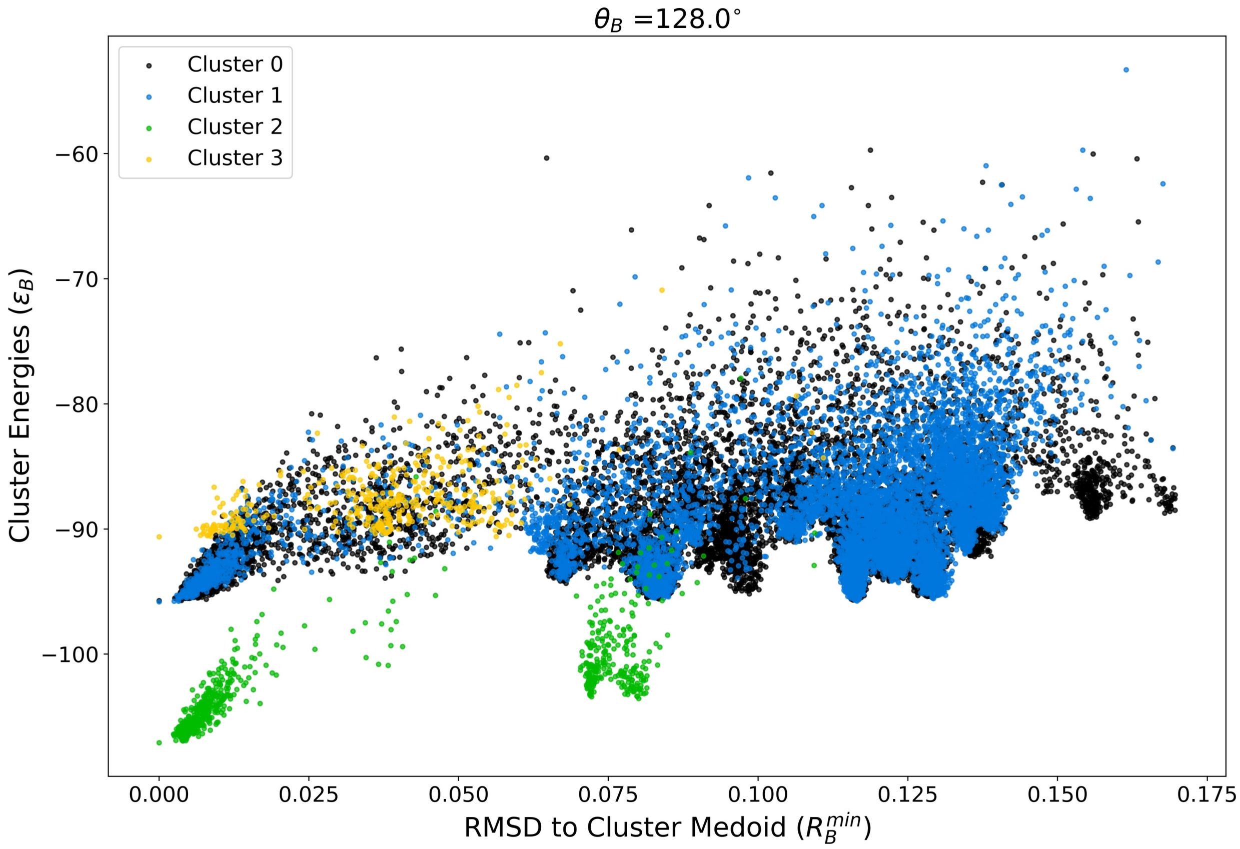


Figure S26: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 128°$ model.

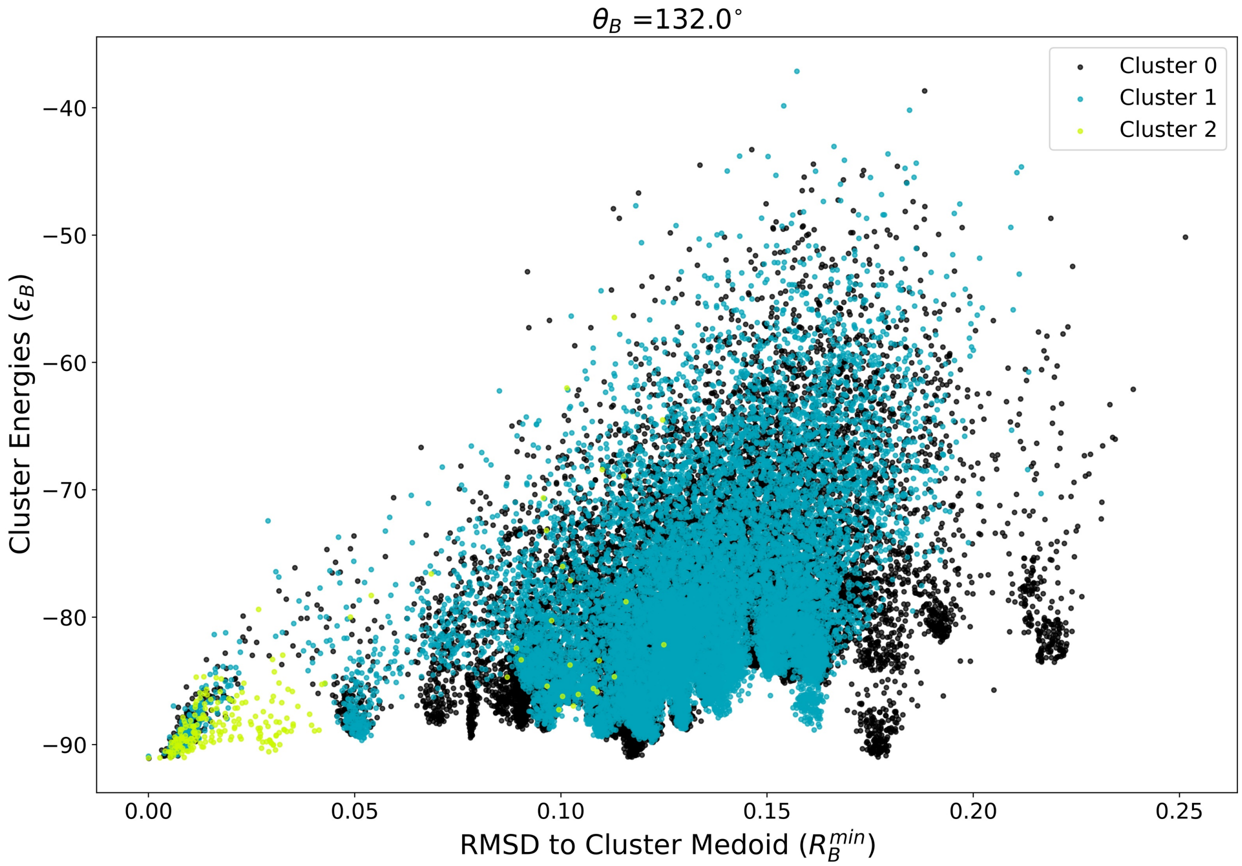


Figure S27: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 132°$ model.

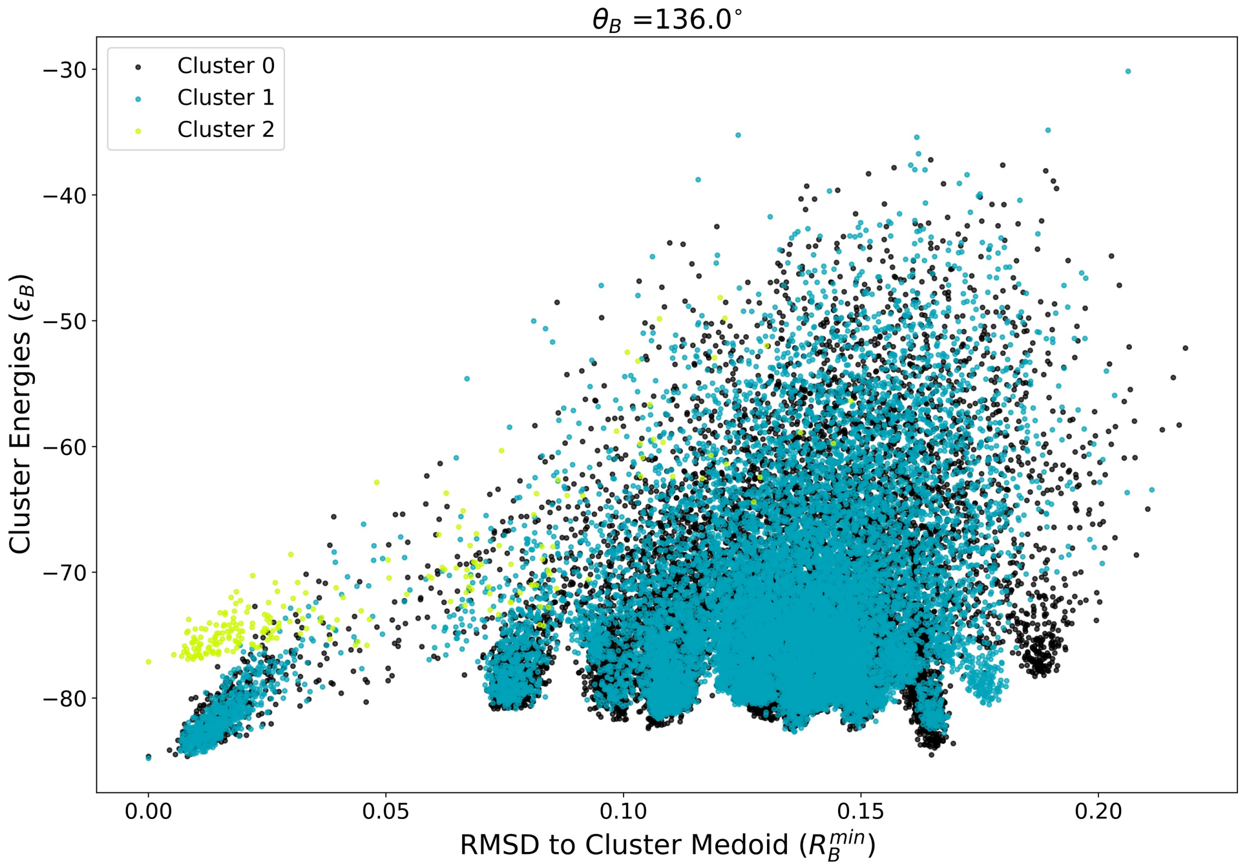


Figure S28: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 136°$ model.

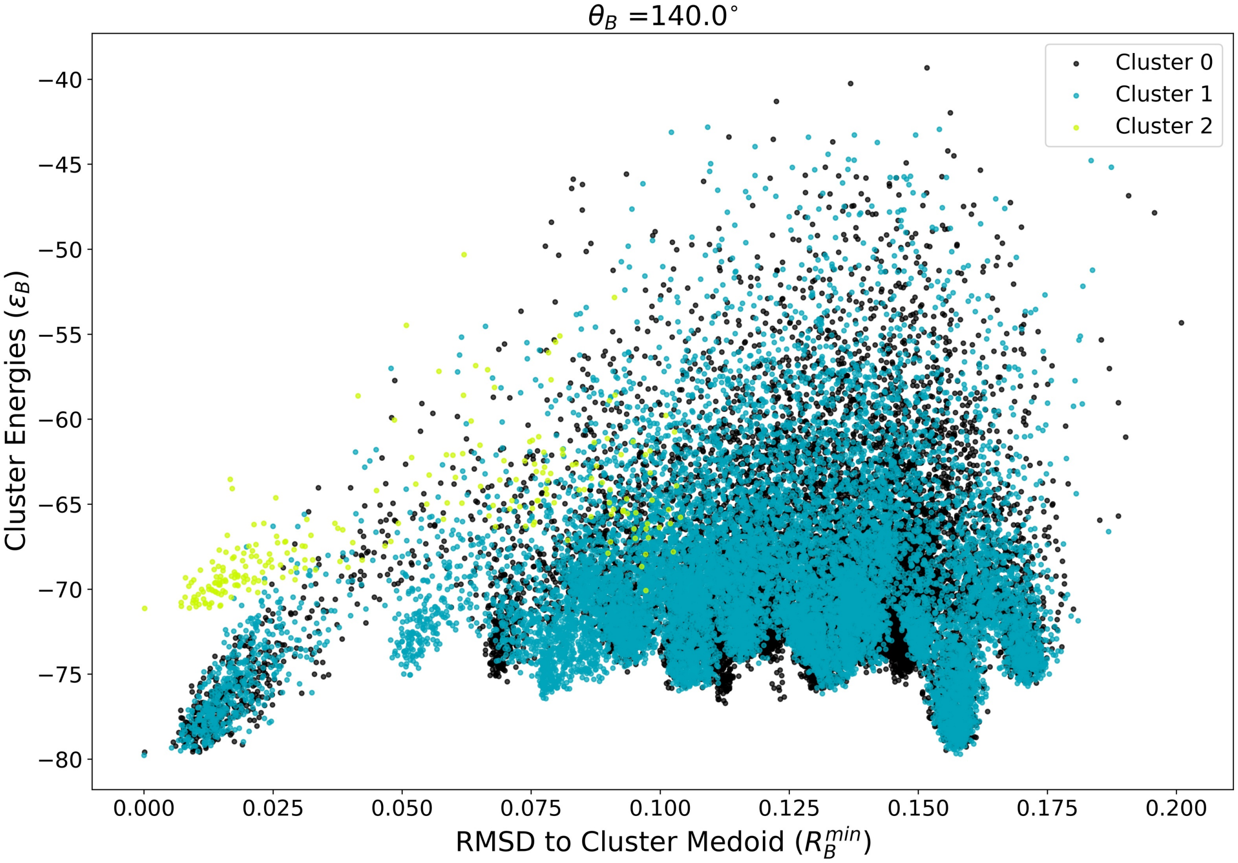


Figure S29: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 140°$ model.

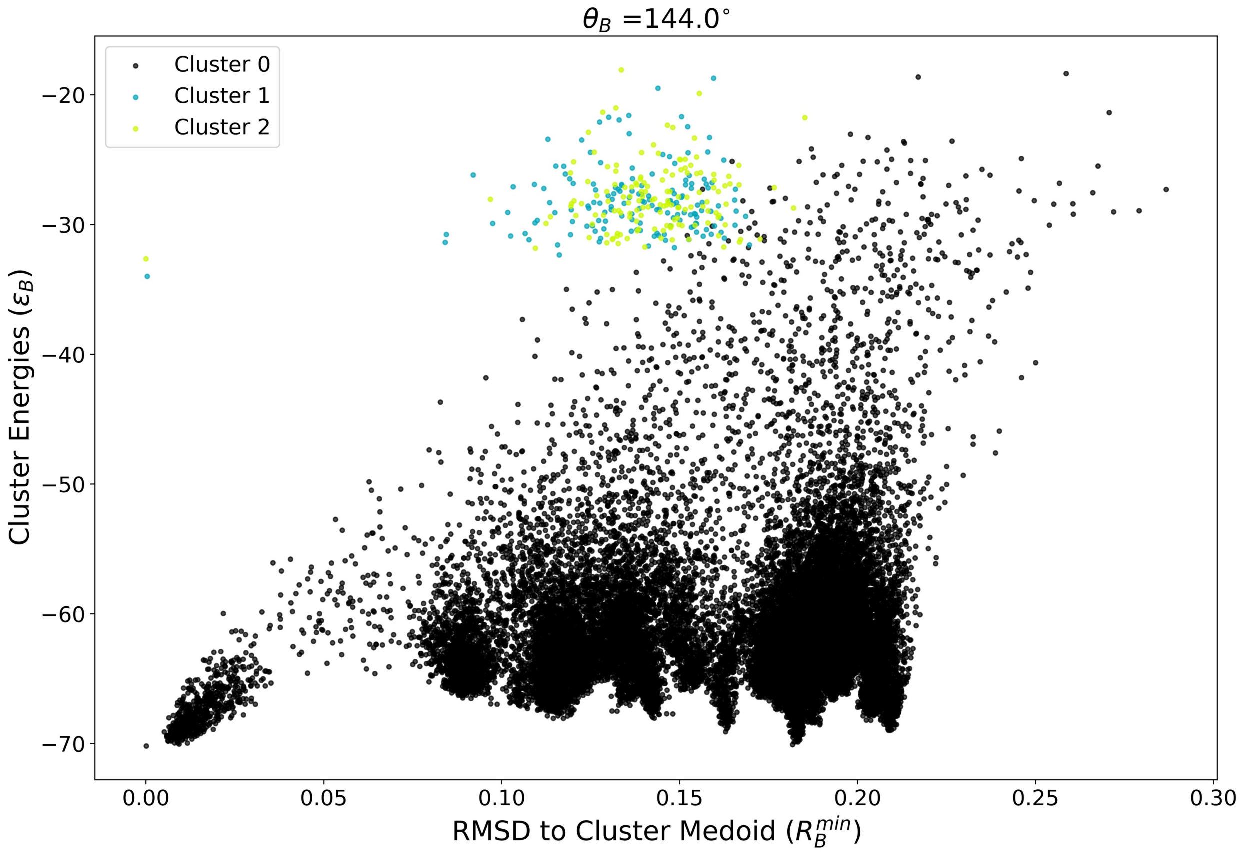


Figure S30: Energy vs. RMSD to minimum energy structure for all clusters identified in for the $\theta_B = 144°$ model.