Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2026, **5**, 214

Computational design of polypeptide-based compartments for synthetic cells

Jianming Mao,[a] Yongkang Xi,[b] Armin Shayesteh Zadeh,[c] Allen P. Liu[bdef] and Andrew L. Ferguson *[ac]

Received 1st July 2025
Accepted 6th November 2025

DOI: 10.1039/d5dd00291e

rsc.li/digitaldiscovery

Synthetic cells are prevalent models for understanding and recapitulating complicated functions of natural cells such as DNA replication and protein expression. Lipid-based vesicles are widely employed but are limited by their fragility under mechanical forces or osmotic pressure. Elastin-like polypeptides (ELPs) composed of repetitive (VPGXG) sequences present alternative building blocks with which to construct the delimiting membrane of synthetic cells possessing high structural stability and tolerance of harsh environmental stress. In this work, we present a high-throughput virtual screening pipeline combining coarse-grained simulations, alchemical free energy calculations, Gaussian process regression, and Bayesian optimization to traverse a library of amphiphilic diblock ELPs for mutant sequences predicted to form thermodynamically stable bilayer vesicles. From our screening campaign, we have identified a range of novel ELP candidates with enhanced predicted stability. Analysis of our screening data exposes new rational design principles that suggest incorporating particular guest residues in hydrophilic blocks – including histidine, tyrosine, and threonine – and in hydrophobic blocks – including alanine, phenylalanine, cysteine, and isoleucine – to enhance the thermodynamic stability of ELP bilayer vesicles. The computational pipeline greatly accelerates the discovery of ELP building blocks for synthetic cells, exposes new design principles for these molecules, and furnishes a transferable framework for designing peptides with desirable structural or functional properties.

1 Introduction

Natural cells employ molecular compartments evolved by natural selection to realize sophisticated biological functions in living systems.[1–3] Inspired by these natural systems, synthetic cells have been proposed to recapitulate one or more cell functions in a compartmentalized volume. These artificial systems have found fundamental applications in understanding the prebiotic origin of life[4,5] and technological applications in engineering synthetic cells capable of operating in harsh environments[6] or carrying and targeted delivery of drug payloads.[7] Synthetic cells can be constructed by bottom-up or top-down methods. Top-down methods typically commence with natural cells followed by removal of the unnecessary genes and organelles or replacement of the intrinsic components (*e.g.*, genome, proteins) with synthetic substitutions to realize a minimal viable system.[8,9] Bottom-up approaches, by contrast, construct a synthetic cell from scratch through assembling different biological machinery into nano- or micro-sized vesicles.[10,11] Lipids have been widely deployed in the construction of the delimiting membrane of synthetic cells but are limited in further development and wider application due to the relatively weak mechanical stability of the resulting vesicles and the requirement for quite harsh and destabilizing chemical conditions for lipid modification or functionalization.[12] Alternative non-lipid materials, such as amphiphilic block copolymers, emerged as one of the early substitutes to assemble polymer vesicles or polymersomes with superior toughness and greater property control,[12–14] and other inorganic building blocks have also been explored.[15,16]

Biopolymers such as polypeptides present an alternative class of potential building blocks for synthetic cells or organelles.[17] Elastin-like polypeptides (ELPs) have emerged as particularly promising candidates that have been used to form vesicles with diameters ranging from 50 nm to 50 μm through self- or templated-assembly.[10,18,19] ELPs are derived from intrinsically disordered proteins such as tropoelastin and share a sequence pattern $(\mathrm{VPGXG})_n$ possessing a VPGXG pentamer repeat, where V is valine, P is proline, G is glycine and X is

[a] Department of Chemistry, University of Chicago, Chicago, IL, 60637, USA. E-mail: andrewferguson@uchicago.edu

[b] Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109, USA

[c] Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL, 60637, USA

[d] Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI 48109, USA

[e] Cellular and Molecular Biology Program, University of Michigan, Ann Arbor, MI 48109, USA

[f] Department of Biophysics, University of Michigan, Ann Arbor, MI 48109, USA

a guest residue that can be any amino acid except proline.[20,21] In solution, ELPs tend to exhibit a lower critical solution temperature (LCST)-like behaviors.[22] Below a transition temperature, which can vary with the prevailing thermodynamic conditions, the chains present a random coil conformation and remain soluble, while above the transition temperature, the phase transition produces liquid droplets or coacervates. Experimental and computational work have exposed the important role of partial structural ordering associated with this phase transition, with a typically increased propensity for β-turns or β-spiral secondary structures,[23–25] but a growing body of work suggests that these transformations are sparse and transient with sub-nanosecond lifetimes, and that the ELP remains highly dynamic and disordered in the condensed form.[26–31]

The number of repeats n and the identity of guest residue X are the most important factors dictating the thermal and mechanical properties of ELPs. Empirically, a longer repeat and higher hydrophobicity of X tends to reduce the transition temperature whereas a shorter repeat and higher hydrophilicity of X tends to elevate the transition temperature.[32] This tunability resulting from the sequence control has encouraged the exploration of amphiphilic diblock or multi-block ELPs in which each block can have distinct transition temperature and thus control the assembly behaviors. For example, Huber *et al.*[33] showed that amphiphilic ELPs comprising phenylalanines and glutamic acids expressed inside *Escherichia coli* were able to form organelle-like compartments. Vogele *et al.*[10] and Frank *et al.*[34] synthesized self-growing ELP-based synthetic cells. Schreiber *et al.*[18] investigated the assembly efficiency of ELPs with different number of repeats and guest residues into unilamellar prebiotic synthetic cells and the dynamics of the resulting membranes. Recently, Sharma *et al.*[19] demonstrated the synthesis of giant ELP vesicles consisting of $(\text{VPGSG})_{48}$ and $(\text{VPGIG})_{48}$ blocks using an inverse emulsion approach (Fig. 1).

In this work, we consider the rational design of amphiphilic ELP diblock polymers capable of forming thermodynamically stable bilayer vesicles to resist dissociation for long time periods and/or under harsh conditions. The ELP sequence space within which we screen is defined by the repeat number and identity of the guest residues X_1 and X_2 in the sequence $(\text{VPGX}_1\text{G})_m(\text{VPGX}_2\text{G})_n$.[21] Selecting guest residue X_1 to be hydrophilic and X_2 to be hydrophobic promotes an ELP bilayer vesicle architecture with the X_1 block oriented towards the aqueous environments on either side of the bilayer, and the X_2 hydrophobic blocks sequestered within the hydrophobic core of the bilayer (Fig. 1). This design strategy has been employed by Schreiber *et al.*,[18] Sharma *et al.*,[19] and Vogele *et al.*[10] to experimentally realize synthetic ELP bilayer vesicles with m and n in the range of 5–100 repeat units.[35] Chemical intuition can help focus the search towards promising choices for the X_1 and X_2 guest residues, but can also introduce bias that might impede the discovery of non-intuitive but high-performing materials. Data-driven methods and computational modeling offer opportunities to systematically search and accelerate the discovery of ELP sequences capable of forming stable vesicles and to infer design rules linking ELP sequence to emergent physicochemical properties. Machine learning techniques present powerful tools for inferring design rules and engineering peptides and proteins with desired chemical properties.[36–39] For example, Lee *et al.* employed support vector machines (SVMs) to aid in the design of membrane active peptides and antimicrobial peptides.[40] Zhang *et al.* built a LSTM model to generate *de novo* peptides with specific prospective therapeutic benefits.[41] Guntuboina *et al.* utilized large language models (LLMs) for the prediction of peptide properties from sequences.[42] Bell *et al.* applied Gaussian process regression (GPR) to an experimental data set for the prediction of binding affinity of antigens to MHC class II.[43]

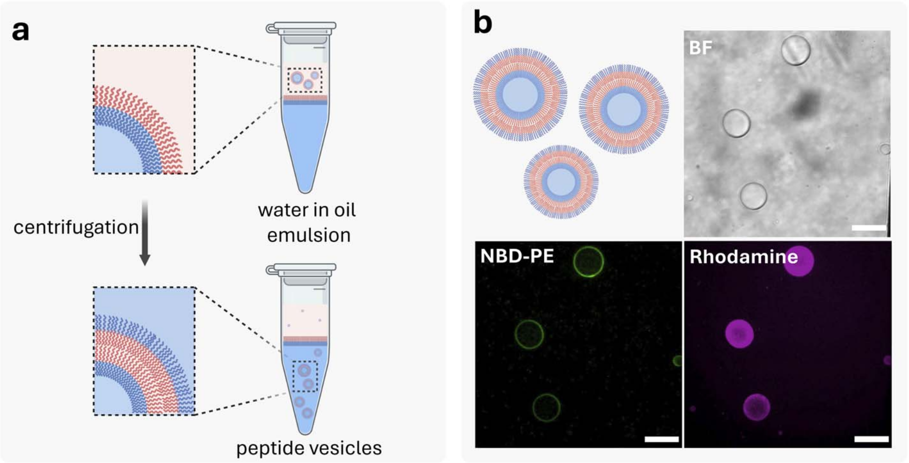

Fig. 1 Construction of ELP vesicles by inverse emulsion transfer. (a) Schematic illustration showing the process of constructing ELP peptide vesicles by the emulsion transfer approach reported by Sharma *et al.*[19] The red portion of the sequence represents the hydrophobic block and the blue portion the hydrophilic block. (b) Schematic illustration and brightfield and fluorescence images of labelled $S_{48}I_{48}$ peptide vesicles. The green fluorescence shows the membrane structure labeled by a NBD-PE lipid and Rhodamine dye encapsulated within the vesicles. Scale bars are 20 μm. Cartoons used to construct the schematics are obtained from BioRender (https://biorender.com).

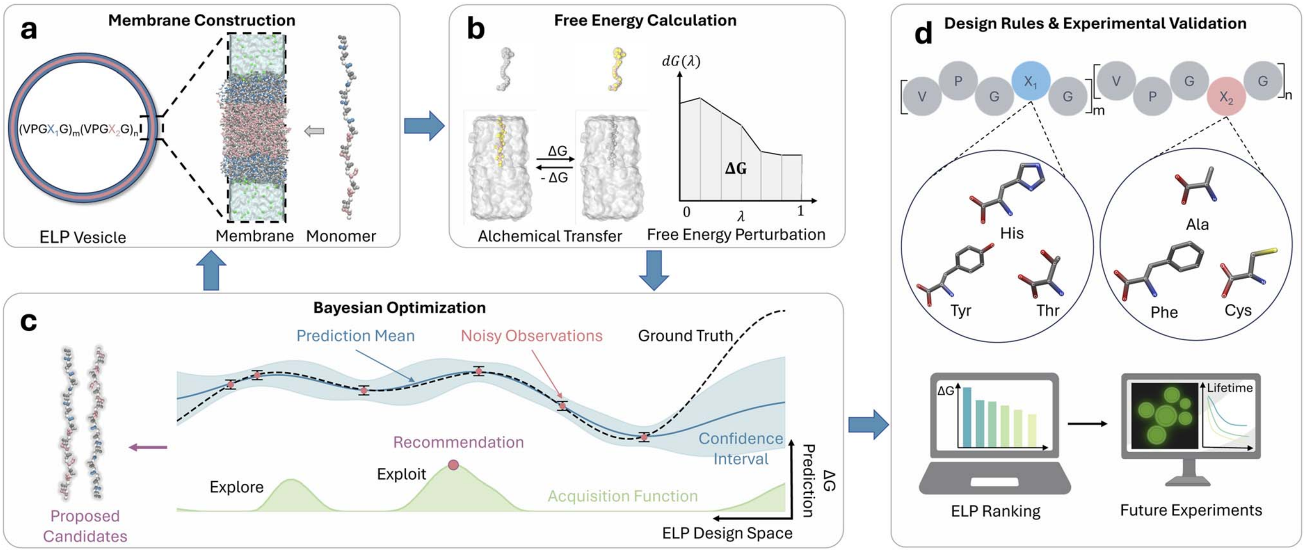

Fig. 2 Schematic illustration of the active learning strategy for data-driven screening of ELPs capable of forming thermodynamically stable vesicles. (a) Molecular modeling of a bilayer ELP vesicle. The simulated region of a vesicle is approximated as a planar sheet assembled by amphiphilic diblock ELPs with a generic sequence of $(\mathrm{VPGX_1G})_m(\mathrm{VPGX_2G})_n$, in which X_1 is a hydrophilic guest amino acid shown in blue, X_2 is a hydrophobic guest amino acid shown in pink, and m and n are the degree of hydrophilic and hydrophobic block repeats. (b) Alchemical free energy calculations are performed to quantify the stability of the bilayer *via* the transfer free energy ΔG of an ELP candidate from the membrane into solvent. The colored ELP denotes the coupled state while the gray ELP denotes decoupled state, which are connected through a reversible alchemical pathway governed by the coupling parameter λ. Calculations are performed by free energy perturbation (FEP) using the mbar tool in the alchemlyb software library.[45] (c) The free energy simulation data are used to train a Gaussian process regression (GPR) surrogate model which is then interrogated by Bayesian optimization (BO) to identify and prioritize unsampled ELP candidates within the design space most likely to possess high values of the objective function for the next round of active learning. The green line indicates the BO acquisition function used to rank candidates and magenta point is the selected unsampled candidate for the next free energy calculation. The iterative loop is terminated when no further improvements in the top performing candidates identified in consecutive rounds are observed and/or the posterior of the GPR model stabilizes and no longer updates with additional rounds of data collection and model retraining. (d) Inference of design rules from the terminal GPR surrogate model and schematic illustration of down-selection of the ELP candidate space for future experimental validation. The simulated ELP molecules and bilayer are visualized using VMD,[46] structures of amino acids are drawn with Avogadro,[47] and cartoons are obtained from BioRender (https://biorender.com).

Herein, we present a high throughput virtual screening (HTVS) pipeline that integrates coarse-grained (CG) simulations, alchemical free energy calculations, Gaussian process regression (GPR), and Bayesian optimization (BO) to identify top candidates from a library of putative amphiphilic diblock ELPs that are predicted to form thermodynamically stable and mechanically robust vesicles (Fig. 2). Although we focus on bilayer vesicles, this modular pipeline can be readily retargeted to optimize alternative assembled morphologies such as micelles or gels. This work builds upon our previous studies[19,44] by introducing a computationally efficient alchemical protocol to evaluate vesicle stability, expanding the space of ELP candidates with calculated stabilities by over 200%, and identifying novel ELP candidates with predicted bilayer stabilities up to 140% higher than any previously identified candidate. The active learning-guided screen efficiently filters the high-performance ELP sequences to refine the large design space into a smaller number of top performing candidates to guide and focus future experimental synthesis and characterization efforts. Moreover, the predictive capability of the model enables large-scale analysis of amino acid residue preferences in the hydrophilic and hydrophobic blocks of amphiphilic ELPs to expose novel design principles that can inform new understanding and rational engineering of these molecules. This HTVS pipeline is broadly transferable and we make it freely available as an open source tools to accelerate the design and optimization of peptide-based biomaterials with desired structural or functional properties.

2 Methods

2.1 Computational modeling of ELP vesicle bilayers

The first step in our screening pipeline is to develop a computational framework for simulating the bilayer vesicles comprising amphiphilic ELPs. Instead of simulating the entire ELP vesicle, we follow a protocol similar to that used for lipid bilayers[48] by commencing from a preassembled vesicle and focusing on a local region that can be reasonably approximated as a flat bilayer. Conceptually, our *in silico* modeling approach endeavors to evaluate the stability of ELP bilayers that may be assembled by a variety of experimental techniques including emulsion transfer,[19] templated synthesis,[10] and self-assembly.[18,49] We focus on the bilayer vesicle chassis because it presents prevalent experimental practice in which vesicles are preassembled (*e.g.*, by emulsion transfer) and reside in a metastable state, which make membrane lifetime a question of thermodynamic stabilization. Thus, our design objective is to use this computational model to quantify the thermodynamic driving force for extraction of an ELP peptide from the bilayer and discover ELP sequence variants that maximize this driving force to stabilize these vesicles against dissociation. Our active learning campaign, therefore, seeks to engineer ELP sequences to promote the stability of a preassembled bilayer, not, necessarily, to spontaneously self-assemble into a bilayer vesicle geometry against other competing morphologies such as micelles or gels.

The ELP candidate space in this work is defined by the mother sequence $(\text{VPGX}_1\text{G})_m(\text{VPGX}_2\text{G})_n$, where X_1 and X_2 are guest residues. We restrict residue X_1 to be hydrophilic and X_2 to be hydrophobic to promote an ELP bilayer architecture with the X_1 block oriented towards the aqueous environments to form either side of the bilayer and the X_2 hydrophobic blocks sequestered within the hydrophobic core.[10,18,21,35] We restrict the hydrophilic and hydrophobic pentamer repeat numbers to be such that $m + n = 9$ or 10 to limit the computational cost of simulations and facilitate HTVS. Following previous work,[44] all-atom ELP structures were built using PyMOL[50] and then coarse-grained with the Martini 2.2 force field.[51] While all-atom simulations offer higher accuracy, they are computationally expensive for such large systems within a HTVS framework. The coarse-grained modeling approach permits the simulation of ~100 nm^2 membrane patches and efficient convergence of the free energy calculations used to assess stability. Studies have shown that ELPs can adopt ordered structures, such as β-turns and β-spirals.[23–25] However, these conformational changes are rare and transient, with lifetimes on the sub-nanosecond scale, and growing evidence supports a picture in which ELPs remain highly dynamic and largely disordered in the condensed phase.[26–30] As such, we chose not to assign specific secondary structures during coarse-graining and modeled the entire ELP polymer as a random coil. It would be computationally laborious to estimate the transition temperature of each ELP considered in the active learning cycle, so we simulate all systems in the random coil state at a temperature of 300 K and pressure of 1 bar. We assume that the correct rank ordering of the thermodynamic stability of ELP bilayers is preserved under these assumptions of our computational model and approach, and that this strategy can permit the identification of high performance ELP candidates predicted to assemble stable bilayer vesicles. To test the effect of secondary structure on stability, we assigned β-turn structure across the entire hydrophobic block for four representative candidate sequences. We observed that poor candidates such as R_6M_3 and E_4L_6 exhibited some limited additional stabilization under the imposition of the β-turn, whereas good candidates such as H_5F_5 and H_6F_3 showed no change in stability within error bars (Fig. S10). Furthermore, the overall rank ordering of stability among these four test sequences was preserved.

A bilayer patch was constructed by arranging 100 ELP monomers in each leaflet on a 10 × 10 grid with ~0.9 nm intermolecular spacing. The initial grid arrangement served only as a starting configuration, and the bilayer models were subsequently simulated and equilibrated under periodic boundary conditions and a barostat to reach its equilibrium density given the chosen thermodynamic conditions. A hydrophobic core was formed by orienting the hydrophobic block of each leaflet of ELPs toward the bilayer interior.[44] The system was then placed in a cuboidal box with x and y dimensions of ~9 nm, a z dimension of ~38 nm, and periodic boundaries were employed in all dimensions. The bilayer was oriented in the x–y plane, and the extended z dimension is designed to eliminate artifacts associated with interactions between periodic copies of the bilayer in this dimension (Fig. 2a). The box was then solvated with non-polarizable Martini water beads to a density of 1.0 g cm^{-3}. In the ELP bilayer model, the C-terminus is buried within the hydrophobic core, and so we represent it in a neutral form (–COOH) represented by a Martini P5 bead. The N-terminus, in contrast, is exposed in a solvent environment and remains a charged state ($-\text{NH}_3^+$) represented by a Qd bead. Charges of ionizable residues were assigned under a physiological pH = 7.4 (E: −1, D: −1, K: +1, R: +1). In each system, counterions were added to maintain charge neutrality by randomly inserting Na^+ (Qd) or Cl^- (Qa) ions into the water region.

The system was then equilibrated using classical molecular dynamics simulations. Energy minimization was performed using the steepest descent algorithm to eliminate forces larger than 1000 kJ mol^{-1} nm^{-2}. After minimization, the system was equilibrated by a 10 ps *NVT* simulation followed by a 10 ns *NPT* equilibration at 300 K and 1 bar. Finally, an *NPT* production run of 200 ns at 300 K and 1 bar was conducted to relax the system prior to subjecting it to free energy calculations (Section 2.2). For *NPT* simulations, semi-isotropic pressure coupling was employed as in lipid membrane simulations.[48] The Berendsen barostat[52] was used for equilibration with a time constant of 12 ps and a compressibility of 3.0×10^{-4} bar^{-1}. The Parrinello–Rahman barostat[53] was employed for production runs with a time constant of 12 ps and a compressibility of 3.0×10^{-4} bar^{-1}. The temperature was coupled by a velocity-rescale thermostat[54] with a time constant of 1 ps with separate coupling to the ELPs and the rest of the system (*i.e.*, water and ions). The time step was set to 20 fs and Newton's equations of motion integrated by the leap-frog algorithm.[55] Lennard-Jones interactions were smoothly shifted to zero at a cutoff of 1.1 nm and electrostatics were treated using the reaction field method with $\epsilon_{\text{rf}} = \infty$ and $\epsilon_{\text{r}} = 15$ as recommended for the non-polarizable Martini 2.2 water model.[56,57] All simulations were performed using the Gromacs 2023.1 simulation suite.[58] The visualization and rendering of simulation trajectories were conducted using VMD.[46] We make the simulation codes and screening data available as a public Github repository at **https://github.com/Ferg-Lab/ELP_Simulation** that is also accessible *via* a persistent doi at **https://doi.org/10.5281/zenodo.15778533**.

2.2 Free energy calculation by alchemical transfer

We employ the free energy cost Δ*G* to extract a single ELP monomer from the membrane into the aqueous phase as a quantitative measure of the thermodynamic stability of the ELP bilayer vesicle (Fig. 2b). We previously employed umbrella sampling and the weighted histogram analysis method

(WHAM) for this purpose,[44,59,60] but this approach requires extensive sampling, large numbers of umbrella windows, and carefully constructed pathways to handle strong pulling forces, avoid hysteresis, and assure good convergence.[61] Free energy calculations along a fictitious pathway, such as free energy perturbation (FEP)[62,63] and thermodynamic integration (TI),[64–66] are widely employed to compute solvation[67] and binding free energies.[68–70] The double decoupling method (DDM)[71–73] is a common approach, but it requires co-alchemical ions for charge balance when decoupling charged molecules,[74–77] which becomes challenging for ELPs with multiple charged residues in the hydrophilic block. Instead, we leverage the recently developed alchemical transfer method,[78–81] to simultaneously decouples an ELP from the membrane and couples its "ghost" copy in the water phase. This ensures charge neutrality by construction and enables efficient free energy calculation within a single set of simulations. Benefiting from the alchemical transfer approach, the number of simulations was reduced from over 400 to 76 windows and led to a ∼4-fold speedup in the free energy calculation. Moreover, the computed per-residue free energies for various ELPs showed better agreement with the expected partition free energy of amino acid residues from water to protein condensates.[82]

The alchemical simulations employed the stochastic dynamics integrator with a friction constant of 1.0 ps^{-1}. Each window was equilibrated for 40 ns followed by 100 ns production at 300 K and 1 bar using a Langevin thermostat[83] and semi-isotropic pressure coupling by a Parrinello–Rahman barostat.[53] Free energy changes were estimated using the multistate Bennett acceptance ratio (MBAR)[84] as implemented in the alchemlyb software library,[45] with uncertainties evaluated by five-fold block averaging. Full details of the alchemical calculations are provided in Section S1 of the SI and the thermodynamic cycle for computing the free energy of extracting an ELP monomer from the membrane to the water region ΔG is represented in Fig. S1. MBAR was selected for its superior convergence behavior compared to thermodynamic integration (TI)[64] and Bennett acceptance ratio (BAR)[85] in our systems, which was particularly pronounced for restraint-related terms (Fig. S2). Illustrations of the representative cumulative free energy at different stages and the convergence profiles for a bilayer membrane composed of $\mathrm{T}_5\mathrm{I}_5$ ELPs are shown in Fig. S3 and S4. The evaluation of ΔG for each ELP requires approximately 95 GPU-h on an NVIDIA A100 GPU.

We validated our alchemical transfer free energy pipeline in five benchmark validations, the full details of which are provided in Section S2 of the SI. In the first three validations, we computed the solvation free energies of three small molecules – 1,2,3-trichloro-5-(2,5-dichlorophenyl)benzene (TCDP), decane, and F-uracil – using all-atom simulations in which we transferred the molecules from a water slab to a vacuum slab employing the alchemical transfer approach (Fig. S5a). In the fourth validation, we computed the partition free energy of a guanine molecule from water into a POPC membrane using the coarse-grained Martini 2.2 force field[51,57] (Fig. S5b). In all four cases, we obtained agreement with previously reported values computed using DDM schemes or umbrella sampling with a mean average error of 0.95 kcal mol^{-1} (Fig. S5c). In our fifth validation, which is a direct test of the approach for the present ELP system, we performed a symmetric decoupling of one ELP monomer within a $(\mathrm{VPGHG})_2(\mathrm{VPGAG})_2$ ELP bilayer and coupling of another ELP monomer within the same bilayer (Fig. S6). The expected value of this free energy change is zero, by construction, and the calculated value of $\Delta G = (0.02 \pm 0.05)$ kcal per mol per residue is consistent with the expected value within error bars. These five validations lend confidence that our alchemical transfer method presents a robust and accurate means to compute transfer free energies.

2.3 Closed-loop optimization of ELP membrane stability

Having defined a computational measure of membrane stability *via* alchemical free energy calculations (Section 2.2), we aim to maximize the free energy ΔG of extracting an ELP monomer out of the membrane so as to maximize the stability of the membrane with respect to the sequence of the constituent ELPs. This process can be regarded as optimization of a black box function $y = f(\mathbf{x})$, where the input $\mathbf{x}$ is the sequence of the ELP dictated by the identity of the guest residues and the number of their repeats $\{\mathrm{X}_1, \mathrm{X}_2, m, n\}$, and the target output y is the free energy ΔG. A closed-loop active learning pipeline was employed to traverse the ELP space more efficiently and minimize the number of simulations required to discover high-performing candidates. A general scheme is shown in Fig. 2a–c, where we employ alchemical transfer free energy calculations to evaluate the performance of ELP candidates, train a data-driven surrogate model employing GPR to predict the performance of all remaining candidates within the design space, then employ BO to prioritize the next candidates for free energy calculations. We terminate the active learning search when we cease to see improvements in ΔG with successive active learning rounds or the GPR posterior stabilizes and no longer changes with additional rounds of data collection and retraining. Simulation tools, codes, and Jupyter Notebooks implementing our pipeline are hosted on GitHub **https://github.com/Ferg-Lab/ELP_Simulation** that is also accessible *via* a persistent doi at **https://doi.org/10.5281/zenodo.15778533**.

2.3.1 Gaussian process regression. Given the ΔG values computed for all ELP candidates considered to date, we wish to predict the ΔG values for the remaining candidates $\hat{f}(\mathbf{x})_j$, $j \in$ unsampled in our design space by constructing a surrogate model. In active learning, the surrogate model $\hat{y} = \hat{f}(\mathbf{x})$ is frequently furnished by GPR models that intrinsically provide the predictive means and uncertainties that are required for BO.[86]

A GPR model is defined by a kernel function defining the similarity of any pair of candidates.[87] In the present case, we therefore require a means to define the similarity between ELP sequences. String kernels, such as the local-alignment kernel,[88] oligo kernel,[89] and physico-chemical descriptors based kernels[90] are all possible options, but, following prior work,[44] we adopt the generic string (GS) kernel proposed by Giguère *et al.*[91] The GS kernel defines the similarity between any pair $(\mathbf{x}, \mathbf{x}')$ of strings of amino acids with lengths $|\mathbf{x}|$ and $|\mathbf{x}'|$ as,

$$\mathrm{GS}\left(\mathbf{x},\mathbf{x}',L,\sigma_{\mathrm{p}},\sigma_{\mathrm{c}}\right):=\sum_{l=1}^{L}\sum_{i=0}^{|\mathbf{x}|-l}\sum_{j=0}^{\left|\mathbf{x}'\right|-l}\exp\left(-\frac{(i-j)^2}{2\sigma_{\mathrm{p}}^2}\right)\times \exp\left(-\frac{\left\|\boldsymbol{\psi}^l(x_{i+1},\ldots,x_{i+l})-\boldsymbol{\psi}^l\left(x'_{j+1},\ldots,x'_{j+l}\right)\right\|^2}{2\sigma_{\mathrm{c}}^2}\right), \quad (1)$$

where $\boldsymbol{\psi}^l(x_1, \ldots, x_l) = (\boldsymbol{\psi}(x_1), \ldots, \boldsymbol{\psi}(x_l))$ and $\boldsymbol{\psi}(x_i)$ is the column corresponding to amino acid x_i in the BLOSUM62 substitution matrix.[92] Mathematically, the GS kernel compares all substrings of length l between two strings under a product of two Gaussian kernels. The parameter σ_{p} in the first Gaussian controls the shifting contribution term encoding the shift in starting positions of the substrings, and the σ_{c} parameter in the second Gaussian controls the characteristic bandwidth on the BLOSUM62 similarity of the two substrings. Conceptually, the product of the two Guassian kernels functions like a logical AND gate,[93] such that a pair of substrings are judged to be similar if both their positional offset is small relative to σ_{p} and their BLOSUM62 distance is small relative to σ_{c}. The overall kernel sums over all substrings of length $l = 1 \ldots L$ and all substring offsets. The GS kernels can be viewed as a generalization of other well-known kernels.[91] For instance, the GS kernel reduces to the oligo kernel[89] when $\sigma_{\mathrm{c}} \rightarrow 0$ and to the radial basis function (RBF)[94] when $L \rightarrow \infty$ and $\sigma_{\mathrm{p}} \rightarrow 0$. In this work, the kernel parameters, σ_{p} and σ_{c}, were optimized during each training round by maximizing the log-likelihood of the training data. The maximum string length parameter $L = 50$ was selected by grid search in the first active learning round and fixed for all subsequent rounds. We implemented the GPR using the BoTorch libraries.[95]

2.3.2 Bayesian optimization. The final step in the active learning cycle is to pass the predictions along with the uncertainties from the surrogate GPR model to a BO routine to prioritize as yet unsampled ELP sequences for alchemical transfer free energy calculations. Candidates are rank ordered by the BO under a so-called acquisition function. A number of such functions are available that balance different degrees of exploitation – prioritizing candidates likely to possess high values of the objective function – and exploration – prioritizing candidates in poorly sampled regions of the search space. The surrogate model is used primarily to guide sequence selection toward promising regions of design space rather than to achieve globally uniform predictive accuracy across the entire library. In the initial rounds of this work, we employed the batched quasi-Monte Carlo (qMC) batch noisy expected improvement (qNEI) acquisition function[95] as a balanced exploit-explore search strategy. After we observed no further improvements in ΔG for 20 successive rounds, we switched to the batched qMC batch upper confidence bound (qUCB) acquisition function,[96] with the exploit-explore trade-off selected to pure exploit. We implemented the BO using the BoTorch libraries.[95] In all rounds we selected a batch size of $q = 2$ ELP candidates that enabled us to make use of our parallel compute resources in conducting our alchemical free energy calculations.

3 Results and discussion

3.1 Definition of ELP sequence space for active learning search

The candidate space of ELP sequences comprises the family of diblocks $(\mathrm{VPGX_1G})_m(\mathrm{VPGX_2G})_n$, or, for brevity, $\mathrm{X}_{1m}\mathrm{X}_{2n}$, where X_1 is one of the hydrophilic amino acid residues (except proline) classified by the Kyte–Doolittle hydropathy scale[97] {G, T, S, W, Y, H, E, Q, D, N, K, R}, and X_2 is one of the hydrophobic residues {I, V, L, F, C, M, A}. We acknowledge that using a single hydropathy scale inevitably introduces bias into residue classification, particularly for borderline amino acids whose classification may vary under alternative hydrophobicity scales, and cause our library to miss potentially high performing candidates. In principle, we could conduct a more comprehensive search of ELP space by constructing the design library as the union of ELP sequences designed under a number of scales, but in this work we elect to employ the Kyte–Doolittle scale as one of the most commonly used hydrophobicity scales and, as a scale based on the relative tendencies of amino acids to be buried within protein interiors *versus* exposed to solvent, an appropriate scale for amphiphilic ELP systems where the interplay between solvent exposure and burial within hydrophobic domains is central to membrane stability.[98–102] The sequence length ($m + n$) is a key factor contributing to the stability of resulting vesicles, with longer sequences typically employed in experimental work. For example, Schreiber *et al.*[18] tested ELPs of different lengths with a repeat number of up to 70, while Sharma *et al.*[19] used a diblock ELP with a repeat number of 96. Free energy calculations become extremely expensive for large molecules due to the computational costs of simulating large systems, slow equilibration times, and the need for increasing numbers of windows to achieve good overlap in the alchemical path. This prompts the question of whether calculations of ΔG for shorter chains ($m + n \approx 10$) can reflect the trends of ΔG for longer chains ($m + n \approx 100$) typically considered in experimental settings.

To probe this question, we considered ELPs consisting of equal-length hydrophobic and hydrophilic blocks ($m = n$) and considered three ELP sequences, $\mathrm{H}_n\mathrm{A}_n$, $\mathrm{K}_n\mathrm{I}_n$, and $\mathrm{T}_n\mathrm{I}_n$, with $n =$ 2–6. The guest residues in these sequences were selected to span both charged and neutral hydrophilic residues and large and small hydrophobic residues. We assembled bilayer membranes for all 15 sequences (Section 2.1) then subjected the relaxed membranes to alchemical free energy calculations to measure the free energy cost ΔG to extract an ELP from the membrane into solvent (Section 2.2). We observe a strongly linear relationship in ΔG as a function of n for each of the three series (Fig. 3a), such that the ΔG per amino acid is approximately constant for a particular choice of X_1 and X_2 (Fig. 3b). While we cannot rule out the possibility that these linear trends may not persist up to the very long sequences employed in experiments due to potentially new folding dynamics, aggregation modes, or entanglement effects, they lend confidence that the stability trends computed for short sequences are representative of the behaviors of longer sequences, and that

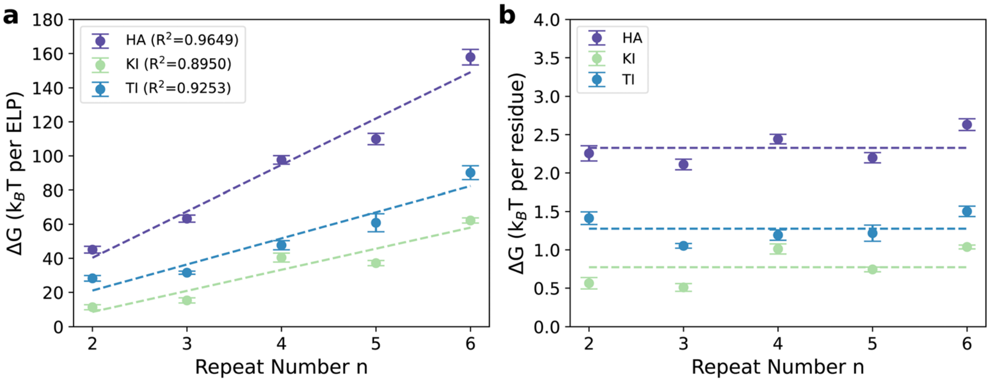

Fig. 3 The free energy ΔG to extract an ELP (VPGX$_1$G)$_n$(VPGX$_2$G)$_n$ from an equilibrated membrane into solvent as a function of repeat number n. We consider three chemically distinct sequences ELPs families – H$_n$A$_n$, K$_n$I$_n$, and T$_n$I$_n$ – with $n = 2$–6. (a) The ΔG to extract the full ELP molecule scales linearly with the repeat number n for each of the three series. (b) As a corollary, the ΔG per residue is approximately constant for each choice of X$_1$ and X$_2$. Dashed lines in (a) correspond to best least squares linear fits and those in (b) to means.

the rank ordering of the short sequence trends can provide a proxy for identifying promising choices of the X$_1$ and X$_2$ guest residues that can guide and focus subsequent experimental synthesis and testing. We observe that similarly strongly linear trends have been previously reported for the self-partition free energies of Nup98 FG domains during phase separation.[103] While the molecular details of FG domain condensation and ELP bilayer stability are distinct molecular processes, both systems share the principle that once the polymer reaches a sufficient length, each additional repeat unit experiences a similar physicochemical environment and therefore contributes approximately additively to the free energy.

On the basis of the linear relationship between ΔG and ELP length, we restrict our virtual screen to ELPs of a total pentamer repeat of $m + n = 9$ or 10 as a suitable regime for high throughput screening, and consider specific combinations of $(m, n) = [(3, 6), (4,5), (4, 6), (5, 5), (6, 4), (5, 4), (6, 3)]$. These seven (m, n) combinations together with the twelve options for the X$_1$ hydrophilic guest residue and seven options for the hydrophobic guest residue define an ELP candidates space of 12 × 7 × 7 = 588 amphiphilic diblocks of the form of (VPGX$_1$G)$_m$(VPGX$_2$G)$_n$. This represents a 2.5-fold expansion in the search space relative to prior work,[44] and we also note that all candidate sequences are, in principle, accessible to recombinant synthesis.[104]

3.2 High throughput screening of amphiphilic ELPs

The primary goal of the HTVS is to discover novel ELPs that maximize thermodynamic stability of peptidic vesicles for synthetic cells. Experimentally, the vesicles are fabricated by emulsion transfer[19] or templated assembly.[10,18,49] Computationally, we seek to mimic this experimental protocol by pre-assembling the ELPs into a bilayer rather than building the membrane patch from self-assembly. With these pre-assembled models, we then aim to search for sequences that maximize the thermodynamic cost to extract a single ELP from the membrane as a proxy for membrane stability. A limitation of this protocol is that competing aggregate structures (*e.g.*, micelles, gels) are not considered, but our rationale is that placing a pre-assembled vesicle into a deep thermodynamic free energy well maximizes its lifetime by minimizing its propensity to disaggregate or transition into alternative structures.

We commenced our active learning campaign by selecting an initial set of the 24 ELP sequences to subject to alchemical free energy calculations and initialize the first round of the active learning search (Fig. 2). This initial set of sequences was designed such that each X$_1$ and X$_2$ guest residue appeared at least once, in order to provide broad initial coverage of the 588-candidate sequence space. We conducted 24 rounds of active learning with the initial 20 rounds employing a batched noisy expected improvement acquisition function to balance exploitation and exploration and propose diverse candidates, while subsequent 4 rounds employed a pure exploit batched upper confidence bound acquisition function to extract the top performing candidate within the explored regions of sequence space. The 24 sequences considered in the initial round of the search together with batches of two candidates considered in each of the subsequent 24 rounds led us to consider a total of 24 + 2 × 24 = 72 ELP sequences, comprising a little over 12% of the 588-candidate search space. A comprehensive listing of the particular ELP sequences considered in each round of the active learning search along with their calculated ΔG values is provided in Table S1 in the SI, and hosted as a machine readable csv file along with the simulation codes on a public Github repository at **https://github.com/Ferg-Lab/ELP_Simulation** that is also accessible *via* a persistent doi at **https://doi.org/10.5281/zenodo.15778533**.

We present in Fig. 4a a summary of the active learning search showing the calculated ΔG values for the ELP sequences sampled in each round over the course of the active learning campaign. The most stable candidate in the initial round of the search (H$_6$I$_3$) possessed a free energy of extraction from the membrane bilayer of $\Delta G = 2.11\ k_BT$ per residue. The most stable ELP sequence identified by the search was discovered in round 10 (H$_6$F$_3$) with a $\Delta G = 2.87\ k_BT$ per residue after consideration of just 44/588 ELP candidates. Although we cannot rule out the possibility that higher performing sequences remain to be

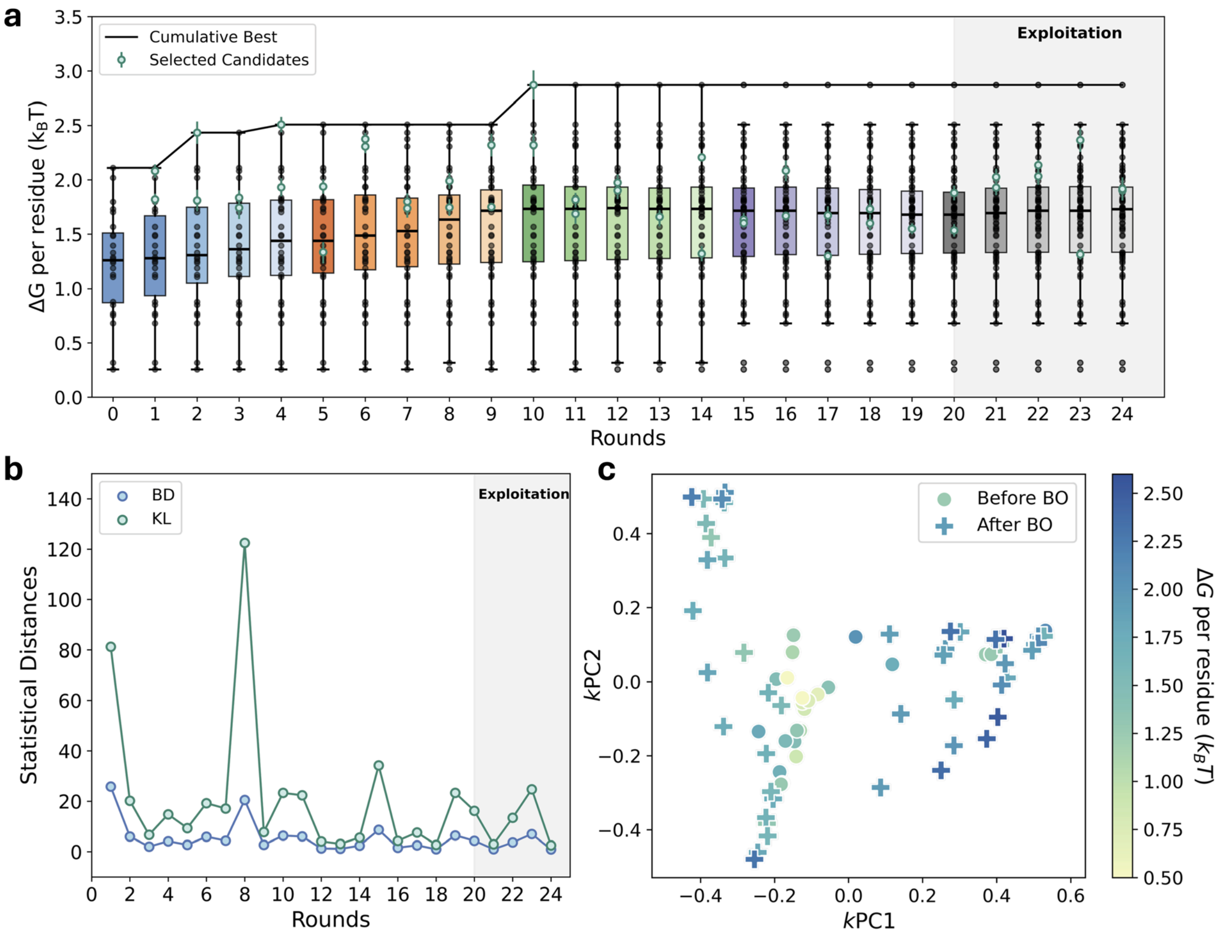

Fig. 4 Active learning screen for amphiphilic diblock ELP sequences forming highly stable bilayer vesicles. (a) Distribution of the transfer free energy ΔG of an ELP candidate from the membrane into solvent over the 24 rounds of the active learning campaign. More positive ΔG values correspond to more stable membranes. The cumulative best ELP candidate is denoted by the black line. The middle line in the box plot represents the median value of the ΔG values and the box includes the middle 50% of the data. The whisker is defined as 1.5 times the inter-quartile range. The initial round comprises 24 initial ELP sequences. Each successive round employs BO to select two ELPs for subsequent alchemical free energy calculations (green points), which are then fed into the loop to retrain and update the GPR model for subsequent predictions. In the first 20 rounds, we employ a qNEI acquisition function that balances exploitation and exploration. In the final five rounds, we employ a pure exploit qUCB acquisition function. (b) The Bhattacharyya distance and Kullback–Leibler divergence between the posterior distributions of GPR models over the whole design space in successive rounds, i and (i + 1), indicate convergence of the GPR posterior distribution after round 10. (c) A two-dimensional kPCA projection of the 72 ELPs sampled over the course of the active learning campaign. The circles correspond to the 24 candidates from the initial round, and the crosses to those selected over the subsequence 24 rounds by Bayesian optimization. Points are colored by the calculated ΔG. The most stable ELP sequence identified by the search was discovered in round 10 (H_6F_3) with a $\Delta G = 2.87\ k_BT$ per residue.

discovered within the candidate space, we observed no further improvements in ΔG over the remaining 14 rounds of the active learning screen. Calculation of the Bhattacharyya distance and Kullback–Leibler divergence between successive GPR models indicates that the GPR posterior converges after round 10, and that the incorporation of additional measurements into the GPR training does not lead to substantial updates to the posterior predictions of the model (Fig. 4b). To gain understanding of the progression of the active learning search over ELP candidate space, we applied kernel principal components analysis (kPCA)[105] to the 72 sequences considered over the active learning campaign using the same kernel as that employed in the terminal GPR. A 2D projection of the 72 ELP sequences into the two leading kPCs reveals the design space to comprise multiple distinct pockets accommodating high-performance ELPs (Fig. 4c). That the top performing sequences span diverse regions of this projection of the design space indicates that the GPR/BO has identified sequence-diverse top performing candidates and highlights the complex structure–property relationships that emerge from the combinatorial diversity of amino acid residues within the ELP candidate sequences.

3.3 Thermal stability of high-performing ELP membranes

Having identified top-performing ELP sequences under the active learning campaign using the transfer free energy ΔG as our measure of membrane stability, we subsequently sought an independent *in silico* verification that ELPs with large ΔG indeed exhibit greater thermodynamic stability. To do so, we

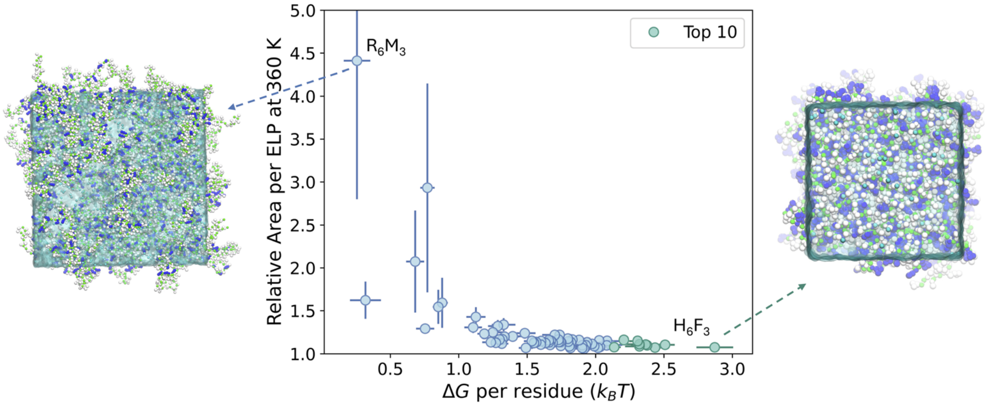

Fig. 5 Evaluation of thermal stability of the ELP membranes by computing the change in membrane area between 300 K and 360 K. Higher transfer free energies ΔG correspond to smaller relative area changes upon heating, with a Spearman correlation coefficient of $\rho_S = -0.69$ ($p = 1.5 \times 10^{-11}$). We visualize the membranes at 360 K of the most stable (H_6F_3) and least stable (R_6M_3) ELPs as ranked by ΔG.

simulated the assembled ELP bilayer membrane at a higher temperature of 360 K and assessed the relative area of the membrane per ELP under the expectation that more thermostable membranes are able to maintain their membrane integrity and exhibit more limited thermal expansion in the lateral dimension due to the stronger intermolecular interactions.[106] We acknowledge the limited temperature transferability of the Martini model and its inability to capture the LCST-like behaviors of ELPs, but assume that while the predictions of the coarse-grained model may not be quantitatively accurate, that the rank ordered stability of the various sequences should be preserved. We present in Fig. 5 the calculated area per ELP at 360 K relative to that at 300 K ($A_{\text{relative}} = A_{\text{360K}}/A_{\text{300K}}$) for all 72 ELPs considered over the active learning campaign. The results show the ELPs possessing large ΔG values maintain intact, compact membrane structures with limited lateral expansion, while the low ΔG candidates undergo significant lateral expansion and, in extreme cases, membrane dissociation. The moderately strong and statistically significant Spearman correlation coefficient $\rho_S = -0.69$ ($p = 1.5 \times 10^{-11}$) between A_{relative} and ΔG further supports the notion that the selected top ELP candidates offer promising candidates for experimental testing in the development of highly stable vesicles for synthetic cells and also suggests that A_{relative} can serve as a computationally efficient alternative to expensive extraction free-energy evaluations and a practical proxy for membrane stability.

3.4 Performance and residue preferences in stable ELP membranes

Relatively little experimental data for the thermodynamic stability of short ELP sequences exist, but Schreiber *et al.* experimentally obtained vesicles of an H_5L_4 ELP—along with some other short chain ELPs $H_{10}L_4$, H_5A_5, H_5V_5, and $D_{10}V_5$—and demonstrated these vesicles to be stable for hours.[18] This particular sequence was considered in our initial round of the active learning search and has an extraction free energy of $\Delta G = 1.2\ k_BT$ per residue. We present in Fig. 6a a graphical representation of the ΔG values of the top ten most stable ELP sequences identified in our active learning screen. These sequences identified by our search possess stabilities in the range $\Delta G = 2.14–2.87\ k_BT$ per residue, representing an increase of 78–139% beyond the $\Delta G = 1.2\ k_BT$ per residue stability of H_5L_4. This suggests that vesicles formed by these sequences may be more stable and longer lived than the experimentally validated H_5L_4 ELP.

We then sought to understand which combinations of hydrophobic and hydrophilic guest residues tend to produce highly stable bilayer membranes with the goal of extracting interpretable design rules from the analysis of our active learning screen. Intuitively, hydrophobicity can be an important factor dominating the stability ELP chassis membranes, with higher hydrophobicity potentially correlating with a higher ΔG. We computed the Grand Average of Hydropathy (GRAVY) values for the whole sequence, the hydrophilic block, and the hydrophobic block of ELPs using 28 hydropathy scales[107] and correlated them with ΔG of the evaluated candidates. However, we observed only weak to moderate mutual information (MI) values between the extraction free energy and sequence hydrophobicity, with the highest MI = 0.57 nats ($p = 0.002$) for the Roseman scale applied to the hydrophilic block (Fig. S7). These results expose relatively moderate shared information between the hydrophobicity of the ELP sequence and the extraction free energy ΔG, and that the stability is only partially explained by simple measures of hydropathy. This motivated us to search for more nuanced, multi-body design rules for engineering vesicle stability.

Accordingly, we next analyzed the influence of particular X_1 and X_2 guest residues on the ΔG from the 72 ELP sequences evaluated in the active learning campaign. Specifically, we classified each ELP as high or low stability using a cutoff of $\Delta G = 1.2\ k_BT$ per residue corresponding to the value of the experimentally demonstrated H_5L_4 sequence that is known to form vesicles stable for hours at room temperature.[18] We illustrate in Fig. 6b the counts of the number of ELP candidates falling into the stable ($\Delta G > 1.2\ k_BT$ per residue) and unstable ($\Delta G \leq 1.2\ k_BT$ per residue) categories partitioned by the identity of the

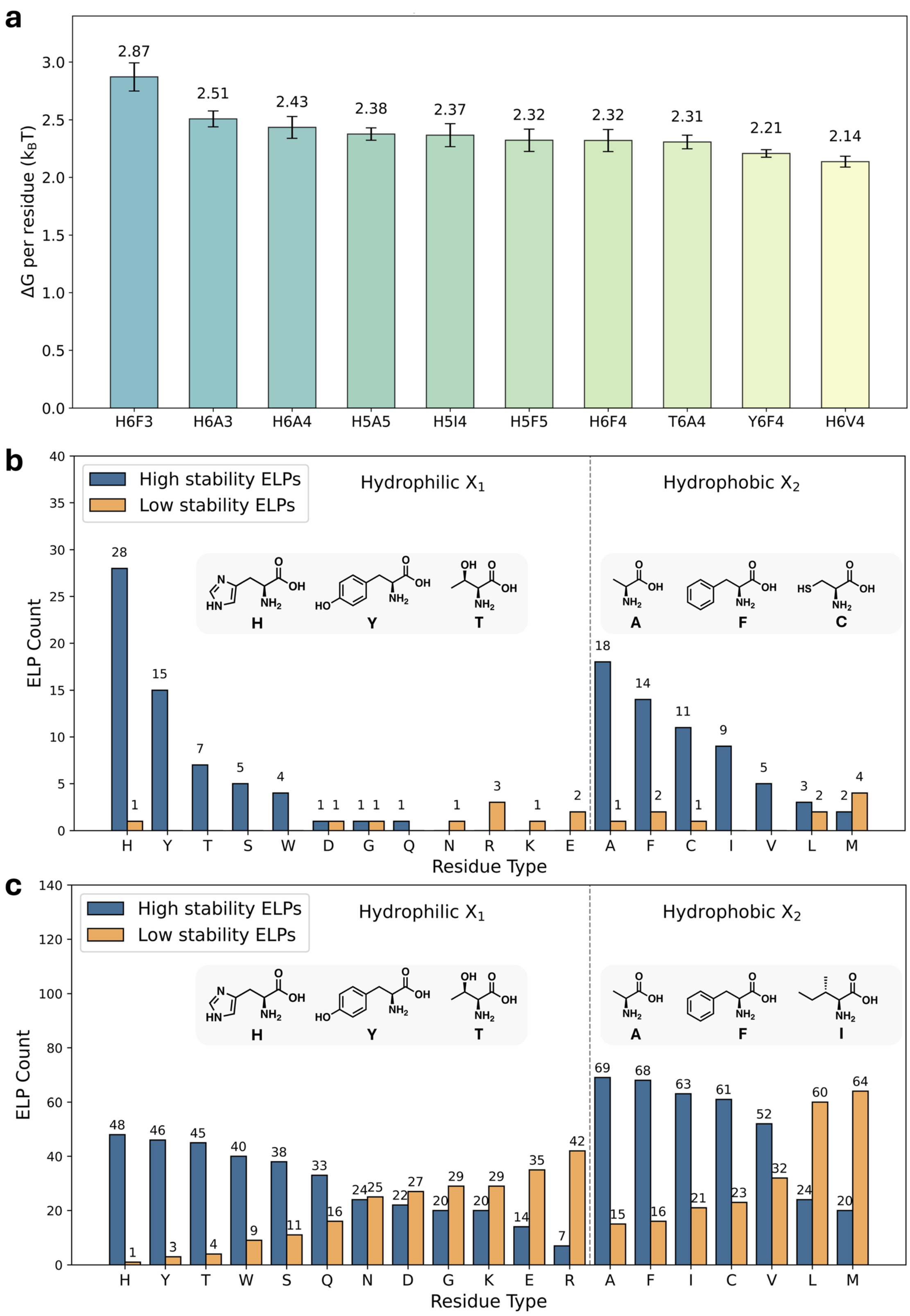

Fig. 6 Top ELPs and design rules obtained from the active learning-guided computational screening. (a) Bar plots showing the ΔG of the top 10 ELPs filtered from the ELP library. The statistical uncertainty is estimated by five-fold block averaging. (b) Analysis of residue preferences at the hydrophilic and hydrophobic guest residue positions X_1 and X_2 among the 72 diblock amphiphilic ELPs considered over the active learning campaign. (c) Analogous plot to panel (b) but pertaining to the predictions of the terminal GPR model over all 588 ELPs in the candidate space. The ELPs are classified into two categories (high stability and low stability) using a threshold of 1.2 k_BT per residue corresponding to a previously reported ELP H_5L_4 capable of forming stable vesicles. The blue and orange bars represent the occurrences of ELPs appearing in high and low stability ELPs, respectively. The number on each bar denotes the number of ELP candidates falling into each class. The top three residues in the X_1 and X_2 guest positions promoting membership of the stable ELP class are illustrated using ChemDraw in the inset.

hydrophilic guest residue X_1 or hydrophobic guest residue X_2. Essentially, we ask the questions: (i) of all the ELP sequences considered with a particular hydrophilic guest residue X_1, how many are more and less stable than H_5L_4 regardless of the identity of the hydrophobic guest residue X_2 (Fig. 6b, left), and (ii) of all the ELP sequences considered with a particular hydrophobic guest residue X_2, how many are more and less stable than H_5L_4 regardless of the identity of the hydrophilic guest residue X_1 (Fig. 6b, right)? In doing so, we deliberately exclude consideration of correlations between the residue identity in the X_1 and X_2 positions in favor of a simple and interpretable appraisal of residue preferences in one guest residue position marginalized over the identity of the residue in the other guest position. In regards to hydrophilic block guest residue, we observe that the active learning screen has prioritized the selection of ELP sequences with histidine (H) and tyrosine (Y) occupying the X_1 hydrophilic guest residue position, with 29 and 15, respectively, of the 72 sequences screened over the course of the active learning campaign possessing H and Y hydrophilic guest residues, all of which fall into the high stability category. Smaller numbers of threonine (T), serine (S), and tryptophan (W) residues were selected, numbering 7, 5, and 4, respectively, all of which were also classified as high stability. Turning to the hydrophobic block guest residue, the active learning screen prioritized the selection of ELP sequences with alanine (A), phenylalanine (F), cysteine (C) and isoleucine (I) occupying the X_2 hydrophobic guest residue position, with 19, 16, 12, and 9, respectively, of the 72 sequences in the active learning campaign possessing A, F, C, and I hydrophobic guest residues, of which 89%, 88%, 92%, and 100%, respectively, fall into the high stability category.

An attractive feature of an analysis of residue preferences founded on the 72 sequences considered within the active learning screen is that ΔG values are available for all sequences from the alchemical free energy calculations. Deficiencies of this analysis include that these 72 sequences represent only 12% of the 588 sequences comprising the ELP design space, the number of ELPs with a particular guest residue are unequally sampled, and the analysis does not account for temporal trends in the active learning screen associated with the changing predictions of the GPR surrogate model as it is exposed to more training data, and changes in candidate prioritization by the BO routine as the search space becomes increasingly explored. As such, we conducted a second analysis in which we analyzed the posterior predictions of the terminal GPR model that allows us to predict the ΔG values for all remaining $(588 - 72) = 516$ ELP sequences that were not subjected to alchemical free energy calculations over the course of the active learning campaign. This allows us to exhaustively analyze all possible combinations of guest residues and diblock sequence lengths and compare residue preferences on an equal footing regardless of the number of each guest residue actually sampled over the course of the active learning campaign, albeit under the caveat that ΔG predictions of the GPR surrogate model may carry significant uncertainties in regions of ELP sequence space where the model lacks substantial training data.

An analysis of the classification of ELP sequences into the high stability and low stability categories under the predictions of the terminal GPR model is presented in Fig. 6c. In the analysis of the hydrophilic guest residue X_1 preferences (Fig. 6c, left), for each selection of the X_1 residue identity there are a total of seven options for the hydrophobic guest residue X_2 and seven different diblock sequence length combinations, meaning that the total number of ELPs in the candidate space with a particular X_1 residue identity number $7 \times 7 = 49$. In the analysis of the hydrophobic guest residue X_2 preferences (Fig. 6c, right), for each selection of the X_2 residue identity there are a total of 12 options for the hydrophobic guest residue X_2 and seven different diblock sequence length combinations, meaning that the total number of ELPs in the candidate space with a particular X_2 residue identity number $12 \times 7 = 84$. Analyzing the posterior GPR predictions for guest residue preferences, we observe that the hydrophilic guest residue X_1 strongly favors histidine (H), tyrosine (Y), and threonine (T), with 48 (98%), 46 (94%), and 45 (92%), respectively, of ELP candidates possessing each of these residues in this position falling into the high stability class. Similarly, the hydrophobic guest residue X_2 strongly favors alanine (A), phenylalanine (F), isoleucine (I), and cysteine (C), with 69 (82%), 68 (81%), 63 (75%), and 61 (73%), respectively, of ELP candidates possessing each of these residues in this position falling into the high stability class.

The residue preference analyses using the 72 sampled sequences with calculated ΔG values (Fig. 6b) and predictions of the terminal GPR surrogate model over all 588 candidate ELPs (Fig. 6c) are in good agreement, with both analyses suggesting a preference for histidine (H), tyrosine (Y), and threonine (T) in the hydrophilic guest residue X_1, and for alanine (A), phenylalanine (F), cysteine (C), and isoleucine (I) in the hydrophobic guest residue X_2 to promote high stability ELP bilayers. These trends are consistent with physicochemical intuition in reflecting the role of hydrogen bonding and polar interactions in the hydrophilic region and non-polar hydrophobic aggregation and π–π stacking in the hydrophobic region in promoting stable bilayer membranes and, notably, these residues are commonly employed in the design of ELP vesicles in experimental studies. For example, histidine (H) is frequently used as the hydrophilic residue while isoleucine (I) and phenylalanine (F) are most often incorporated in the hydrophobic residues (Table S2). In contrast, charged residues consistently performed poorly, likely due to electrostatic repulsion between charged chains within the bilayer and their strong hydrophilicity, which may facilitate water permeation into the membrane and therefore reduce the membrane stability. To gain insight into the role of hydrogen bonding and π–π stacking, we backmapped the membranes of three selected top candidates – H_5A_5, H_6F_3, Y_6F_4 – to all-atom resolution using the Martini backmap tool,[51] relaxed the structures under all-atom molecular dynamics using the CHARMM36 force field,[108] and quantified the presence of ~0.2 hydrogen bonds and ~0.05 π–π stacking interactions within the membranes per residue in the ELP chain (Fig. S8). We observe that the overall membrane stability results from a complex interplay of many-body interactions and a variety of molecular forces, but these all-atom calculations are consistent with a role for hydrogen bonds and π interactions in mediating membrane stability.

4 Conclusions

In this work, we present an active learning computational screen that integrates CG molecular simulations, alchemical free energy calculations, GPR, and BO to guide the rational design of amphiphilic ELP diblock sequences capable of forming stable membrane bilayers for synthetic cells. Our approach employs molecular simulation to give a quantitative measure of the stability of ELP bilayer vesicles and leverages BO to maximize the thermodynamic stability of the vesicles. The iterative screening process converged within 24 cycles after sampling 72 ELP sequences corresponding to ∼12% of the 588-member candidate space defined by variations in guest residues and block lengths and requiring ∼6840 GPU-h of compute. From our screen, we identified a number of novel ELP sequences with predicted bilayer stabilities up to 140% higher than previously reported experimental systems[10,18,19] and which are strong candidates for experimental validation in the construction of robust synthetic cells. Additionally, we observed that high-stability ELPs tend to incorporate hydrophilic blocks with particular guest residues including histidine, tyrosine, and threonine, and hydrophobic blocks with alanine, phenylalanine, cysteine, and isoleucine. These trends align with chemical intuition but also expose the value of combining physical modeling with active learning to extract new understanding and predictive design rules for accelerated materials discovery.[18] Moreover, the framework employed in this work can be readily extended to optimize other measurable properties of interest that may be determined computationally or experimentally.[109–111] Altogether, this work presents a generalizable computational framework for the rational design of peptide-based materials and extracts interpretable design rules for ELP sequence–property relationships to guide future experimental design, synthesis, and testing.

We envisage a number of future directions for this work. ELPs are known to exhibit LCST-like transitions, enabling them to undergo phase transitions and structural changes in response to environmental stimuli such as temperature, salt concentration, and pH.[35] A compelling future direction would be to assess the stability of ELP-based vesicles under a broader range of thermodynamic conditions and probe the intersection of thermodynamic stability and phase behavior. In our proof-of-concept study, we observed decreasing stability of an H_6F_3 ELP membrane with increasing fraction of protonated histidines mimicking decreasing pH environment (Fig. S9). In addition, future studies could focus on the feasibility of decorating ELP membranes with functional proteins, channels, or signaling moieties through engineered ELP-protein fusions, therefore enabling more complex architectures and functions of synthetic cells.[10,34,112] Another important extension relates to the exploration of additional ELP vesicle properties, such as transport properties. In particular, the permeability of solutes or small molecules across ELP membranes is of great interest because cross-membrane transportation plays a central role in engineering synthetic cells to allow for exchange of mass and chemical signals with the external environment.[113,114] The concurrent optimization of both membrane stability and relevant biological properties or functions can lead to the design of ELP-based synthetic cells with desired functions. ELP sequences comprising different types of blocks may also display multi-phase transition behaviors, offering opportunities to fine control the physicochemical properties of vesicles and engineer more complex cellular functions. For example, Ibrahimova *et al.*[115] designed temperature-sensitive lipo-proteinosomes by incorporating ELPs into the membrane architecture, enabling thermally triggered cargo release from vesicles. Investigating such dynamic membrane behaviors may necessitate higher-resolution models, such as all-atom simulations, which are often prohibitively expensive for these large systems. However, recent advancements in coarse-grained force fields tailored for intrinsically disordered proteins, such as hydropathy-scale models,[116,117] Mpipi,[118] and Mpipi-T,[119] offer promising alternatives. Leveraging these coarse-grained models, we also aim to broaden the ELP design space by incorporating longer sequences and more complex architectures such as multi-block copolymers.[120] In parallel, future work could also incorporate a consensus classification of amino acids across multiple hydropathy scales to mitigate potential biases introduced by dependence on a single scale and further validate residue preference trends. Finally, we envision our active learning-guided screening framework as a versatile platform for broader applications in biopolymer and biomolecular design. By coupling GPR and BO with diverse computational or experimental performance metrics, this approach can facilitate the efficient exploration of large sequence spaces in domains ranging from peptide therapeutics to smart biomaterials.[121,122]

Conflicts of interest

A. L. F. is a co-founder and consultant of Evozyne, Inc. and a co-author of US Patent Applications 16/887 710 and 17/642 582, US Provisional Patent Applications 62/853 919, 62/900 420, 63/314 898, 63/479 378, 63/521 617, and 63/669 836, and International Patent Applications PCT/US2020/035206, PCT/US2020/050466, and PCT/US24/10805.

Data availability

Simulation tools, codes, and Jupyter Notebooks implementing our pipeline along with screening data listing of the particular ELP sequences considered in each round of the active learning search along with their calculated ΔG values are available as a public GitHub repository at **https://github.com/Ferg-Lab/ELP_Simulation** that is also accessible *via* a persistent doi at **https://doi.org/10.5281/zenodo.15778533**.

Supplementary information: additional computational details of the free energy calculations, validations of the alchemical transfer method, and SI figures and tables (PDF). accounting of the particular ELP sequences considered in each round of the active learning campaign along with their calculated Δ*G* values (CSV). See DOI: **https://doi.org/10.1039/d5dd00291e**.

Acknowledgements

Research was sponsored by the Army Research Office and was accomplished under Award Numbers: W911NF-23-1-0050 and W911NF-23-1-0084. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This work was completed in part with resources provided by the University of Chicago Research Computing Center. We gratefully acknowledge computing time on the University of Chicago high-performance GPU-based cyberinfrastructure supported by the National Science Foundation under Grant No. DMR-1828629. This work used the Delta system at the National Center for Supercomputing Applications through allocation CHE240069 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program,[123] which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

1 A. F. Mason, N. A. Yewdall, P. L. W. Welzen, J. Shao, M. van Stevendaal, J. C. M. van Hest, D. S. Williams and L. K. E. A. Abdelmohsen, Mimicking Cellular Compartmentalization in a Hierarchical Protocell through Spontaneous Spatial Organization, *ACS Cent. Sci.*, 2019, **5**, 1360–1365.

2 M. Nijemeisland, L. K. E. A. Abdelmohsen, W. T. S. Huck, D. A. Wilson and J. C. M. van Hest, A Compartmentalized Out-of-Equilibrium Enzymatic Reaction Network for Sustained Autonomous Movement, *ACS Cent. Sci.*, 2016, **2**, 843–849.

3 T. Gabaldón and A. A. Pittis, Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes, *Biochimie*, 2015, **119**, 262–268.

4 S. S. Mansy, J. P. Schrum, M. Krishnamurthy, S. Tobé, D. A. Treco and J. W. Szostak, Template-directed synthesis of a genetic polymer in a model protocell, *Nature*, 2008, **454**, 122–125.

5 B. C. Buddingh' and J. C. M. van Hest, Artificial Cells: Synthetic Compartments with Life-like Functionality and Adaptivity, *Acc. Chem. Res.*, 2017, **50**, 769–777.

6 J. Ye, Y. Lin, X. Yi, Z. Yu, X. Liu and G. Chen, Synthetic biology of extremophiles: a new wave of biomanufacturing, *Trends Biotechnol.*, 2023, **41**, 342–357.

7 S. Emir Diltemiz, M. Tavafoghi, N. R. de Barros, M. Kanada, J. Heinämäki, C. Contag, S. K. Seidlits and N. Ashammakhi, Use of artificial cells as drug carriers, *Mater. Chem. Front.*, 2021, **5**, 6672–6692.

8 J. F. Pelletier, L. Sun, K. S. Wise, N. Assad-Garcia, B. J. Karas, T. J. Deerinck, M. H. Ellisman, A. Mershin, N. Gershenfeld, R. Chuang, J. I. Glass and E. A. Strychalski, Genetic requirements for cell division in a genomically minimal cell, *Cell*, 2021, **184**, 2430–2440.

9 D. G. Gibson, *et al.*, Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome, *Science*, 2010, **329**, 52–56.

10 K. Vogele, T. Frank, L. Gasser, M. A. Goetzfried, M. W. Hackl, S. A. Sieber, F. C. Simmel and T. Pirzer, Towards synthetic cells using peptide-based reaction compartments, *Nat. Commun.*, 2018, **9**, 3862.

11 K. Göpfrich, I. Platzman and J. P. Spatz, Mastering Complexity: Towards Bottom-up Construction of Multifunctional Eukaryotic Synthetic Cells, *Trends Biotechnol.*, 2018, **36**, 938–951.

12 W. Jiang, Z. Wu, Z. Gao, M. Wan, M. Zhou, C. Mao and J. Shen, Artificial Cells: Past, Present and Future, *ACS Nano*, 2022, **16**, 15705–15733.

13 V. Maffeis, L. Heuberger, A. Nikoletić, C. Schoenenberger and C. G. Palivan, Synthetic Cells Revisited: Artificial Cell Construction Using Polymeric Building Blocks, *Adv. Sci.*, 2024, **11**, 2305837.

14 B. C. Paruchuri, V. Gopal, S. Sarupria and J. Larsen, Toward Enzyme-Responsive Polymersome Drug Delivery, *Nanomedicine*, 2021, **16**, 2679–2693.

15 Z. Xu, T. Hueckel, W. T. M. Irvine and S. Sacanna, Transmembrane transport in inorganic colloidal cell-mimics, *Nature*, 2021, **597**, 220–224.

16 M. Li, D. C. Green, J. L. R. Anderson, B. P. Binks and S. Mann, In vitro gene expression and enzyme catalysis in bio-inorganic protocells, *Chem. Sci.*, 2011, **2**, 1739.

17 F. Sheehan, D. Sementa, A. Jain, M. Kumar, M. Tayarani-Najjaran, D. Kroiss and R. V. Ulijn, Peptide-Based Supramolecular Systems Chemistry, *Chem. Rev.*, 2021, **121**, 13869–13914.

18 A. Schreiber, M. C. Huber and S. M. Schiller, Prebiotic Protocell Model Based on Dynamic Protein Membranes Accommodating Anabolic Reactions, *Langmuir*, 2019, **35**, 9593–9610.

19 B. Sharma, Y. Ma, H. L. Hiraki, B. M. Baker, A. L. Ferguson and A. P. Liu, Facile formation of giant elastin-like polypeptide vesicles as synthetic cells, *Chem. Commun.*, 2021, **57**, 13202–13205.

20 D. H. T. Le and A. Sugawara-Narutaki, Elastin-like polypeptides as building motifs toward designing functional nanobiomaterials, *Mol. Syst. Des. Eng.*, 2019, **4**, 545–565.

21 A. K. Varanko, J. C. Su and A. Chilkoti, Elastin-Like Polypeptides for Biomedical Applications, *Annu. Rev. Biomed. Eng.*, 2020, **22**, 343–369.

22 A. Prhashanna, P. A. Taylor, J. Qin, K. L. Kiick and A. Jayaraman, Effect of Peptide Sequence on the LCST-Like Transition of Elastin-Like Peptides and Elastin-Like Peptide–Collagen-Like Peptide Conjugates: Simulations and Experiments, *Biomacromolecules*, 2019, **20**, 1178–1189.

23 H. Reiersen, A. R. Clarke and A. R. Rees, Short elastin-like peptides exhibit the same temperature-induced structural transitions as elastin polymers: implications for protein engineering, *J. Mol. Biol.*, 1998, **283**, 255–264.

24 S. E. Reichheld, L. D. Muiznieks, F. W. Keeley and S. Sharpe, Direct observation of structure and dynamics during phase separation of an elastomeric protein, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, E4408–E4415.

25 N. K. Li, F. G. Quiroz, C. K. Hall, A. Chilkoti and Y. G. Yingling, Molecular Description of the LCST Behavior of an Elastin-Like Polypeptide, *Biomacromolecules*, 2014, **15**, 3522–3530.

26 D. López Barreiro, I. J. Minten, J. C. Thies and C. M. J. Sagt, Structure–Property Relationships of Elastin-like Polypeptides: A Review of Experimental and Computational Studies, *ACS Biomater. Sci. Eng.*, 2021, **9**, 3796–3809.

27 K. N. Greenland, M. F. C. A. Carvajal, J. M. Preston, S. Ekblad, W. L. Dean, J. Y. Chiang, R. L. Koder and R. J. Wittebort, Order, Disorder, and Temperature-Driven Compaction in a Designed Elastin Protein, *J. Phys. Chem. B*, 2018, **122**, 2725–2736.

28 M. F. C. A. Carvajal, J. M. Preston, N. M. Jamhawi, T. M. Sabo, S. Bhattacharya, J. M. Aramini, R. J. Wittebort and R. L. Koder, Dynamics in natural and designed elastins and their relation to elastic fiber structure and recoil, *Biophys. J.*, 2021, **120**, 4623–4634.

29 N. M. Jamhawi, R. L. Koder and R. J. Wittebort, Elastin recoil is driven by the hydrophobic effect, *Proc. Natl. Acad. Sci. U. S. A.*, 2024, **121**, e2304009121.

30 S. Rauscher and R. Pomés, The liquid structure of elastin, *eLife*, 2017, **6**, e26526.

31 Y. Zhang, V. Zai-Rose, C. J. Price, N. A. Ezzell, G. L. Bidwell, J. J. Correia and N. C. Fitzkee, Modeling the Early Stages of Phase Separation in Disordered Elastin-like Proteins, *Biophys. J.*, 2018, **114**, 1563–1578.

32 Y. Guo, S. Liu, D. Jing, N. Liu and X. Luo, The construction of elastin-like polypeptides and their applications in drug delivery system and tissue repair, *J. Nanobiotechnol.*, 2023, **21**, 418.

33 M. C. Huber, A. Schreiber, P. von Olshausen, B. R. Varga, O. Kretz, B. Joch, S. Barnert, R. Schubert, S. Eimer, P. Kele and S. M. Schiller, Designer amphiphilic proteins as building blocks for the intracellular formation of organelle-like compartments, *Nat. Mater.*, 2014, **14**, 125–132.

34 T. Frank, K. Vogele, A. Dupin, F. C. Simmel and T. Pirzer, Growth of Giant Peptide Vesicles Driven by Compartmentalized Transcription–Translation Activity, *Chem.–Eur. J.*, 2020, **26**, 17356–17360.

35 S. Roberts, M. Dzuricky and A. Chilkoti, Elastin-like polypeptides as models of intrinsically disordered proteins, *FEBS Lett.*, 2015, **589**, 2477–2486.

36 S. Yin, X. Mi and D. Shukla, Leveraging machine learning models for peptide–protein interaction prediction, *RSC Chem. Biol.*, 2024, **5**, 401–417.

37 M. Goles, A. Daza, G. Cabas-Mora, L. Sarmiento-Varón, J. Sepúlveda-Yañez, H. Anvari-Kazemabad, M. D. Davari, R. Uribe-Paredes, A. Olivera-Nappa, M. A. Navarrete and D. Medina-Ortiz, Peptide-based drug discovery through artificial intelligence: towards an autonomous design of therapeutic peptides, *Briefings Bioinf.*, 2024, **25**, bbae275.

38 Y. Xu, D. Verma, R. P. Sheridan, A. Liaw, J. Ma, N. M. Marshall, J. McIntosh, E. C. Sherer, V. Svetnik and J. M. Johnston, Deep Dive into Machine Learning Models for Protein Engineering, *J. Chem. Inf. Model.*, 2020, **60**, 2773–2790.

39 T. E. Gartner, A. L. Ferguson and P. G. Debenedetti, Data-driven molecular design and simulation in modern chemical engineering, *Nat. Chem. Eng.*, 2024, **1**, 6–9.

40 E. Y. Lee, G. C. Wong and A. L. Ferguson, Machine learning-enabled discovery and design of membrane-active peptides, *Bioorg. Med. Chem.*, 2018, **26**, 2708–2718.

41 H. Zhang, K. M. Saravanan, Y. Wei, Y. Jiao, Y. Yang, Y. Pan, X. Wu and J. Z. H. Zhang, Deep Learning-Based Bioactive Therapeutic Peptide Generation and Screening, *J. Chem. Inf. Model.*, 2023, **63**, 835–845.

42 C. Guntuboina, A. Das, P. Mollaei, S. Kim and A. B. Farimani, PeptideBERT: A Language Model Based on Transformers for Peptide Property Prediction, *J. Phys. Chem. Lett.*, 2023, **14**, 10427–10434.

43 D. R. Bell and S. H. Chen, Toward Guided Mutagenesis: Gaussian Process Regression Predicts MHC Class II Antigen Mutant Binding, *J. Chem. Inf. Model.*, 2021, **61**, 4857–4867.

44 Y. Ma, R. Kapoor, B. Sharma and A. P. Liu, A. L. Ferguson Computational design of self-assembling peptide chassis materials for synthetic cells, *Mol. Syst. Des. Eng.*, 2023, **8**, 39–52.

45 Z. Wu, *et al.*, alchemlyb: the simple alchemistry library, *J. Open Source Softw.*, 2024, **9**, 6934.

46 W. Humphrey, A. Dalke and K. Schulten, VMD: Visual molecular dynamics, *J. Mol. Graphics*, 1996, **14**, 33–38.

47 M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek and G. R. Hutchison, Avogadro: an advanced semantic chemical editor, visualization, and analysis platform, *J. Cheminf.*, 2012, **4**, 17.

48 D. J. Smith, J. B. Klauda and A. J. Sodt, Simulation Best Practices for Lipid Membranes [Article v1.0], *LiveCoMS*, 2019, **1**, 5966.

49 H. R. Marsden, L. Gabrielli and A. Kros, Rapid preparation of polymersomes by a water addition/solvent evaporation method, *Polym. Chem.*, 2010, **1**, 1512.

50 Schrödinger, *LLC The PyMOL Molecular Graphics System, Version 3.0*, http://www.pymol.org/pymol.

51 D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schäfer, X. Periole, D. P. Tieleman and S. J. Marrink, Improved Parameters for the Martini Coarse-Grained Protein Force Field, *J. Chem. Theory Comput.*, 2012, **9**, 687–697.

52 H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, Molecular dynamics with

coupling to an external bath, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

53 M. Parrinello and A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method, *J. Appl. Phys.*, 1981, **52**, 7182–7190.

54 G. Bussi, D. Donadio and M. Parrinello, Canonical sampling through velocity rescaling, *J. Chem. Phys.*, 2007, **126**, 014101.

55 R. Hockney and J. Eastwood, *Computer Simulation Using Particles*, CRC Press, 2021.

56 A. z. Kubincová, S. Riniker and P. H. Hünenberger, Reaction-field electrostatics in molecular dynamics simulations: development of a conservative scheme compatible with an atomic cutoff, *Phys. Chem. Chem. Phys.*, 2020, **22**, 26419–26437.

57 S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman and A. H. de Vries, The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations, *J. Phys. Chem. B*, 2007, **111**, 7812–7824.

58 M. Abraham, *et al.*, *GROMACS 2023.1 Source Code*, 2023, **https://zenodo.org/record/7852175**.

59 G. Torrie and J. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling, *J. Comput. Phys.*, 1977, **23**, 187–199.

60 S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method, *J. Comput. Chem.*, 1992, **13**, 1011–1021.

61 J. Cruz, L. Wickstrom, D. Yang, E. Gallicchio and N. Deng, Combining Alchemical Transformation with a Physical Pathway to Accelerate Absolute Binding Free Energy Calculations of Charged Ligands to Enclosed Binding Sites, *J. Chem. Theory Comput.*, 2020, **16**, 2803–2813.

62 A. S. Mey, B. K. Allen, H. E. Bruce Macdonald, J. D. Chodera, D. F. Hahn, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, A. Rizzi, J. Scheen, M. R. Shirts, G. Tresadern and H. Xu, Best Practices for Alchemical Free Energy Calculations [Article v1.0], *LiveCoMS*, 2020, **2**, 18378.

63 T. P. Straatsma and J. A. McCammon, Computational Alchemy, *Annu. Rev. Phys. Chem.*, 1992, **43**, 407–435.

64 J. G. Kirkwood, Statistical Mechanics of Fluid Mixtures, *J. Chem. Phys.*, 1935, **3**, 300–313.

65 M. Jorge, N. M. Garrido, A. J. Queimada, I. G. Economou and E. A. Macedo, Effect of the Integration Method on the Accuracy and Computational Efficiency of Free Energy Calculations Using Thermodynamic Integration, *J. Chem. Theory Comput.*, 2010, **6**, 1018–1027.

66 T. P. Straatsma and H. J. C. Berendsen, Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations, *J. Chem. Phys.*, 1988, **89**, 5876–5886.

67 G. Duarte Ramos Matos, D. Y. Kyu, H. H. Loeffler, J. D. Chodera, M. R. Shirts and D. L. Mobley, Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database, *J. Chem. Eng. Data*, 2017, **62**, 1559–1569.

68 D. M. York, Modern Alchemical Free Energy Methods for Drug Discovery Explained, *ACS Phys. Chem. Au*, 2023, **3**, 478–491.

69 M. A. La Serra, P. Vidossich, I. Acquistapace, A. K. Ganesan and M. De Vivo, Alchemical Free Energy Calculations to Investigate Protein–Protein Interactions: the Case of the CDC42/PAK1 Complex, *J. Chem. Inf. Model.*, 2022, **62**, 3023–3033.

70 G. A. Ross, C. Lu, G. Scarabelli, S. K. Albanese, E. Houang, R. Abel, E. D. Harder and L. Wang, The maximal and current accuracy of rigorous protein-ligand binding free energy calculations, *Commun. Chem.*, 2023, **6**, 222.

71 D. Hamelberg and J. A. McCammon, Standard Free Energy of Releasing a Localized Water Molecule from the Binding Pockets of Proteins: Double-Decoupling Method, *J. Am. Chem. Soc.*, 2004, **126**, 7683–7689.

72 M. Aldeghi, A. Heifetz, M. J. Bodkin, S. Knapp and P. C. Biggin, Accurate calculation of the absolute free energy of binding for drug molecules, *Chem. Sci.*, 2016, **7**, 207–218.

73 Y. Qian, I. Cabeza de Vaca, J. Z. Vilseck, D. J. Cole, J. Tirado-Rives and W. L. Jorgensen, Absolute Free Energy of Binding Calculations for Macrophage Migration Inhibitory Factor in Complex with a Druglike Inhibitor, *J. Phys. Chem. B*, 2019, **123**, 8675–8685.

74 W. Chen, Y. Deng, E. Russell, Y. Wu, R. Abel and L. Wang, Accurate Calculation of Relative Binding Free Energies between Ligands with Different Net Charges, *J. Chem. Theory Comput.*, 2018, **14**, 6346–6358.

75 C. Öhlknecht, J. W. Perthold, B. Lier and C. Oostenbrink, Charge-Changing Perturbations and Path Sampling via Classical Molecular Dynamic Simulations of Simple Guest–Host Systems, *J. Chem. Theory Comput.*, 2020, **16**, 7721–7734.

76 J. E. Hernández González and A. S. de Araujo, Alchemical Calculation of Relative Free Energies for Charge-Changing Mutations at Protein–Protein Interfaces Considering Fixed and Variable Protonation States, *J. Chem. Inf. Model.*, 2023, **63**, 6807–6822.

77 R. Zhou, P. Das and A. K. Royyuru, Single Mutation Induced H3N2 Hemagglutinin Antibody Neutralization: A Free Energy Perturbation Study, *J. Phys. Chem. B*, 2008, **112**, 15813–15820.

78 F. Sabanés Zariquiey, A. Pérez, M. Majewski, E. Gallicchio and G. De, Fabritiis Validation of the Alchemical Transfer Method for the Estimation of Relative Binding Affinities of Molecular Series, *J. Chem. Inf. Model.*, 2023, **63**, 2438–2444.

79 J. Z. Wu, S. Azimi, S. Khuttan, N. Deng and E. Gallicchio, Alchemical Transfer Approach to Absolute Binding Free Energy Estimation, *J. Chem. Theory Comput.*, 2021, **17**, 3309–3319.

80 S. Azimi, S. Khuttan, J. Z. Wu, R. K. Pal and E. Gallicchio, Relative Binding Free Energy Calculations for Ligands with Diverse Scaffolds with the Alchemical Transfer Method, *J. Chem. Inf. Model.*, 2022, **62**, 309–323.

81 L. Chen, Y. Wu, C. Wu, A. Silveira, W. Sherman, H. Xu and E. Gallicchio, Performance and Analysis of the Alchemical Transfer Method for Binding-Free-Energy Predictions of Diverse Ligands, *J. Chem. Inf. Model.*, 2023, **64**, 250–264.

82 S. Rekhi and J. Mittal, Amino acid transfer free energies reveal thermodynamic driving forces in biomolecular condensate formation, *Proc. Natl. Acad. Sci. U. S. A.*, 2025, **122**, e2425422122.

83 N. Goga, A. J. Rzepiela, A. H. de Vries, S. J. Marrink and H. J. C. Berendsen, Efficient Algorithms for Langevin and DPD Dynamics, *J. Chem. Theory Comput.*, 2012, **8**, 3637–3649.

84 M. R. Shirts and J. D. Chodera, Statistically optimal analysis of samples from multiple equilibrium states, *J. Chem. Phys.*, 2008, **129**, 124105.

85 C. H. Bennett, Efficient estimation of free energy differences from Monte Carlo data, *J. Comput. Phys.*, 1976, **22**, 245–268.

86 E. Schulz, M. Speekenbrink and A. Krause, A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions, *J. Math. Psychol.*, 2018, **85**, 1–16.

87 C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, 2006, vol. 2, pp. 63–71.

88 H. Saigo, J. Vert and T. Akutsu, Optimizing amino acid substitution matrices with a local alignment kernel, *BMC Bioinf.*, 2006, **7**, 246.

89 P. Meinicke, M. Tech, B. Morgenstern and R. Merkl, Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites, *BMC Bioinf.*, 2004, **5**, 169.

90 N. C. Toussaint, C. Widmer, O. Kohlbacher and G. Rätsch, Exploiting physico-chemical properties in string kernels, *BMC Bioinf.*, 2010, **11**, S7.

91 S. Giguère, M. Marchand, F. c. Laviolette, A. Drouin and J. Corbeil, Learning a peptide-protein binding affinity predictor with kernel ridge regression, *BMC Bioinf.*, 2013, **14**, 1471–2105.

92 S. Henikoff and J. G. Henikoff, Amino acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci. U. S. A.*, 1992, **89**, 10915–10919.

93 D. Duvenaud, Automatic model construction with Gaussian processes, PhD thesis, Pembroke College, University of Cambridge, 2014.

94 K. Thurnhofer-Hemsi, E. López-Rubio, M. A. Molina-Cabello and K. Najarian, Radial basis function kernel optimization for Support Vector Machine classifiers, *arXiv*, 2020, preprint, arXiv:2007.08233, DOI: 10.48550/arXiv.2007.08233.

95 M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson and E. Bakshy, BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 21524–21538.

96 J. T. Wilson, R. Moriconi, F. Hutter and M. P. Deisenroth, The reparameterization trick for acquisition functions, *arXiv*, 2017, preprint, arXiv:1712.00424, DOI: 10.48550/arXiv.1712.00424.

97 J. Kyte and R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, 1982, **157**, 105–132.

98 Z. Sun, J. Xu, Y. Zhang, Y. Zhang, Z. Wang, X. Wang, S. Li, Y. Guo, H. H. Shen and J. Song, Multimodal geometric learning for antimicrobial peptide identification by leveraging alphafold2-predicted structures and surface features, *Briefings Bioinf.*, 2025, **26**, bbaf261.

99 M. Frank, P. Ni, M. Jensen and M. B. Gerstein, Leveraging a large language model to predict protein phase transition: A physical, multiscale, and interpretable approach, *Proc. Natl. Acad. Sci. U. S. A.*, 2024, **121**, e2320510121.

100 G. Duart, R. Graña-Montes, N. Pastor-Cantizano and I. Mingarro, Experimental and computational approaches for membrane protein insertion and topology determination, *Methods*, 2024, **226**, 102–119.

101 Z. Ahmed, K. Shahzadi, R. Li, Y. Jiang, Y. Jin, M. Arif and J. Feng, An artificial intelligence-based approach for identifying the proteins regulating liquid-liquid phase separation, *Briefings Bioinf.*, 2025, **26**, bbaf313.

102 A. Marchand, *et al.*, Targeting protein-ligand neosurfaces with a generalizable deep learning tool, *Nature*, 2025, **639**, 522–531.

103 S. C. Ng and D. Görlich, A simple thermodynamic description of phase separation of Nup98 FG domains, *Nat. Commun.*, 2022, **13**, 6172.

104 J. Despanie, J. P. Dhandhukia, S. F. Hamm-Alvarez and J. A. MacKay, Elastin-like polypeptides: Therapeutic applications for an emerging class of nanomedicines, *J. Contr. Release*, 2016, **240**, 93–108.

105 B. Schölkopf, A. Smola and K. Müller, Kernel principal component analysis, *Artificial Neural Networks—ICANN'97*, Berlin, Heidelberg, 1997, pp. 583–588.

106 X. Zhuang, J. R. Makover, W. Im and J. B. Klauda, A systematic molecular dynamics simulation study of temperature dependent bilayer structural properties, *Biochim. Biophys. Acta, Biomembr.*, 2014, **1838**, 2520–2529.

107 P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski and M. J. L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics, *Methods Biochem. Anal.*, 2009, **25**, 1422–1423.

108 J. Huang and A. D. MacKerell Jr, CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data, *J. Comput. Chem.*, 2013, **34**, 2135–2145.

109 Z. E. Hughes, M. A. Nguyen, J. Wang, Y. Liu, M. T. Swihart, M. Poloczek, P. I. Frazier, M. R. Knecht and T. R. Walsh, Tuning Materials-Binding Peptide Sequences toward Gold- and Silver-Binding Selectivity with Bayesian Optimization, *ACS Nano*, 2021, **15**, 18260–18269.

110 Y. N. Talluri, S. K. Sankaranarayanan, H. C. Fry and R. Batra, Discovery of unconventional and nonintuitive self-assembling peptide materials using experiment-driven machine learning, *Sci. Adv.*, 2025, **11**, eadt9466.

111 A. S. Zadeh, A. J. Winton, J. M. Palomba and A. L. Ferguson, High-throughput virtual screening of protein-catalyzed capture agents for novel hydrogel-nanoparticle fentanyl sensors, *J. Phys. Chem. B*, 2025, **129**, 10568–10583.

112 J. Yamaguchi, K. Nishida, E. Kobatake and M. Mie, Functional decoration of elastin-like polypeptides-based nanoparticles with a modular assembly via isopeptide bond formation, *Biotechnol. Lett.*, 2024, **47**, 6.

113 V. Mukwaya, S. Mann and H. Dou, Chemical communication at the synthetic cell/living cell interface, *Commun. Chem.*, 2021, **4**, 161.

114 C. Presutti, E. Vreeker, S. Sasidharan, Z. Ferdinando, M. Stuart, J. Juhaniewicz-Dębińska, G. Maglia, W. H. Roos and B. Poolman, Balancing Permeability and Stability: A Study of Hybrid Membranes for Synthetic Cells Using Lipids and PBd-b-PEO Block Copolymers, *Biomacromolecules*, 2025, 2868–2881.

115 V. Ibrahimova, H. Zhao, E. Ibarboure, E. Garanger and S. Lecommandoux, Thermosensitive Vesicles from Chemically Encoded Lipid-Grafted Elastin-like Polypeptides, *Angew. Chem., Int. Ed.*, 2021, **60**, 15036–15040.

116 A. Rizuan, N. Jovic, T. M. Phan, Y. C. Kim and J. Mittal, Developing Bonded Potentials for a Coarse-Grained Model of Intrinsically Disordered Proteins, *J. Chem. Inf. Model.*, 2022, **62**, 4474–4485.

117 R. M. Regy, J. Thompson, Y. C. Kim and J. Mittal, Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins, *Protein Sci.*, 2021, **30**, 1371–1379.

118 J. A. Joseph, A. Reinhardt, A. Aguirre, P. Y. Chew, K. O. Russell, J. R. Espinosa, A. Garaizar and R. Collepardo-Guevara, Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy, *Nat. Comput. Sci.*, 2021, **1**, 732–743.

119 A. Chakravarti and J. A. Joseph, Accurate prediction of thermoresponsive phase behavior of disordered proteins, *bioRxiv*, 2025, preprintDOI: 10.1002/pro.70284.

120 L. MartÃn, E. Castro, A. Ribeiro, M. Alonso and J. C. RodrÃguez-Cabello, Temperature-Triggered Self-Assembly of Elastin-Like Block Co-Recombinamers:The Controlled Formation of Micelles and Vesicles in an Aqueous Medium, *Biomacromolecules*, 2012, **13**, 293–298.

121 R. Barrett and A. D. White, Investigating Active Learning and Meta-Learning for Iterative Peptide Design, *J. Chem. Inf. Model.*, 2020, **61**, 95–105.

122 K. Shmilovich, R. A. Mansbach, H. Sidky, O. E. Dunne, S. S. Panda, J. D. Tovar and A. L. Ferguson, Discovery of Self-Assembling π-Conjugated Peptides by Active Learning-Directed Coarse-Grained Molecular Simulation, *J. Phys. Chem. B*, 2020, **124**, 3873–3891.

123 T. J. Boerner, S. Deems, T. R. Furlani, S. L. Knuth and J. Towns, *ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support*, PEARC, 2023, pp. 173–176.