Understanding protein-complex assembly through grand canonical maximum entropy modeling

Andrei G. Gasic,[1,2] Atrayee Sarkar,[1,2] and Margaret S. Cheung[1,2,3,*]

[1]*Department of Physics, University of Houston, Houston, Texas 77204, USA*
[2]*Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, USA*
[3]*Pacific Northwest National Laboratory, Seattle Research Center, Seattle, Washington 98109, USA*

(Received 15 March 2021; accepted 30 July 2021; published 7 September 2021)

Inside a cell, heterotypic proteins assemble in inhomogeneous, crowded systems where the abundance of these proteins vary with cell types. While some protein complexes form putative structures that can be visualized with imaging, there are far more protein complexes that are yet to be solved because of their dynamic associations with one another. Nevertheless, it is possible to infer these protein complexes through a physical model. However, it is often not clear to physicists what kind of data from biology is necessary for such a modeling endeavor. Here, we aim to model these clusters of coarse-grained protein assemblies from multiple subunits through the constraints of interactions among the subunits and the chemical potential of each subunit. We obtained the constraints on the interactions among subunits from the known protein structures. We inferred the chemical potential that dictates the particle number distribution of each protein subunit from the knowledge of protein abundance from experimental data. Guided by the maximum entropy principle, we formulated an inverse statistical mechanical method to infer the distribution of particle numbers from the data of protein abundance as chemical potentials for a grand canonical multicomponent mixture. Using grand canonical Monte Carlo simulations, we captured a distribution of high-order clusters in a protein complex of succinate dehydrogenase with four known subunits. The complexity of hierarchical clusters varies with the relative protein abundance of each subunit in distinctive cell types such as lung, heart, and brain. When the crowding content increases, we observed that crowding stabilizes emergent clusters that do not exist in dilute conditions. We, therefore, proposed a testable hypothesis that the hierarchical complexity of protein clusters on a molecular scale is a plausible biomarker of predicting the phenotypes of a cell.

DOI: 10.1103/PhysRevResearch.3.033220

I. INTRODUCTION

Living cells can contain on the order of 10^4 [1] distinct types of proteins and other macromolecules at a given time. In this many-component mixture environment, macromolecules like proteins fold, unfold, and assemble into complexes and organize hierarchically into spatial networks [2–4]. In fact, these unfathomably complex networks give rise to the emergence of all biological functions and ultimately the properties of life [2,3,5,6]. The specific arrangements of macromolecules are thought to emerge from the vast amount of weak "quinary" and entropic interactions [7–10]. Of these types of interactions, the most intuitive conception of protein biophysics in this crowded environment is that of volume exclusion [11–13] exerted on a given protein by surrounding macromolecules, so called the *macromolecular crowding effects* [14]. Proteins interact weakly and form higher-order complexes [15] through quinary interactions [16], where counteracting forces between favorable electrostatic interactions and unfavorable solvation energies are provided by their metabolites [17]. Computer simulations have investigated the mechanism of these molecules forming small clusters often under the constraint of a fixed number of particles N in a closed system [i.e., a canonical ensemble in Fig. 1(a)]. This does not allow for particle fluctuations—a key free energy term. Nevertheless, the physical mechanism of these complex assemblies and organization in an open system where N varies is still unclear.

To resolve this issue, one may use semigrand canonical ensembles (total N is constant, but particle number of specific species fluctuates), which is especially important for studying phase separation of multicomponent mixtures. However, there remains the issue of setting the correct chemical potentials for each particle species that produces the correct stoichiometry or relative abundance [18,19]. Another way is to identify the constraint of mean particle numbers through a "chemical potential" in a grand canonical ensemble, thus allowing the N to fluctuate [Fig. 1(b)]. However, it is challenging to establish constraints for chemical potentials for proteins in a multicomponent mixture in an open system like cytoplasm. Here, we formulate a method to solve the inverse statistical mechanics problem of finding the correct chemical potentials for a multicomponent mixture from the database of protein abundance. Current high-throughput experiments such as mass spectrometry quantify the protein abundance of a cell

[*] margaret.cheung@pnnl.gov

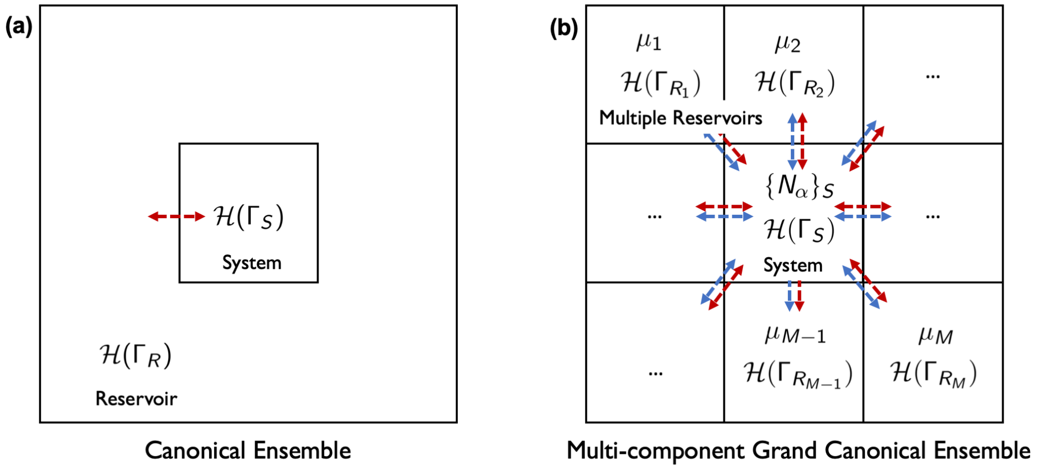

FIG. 1. (a) Canonical ensemble: the system S is maintained at a constant temperature through energy exchange (red arrow) with heat reservoir R. (b) Multicomponent grand canonical ensemble: the system S is maintained at constant temperature and chemical potentials $\{\mu_\alpha\}$ through energy (red arrow) and particle (blue arrows) exchanges with multiple particle reservoirs R_α. $\mathcal{H}$ is the Hamiltonian with the set of positions and momenta Γ of the system or reservoir(s).

with high accuracy. It is often shown that the measurement of protein abundance is indicative of the state of a cell [20–22].

There are several important reasons for using the grand canonical ensemble to study biological many-component systems instead of using the canonical ensemble:

(1) The assemblies of these macromolecular complexes are inherently finite-sized processes where the abundance of a single protein species is experimentally measured in parts per million. Despite that a cellular mixture is highly crowded with biological molecules, the number of the individual species in this many-component mixture is still far from the thermodynamic limit. Without the correct ensemble, key free energy contributions will be missed [23,24]. Frequently, finite-sized corrections need to be used in canonical ensembles of protein binding for this reason [25,26].

(2) In this paper, we do not study phase transitions; however, our grand canonical method would be more advantageous in sampling various phases and mapping phase diagrams than an analogous canonical method. The grand canonical ensemble allows a single phase to occupy the entire simulation volume and avoids the costly interfaces between phases [27–29]. Also, with the grand canonical ensemble, it may reveal metastable free energy basins that are unstable phases in the corresponding canonical phase diagram [30].

(3) Information from the protein sequences provides a way of understanding protein-protein interaction networks [31,32], and chemical potential provides another source of rich information which dictates the state of a cell. While abundance may provide similar information, chemical potential may be more useful because the latter extends these systems into nonequilibrium or steady-state settings.

With the knowledge of protein abundance, we used grand canonical Monte Carlo (GCMC) simulations and the principle of maximum entropy to model the distinct features of protein assemblies. We use the mitochondrial respiratory Complex II, also called the *succinate dehydrogenase (SDH) complex* [Fig. 2(a)] to establish our model. It is a tetraprotein complex composed of four subunits SDHA, SDHB, SDHC, and SDHD, and is well studied in a metabolic pathway that breaks down carbohydrates and produces energy [33–36]. Its integrative structure has also been constructed with constraints from the interactome through cross-linking mass spectrometry [37]. SDH is an integral membrane protein complex, where SDHC and SDHD are intermembrane proteins, and is in both the tricarboxylic acid cycle and aerobic respiration [35,36]. To perform its various functions, the SDH form heterodimers and in some cases trimers before completely forming tetramer. Many other proteins and small molecules regulate the assembly and function of the SDH complex depending on the state or type of cell. As such, different cell types contain different abundances of each subunit [20–22].

With the constraints of interaction topology among subunits and the constraints of abundance in the multicomponent mixture, we can integrate this information into protein

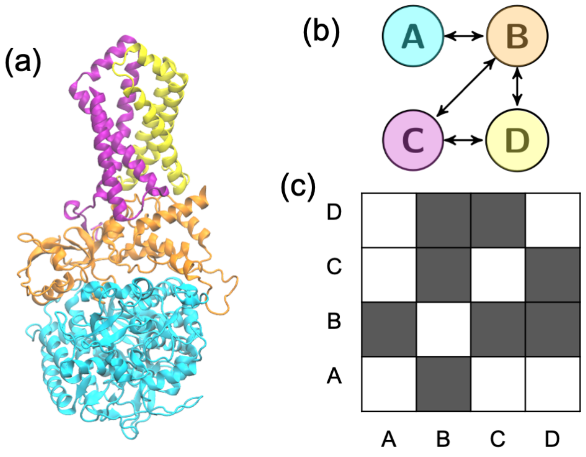

FIG. 2. (a) Crystal structure of succinate dehydrogenase (SDH) complex (protein data bank ID: 1NEN) with subunits SDHA–SDHD labeled A–D, respectively. (b) Interaction topology of the SDH complex. Arrows indicate an interaction between the two subunits. (c) Contact map of the SDH complex, where white represents no interaction and gray represents an interaction between two of the subunits.

TABLE I. Protein abundance x_α^{exp} in parts per million (ppm) and relative abundance $\tilde{x}_\alpha^{\text{exp}}$ (maximum of 1) for different mouse cell types from PaxDB [39].

	Abundance x_α^{exp} (ppm)				Relative abundance $\tilde{x}_\alpha^{\text{exp}}$			
Type	A	B	C	D	A	B	C	D
Whole	196	77.9	16.1	8.65	0.66	0.26	0.05	0.03
Heart	2358	637	108	2.88	0.76	0.21	0.03	0.001
Brain	853	87.2	56.7	2.9	0.85	0.09	0.06	0.003
Lung	495	2229	30.1	2.95	0.18	0.81	0.01	0.001

assemblies with tunable topological properties and expandable cluster sizes. By varying the crowding content, emergent clusters are formed where the transient complex exists at high crowding but not in lower crowding content. We have applied the Jaccard index to characterize the higher order of these complexes and note that the emergent cluster for lung and brain is prominent, and the representation of whole cell is not adequate. This integrative model shows how complex matter connects to the phenotype of a cell.

II. THEORETICAL MODEL AND METHODS

A. Abundance for SDH

The SDH complex from the Protein Data Bank (PDB) reveals four subunits; interaction maps are shown in Figs. 2(b) and 2(c). The stoichiometry of each subunit is the same in an SDH complex [31]; however, the abundance of each unit varies significantly among cell types [38]. To understand how varying the abundances of the subunits affects the assembly of the complex, we study four different cell types in Table I. It shows these abundances in parts per million for each of the subunits for whole (integrated average of all cell types over the whole organism), heart, brain, and lung cell types of a mouse from the Protein Abundances Across Organisms Database (PaxDB) [39]. Table I also shows the relative abundance between the four subunits among these four cell types.

B. Applying the principle of maximum entropy to the grand canonical ensemble

Our approach is to infer a protein cluster model for the probability distribution of particle numbers over system states that are consistent with these experimental results with as little bias as possible. Using the principle of maximum entropy [40,41] is a way to best guess the distribution which agrees with an average observable of the data. In our case, the probability distribution we seek is of the set of particle numbers N_α of species α (or subunit), which is consistent with the experimentally derived protein abundance for α.

The distribution $P(\{N_\alpha\})$ is estimated by maximizing the entropy:

$$S[P] = -\sum_{\{N_\alpha\}} P(\{N_\alpha\}) \ln P(\{N_\alpha\}), \tag{1}$$

with constraints. This will give an exponential distribution for the energy landscape of the system that allows particle number fluctuations, i.e., a grand canonical distribution [see Fig. 2(b)]:

$$P(\{N_\alpha\}) = \frac{Z(\{N_\alpha\})}{\Xi(\{\mu_\alpha\})} \exp\left(\frac{1}{k_B T}\sum_\alpha^M \mu_\alpha N_\alpha\right). \tag{2}$$

Here, the Lagrange multipliers for the constraints, i.e., the protein abundances, are chemical potentials μ_α times inverse temperature ($\beta = 1/k_B T$). Here, $Z(\{N_\alpha\})$ is the N-particle canonical partition function of a M-component protein mixture:

$$Z(\{N_\alpha\}) = \prod_i^N \int \exp\left[-\frac{\mathcal{H}_N(\{\boldsymbol{r}_i\})}{k_B T}\right] d\boldsymbol{r}_i, \tag{3}$$

where $\mathcal{H}_\mathcal{N}$ is the N-particle Hamiltonian and $N = \sum_\alpha^M N_\alpha$. The distribution in Eq. (2) is normalized by the grand canonical partition function:

$$\Xi(\boldsymbol{\mu}) = \sum_{N_1=0}^{\infty} \cdots \sum_{N_M=0}^{\infty} \left[\prod_\alpha^M \exp\left(\frac{\mu_\alpha N_\alpha}{k_B T}\right)\right] Z(\{N_\alpha\}). \tag{4}$$

Additionally, this model does not require parameter tuning; all parameters are completely determined by the experimental data.

C. Parametrizing chemical potential using max entropy

There exists a unique set of chemical potential $\{\mu_\alpha\}$ that produces observable mean particle numbers $\langle N_\alpha \rangle$ that are consistent with the experimentally measured $\langle N_\alpha \rangle^{\text{exp}}$ (from PaxDB [39]), but finding them is a computationally difficult inverse statistical mechanics problem.

To solve for the correct chemical potential μ_α for each particle type α, we minimize the Kullback-Leibler (KL) divergence between the "real" distribution Q and the model distribution [Eq. (2)], defined as

$$D_{\text{KL}}(Q\,||P) = \sum_{\{N_\alpha\}} Q(\{N_\alpha\}) \ln \frac{Q(\{N_\alpha\})}{P(\{N_\alpha\})}. \tag{5}$$

The real distribution is the Boltzmann-like distribution containing $\langle N_\alpha \rangle^{\text{exp}}$. The KL divergence becomes

$$D_{\text{KL}}(Q\,||P) = \ln \frac{\Xi(\boldsymbol{\mu})}{\Xi_0} - \frac{1}{k_B T}\sum_\alpha^M \mu_\alpha \langle N_\alpha \rangle^{\text{exp}}, \tag{6}$$

where $\Xi_0 \equiv \Xi(\boldsymbol{\mu} = 0)$ and $\boldsymbol{\mu} \equiv \{\mu_\alpha\}$.

The partition function ratio in the expression can be further simplified using a cumulant expansion as

$$\ln \frac{\Xi(\boldsymbol{\mu})}{\Xi_0} = \ln \frac{\sum_{\{N_\alpha\}} \prod_i^N \int \exp\left(-\frac{\mathcal{H}_N}{k_B T}\right) \exp\left(\frac{\sum_\alpha^M \mu_\alpha N_\alpha}{k_B T}\right) d\boldsymbol{r}_i}{\sum_{\{N_\alpha\}} \prod_i^N \int \exp\left(-\frac{\mathcal{H}_N}{k_B T}\right) d\boldsymbol{r}_i} \tag{7}$$

$$= \ln \left\langle \exp\left(\frac{\sum_\alpha^M \mu_\alpha N_\alpha}{k_B T}\right) \right\rangle_0 \tag{8}$$

$$= \sum_{\nu=1}^{\infty} \frac{1}{\nu!} \left\langle\!\left\langle \left(\frac{1}{k_B T}\sum_\alpha^M \mu_\alpha N_\alpha\right)^\nu \right\rangle\!\right\rangle_c, \tag{9}$$

Algorithm 1 Self-consistent algorithm

```
Set $\langle N_\alpha\rangle^{\rm exp}$ and $\mathcal{H}_N$ for system
Initialize: performing simulations with $\{\mu_\alpha\} = 0$
Estimate $\langle N_\alpha\rangle$ and $C_{\alpha\beta}$
while (error $= \sum_\alpha |\langle N_\alpha\rangle - \langle N_\alpha\rangle^{\rm exp}| / \sum_\alpha \langle N_\alpha\rangle^{\rm exp}) >$ tolerance value, do
    Update: $\boldsymbol{\mu} \rightarrow \boldsymbol{\mu} - \frac{1}{\beta} C^{-1} \cdot (\langle \boldsymbol{N}\rangle - \langle \boldsymbol{N}\rangle^{\rm exp})$
    Repeat simulations with updated $\{\mu_\alpha\}$
    Estimate new $\langle N_\alpha\rangle$ and $C_{\alpha\beta}$
end while
```

where $\langle\ldots\rangle_c$ signifies the cumulant. To the second order, the cumulant expansion is

$$\ln\frac{\Xi(\boldsymbol{\mu})}{\Xi_0} = \frac{1}{k_BT}\sum_\alpha \mu_\alpha\langle N_\alpha\rangle + \frac{1}{2(k_BT)^2} \times \sum_{\alpha\beta}\mu_\alpha\mu_\beta(\langle N_\alpha N_\beta\rangle - \langle N_\alpha\rangle\langle N_\beta\rangle). \qquad (10)$$

Plugging this back into the KL divergence in Eq. (6),

$$D_{\rm KL} = \frac{1}{k_BT}\boldsymbol{\mu}\cdot(\langle\boldsymbol{N}\rangle - \langle\boldsymbol{N}\rangle^{\rm exp}) + \frac{1}{2(k_BT)^2}\boldsymbol{\mu}^T\cdot C\cdot\boldsymbol{\mu}, \qquad (11)$$

where $C_{\alpha\beta} = \langle N_\alpha N_\beta\rangle - \langle N_\alpha\rangle\langle N_\beta\rangle$. The solution for μ that minimizes $D_{\rm KL}$ (i.e., $\frac{\delta D_{\rm KL}}{\delta\mu} = 0$) is

$$\boldsymbol{\mu} = -k_BT\, C^{-1}\cdot(\langle\boldsymbol{N}\rangle - \langle\boldsymbol{N}\rangle^{\rm exp}). \qquad (12)$$

Since this is an approximate solution, when $\langle N_\alpha\rangle \neq \langle N_\alpha\rangle^{\rm exp}$, we use the following iterative self-consistent procedure to find more accurate values that eventually would converge to reproduce the experimental measurements. The algorithm is as follows in Algorithm 1.

Additionally, since PaxDB records protein abundance $x_\alpha^{\rm exp}$ in units of part per million (ppm) and our simulation box size is a small fraction of the size of a cell, we determine $\langle N_\alpha\rangle^{\rm exp}$ from the relative abundance for a certain volume fraction ϕ ($\equiv \frac{Nv}{V_{\rm box}}$, where v is the volume of the protein and $V_{\rm box}$ is the volume of the box). That is,

$$\langle N_\alpha\rangle^{\rm exp} \equiv \frac{x_\alpha^{\rm exp}}{\sum_\alpha^M x_\alpha^{\rm exp}}\frac{\phi V_{\rm box}}{v}. \qquad (13)$$

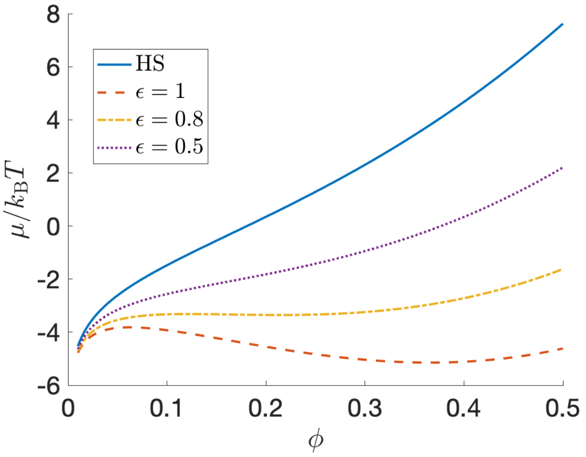

FIG. 3. Chemical potential as a function of total volume fraction (ϕ) of a single-component fluid with Lennard-Jones (LJ) interactions at various interaction strengths. HS is hard spheres.

D. Grand canonical Hamiltonian

To model protein assemblies *in vivo*, we will join aspects of both models used in chromosomes [42,43] and molecule self-assembly modeling [44]. We used a structure-based Hamiltonian, i.e., the model has attractive interactions between proteins that are in contact in the crystal structure [Fig. 2(a)], and there are volume exclusion interactions for proteins not in contact. The N-particle Hamiltonian is a function of the particle positions $\{r_i^\alpha\}$ of particle species α having the following form

$$\mathcal{H}_N(\{\boldsymbol{r}_i^\alpha\}) = \sum_{\alpha<\beta}^{M}\sum_i^{N_\alpha}\sum_j^{N_\beta} U_{\rm LJ}(\boldsymbol{r}_i^\alpha, \boldsymbol{r}_j^\beta)\Theta(\Delta_{\alpha\beta} - |\boldsymbol{r}_i^\alpha - \boldsymbol{r}_j^\beta|), \qquad (14)$$

where our effective pairwise protein potential is the Lennard-Jones (LJ) potential $U_{\rm LJ}$:

$$U_{\rm LJ}(\boldsymbol{r}_i^\alpha, \boldsymbol{r}_j^\beta) = 4\epsilon\left[\left(\frac{\sigma_{\alpha\beta}}{|\boldsymbol{r}_i^\alpha - \boldsymbol{r}_j^\beta|}\right)^{12} - \left(\frac{\sigma_{\alpha\beta}}{|\boldsymbol{r}_i^\alpha - \boldsymbol{r}_j^\beta|}\right)^{6}\right]. \qquad (15)$$

For simplicity, $\sigma_{\alpha\beta} = \sigma$. The term $\Delta_{\alpha\beta}$ in the Heaviside function Θ is the cutoff value between particle species α and β. Since our contact data are binary (probability of 1 or 0 of being connected), we use

$$\Delta_{\alpha\beta} = \begin{cases} 2.5\sigma & \text{if } \alpha, \beta \text{ interact specifically} \\ 1\sigma & \text{otherwise} \end{cases}. \qquad (16)$$

The interaction strength $\epsilon = 0.5$ is used to ensure the stability of the system. To ensure a one-to-one correspondence of μ to ϕ, we need to select the best interaction strength ϵ. For $\epsilon \gtrsim 0.8$, certain values of μ correspond to multiple ϕ values (see Fig. 3), which signify multiple phases. Thus, a one-to-one correspondence of μ to ϕ reduces the chance of having large fluctuation in energy and density from phase transitions.

E. Monte Carlo simulation details

We conducted GCMC simulations on LAMMPS [45] using a box size $V_{\rm box} = 1000\sigma^3$ for 120 000 GCMC steps, where each GCMC step attempted 800 insertions or deletions and 800 translations per particle type. The mean energy and variance of particle number plateaus at ∼20 000 steps, and the errors for the mean energy and variance of particle number plateaus at ∼100 000 steps. Thus, 120 000 steps are sufficient for data analysis. The number for insertions or deletions and translations ensures that the autocorrelation time for energy and density is 1 GCMC step. The attempts in the GCMC steps follow the metropolis criterion:

$$p(\mathcal{H}_N \to \mathcal{H}_N^{'}) = \min\left[1, \exp\left(-\frac{\mathcal{H}_N^{'} - \mathcal{H}_N}{k_B T}\right)\right],$$

for translation, and

$$p\,(N \to N')$$

$$= \begin{cases} \min\left[1, \frac{V}{(N+1)} \exp\left(-\frac{\mathcal{H}_{N+1}-\mathcal{H}_N-\mu}{k_B T}\right)\right] & \text{for } N' = N+1 \\ \min\left[1, \frac{N}{V} \exp\left(-\frac{\mathcal{H}_{N-1}-\mathcal{H}_N+\mu}{k_B T}\right)\right] & \text{for } N' = N-1 \end{cases},$$

for insertion and deletion of particles.

F. Jaccard index

We followed the analysis by Sardiu *et al.* [46,47], who have used the Jaccard index J to analyze and predict core components and modules in higher-order protein complexes. The Jaccard index J measures similarity among sample sets. Here, J is defined as the size of the intersection divided by the size of the union of the sample sets. In our case, we have four sample sets corresponding to the four subunits A, B, C and D. Each set contains all the subunit pairs interacting with the subunit corresponding to that set. For instance, the sample set corresponding to A contains subunits pairs AA, AB, AC, and AD, while the set corresponding to B contains AB, BB, BC, and BD. We follow the number of contacts formed by a pair of subunits α and β, $n_{\alpha\beta}$, among the four subunits A, B, C, and D. A pair of subunits α and β are considered interacting (i.e., forming a pair) if their Euclidean distance is less than a predefined cutoff value of 1.1σ. The Jaccard index $J(\alpha,\beta)$ between subunits α and β is below:

$$J(\alpha,\ \beta) = \frac{n_{\alpha\beta}}{\sum_\gamma (n_{\ \alpha\gamma} + n_{\beta\gamma}) - n_{\alpha\beta}}. \tag{19}$$

We then calculated the time average of $\langle \mathrm{J}(\alpha,\beta)\rangle$ for each cell type.

III. RESULTS

A. A self-consistent algorithm converges to chemical potentials according to mean abundances of multicomponent protein complex

In accordance with the principle of maximum entropy (discussed in Sec. II B), we inferred the set of chemical potentials for each of the subunits of the SDH complex $\{\mu_\alpha\}$ (where $\alpha \equiv$ A, B, C, and D corresponding to SDHA, SDHB, SDHC, and SDHD, respectively), such that the mean particle numbers $\langle N_\alpha\rangle$ match with the relative abundance from PaxDB in Table I. We used our self-consistent algorithm (Algorithm 1) to compute $\{\mu_\alpha\}$ for each cell type at three different volume fractions $\varphi = 0.1$, 0.2, and 0.3. These three volume fractions represent the range in the fraction of macromolecular volumes in a cellular volume where macromolecular crowding [14] affects protein dynamics [12,48] and assemblies. On average, the algorithm converges to $\{\mu_\alpha\}$ of each system in under three iterations.

As a demonstration of our self-consistent algorithm (Algorithm 1), we present the evolution of the sets $\{\mu_\alpha\}$ and $\{\langle N_\alpha\rangle\}$ for the whole cell type at a volume fraction $\varphi = 0.1$. Figure 4(a) shows $\{\mu_\alpha\}$ changing with each iteration of the algorithm, resulting in a $\langle N_\alpha\rangle$ [Fig. 4(b)]. Starting with $\{\mu_\alpha\}$ $= 0$, the error [Fig. 4(a) right y axis] is large due to the large deviations of $\langle N_\alpha\rangle$ from $\langle N\rangle^{\mathrm{exp}}$ [Fig. 4(b)]. By the third iteration of Algorithm 1, $\{\mu_\alpha\}$ results in a simulation that captures the correct experimental particle mean $\langle N\rangle^{\mathrm{exp}}$ values for each subunit.

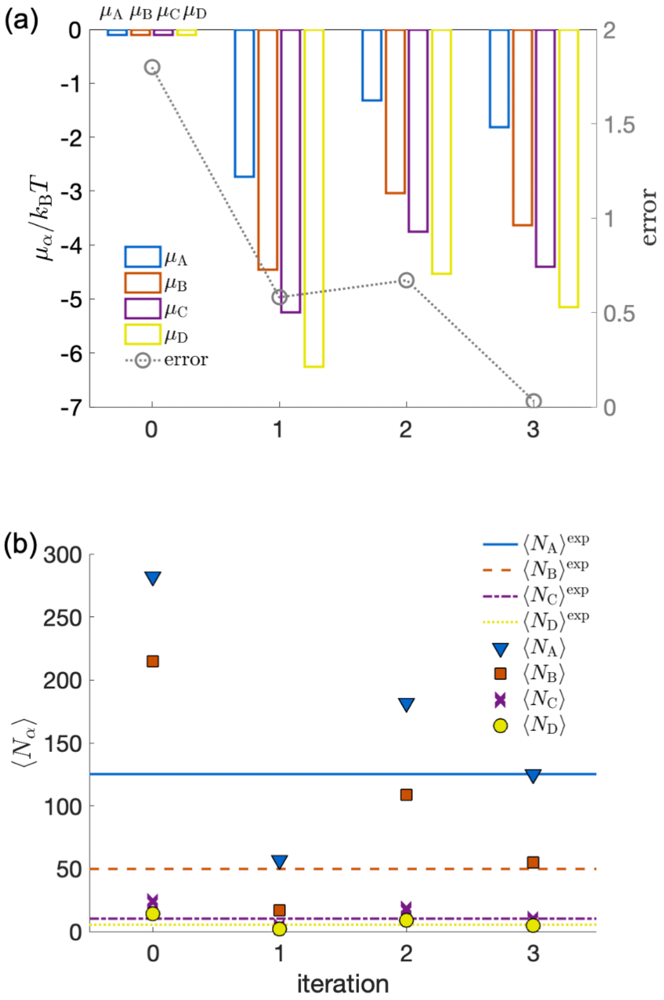

FIG. 4. (a) Values of μ_α and (b) $\langle N_\alpha\rangle$, where $\alpha \equiv$ A–D corresponding to subunits succinate dehydrogenase (SDH) SDHA–SDHD, respectively, vs iterations of the self-consistent algorithm for SDH complex of whole cell type at $\phi = 0.1$. Dashed lines indicate experimentally observed average particle numbers. (a, right axis) Error vs iterations in gray.

This algorithm converges to $\{\mu_\alpha\}$ of the different cell types (Table I) and different volume fractions in a similar fashion. The values of the converged $\{\mu_\alpha\}$ are shown in the next section.

B. Cell type is distinguished by chemical potentials of subunits

Since each cell type has varying protein abundances for each subunit of the SDH complex (as shown in Table I), we hypothesized that the chemical potential will change to capture the correct statistics accordingly. To show the effects on the μ_α values when cell type varies, we calculate the μ_α

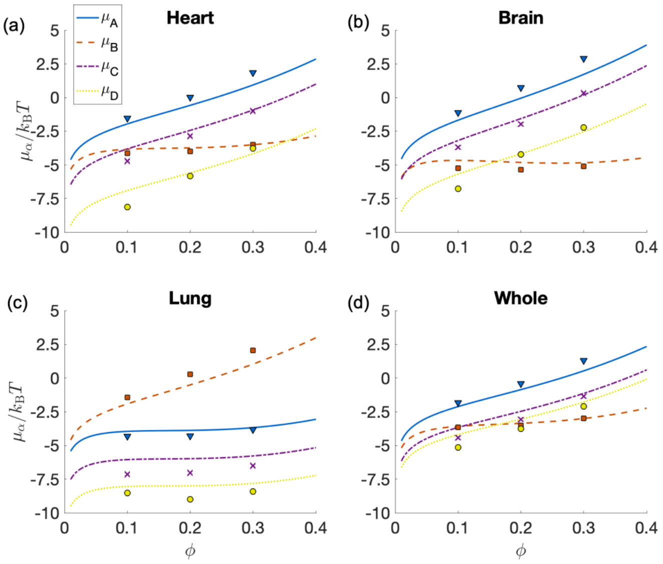

FIG. 5. Chemical potential as a function of total volume fraction (ϕ) using for various cell types: (a) heart, (b) brain, (c) lung, and (d) whole. Solid lines are the analytical approximations from Eq. (B1), and symbols (triangle: A, square: B, cross: C, and circle: D) are values of μ calculated from the self-consistent algorithm (Algorithm 1).

using Algorithm 1 for cell types whole, lung, heart, and brain at $\phi = 0.1$, 0.2, and 0.3. The converged μ_α values are shown in Fig. 5 (presented as symbols) as a function of ϕ. The solid line curves are analytical approximations (see Appendix A) of Eq. (B1). The μ_α values calculated via Algorithm 1 match well with the analytical results of Eq. (B1). This agreement confirms that Algorithm 1 works correctly. The slight deviations of the algorithm computed values from Eq. (B1) may be due to the nonzero error of Algorithm 1 or that Eq. (B1) is not an exact result.

Indeed, each cell type gives a unique set $\{\mu_\alpha\}$ shown in Fig. 5. These chemical potentials are the subunit free energy, resulting in a unique landscape for each cell type. Since subunit B interacts with all the other subunits (see Fig. 2), $\mu_B(\phi)$ will behave differently from the others. The general trends are that μ_B is flat for all values of ϕ, while the chemical potentials of the other subunits are monotonically increasing.

The cell type with chemical potential trends that is the most strikingly different from the others is that of the lung cell type in Fig. 5(c). In the lung cell type, the μ_B curve (orange) monotonically increases unlike that of the other cell types. Such an effect is due to the large relative abundance of SDHB ($\tilde{x}_B^{\text{exp}}$) dominating in the lung cell type by a factor of 4.5 times the next largest relative abundance, which is SDHA. In the other cases, $\tilde{x}_A^{\text{exp}}$ is at least 2.5 times larger than $\tilde{x}_B^{\text{exp}}$ (see Table I for exact numbers).

C. The radial distribution of the contact hub

As the results from the previous section showed that $\mu_B(\phi)$ behaves differently from the other subunits, we focus our attention to subunit B. To further understand the effect of changing cell types, we extend our analysis to calculating the radial distribution $g(r)$ between subunit pairs B and itself (g_{B-B}), B and the other non-B subunits ($g_{B-!B}$), and all pairs (g_{tot}) for $\phi = 0.1$ and 0.3, shown in Figs. 6 and 7, respectively.

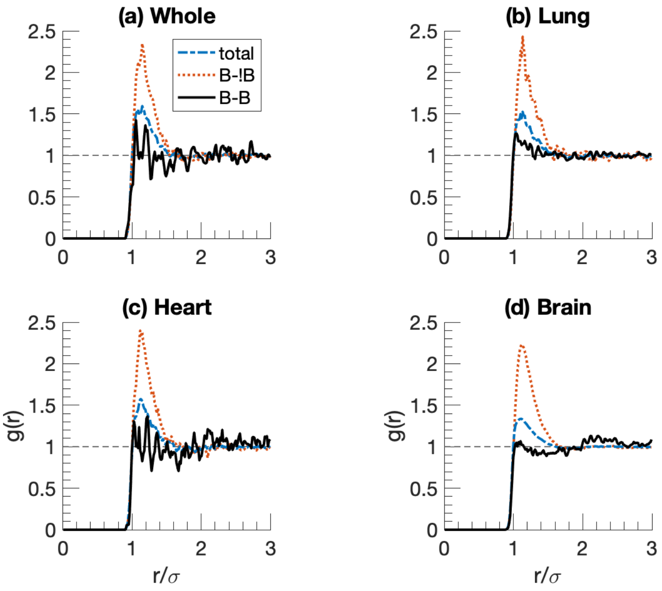

FIG. 6. Radial distribution function $g(r)$ between all particles (blue dash-dot, total), B and not B (orange dotted, B-!B), and B and itself (black solid line, B-B) for various cell types at $\phi = 0.1$: (a) lung, (b) heart, (c) brain, and (d) whole. The graphs are smoothed using mean over a window of three timeframes.

At $\phi = 0.1$ (Fig. 6), the radial distribution function of interparticle distance r between any two pairs of subunits $g_{\text{tot}}(r)$ resembles that of a gas or dilute liquid for all four cell types. Its first peak at $r \approx 1.2\sigma$ is ~1.5 times the average density of

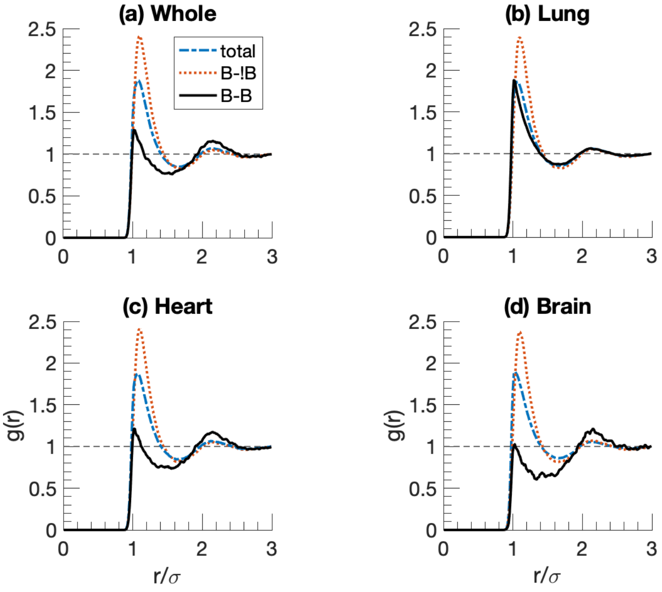

FIG. 7. Radial distribution function $g(r)$ between all particles (blue dash-dot line, total), B and not B (orange dotted line, B-!B), and B and itself (black solid line, B-B) for various cell types at $\phi = 0.3$: (a) lung, (b) heart, (c) brain, and (d) whole. The graphs are smoothed using mean over a window of three timeframes.

the system. The change in the first peak of the curves can be thought of as a change in an "osmotic pressure" between two particle types, which is dictated by both entropic and energetic effects. Since B is the contact hub (all subunits interact with B; see Fig. 2), the first peak of $g_{B-!B}(r)$ is ~70% larger in amplitude than that of $g_{\text{tot}}(r)$ regardless of the cell type. This first peak amplitude increase is due to the preferred interaction with subunit B rather with any other random particle of the multicomponent mixture, whereas the first peak of the $g_{B-B}(r)$ curve is only slightly less than that of $g_{\text{tot}}(r)$, resembling a hard sphere (HS) gas.

At $\phi = 0.3$ (Fig. 7), the radial distribution function plots have emerged more pronounced peaks than those curves at $\phi = 0.1$ in Fig. 6. Here, a distinct second peak appears, characterizing a higher-order structure. The first peak of $g_{B-!B}(r)$ is ~25% larger in amplitude than that of $g_{\text{tot}}(r)$ regardless of the cell type. Again, this first peak amplitude increase is due to the preferred interaction with subunit B. However, the amplitude increases of the first peak of $g_{B-!B}(r)$ is less than that of $g_{\text{tot}}(r)$ at $\phi = 0.1$. This outcome is due to the more prominent entropic forces at $\phi = 0.3$ than at $\phi = 0.1$.

Interestingly, unlike at $\phi = 0.1$, at $\phi = 0.3$, the $g_{B-B}(r)$ curve differs between the cell types. For instance, the effect of changing the cell type either increases the osmotic pressure for lung in Fig. 7(c) or decreases it for brain in Fig. 7(b). The highest first peak value for $g_{B-B}(r)$ is that of the lungs. Here, B is closer to itself than in the other cell types due to the relative increase in μ_B in Fig. 5. Even though subunit B does have energetic self-interactions other than volume exclusion, the growth in μ_B increases the entropic forces (or osmotic pressure) between itself due to macromolecular crowding. The lowest first peak value of $g_{B-B}(r)$ is that of the brain, signifying a decrease in entropic interaction. Interestingly, the second peak of $g_{B-B}(r)$ for the brain is the largest in amplitude, signifying that the B subunits are assembling with another subunit type in between. This points toward subunits forming emergent higher-order assemblies under high macromolecular crowding content.

D. Higher-order assemblies built from pair-wise information

Because of crowding and the certain preferential interactions between subunits, the mixture is highly inhomogeneous, filling with particles in lumps and clusters. The variation in the profile in $g(r)$ signifies the presence of emergent clustering of the subunits around B at high crowding content at $\phi = 0.3$, which does not exist at low crowding content at $\phi = 0.1$. Next, we gained more insight into the higher-order structure of the assemblies by calculating the Jaccard index J using Eq. (19). The use of J is a well-established computational-topology algorithm to identify a new protein family or a gene from the existing database by comparing similarities in their interacting networks [49]. Sardiu *et al.* [46,47] have applied this index to determine whether a protein belongs to a core of protein networks in an interactome database as a justification of forming physical assemblies. Here, we leveraged the Jaccard index to further measure whether a subunit belongs to an emergent cluster of heterotypic particles. If $J(\alpha,\ \beta) = 1$, the subunits α and β only associate with each other and no other subunit. If $J(\alpha,\ \beta) = 0$, the subunits α and β only associate with other subunits, or they are not found in the system for the specific time. Thus, Eq. (19) is the probability of subunits α and β associating with each other, given that the two subunits are not isolated and part of a complex assembly. Comparing this with the contact matrix in Fig. 2(c), we can identify the emergent properties of forming a cluster that stems from the chemical potentials or macromolecular crowding instead of specific protein-protein interactions. Since the heart cell type has a similar $g(r)$ profile as the whole cell type, we focused this analysis on whole, brain, and lung cell types.

First, we examine the time average $\langle J(\alpha,\ \beta)\rangle$ at a low crowding content at $\phi = 0.1$ in Figs. 8(a)–8(c). The J index highest values are the A-B pair in whole, whereas in lung and brain, it is in the self-interaction B-B and A-A, respectively. Since there are no self-interactions of the subunits on the contact matrix [Fig. 2(c)], these highest J values in lung and brain are purely driven by the relative abundance and hence chemical potentials. A representative snapshot with the largest cluster highlight is shown in Figs. 8(d)–8(f). In these plots, A, B, C and D subunit types are color coded as cyan, orange, purple, and yellow, respectively. Furthermore, in comparison with the whole cell type [Figs. 8(a)], the lung and brain J matrices do not average to that of whole. The clusters formed for the whole cell type resemble the structure in Fig. 2(b), but the lung and brain have clusters of mainly B and A subunits, respectively.

Beyond the diagonal of the J matrices, we also see nontrivial off-diagonal variations in J values between the cell types. The largest off-diagonal values for all cell types is the J(A,B), which can be attributed to the fact that subunits A and B have the two largest abundances (Table I) and chemical potentials (Fig. 5) in all cell types. A key difference, though, between lung and brain is the increase in the J(A,C), J(B,C), and J(C,C) values for $\phi = 0.1$ [Figs. 8(b) and 8(c)].

Next, we increase the crowding content from $\phi = 0.1$ to $\phi = 0.3$. Similar trends on the J index persist at $\phi = 0.3$ for all cell types [Figs. 9(a)–9(c)]. However, the higher volume fraction gives rise to emergent higher-order assemblies seen in Figs. 9(d)–9(f). Again, in lung, J(B,B) has the maximal value [Fig. 9(b)], and J(A,A) has the maximal value [Fig. 9(c)] in brain. Since entropic forces are heightened by increasing N, at $\phi = 0.3$ [Figs. 9(b) and 9(c)], the maximum J value has increased to 0.51 for lung and 0.57 for brain.

Lastly, comparing the J values between $\phi = 0.1$ and $\phi = 0.3$, the general shifts in all cell types are toward the subunit pairs which have no attractive interaction [the pairs in white in Fig. 2(c)] in the Hamiltonian [Eqs. (14) and (15)] and away from the subunit pairs that do [the pairs in gray in Fig. 2(c)]. The increase in entropic interactions (i.e., depletion forces from crowding), due to the increase in N, causes this shift. Even though the main changes in relative abundance between the cell types are of subunits A and B (Table I), the entropic interactions affect the J values of all the subunits nontrivially. At $\phi = 0.3$, higher-order clusters emerge with distinct features shown by the representative snapshots in Figs. 9(d)–9(f). In lung, the dominant cluster with transient stability is composed of subunit B [Fig. 9(e)]. As crowding content reduces, such a cluster breaks down [Fig. 8(e)]. Such features permit the development of hypotheses connecting the properties of molecular assemblies to cell phenotypes.

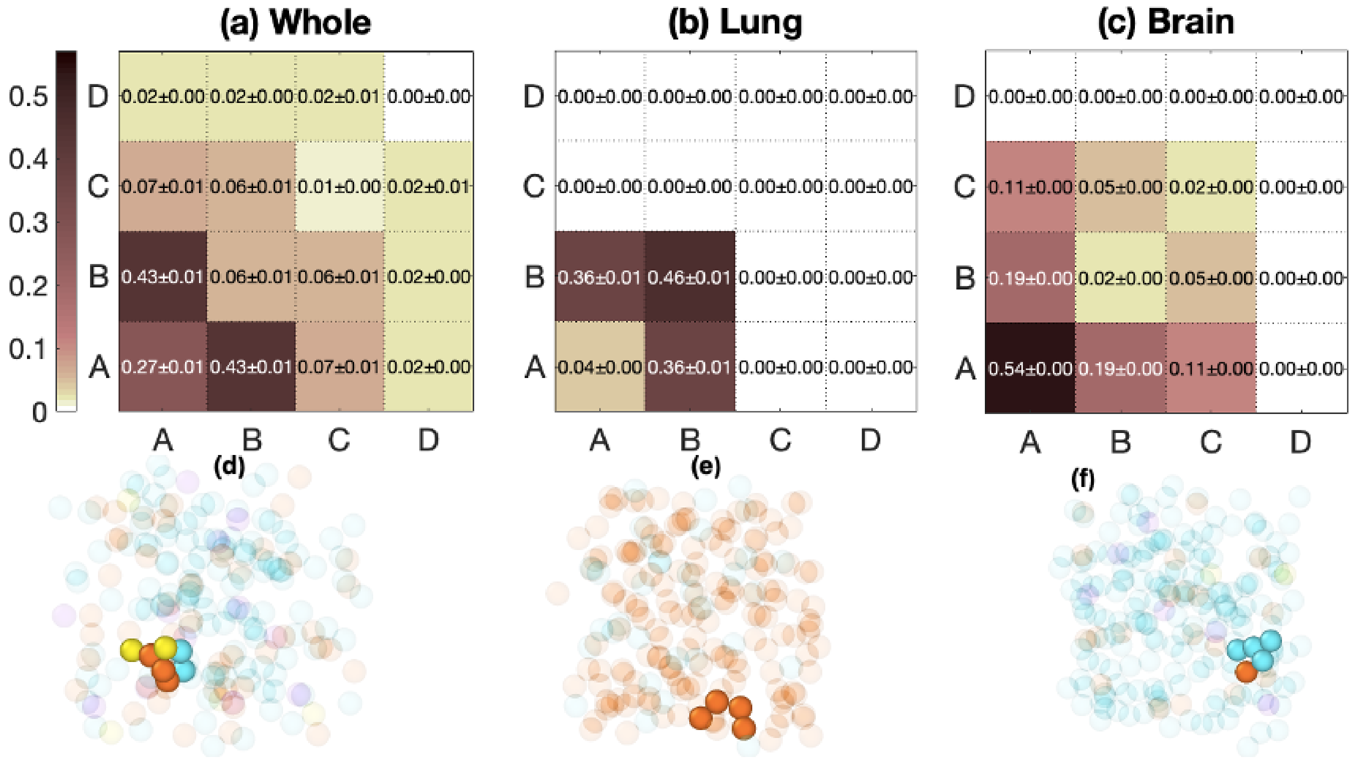

FIG. 8. J matrix for (a) whole, (b) lung, and (c) brain cell types for $\phi = 0.1$. Panels (d)–(f) show snapshots from randomly selected timeframes of the clusters for respective cell types. The largest clusters have been highlighted. Subunit types A–D are color coded as cyan, orange, purple, and yellow, respectively. Contact cutoff is 1.1σ, and visualization of the system is obtained using OVITO [50].

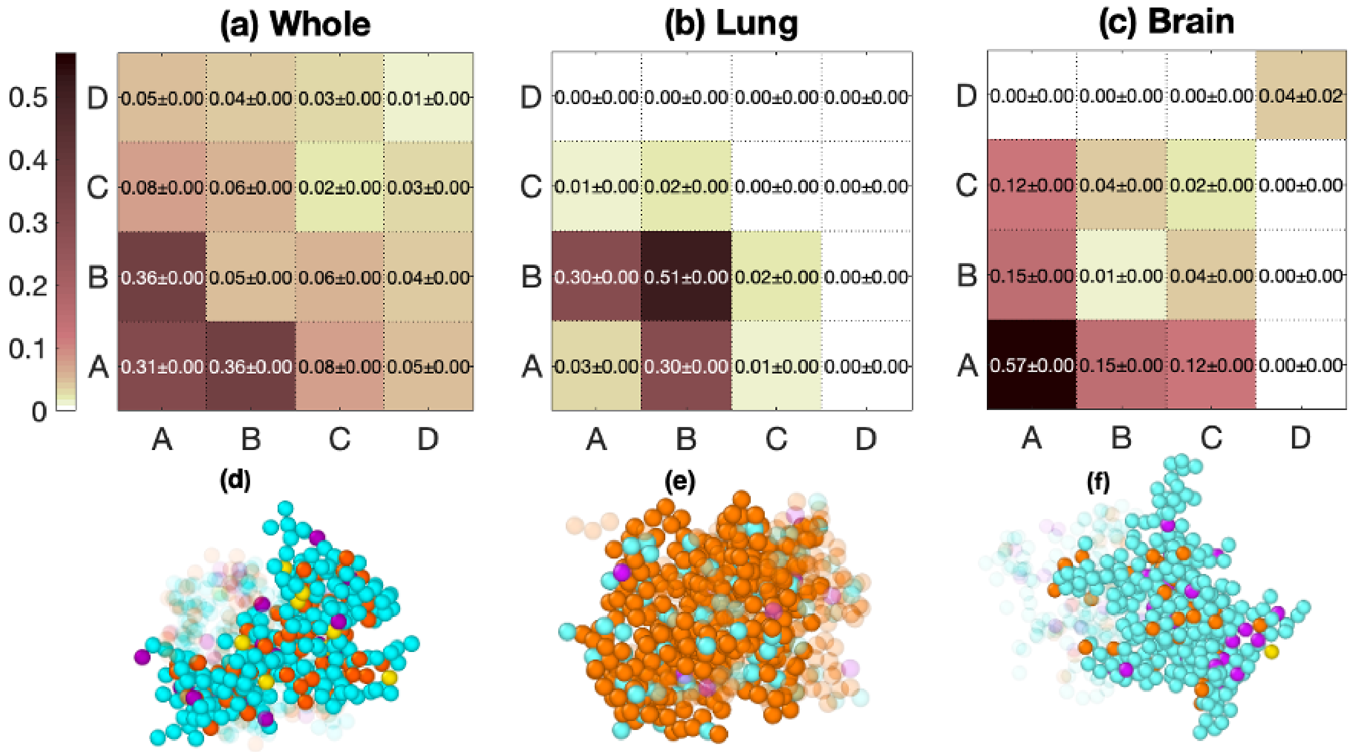

FIG. 9. J matrix for (a) whole, (b) lung, and (c) brain cell types for $\phi = 0.3$. Panels (d)–(f) show snapshots from randomly selected timeframes of the clusters for respective cell types. The largest clusters have been highlighted. Subunit types A–D are color coded as cyan, orange, purple, and yellow, respectively. Contact cutoff is $1.1\ \sigma$, and visualization of the system is obtained using OVITO [50].

IV. DISCUSSION

A. Leverage protein abundance as a constraint on the particle numbers by combining GCMC simulations with max entropy inference

Protein assembly in the cell is thermodynamically governed by various enthalpic and entropic factors. In this paper, we have allowed the particle number to fluctuate (i.e., the grand canonical ensemble; see Fig. 1) as an additional free energy contribution to understand the effects on protein assembly. The difficulty in using the grand canonical ensemble is that the chemical potentials for specific particle types are unknown. To avoid this problem, previous efforts in studying protein or synthetic particle assembly have either defined the chemical potentials as known [28–30,44,51,52] or used semigrand canonical ensembles [53], where particle types may vary in particle number with the total N fixed.

In this paper, we have leveraged the protein abundance attained from PaxDb [39] as a constraint to solve for the chemical potential of a protein type in an open, crowded environment with the tools of inverse statistical mechanics [54], GCMC simulations, and the maximum entropy principle [40,41]. The grand canonical ensemble allows particle numbers to fluctuate by placing particle reservoirs with fixed chemical potentials of each particle type in contact with the multicomponent system (Fig. 1).

By using the grand canonical ensemble, the fixed particle number is switched for a fixed chemical potential, trading one unknown for another. Analytically calculating the chemical potential is possible for simple systems such as HSs; however, this becomes increasingly difficult or impossible with varying protein-protein interactions, polydisperse mixture, or flexible polymers. To our knowledge, this approach is the only method that will approximate the correct chemical potentials, given the particle number distribution. Our self-consistent algorithm (Algorithm 1) has proven to be a useful way to calculate the chemical potentials of the particles in a multicomponent mixture from protein abundance. In principle, our method may be used for more complex systems such as a mixture differing particle shapes or macromolecules (or polymers) instead of simple spheres with the same radius.

B. GCMC allows for the investigation of emergent complex formation in crowded media

Here, in this paper, the use of the grand canonical ensemble allows variation in particle number which in turn allows the possibility of having different chemical potentials for different cell types. This difference in chemical potential and variation in the number of particles in the simulation leads to the formation of emergent, higher-order complex structures. This feature is different from the simulation approaches based on a conventional canonical ensemble where the total particle number in a simulation box is fixed [18,19]. Previous inverse design studies have learned interaction potentials using radial distribution function $g(r)$ in a canonical ensemble. The interactions of the protein complex in a canonical ensemble become the main driving force that guides the formation of the complex structure.

By increasing the volume fraction of the system, the entropically favored assembly increases as well. Increasing the volume fraction (or crowding) of the system and keeping other physical properties constant will only change the entropy since the crowding effect or depletion force is an entropic effect [11,12,55]. In doing so, specific protein-protein entropic forces or “osmotic pressures” are created and varied by the level of crowding [12,48]. This osmotic pressure can be seen with the radial distribution function $g(r)$ since the integral of $g(r)$ gives the corrections to the ideal gas pressure in Figs. 6 and 7, dictated by a combination of entropic forces such as macromolecular crowding and specific interactions. Interestingly, even at low crowding content (Fig. 6) the $g(r)$ between subunits B and not B subunits (i.e., B and !B) shows deviations from the total system regardless of the cell type by comparing with the contact map in Fig. 2, which signifies specific protein-protein osmotic pressure that is controlled by the chemical potentials in a crowded environment where particle numbers are allowed to fluctuate.

Because of differences in abundance and interactions among subunits, the mixture is highly inhomogeneous with small lumps and clusters. This emergence of new cluster formation in a grand canonical ensemble is heightened when the crowding content increases to $\phi = 0.3$ in Figs. 9(d)–9(f) from 0.1 in Figs. 8(d)–8(f). We show that, particularly for the lung, where the abundance of subunit B is much higher than those for other cell types, the emergent complex with lumps of B subunits is most significant at $\phi = 0.3$, while other cell types do not show prominent lumps of B subunits.

C. Our investigations allow development of hypotheses connecting the high-order protein complexes with the cell phenotype

Cell types may have the same gene sequences, but the cell state will differ in protein abundance [56]. As the mixture with heterotypic particles is inhomogeneous, it is challenging to characterize the lumps or clusters in the system, as densities of heterotypic particles alone is not sufficient to address their association with one another in space. In this paper, we use the Jaccard index to elucidate the higher-order complex structures from particle interactions. This measurement is key to uncover the topology of these complex structures, reflecting the differences in cell type.

From the J matrices (Figs. 8 and 9), we have seen that the whole cell type does not resemble the specialized cell types (lung and brain). These distinct behaviors between the cell types may be the reason why proteins that are similar in structure and sequence form different complexes. For example, SDH forms trimers of the complex in *E. coli* [33] but is only a single monomer in the porcine heart [34]. Our method may be used to study the various cell phenotypes leading from the higher-order complex assembly that is dictated from protein abundance or the chemical potentials of the protein species. The unique state or phenotype is connected to the unique chemical potential landscape (Fig. 5), which gives rise to emergent molecular topologies of higher-order complexes, depending on the crowding content.

V. CONCLUSIONS

We have developed a self-consistent algorithm, in accordance with the principle of maximum entropy, that calculates the chemical potentials that produce experimentally observed mean particle numbers. With this method and the GCMC simulations, we have gained insight into the mechanism and underlying principles of hierarchical assembly of macromolecular complexes, with emergent features varying with the crowding content.

Our method is a framework to connect the growing proteomic (or other "-omic") information [20,57] to physical models. We attained the protein abundance from PaxDb [39]; however, any other methods for extracting the experimentally observed relative abundance can be used. To establish this method of finding the correct chemical potentials given the cell type, we used the simplest protein-protein interactions in our structure-based Hamiltonian (Sec. II D) that contributes to the quinary interactions in the formation of higher-order protein complexes. Many studies have focused on understanding the protein-protein interaction networks [31,32,58], and here, we bring attention to the importance of the chemical potential for protein complex assembly in the spatial arrangement of proteins in quinary complexes. Since the state of the cell may change both the interaction between proteins and chemical potentials, understanding the relationship between both aspects will be an important future work. Our method lays the foundation to create physical models that are bioinformatically consistent.

ACKNOWLEDGMENTS

We thank the computing resources from the Hewlett Packard Enterprise Data Science Institute at UH. A.G.G. was supported by a training fellowship on the Houston Area Molecular Biophysics Program (No. T32 GM008280). We also thank the National Science Foundation (MCB-1412532, OAC-1531814) for funding.

APPENDIX A: SINGLE-COMPONENT FLUID INTERACTION STRENGTH ANALYSIS

We chose $\epsilon = 0.5$ for the LJ potential interaction strength because, > 0.8, certain values of the chemical potential corresponded to multiple volume fractions, which signified multiple phases (shown in Fig. 3). Thus, a one-to-one correspondence of chemical potential to volume fraction reduces the chance of having large fluctuation in energy and/or density.

APPENDIX B: ANALYTICAL APPROXIMATION OF μ FOR MULTICOMPONENT MIXTURES

In Fig. 5, we compared the calculated chemical potential from our self-consistent algorithm with an analytical approximation of μ. Here, we derive the equations used for those curves. In an ideal gas, $\mu = k_B T \ln\phi$. However, at nondilute conditions such as the crowded cell,

$$\frac{\mu}{k_B T} = \ln\phi + \sum_{k=2}^{\infty} B^{(k)}\phi^{k-1},$$

where, $B^{(k)}$ is the kth virial coefficient. For multicomponent mixture, the second coefficient between particles α and β can be calculated by

$$B^{(2)}_{\alpha\beta} = -2\pi \int r^2_{\alpha\beta}[e^{-\beta U_{\mathrm{LJ}}(r_{\alpha\beta})} - 1]dr_{\alpha\beta}.$$

For higher-order terms, the integrals become increasingly more complex. Since the repulsive terms become more dominant as $\phi \to 1$, we can assume the potential of HSs and ignore the attractive terms:

$$\frac{\mu_\alpha}{k_B T} = \ln[\phi_\alpha(1 - \ln x_\alpha)] + \sum_j B^{(2)}_{\alpha\beta}\phi_\alpha + B^{(3)}_{\mathrm{HS}}\phi^2 + B^{(4)}_{\mathrm{HS}}\phi^3 + \mathcal{O}(\phi^4), \qquad \text{(B1)}$$

where $\phi_\alpha \equiv \phi\tilde{x}^{\mathrm{exp}}_\alpha$, and $B^{(k)}_{\mathrm{HS}}$ is the HS virial coefficient found from the lookup table [59].

[1] R. Milo and R. Phillips, *Cell Biology by the Numbers* (Garland Science, New York, 2015).

[2] S. Kühner *et al.*, Proteomic organization in a genome-reduced bacterium, Science **326**, 1235 (2009).

[3] H. Wu and M. Fuxreiter, The structure and dynamics of higher-order assemblies: Amyloids, signalosomes, and granules, Cell **165**, 1055 (2016).

[4] P. C. Havugimana *et al.*, A census of human soluble protein complexes, Cell **150**, 1068 (2012).

[5] M. E. Cusick, N. Klitgord, M. Vidal, and D. E. Hill, Interactome: Gateway into systems biology, Hum. Mol. Genet. **14**, R171 (2005).

[6] J. D. O' Connell, A. Zhao, A. D. Ellington, and E. M. Marcotte, Dynamic reorganization of metabolic enzymes into intracellular bodies, Annu. Rev. Cell Dev. Biol. **28**, 89 (2012).

[7] A. J. Wirth and M. Gruebele, Quinary protein structure and the consequences of crowding in living cells: Leaving the test-tube behind, Bioessays **35**, 984 (2013).

[8] P. Chien and L. M. Gierasch, Challenges and dreams: Physics of weak interactions essential to life, Mol. Biol. Cell **25**, 3474 (2014).

[9] Y. Shin and C. P. Brangwynne, Liquid phase condensation in cell physiology and disease, Science **357**, eaaf4382 (2017).

[10] M. S. Cheung and A. G. Gasic, Towards developing principles of protein folding and dynamics in the cell, Phys. Biol. **15**, 063001 (2018).

[11] A. P. Minton, Models for excluded volume interaction between an unfolded and rigid macromolecular cosolutes: Macromolecular crowding and protein stability revisited, Biophys. J. **88**, 971 (2005).

[12] M. S. Cheung, D. Klimov, and D. Thirumalai, Molecular crowding enhances native state stability and refolding rates, Proc. Natl. Acad. Sci. USA **102**, 4753 (2005).

[13] H. X. Zhou, G. Rivas, and A. P. Minton, Macromolecular crowding and confinement: Biochemical, biophysical, and

potential physiological consequences, Annu. Rev. Biophys. **37**, 375 (2008).

[14] A. P. Minton and J. Wilf, Effect of macromolecular crowding upon the structure and function of an enzyme: Glyceraldehyde-3-phosphate dehydrogenase, Biochemistry **20**, 4821 (1981).

[15] D. Guin and M. Gruebele, Weak chemical interactions that drive protein evolution: Crowding, sticking, and quinary structure in folding and function, Chem. Rev. **119**, 10691 (2019).

[16] E. H. McConkey, Molecular evolution, intracellular organization and the quinary structure of proteins, Proc. Nat. Acad. Sci. USA **79**, 3236 (1982).

[17] I. Yu, T. Mori, T. Ando, R. Harada, J. Jung, Y. Sugita, and M. Feig, Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm, eLife **5**, e19274 (2016).

[18] B. A. Lindquist, R. B. Jadrich, M. P. Howard, and T. M. Truskett, The role of pressure in inverse design for assembly, J. Chem. Phys. **151**, 104104 (2019).

[19] W. D. Pineros, B. A. Lindquist, R. B. Jadrich, and T. M. Truskett, Inverse design of multicomponent assemblies, J. Chem. Phys. **148**, 104509 (2018).

[20] M. Larance and A. I. Lamond, Multidimensional proteomics for cell biology, Nat. Rev. Mol. Cell Biol. **16**, 269 (2015).

[21] B. Ho, A. Baryshnikova, and G. W. Brown, Unification of protein abundance datasets yields a quantitative saccharomyces cerevisiae proteome, Cell Syst. **6**, 192 (2018).

[22] J. W. Harper and E. J. Bennett, Proteome complexity and the forces that drive proteome imbalance, Nature (London) **537**, 328 (2016).

[23] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, Extracting bulk properties of self-assembling systems from small simulations, J. Phys. Condens. Matter **22**, 104102 (2010).

[24] T. E. Ouldridge, Inferring bulk self-assembly properties from simulations of small systems with multiple constituent species and small systems in the grand canonical ensemble, J. Chem. Phys. **137**, 144105 (2012).

[25] H. Reiss and R. K. Bowles, Some fundamental statistical mechanical relations concerning physical clusters of interest to nucleation theory, J. Chem. Phys. **111**, 7501 (1999).

[26] W. Zheng, M.-Y. Tsai, M. Chen, and P. G. Wolynes, Exploring the aggregation free energy landscape of the amyloid-β protein (1–40), Proc. Nat. Acad. Sci. USA **113**, 11835 (2016).

[27] P. R. ten Wolde and D. Frenkel, Enhancement of protein crystal nucleation by critical density fluctuations, Science **277**, 1975 (1997).

[28] W. M. Jacobs and D. Frenkel, Predicting phase behavior in multicomponent mixtures, J. Chem. Phys. **139**, 024108 (2013).

[29] P. Sartori and S. Leibler, Lessons from equilibrium statistical physics regarding the assembly of protein complexes, Proc. Natl. Acad. Sci. USA **117**, 114 (2020).

[30] W. M. Jacobs and D. Frenkel, Phase transitions in biological systems with many components, Biophys. J. **112**, 683 (2017).

[31] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing, Proc. Nat. Acad. Sci. USA **106**, 67 (2009).

[32] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Nat. Acad. Sci. USA **108**, E1293 (2011).

[33] V. Yankovskaya, R. Horsefield, S. Tornroth, C. Luna-Chavez, H. Miyoshi, C. Leger, B. Byrne, G. Cecchini, and S. Iwata, Architecture of succinate dehydrogenase and reactive oxygen species generation, Science **299**, 700 (2003).

[34] F. Sun, X. Huo, Y. Zhai, A. Wang, J. Xu, D. Su, M. Bartlam, and Z. Rao, Crystal structure of mitochondrial respiratory membrane protein complex II, Cell **121**, 1043 (2005).

[35] E. Gottlieb and I. P. Tomlinson, Mitochondrial tumour suppressors: A genetic and biochemical update, Nat. Rev. Cancer **5**, 857 (2005).

[36] B. Moosavi, E. A. Berry, X. L. Zhu, W. C. Yang, and G. F. Yang, The assembly of succinate dehydrogenase: A key enzyme in bioenergetics, Cell. Mol. Life Sci. **76**, 4023 (2019).

[37] D. K. Schweppe, J. D. Chavez, C. F. Lee, A. Caudal, S. E. Kruse, R. Stuppard, D. J. Marchinek, S. S. Gerald, R. Tian, and J. E. Bruce, Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry, Proc. Nat. Acad. Sci. USA **114**, 1732 (2017).

[38] M. Wang, C. J. Herrmann, M. Simonovic, D. Szklarczyk, and C. von Mering, Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines, Proteomics **15**, 3163 (2015).

[39] https://pax-db.org/.

[40] E. T. Jaynes, Information theory and statistical mechanics. II. Phys. Rev. **108**, 171 (1957).

[41] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. **106**, 620 (1957).

[42] B. Zhang and P. G. Wolynes, Shape Transitions and Chiral Symmetry Breaking in the Energy Landscape of the Mitotic Chromosome, Phys. Rev. Lett. **116**, 248101 (2016).

[43] B. Zhang and P. G. Wolynes, Topology, structures, and energy landscapes of human chromosomes, Proc. Natl. Acad. Sci. USA **112**, 6062 (2015).

[44] A. Murugan, Z. Zeravcic, M. P. Brenner, and S. Leibler, Multifarious assembly mixtures: Systems allowing retrieval of diverse stored structures, Proc. Natl. Acad. Sci. USA **112**, 54 (2015).

[45] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, J. Comput. Phys. **117**, 1 (1995).

[46] M. E. Sardiu, Y. Cai, J. Jin, S. K. Swanson, R. C. Conaway, J. W. Conaway, L. Florens, and M. P. Washburn, Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics, Proc. Nat. Acad. Sci. USA **105**, 1454 (2008).

[47] M. E. Sardiu, J. M. Gilmore, B. D. Groppe, A. Dutta, L. Florens, and M. P. Washburn, Topological scoring of protein interaction networks, Nat. Commun. **10**, 1118 (2019).

[48] A. G. Gasic, M. M. Boob, M. B. Prigozhin, D. Homouz, C. M. Daugherty, M. Gruebele, and M. S. Cheung, Critical Phenomena in the Temperature-Pressure-Crowding Phase Diagram of a Protein, Phys. Rev. X **9**, 041035 (2019).

[49] J. I. F. Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, and A. J. M. Walhout, Using networks to measure similarity between genes: Association index selection, Nat. Methods **10**, 1169 (2013).

[50] A. Stukowski, Visualization and analysis of atomistic simulation data with OVITO—the Open Visualization Tool, Model. Simul. Mater. Sci. Eng. **18**, 015012 (2010).

[51] N. A. Mahynski, R. Mao, E. Pretti, V. K. Shen, and J. Mittal, Grand canonical inverse design of multicomponent colloidal crystals, Soft Matter **16**, 3187 (2020).

[52] N. A. Mahynski, E. Pretti, V. K. Shen, and J. Mittal, Using symmetry to elucidate the importance of stoichiometry in colloidal crystal assembly, Nat. Commun. **10**, 2028 (2019).

[53] Y. Tang, A new method of semigrand canonical ensemble to calculate first-order phase transitions for binary mixtures, J. Chem. Phys. **136**, 034505 (2012).

[54] H. C. Nguyen, R. Zecchina, and J. Berg, Inverse statistical problems: From the inverse Ising problem to data science, Adv. Phys. **66**, 197 (2017).

[55] S. Asakura and F. Oosawa, On interaction between two bodies immersed in a solution of macromolecules, J. Chem. Phys. **22**, 1255 (1954).

[56] Y. Liu, A. Beyer, and R. Aebersold, On the dependency of cellular protein levels on mRNA abundance, Cell **165**, 535 (2016).

[57] B. Zhang and B. Kuster, Proteomics is not an island: Multi-omics integration is the key to understanding biological systems, Mol. Cell. Proteomics **18**, S1 (2019).

[58] J. Zhang, S. Maslov, and E. I. Shakhnovich, Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size, Mol. Syst. Biol. **4**, 210 (2008).

[59] J. P. Hansen and I. R. McDonald, *Theory of Simple Liquids: With Applications to Soft Matter* (Elsevier, Oxford, 2013).