G25.2651: Statistical Mechanics

Notes for Lecture 9

I. OVERVIEW

Our treatment of the classical ensembles makes clear that the free energy is a quantity of particular importance in statistical mechanics. Being related to the logarithm of the partition function, the free energy is the generator through which other thermodynamic quantities are obtained, via differentiation. In many cases, the free energy *difference* between two thermodynamic states is sought. Such differences tell, for example, whether or not a chemical reaction can occur spontaneously or requires input of work and is directly related to the equilbrium constant for the reaction. Thus, for example, from free energy differences, one can compute solvation free energies, acid ionization constants K_a and associated pK_a values, or drug inhibition constants K_i, that quantify the ability of a compound to bind to the active site of an enzyme. Another type of free energy often sought is the free energy as a function of one or more generalized coordinates in a system. An example might be the free energy surface as a function of a pair of Ramachandran angles ϕ and ψ in an oligopeptide. Such a surface would provide a map of the stable conformations of the molecule, the relative stability of such conformations and the heights of barriers that need to be crossed in a conformational change.

II. FREE-ENERGY PERTURBATION THEORY

We begin our treatment of free energy differences by examining the problem of transforming a system from one thermodynamic state to another. Let these states be denoted generically as $\mathcal{A}$ and $\mathcal{B}$. At the microscopic level, these two states are characterized by potential energy functions $U_{\mathcal{A}}(\mathbf{r}_1,...,\mathbf{r}_N)$ and $U_{\mathcal{B}}(\mathbf{r}_1,...,\mathbf{r}_N)$. For example, in a drug-binding study, the state $\mathcal{A}$ might correspond to the unbound ligand and enzyme, while $\mathcal{B}$ would correspond to the bound complex. In this case, the potential $U_{\mathcal{A}}$ would exclude all interactions between the enzyme and the ligand and the enzyme, whereas they would be included in the potential $U_{\mathcal{B}}$.

The Helmholtz free energy difference between the states $\mathcal{A}$ and $\mathcal{B}$ is simply $A_{\mathcal{AB}} = A_{\mathcal{B}} - A_{\mathcal{A}}$. The two free energies $A_{\mathcal{A}}$ and $A_{\mathcal{B}}$ are given in terms of their respective canonical partition functions $Q_{\mathcal{A}}$ and $Q_{\mathcal{B}}$, respectively by $A_{\mathcal{A}} = -kT\ln Q_{\mathcal{A}}$ and $A_{\mathcal{B}} = -kT\ln Q_{\mathcal{B}}$, where

$$\begin{aligned}
Q_{\mathcal{A}}(N,V,T) &= C_N\int d^N\mathbf{p}\,d^N\mathbf{r}\ \exp\left\{-\beta\left[\sum_{i=1}^N\frac{\mathbf{p}_i^2}{2m_i}+U_{\mathcal{A}}(\mathbf{r}_1,...,\mathbf{r}_N)\right]\right\} \\
&= \frac{Z_{\mathcal{A}}(N,V,T)}{N!\lambda^{3N}} \\
Q_{\mathcal{B}}(N,V,T) &= C_N\int d^N\mathbf{p}\,d^N\mathbf{r}\ \exp\left\{-\beta\left[\sum_{i=1}^N\frac{\mathbf{p}_i^2}{2m_i}+U_{\mathcal{B}}(\mathbf{r}_1,...,\mathbf{r}_N)\right]\right\} \\
&= \frac{Z_{\mathcal{B}}(N,V,T)}{N!\lambda^{3N}}
\end{aligned} \tag{1}$$

The free energy difference is, therefore,

$$A_{\mathcal{AB}} = A_{\mathcal{B}} - A_{\mathcal{A}} = -kT\ln\left(\frac{Q_{\mathcal{B}}}{Q_{\mathcal{A}}}\right) = -kT\ln\left(\frac{Z_{\mathcal{B}}}{Z_{\mathcal{A}}}\right) \tag{2}$$

where $Z_{\mathcal{A}}$ and $Z_{\mathcal{B}}$ are the configurational partition functions for states $\mathcal{A}$ and $\mathcal{B}$, respectively,

$$\begin{aligned}
Z_{\mathcal{A}} &= \int d^N\mathbf{r}\ e^{-\beta U_{\mathcal{A}}(\mathbf{r}_1,...,\mathbf{r}_N)} \\
Z_{\mathcal{B}} &= \int d^N\mathbf{r}\ e^{-\beta U_{\mathcal{B}}(\mathbf{r}_1,...,\mathbf{r}_N)}
\end{aligned} \tag{3}$$

The ratio of full partition functions $Q_{\mathcal{B}}/Q_{\mathcal{A}}$ reduces to the ratio of configurational partition functions $Z_{\mathcal{B}}/Z_{\mathcal{A}}$ because the momentum integrations in the former cancel out of the ratio.

Eqn. (2) is difficult to implement in practice because in any numerical calculation via either molecular dynamics or Monte Carlo, we do not have direct access to the partition function only averages of phase-space functions corresponding to physical observables. However, if we are willing to extend the class of phase-space functions whose averages we seek to functions that do not necessarily correspond to direct observables, then the ratio of configurational partition functions can be manipulated to be in the form of such an average. Consider inserting unity into the expression for $Z_{\mathcal{B}}$ as follows:

$$\begin{aligned} Z_{\mathcal{B}} &= \int d^N\mathbf{r}\; e^{-\beta U_{\mathcal{B}}(\mathbf{r}_1,...,\mathbf{r}_N)} \\ &= \int d^N\mathbf{r}\; e^{-\beta U_{\mathcal{B}}(\mathbf{r}_1,...,\mathbf{r}_N)} e^{-\beta U_{\mathcal{A}}(\mathbf{r}_1,...,\mathbf{r}_N)} e^{\beta U_{\mathcal{A}}(\mathbf{r}_1,...,\mathbf{r}_N)} \\ &= \int d^N\mathbf{r}\; e^{-\beta U_{\mathcal{A}}(\mathbf{r}_1,...,\mathbf{r}_N)} e^{-\beta(U_{\mathcal{B}}(\mathbf{r}_1,...,\mathbf{r}_N)-U_{\mathcal{A}}(\mathbf{r}_1,...,\mathbf{r}_N))} \end{aligned} \tag{4}$$

If we now take the ratio $Z_{\mathcal{B}}/Z_{\mathcal{A}}$, we find

$$\begin{aligned} \frac{Z_{\mathcal{B}}}{Z_{\mathcal{A}}} &= \frac{1}{Z_{\mathcal{A}}}\int d^N\mathbf{r}\; e^{-\beta U_{\mathcal{A}}(\mathbf{r}_1,...,\mathbf{r}_N)} e^{-\beta(U_{\mathcal{B}}(\mathbf{r}_1,...,\mathbf{r}_N)-U_{\mathcal{A}}(\mathbf{r}_1,...,\mathbf{r}_N))} \\ &= \left\langle e^{-\beta(U_{\mathcal{B}}(\mathbf{r}_1,...,\mathbf{r}_N)-U_{\mathcal{A}}(\mathbf{r}_1,...,\mathbf{r}_N))}\right\rangle_{\mathcal{A}} \end{aligned} \tag{5}$$

where the notation $\langle\cdots\rangle_{\mathcal{A}}$ indicates an average taken with respect to the canonical configurational distribution of the state $\mathcal{A}$. Substituting eqn. (5) into eqn. (2), we find

$$A_{\mathcal{AB}} = -kT\ln\left\langle e^{-\beta(U_{\mathcal{B}}-U_{\mathcal{A}})}\right\rangle_{\mathcal{A}} \tag{6}$$

Eqn. (6) is known as the *free-energy perturbation* formula; it should be reminiscent of the thermodynamic perturbation formula used to derive the van der Waals equation. Eqn. (6) can be interpreted as follows: We start with microstates $\{\mathbf{r}_1,...,\mathbf{r}_N\}$ selected from the canonical ensemble of state $\mathcal{A}$ and use these to compute $Z_{\mathcal{B}}$ by placing them in the state $\mathcal{B}$ by simply changing the potential energy from $U_{\mathcal{A}}$ to $U_{\mathcal{B}}$. In so doing, we need to "unbias" our choice to sample the configurations from the canonical distribution of state $\mathcal{A}$ by removing the weight factor $\exp(-\beta U_{\mathcal{A}})$ from which the microstates are sample and reweighting the states by the factor $\exp(-\beta U_{\mathcal{B}})$ corresponding to state $\mathcal{B}$. This leads to eqn. (5). The difficulty with this approach is that the microstates corresponding to the canonical distribution of state $\mathcal{A}$ may not be states of high probability in the canonical distribution of state $\mathcal{B}$. If this is the case, then the potential enegy difference $U_{\mathcal{B}} - U_{\mathcal{A}}$ will be large, he exponential factor $\exp[-\beta(U_{\mathcal{B}} - U_{\mathcal{A}})]$ will be negligibly small, and the free energy difference will be very slow to converge in an actual simulation. For this reason, it is clear that the free-energy perturbation formula is only useful for cases in which the two states $\mathcal{A}$ and $\mathcal{B}$ are not that different from each other.

If $U_{\mathcal{B}}$ is not a small perturbation to $U_{\mathcal{A}}$, then the free-energy perturbation formula can still be salvaged by introducing a set of $M-2$ intermediate states with potentials $U_\alpha(\mathbf{r}_1,...,\mathbf{r}_N)$, where $\alpha = 1,...,M$, $\alpha = 1$ corresponds to the state $\mathcal{A}$ and $\alpha = M$ corresponds to the state $\mathcal{B}$. Let $\Delta U_{\alpha,\alpha+1} = U_{\alpha+1} - U_\alpha$. We can now imagine transforming the system from state $\mathcal{A}$ to state $\mathcal{B}$ by passing through these intermediate states and computing the average of $\Delta U_{\alpha,\alpha+1}$ in the state α. Applying the free-energy perturbation formula to this protocol yields the free-energy difference as

$$A_{\mathcal{AB}} = -kT\sum_{\alpha=1}^{M-1}\ln\left\langle e^{-\beta\Delta U_{\alpha,\alpha+1}}\right\rangle_\alpha \tag{7}$$

where $\langle\cdots\rangle_\alpha$ means an average taken over the distribution $\exp(-\beta U_\alpha)$. The key to applying eqn. (7) is choosing the intermediate states so as to achieve sufficient overlap between the intermediate states without requiring a large number of them, i.e. choosing the thermodynamic path between states $\mathcal{A}$ and $\mathcal{B}$ effectively.

III. ADIABATIC SWITCHING AND THERMODYNAMIC INTEGRATION

The free-energy perturbation approach evokes a physical picture in which configurations sampled from the canonical distribution of state $\mathcal{A}$ are immediately "switched" to the state $\mathcal{B}$ by simply changing the potential from $U_{\mathcal{A}}$ to $U_{\mathcal{B}}$.

Such "instantaneous" switching clearly represents an unphysical path from one state to the other, but we need not concern ourselves with this because the free energy is a state function and, therefore, independent of the path connecting the states. Nevertheless, we showed that the free-energy perturbation theory formula, eqn. (6), is only useful if the states $\mathcal{A}$ and $\mathcal{B}$ do not differ vastly from one another, thus naturally raising the question of what can be done if the states are very different.

The use of a series of intermediate states, by which eqn. (7) is derived, exploits the fact that any path between the states can be employed to obtain the free energy difference. In this section, we will discuss an alternative approach in which the system is switched slowly or *adiabatically* from one state to the other, allowing the system to fully relax at each point along a chosen path from state $\mathcal{A}$ to state $\mathcal{B}$, rather than instantaneously switching the system between intermediate states, as occurs in eqn. (7). In order to effect the switching from one state to the other, we will employ a common trick in the form of an "external" switching parameter, λ. This paramter is introduced by defining a new potential energy function

$$U(\mathbf{r}_1,...,\mathbf{r}_N,\lambda) \equiv f(\lambda)U_{\mathcal{A}}(\mathbf{r}_1,...,\mathbf{r}_N) + g(\lambda)U_{\mathcal{B}}(\mathbf{r}_1,...,\mathbf{r}_N) \tag{8}$$

The functions $f(\lambda)$ and $g(\lambda)$ are referred to as *switching functions*, and they required to satisfy the conditions $f(0) = 1$, $f(1) = 0$, corresponding to the state $\mathcal{A}$, and $g(0) = 0$, $g(1) = 1$, corresponding to the state $\mathcal{B}$. Apart from these conditions, $f(\lambda)$ and $g(\lambda)$ are completely arbitrary. The mechanism embodied in eqn. (8) is one in which some imaginary external controlling influence ("hand of God"), represented by the λ parameter, starts the system off in state $\mathcal{A}$ ($\lambda = 0$) and slowly switches off the potential $U_{\mathcal{A}}$ while simultaneously switching on the potential $U_{\mathcal{B}}$. The process is complete when $\lambda = 1$, when the system is in state $\mathcal{B}$. A simple choice for the functions $f(\lambda)$ and $g(\lambda)$ is, for example, $f(\lambda) = 1 - \lambda$ and $g(\lambda) = \lambda$.

In order to see how eqn. (8) can be used to compute the free energy difference $A_{\mathcal{AB}}$, consider the canonical partition function of a system described by the potential of eqn. (8) for a particular choice of λ:

$$Q(N,V,T,\lambda) = C_N \int d^N\mathbf{p}\, d^N\mathbf{r}\ \exp\left\{-\beta\left[\sum_{i=1}^{N}\frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}_1,...,\mathbf{r}_N,\lambda)\right]\right\} \tag{9}$$

This partition function leads to a free energy $A(N,V,T,\lambda)$ via

$$A(N,V,T,\lambda) = -kT\ln Q(N,V,T,\lambda) \tag{10}$$

Recall, however, that the derivatives of the free energy with repsect to N, and V and T lead to the chemical potential, pressure and entropy, respectively. What does the derivative of the free energy $A(N,V,T,\lambda)$ with respect to λ represent? According to eqn. (10)

$$\frac{\partial A}{\partial \lambda} = -\frac{kT}{Q}\frac{\partial Q}{\partial \lambda} = -\frac{kT}{Z}\frac{\partial Z}{\partial \lambda} \tag{11}$$

The reader should check that the expressions involving Q and Z are equivalent. Computing the derivative of Z with respect to λ, we find

$$\begin{aligned}\frac{kT}{Z}\frac{\partial Z}{\partial \lambda} &= \frac{kT}{Z}\frac{\partial}{\partial \lambda}\int d^N\mathbf{r}\ e^{-\beta U(\mathbf{r}_1,...,\mathbf{r}_N,\lambda)} \\ &= \frac{kT}{Z}\int d^N\mathbf{r}\ \left(-\beta\frac{\partial U}{\partial \lambda}\right) e^{-\beta U(\mathbf{r}_1,...,\mathbf{r}_N,\lambda)} \\ &= -\left\langle \frac{\partial U}{\partial \lambda}\right\rangle \end{aligned} \tag{12}$$

Now, the free energy difference $A_{\mathcal{AB}}$ can be obtained trivially from the relation

$$A_{\mathcal{AB}} = \int_0^1 \frac{\partial A}{\partial \lambda} d\lambda \tag{13}$$

Substituting eqns. (11) and (12) into eqn. (13), we obtain the free energy difference as

$$A_{\mathcal{AB}} = \int_0^1 \left\langle \frac{\partial U}{\partial \lambda}\right\rangle_{\lambda} d\lambda \tag{14}$$

where $\langle \cdots \rangle_\lambda$ denotes an average over the canonical ensemble described by the distribution $\exp[-\beta U(\mathbf{r}_1, ..., \mathbf{r}_N, \lambda)]$ with λ fixed at a particular value. The special choice of $f(\lambda) = 1 - \lambda$ and $g(\lambda) = \lambda$ has a simple interpretation. For this choice, eqn. (14) becomes

$$A_{\mathcal{AB}} = \int_0^1 \langle U_{\mathcal{B}} - U_{\mathcal{A}} \rangle_\lambda \, d\lambda \tag{15}$$

The content of eqn. (15) can be understood by recalling the relationship between work and free energy from the second law of thermodynamics. If, in transforming the system from state $\mathcal{A}$ to state $\mathcal{B}$, an amount of work W is performed on the system, then

$$W \geq A_{\mathcal{AB}} \tag{16}$$

where equality holds *only* if the transformation is carried out along a *reversible path.* Since reversible work is related to a change in potential energy, eqn. (15) is actually a statistical version of eqn. (16) for the special case of equality. Eqn. (15) tells us that the free energy difference is the ensemble average of the microscopic reversible work needed to change the potential energy of each configuration from $U_{\mathcal{A}}$ to $U_{\mathcal{B}}$ along the chosen λ-path. Note, however, that eqn. (14), which is known as the *thermodynamic integration* formula, is true independent of the choice of $f(\lambda)$ and $g(\lambda)$, which means that eqn. (14) always yields the reversible work via the free energy difference. The flexibility in the choice of the λ-path, however, can be exploited to design adiabatic switching algorithms of greater efficiency that can be achieved with the simple choice $f(\lambda) = 1 - \lambda$, $g(\lambda) = \lambda$.

In practice, the thermodynamic integration formula is implemented as follows: A set of M values of λ is chosen from the interval $[0, 1]$, and at each chosen value λ_k, a full molecular dynamics or Monte Carlo calculation is carried out in order to generate the average $\langle \partial U / \partial \lambda_k \rangle_{\lambda_k}$. The resulting values of $\langle \partial U / \partial \lambda_k \rangle_{\lambda_k}$, $k = 1, ..., M$ are then substituted into eqn. (14), and the resulted is integrated numerically to produce the free energy difference $A_{\mathcal{AB}}$. Thus, we see that the selected values $\{\lambda_k\}$ can be evenly spaced, for example, or they could be a set of Gaussian quadrature nodes, depending on how $A(N, V, T, \lambda)$ is expected to vary with λ for the chosen $f(\lambda)$ and $g(\lambda)$.

As with free-energy perturbation theory, the thermodynamic integration approach can be implemented very easily. An immediately obvious disadvantage of the method, however, is the same one that applies to eqn. (7): In order to perform the numerical integration, it is necessary to perform many simulations of a system at physically uninteresting intermediate values of λ where the potential $U(\mathbf{r}_1, ..., \mathbf{r}_N, \lambda)$ is, itself, unphysical. Only $\lambda = 0, 1$ correspond to actual physical states and ultimately, we can only attach physical meaning to the free energy difference $A_{\mathcal{AB}} = A(N, V, T, 1) - A(N, V, T, 0)$. Nevertheless, the intermediate averages must be accurately calculated in order for the integration to yield a correct result. The approach to be presented in the next section attempts to reduce the time spent in such unphysical intermediate states and focuses the sampling in the important regions $\lambda = 0, 1$.

IV. JARZYNSKI'S EQUALITY AND NONEQUILIBRIUM METHODS

In this section, the relationship between work and free energy will be explored in greater detail. We have already introduced the inequality in eqn. (16), which states that if an amount of work $W_{\mathcal{AB}}$ is performed on a system, taking from state $\mathcal{A}$ to state $\mathcal{B}$, then $W_{\mathcal{AB}} \geq A_{\mathcal{AB}}$. Here, equality holds only if the work is performed reversibly. The work referred to here is thermodynamic quantity and, as such, must be regarded as an ensemble average. In statistical mechanics, we can also introduce the mechanical or microscopic work $\mathcal{W}_{\mathcal{AB}}(\mathrm{x})$ performed on one member of the ensemble to drive it from state $\mathcal{A}$ to state $\mathcal{B}$. Then, $W_{\mathcal{AB}}$ is simply an ensemble average of $\mathcal{W}_{\mathcal{AB}}$. However, we need to be somewhat careful about how we define this ensemble average because the work is defined along a particular path or trajectory which takes the system from state $\mathcal{A}$ to state $\mathcal{B}$, and equilibrium averages do not refer not to paths but to microstates. This distinction is emphasized by the fact that the work could be carried out irreversibly, such that the system is driven out of equilibrium. Thus, the proper definition of the ensemble average follows along the lines already discussed in the context of the free-energy perturbation approach, namely, averaging over the canonical distribution for the state $\mathcal{A}$. In this case, since we will be discussing actual paths x_t, we let the initial condition x_0 be the phase space vector for the system in the (initial) state $\mathcal{A}$. Recall that $\mathrm{x}_t = \mathrm{x}_t(\mathrm{x}_0)$ is a unique function of the initial conditions. Then

$$W_{\mathcal{AB}} = \langle \mathcal{W}_{\mathcal{AB}}(\mathrm{x}_0) \rangle_{\mathcal{A}} = \frac{C_N}{Q_{\mathcal{A}}(N, V, T)} \int d\mathrm{x}_0 \; e^{-\beta H_{\mathcal{A}}(\mathrm{x}_0)} \mathcal{W}_{\mathcal{AB}}(\mathrm{x}_0) \tag{17}$$

and the Clausius inequality can be stated as $\langle \mathcal{W}_{\mathcal{AB}}(\mathrm{x}_0) \rangle_{\mathcal{A}} \geq A_{\mathcal{AB}}$.

From such an inequality, it would seem that using the work as a method for calculating the free energy is of limited utility, since the work necessarily must be performed reversibly, otherwise one obtains only upper bound on the free energy. It turns out, however, that irreversible work can be used to calculate free energy differences by virtue of a connection between the two quantities first discovered in 1997 by C. Jarzynski that as come to be known as the *Jarzynski equality*. This equality states that if, instead of averaging $\mathcal{W}_{\mathcal{AB}}(\mathrm{x}_0)$ over the initial canonical distribution (that of state $\mathcal{A}$), an average of $\exp[-\beta \mathcal{W}_{\mathcal{AB}}(\mathrm{x}_0)]$ is performed over the same distribution, the result is $\exp[-\beta A_{\mathcal{AB}}]$, i.e.

$$e^{-\beta A_{\mathcal{AB}}} = \left\langle e^{-\beta \mathcal{W}_{\mathcal{AB}}(\mathrm{x}_0)} \right\rangle_{\mathcal{A}} = \frac{C_N}{Q_{\mathcal{A}}(N,V,T)} \int d\mathrm{x}_0 \; e^{-\beta H_{\mathcal{A}}(\mathrm{x}_0)} e^{-\beta \mathcal{W}_{\mathcal{AB}}(\mathrm{x}_0)} \tag{18}$$

This remarkable result not only provides a foundation for the development of nonequilibrium free-energy methods but also has profound implications for thermodynamics, in general.

The Jarzynski equality be proved using different strategies. Here, however, we will present a proof that is most relevant for the finite-sized systems and techniques employed in molecular dynamics calculations. Consider a time-dependent Hamiltonian of the form

$$H(\mathbf{p},\mathbf{r},t) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}_1, ..., \mathbf{r}_N, t) \tag{19}$$

For time-dependent Hamiltonian's, the usual conservation law $dH/dt = 0$ no longer holds, which can be seen by computing

$$\frac{dH}{dt} = \nabla_{\mathrm{x}_t} H \dot{\mathrm{x}}_t + \frac{\partial H}{\partial t} \tag{20}$$

where the phase space vector $\mathrm{x} = (\mathbf{p}_1, ..., \mathbf{p}_N, \mathbf{r}_1, ..., \mathbf{r}_N) \equiv (\mathbf{p}, \mathbf{r})$ has been introduced. Integrating both sides over time from $t = 0$ to a final time $t = \tau$, we find

$$\int_0^\tau dt \; \frac{dH}{dt} = \int_0^\tau dt \; \nabla_{\mathrm{x}_t} H \dot{\mathrm{x}}_t + \int_0^\tau dt \; \frac{\partial H}{\partial t} \tag{21}$$

Eqn. (21) can be regarded as a microscopic version of the first law of thermodynamics, in which the first and second terms represent the heat absorbed by the system and the work done on the system over the trajectory, respectively. Note that the work is actually a function of the initial phase-space vector x_0, which can be seen by writing this term explicitly as

$$W_\tau(\mathrm{x}_0) = \int_0^\tau dt \; \frac{\partial}{\partial t} H(\mathrm{x}_t(\mathrm{x}_0), t) \tag{22}$$

where the fact that the work depends explicitly on τ in eqn. (22) is indicated by the subscript. In the present discussion, we will consider that each initial condition, selected from a canonical distribution in x_0, evolves according to Hamilton's equations in isolation. In this case, the heat term $\nabla_{\mathrm{x}_t} H \cdot \mathrm{x}_t = 0$, and we have the usual addition to Hamilton's equations $dH/dt = \partial H/\partial t$.

With the above condition, we can write the microscopic work as

$$\mathcal{W}_{\mathcal{AB}} = \int_0^\tau \frac{d}{dt} H(\mathrm{x}_t(\mathrm{x}_0), t) dt = H(\mathrm{x}_\tau(\mathrm{x}_0), \tau) - H(\mathrm{x}_0, 0) \tag{23}$$

The last term $H(\mathrm{x}_0, 0)$ is also $H_{\mathcal{A}}(\mathrm{x}_0)$. Thus, the ensemble average of the exponential of the work becomes

$$\left\langle e^{-\beta \mathcal{W}_{\mathcal{AB}}} \right\rangle_{\mathcal{A}} = \frac{C_N}{Q_{\mathcal{A}}(N,V,T)} \int d\mathrm{x}_0 \; e^{-\beta H_{\mathcal{A}}(\mathrm{x}_0)} e^{-\beta [H(\mathrm{x}_\tau(\mathrm{x}_0),\tau) - H_{\mathcal{A}}(\mathrm{x}_0)]}$$

$$\frac{C_N}{Q_{\mathcal{A}}(N,V,T)} \int d\mathrm{x}_0 \; e^{-\beta H(\mathrm{x}_\tau(\mathrm{x}_0),\tau)}$$

The numerator in this expression becomes much more interesting if we perform a change of variables from x_0 to x_τ. Since the solution of Hamilton's equations for the time-dependent Hamiltonian uniquely map the initial condition x_0 onto x_t, when $t = \tau$, we have a new set of phase-space variables, and by Liouville's theorem, the phase-space volume element is preserved

$$dx_\tau = dx_0 \tag{24}$$

When the Hamiltonian is transformed, we find $H(x_\tau, \tau) = H_{\mathcal{B}}(x_\tau)$. Consequently,

$$\begin{aligned}
\left\langle e^{-\beta \mathcal{W}_{\mathcal{AB}}} \right\rangle_{\mathcal{A}} &= \frac{C_N}{Q(N,V,T)} \int dx_\tau \; e^{-\beta H_{\mathcal{B}}(x_\tau)} \\
&= \frac{Q_{\mathcal{B}}(N,V,T)}{Q_{\mathcal{A}}(N,V,T)} \\
&= e^{-\beta A_{\mathcal{AB}}}
\end{aligned}$$

thus proving the equality. The implication of the Jarzynski equality is that the work can be carried out along a reversible or irreversible path, and the correct free energy will still be obtained.

Note that due to Jensen's inequality:

$$\left\langle e^{-\beta \mathcal{W}_{\mathcal{AB}}} \right\rangle_{\mathcal{A}} \geq e^{-\beta \langle \mathcal{W}_{\mathcal{AB}} \rangle_{\mathcal{A}}} \tag{25}$$

Using Jarzynski's equality, this becomes

$$e^{-\beta A_{\mathcal{AB}}} \geq e^{-\beta \langle \mathcal{W}_{\mathcal{AB}} \rangle_{\mathcal{A}}} \tag{26}$$

which implies, as expected, that

$$A_{\mathcal{AB}} \leq \langle \mathcal{W}_{\mathcal{AB}} \rangle_{\mathcal{A}} \tag{27}$$

V. REACTION COORDINATES

It is frequently the case that the progress of some chemical, mechanical, or thermodynamics process can be followed by following the evolution of a small subset of generalized coordinates in a system. When generalized coordinates are used in this manner, they are typically referred to as *reaction coordinates*, *collective variables*, or *order parameters*, often depending on the context and type of system. Whenever referring to these coordinates, we will refer to them as *reaction coordinates*, although the reader should be aware that the other two designations are also used in the literature.

As an example of a useful reaction coordinate, consider a simple gas-phase diatomic dissociation process AB $\longrightarrow$ A+B. If $\mathbf{r}_{\mathrm{A}}$ and $\mathbf{r}_{\mathrm{B}}$ denote the Cartesian coordinates of atom A and B, then a useful generalized coordinate for following the progress of the dissociation is simply the distance $r = |\mathbf{r}_{\mathrm{B}} - \mathbf{r}_{\mathrm{A}}|$. A complete set of generalized coordinates that contains r as one of the coordinates is the set that contains the center of mass $\mathbf{R} = (m_{\mathrm{A}}\mathbf{r}_{\mathrm{A}} + m_{\mathrm{B}}\mathbf{r}_{\mathrm{B}})/(m_{\mathrm{A}} + m_{\mathrm{B}})$, the magnitude of the relative coordinate $r = |\mathbf{r}_{\mathrm{B}} - \mathbf{r}_{\mathrm{A}}|$, and the two angles $\phi = \tan^{-1}(y/x)$ and $\theta = \tan^{-1}(\sqrt{x^2 + y^2}/z)$, where x, y, and z are the components of the relative coordinate $\mathbf{r} = \mathbf{r}_{\mathrm{B}} - \mathbf{r}_{\mathrm{A}}$. Of course, in the gas-phase, where the potential between A and B likely only depends on the distance between A and B, r is really the *only* interesting coordinate. However, if the reaction were to take place in solution, then other coordinate such as θ and ϕ become more relevant as specific orientations might change the mechanism or thermodynamic picture of the process, depending on the complexity of the solvent, and averaging over these degrees of freedom to produce a free energy profile $A(r)$ in r alone will wash out some of this information.

As another example, consider a gas-phase proton transfer reaction A-H$\cdots$B$\longrightarrow$A$\cdots$H-B. Here, although the distance $|\mathbf{r}_{\mathrm{H}} - \mathbf{r}_{\mathrm{A}}|$ can be used to monitor the progress of the proton away from A and the distance $|\mathbf{r}_{\mathrm{H}} - \mathbf{r}_{\mathrm{B}}|$ can be used to monitor the progress of the proton toward B, neither distance alone is sufficient for following the progress of the reaction. However, the difference $\delta = |\mathbf{r}_{\mathrm{H}} - \mathbf{r}_{\mathrm{B}}| - |\mathbf{r}_{\mathrm{H}} - \mathbf{r}_{\mathrm{A}}|$ can be used to follow the progress of the proton transfer from A to B and, therefore, is a potentially useful reaction coordinate. A complete set of generalized coordinates involving δ can be constructed as follows. If $\mathbf{r}_{\mathrm{A}}$, $\mathbf{r}_{\mathrm{B}}$ and $\mathbf{r}_{\mathrm{H}}$ denote the Cartesian coordinates of the three atoms, then first introduce the center-of-mass $\mathbf{R} = (m_{\mathrm{A}}\mathbf{r}_{\mathrm{A}} + m_{\mathrm{B}}\mathbf{r}_{\mathrm{B}} + m_{\mathrm{H}}\mathbf{r}_{\mathrm{H}}/(m_{\mathrm{A}} + m_{\mathrm{B}} + m_{\mathrm{H}})$, the relative coordinate between A and B, $\mathbf{r} = \mathbf{r}_{\mathrm{B}} - \mathbf{r}_{\mathrm{A}}$, and a third relative coordinate $\mathbf{s}$ between H and the center-of-mass of A and B, $\mathbf{s} = \mathbf{r} - (m_{\mathrm{A}}\mathbf{r}_{\mathrm{A}} + m_{\mathrm{B}}\mathbf{r}_{\mathrm{B}}/(m_{\mathrm{A}} + m_{\mathrm{B}})$. Finally, $\mathbf{r}$ is transformed into spherical polar coordinates, (r, θ, ϕ), and from $\mathbf{r}$ and $\mathbf{s}$, three more coordinates are formed:

$$\sigma = |\mathbf{s} + \frac{m_{\mathrm{B}}}{m_{\mathrm{A}} + m_{\mathrm{B}}}\mathbf{r}| + |\mathbf{s} - \frac{m_{\mathrm{A}}}{m_{\mathrm{A}} + m_{\mathrm{B}}}\mathbf{r}| \qquad \delta = |\mathbf{s} + \frac{m_{\mathrm{B}}}{m_{\mathrm{A}} + m_{\mathrm{B}}}\mathbf{r}| - |\mathbf{s} - \frac{m_{\mathrm{A}}}{m_{\mathrm{A}} + m_{\mathrm{B}}}\mathbf{r}| \tag{28}$$

and the angle α, which measures the "tilt" of the plane containing the three atoms from the vertical. The coordinates (σ, δ, α) are known as *confocal elliptic* coordinates. These coordinates could also be used if the reaction takes place in solution. As expected, the generalized coordinates are functions of the original Cartesian coordinates. The alanine-dipeptide example above also employs the Ramachandran angles ϕ and ψ as reaction coordinates, and these can also be expressed as part of a set of generalized coordinates that are functions of the original Cartesian coordinates of a system.

While reaction coordinates or collective variables are potentially very useful constructs, they must be used with care, particularly when enhanced sampling methods are applied to them. Enhanced sampling of a poorly chosen reaction coordinate can bias the system in unnatural ways, leading to erroneous predictions of free energy barriers and associated mechanisms. A dramatic example of this is the autodissociation of liquid water following the classic reaction $2\mathrm{H_2O}(l) \longrightarrow \mathrm{H_3O^+}(aq) + \mathrm{OH^-}(aq)$, which ostensibly only requires transferring a proton from one water molecule to another. If this notion of the reaction is pursued, then a seemingly sensible reaction coordinate would simply be the distance between the oxygen and the transferring proton or the number of hydrogens covalently bonded to the oxygen. These reaction coordinates, as it turns out, are inadequate for describing the true nature of the reaction and, therefore, fail to yield reasonable free energies (and hence, values of the autoionization constant K_{w}). Chandler and coworkers showed that the dissociation reaction can only be considered to have occurred when the $\mathrm{H_3O^+}$ and $\mathrm{OH^-}$ ions are sufficiently far apart that no contiguous or direct path of hydrogen-bonding in the liquid can allow the proton to transfer back to the water or its origin. In order to describe such a process correctly, a very different type of reaction coordinate would clearly be needed.

Keeping in mind such caveats about the use of reaction coordinates, we now proceed to describe a number of popular methods designed to enhance sampling along pre-selected reaction coordinates. All of these methods are designed to generate, either directly or indirectly, the probability distribution function $P(q_1, ..., q_n)$ of a subset of n reaction coordinates of interest in a system. If these reaction coordinates are obtained from a transformation of the Cartesian coordinates $q_\alpha = f_\alpha(\mathbf{r}_1, ..., \mathbf{r}_N)$, $\alpha = 1, ..., n$, then the probability density that these n coordinates will have values $q_\alpha = s_\alpha$ in the canonical ensemble is

$$P(s_1, ..., s_n) = \frac{C_N}{Q(N,V,T)} \int d^N\mathbf{p}\, d^N\mathbf{r} e^{-\beta H(\mathbf{p},\mathbf{r})} \prod_{\alpha=1}^{n} \delta(f_\alpha(\mathbf{r}_1, ..., \mathbf{r}_N) - s_\alpha) \tag{29}$$

where the δ-functions are introduced to fix the reaction coordinates at values $q_1, ..., q_n$ at $s_1, ..., s_n$. Once $P(s_1, ..., s_n)$ is known, the free energy hypersurface in these coordinates is given by

$$A(s_1, ..., s_n) = -kT \ln P(s_1, ..., s_n) \tag{30}$$

VI. THE "BLUE MOON" ENSEMBLE APPROACH

The term "blue moon" in the present context describes rare events, *i.e.* events that happen once in a blue moon. The blue moon ensemble approach was introduced by Ciccotti and coworkers as a technique for computing the free energy profile along a reaction coordinate direction characterized by one or more barriers high enough that they would not likely be crossed in a normal thermostatted molecular dynamics calculation.

Suppose a process of interest can be monitored by a single reaction coordinate $q_1 = f_1(\mathbf{r}_1, ..., \mathbf{r}_N)$ so that eqns. (29) and (30) reduce to

$$\begin{aligned} P(s) &= \frac{C_N}{Q(N,V,T)} \int d^N\mathbf{p}\, d^N\mathbf{r} e^{-\beta H(\mathbf{p},\mathbf{r})} \delta(f_1(\mathbf{r}_1, ..., \mathbf{r}_N) - s) \\ &= \frac{1}{N!\lambda^{3N} Q(N,V,T)} \int d^N\mathbf{r} e^{-\beta U(\mathbf{r})} \delta(f_1(\mathbf{r}_1, ..., \mathbf{r}_N) - s) \\ A(s) &= -kT \ln P(s) \end{aligned} \tag{31}$$

The "1" subscript on the value s of q_1 is superfluous and will be dropped throughout this discussion. In the second line, the integration over the momenta has been performed giving the thermal prefactor factor λ^{3N}. In the blue moon ensemble approach, a holonomic constraint $\sigma(\mathbf{r}_1, ..., \mathbf{r}_N) = f_1(\mathbf{r}_1, ..., \mathbf{r}_N) - s$ is introduced in a molecular dynamics calculation as a means of "driving" the reaction coordinate from an initial value s_i to a final value s_f via a set of intermediate points $s_1, ..., s_n$ between s_i and s_f. Unfortunately, the introduction of a holonomic, constraint does not yield the single δ-function condition $\delta(\sigma(\mathbf{r}) = \delta(f_1(\mathbf{r}) - s)$, where $\mathbf{r} \equiv \mathbf{r}_1, ..., \mathbf{r}_N$ required by eqn. (31) but rather the product of δ-functions $\delta(\sigma(\mathbf{r}))\delta(\dot{\sigma}(\mathbf{r}, \mathbf{p}))$, since both the constraint and its first time derivative are imposed in a constrained dynamics calculation. We will return to this point a bit later in this section. In addition to this, the blue moon ensemble approach does not yield $A(s)$ directly but rather the derivative

$$\frac{dA}{ds} = -\frac{kT}{P(s)} \frac{dP}{ds} \tag{32}$$

from which the free energy profile $A(q)$ along the reaction coordinate and the free energy difference $DeltaA = A(s_f) - A(s_i)$ are given by the integrals

$$A(q) = A(s_i) + \int_{s_i}^{q} ds \frac{dA}{ds} \qquad \Delta A = \int_{s_i}^{s_f} ds \frac{dA}{ds} \tag{33}$$

In the free-energy profile expression $A(s_i)$ is just an additive constant that can be left off. The values $s_1, ..., s_n$ at which the reaction coordinate $q = f_1(\mathbf{r})$ is constrained can be chosen at equally-spaced intervals between s_i and s_f, in which a standard numerical quadrature can be applied for evaluating the integrals in eqn. (33), or they can be chosen according to a more sophisticated quadrature scheme.

We next turn to the evaluation of the derivative in eqn. (32). Noting that $P(s) = \langle \delta(f_1(\mathbf{r}) - s) \rangle$, the derivative can be written as

$$\frac{1}{P(s)} \frac{dP}{ds} = \frac{C_N}{Q(N,V,T)} \frac{\int d^N\mathbf{p}\, d^N\mathbf{r}\, e^{-\beta H(\mathbf{p},\mathbf{r})} \frac{\partial}{\partial s} \delta(f_1(\mathbf{r}) - s)}{\langle \delta(f_1(\mathbf{r}) - s) \rangle} \tag{34}$$

In order to avoid evaluating the derivative of the δ-function, an integration by parts can be used. First, we introduce a complete set of $3N$ generalized coordinates:

$$q_\alpha = f_\alpha(\mathbf{r}_1, ..., \mathbf{r}_N) \tag{35}$$

and their conjugate momenta p_α. Such a transformation has a unit Jacobian so that $d^N\mathbf{p}\, d^N\mathbf{r} = d^{3N}p\, d^{3N}q$. Denoting the transformed Hamiltonian as $\tilde{H}(p, q)$, eqn. (34) becomes

$$\frac{1}{P(s)} \frac{dP}{ds} = \frac{C_N}{Q(N,V,T)} \frac{\int d^{3N}p\, d^{3N}q\, e^{-\beta \tilde{H}(p,q)} \frac{\partial}{\partial s} \delta(q_1 - s)}{\langle \delta(q_1 - s) \rangle} \tag{36}$$

Changing the derivative in front of the δ-function from $\partial/\partial s$ to $\partial/\partial q_1$, which introduces an overall minus sign, and then integrating by parts yields

$$\begin{aligned} \frac{1}{P(s)} \frac{dP}{ds} &= \frac{C_N}{Q(N,V,T)} \frac{\int d^{3N}p\, d^{3N}q \left[\frac{\partial}{\partial q_1} e^{-\beta \tilde{H}(p,q)}\right] \delta(q_1 - s)}{\langle \delta(q_1 - s) \rangle} \\ &= -\frac{\beta C_N}{Q(N,V,T)} \frac{\int d^{3N}p\, d^{3N}q\, \frac{\partial \tilde{H}}{\partial q_1} e^{-\beta \tilde{H}(p,q)} \delta(q_1 - s)}{\langle \delta(q_1 - s) \rangle} \\ &= -\beta \frac{\left\langle \left(\frac{\partial \tilde{H}}{\partial q_1}\right) \delta(q_1 - s) \right\rangle}{\langle \delta(q_1 - s) \rangle} \end{aligned} \tag{37}$$

The last line defines a new ensemble average, specifically an average subject to the condition (not constraint) that the coordinate q_1 have the particular value s. This average will be denoted $\langle \cdots \rangle_s^{\rm cond}$. Thus, the derivative becomes

$$\frac{1}{P(s)} \frac{dP}{ds} = -\beta \left\langle \frac{\partial \tilde{H}}{\partial q_1} \right\rangle_s^{\rm cond} \tag{38}$$

Substituting eqn. (38) yields a free energy profile of the form

$$A(q) = A(s_i) + \int_{s_i}^{q} ds \left\langle \frac{\partial \tilde{H}}{\partial q_1} \right\rangle_s^{\rm cond} \tag{39}$$

from which ΔA can be computed by letting $q = s_f$. Given that -$\langle\partial\tilde{H}/\partial q_1\rangle_s^{\rm cond}$ is the expression for the average of the generalized force on q_1 when $q_1 = s$, the integral represents the work done *on* the system, i.e. the negative of the work done by the system, in moving from s_i to an arbitrary final point q. Since the conditional average implies a full simulation at each fixed value of q_1, the thermodynamic transformation is certainly carried out reversibly, so that eqn. (39) is consistent with the Clausius inequality.

Although eqn. (39) provides a very useful insight into the underlying statistical mechanical expression for the free energy, technically, the need for a full canonical transformaion of both coordinates and momenta is inconvenient since, from the chain rule

$$\frac{\partial\tilde{H}}{\partial q_1} = \sum_{i=1}^{N}\left[\frac{\partial H}{\partial \mathbf{p}_i}\cdot\frac{\partial \mathbf{p}_i}{\partial q_1} + \frac{\partial H}{\partial \mathbf{r}_i}\cdot\frac{\partial \mathbf{r}_i}{\partial q_1}\right] \tag{40}$$

A more useful expression results if the momenta integrations are performed before introducing the transformation to generalized coordinates. Starting again with eqn. (34), we carry out the momentum integrations, yielding

$$\frac{1}{P(s)}\frac{dP}{ds} = \frac{1}{N!\lambda^{3N}Q(N,V,T)}\frac{\int\ d^N\mathbf{r}\ e^{-\beta U(\mathbf{r})}\frac{\partial}{\partial s}\delta(f_1(\mathbf{r})-s)}{\langle\delta(f_1(\mathbf{r})-s)\rangle} \tag{41}$$

Now, we introduce *only* the transformation of the coordinates to generalized coordinates $q_\alpha = f_\alpha(\mathbf{r}_1,...,\mathbf{r}_N)$. However, because there is no corresponding momentum transformation, the Jacobian of the transformation is not unity. Let $J(q) \equiv J(q_1,...,q_{3N}) = \partial(\mathbf{r}_1,...,\mathbf{r}_N)/\partial(q_1,...,q_{3N})$ denote the Jacobian of the transformation. Then, eqn. (41) becomes

$$\begin{aligned}\frac{1}{P(s)}\frac{dP}{ds} &= \frac{1}{N!\lambda^{3N}Q(N,V,T)}\frac{\int\ d^{3N}q\ J(q)e^{-\beta\tilde{U}(q)}\frac{\partial}{\partial s}\delta(q_1-s)}{\langle\delta(q_1-s)\rangle}\\ &= \frac{1}{N!\lambda^{3N}Q(N,V,T)}\frac{\int\ d^{3N}q\ e^{-\beta\left(\tilde{U}(q)-kT\ln J(q)\right)}\frac{\partial}{\partial s}\delta(q_1-s)}{\langle\delta(q_1-s)\rangle}\end{aligned} \tag{42}$$

where, in the last line, the Jacobian has been exponentiated. Changing the derivative $\partial/\partial s$ to $\partial/\partial q_1$ and performing the integration by parts as was done in eqn. (37), we obtain

$$\begin{aligned}\frac{1}{P(s)}\frac{dP}{ds} &= \frac{1}{N!\lambda^{3N}Q(N,V,T)}\frac{\int\ d^{3N}q\ \frac{\partial}{\partial q_1}e^{-\beta\left(\tilde{U}(q)-kT\ln J(q)\right)}\delta(q_1-s)}{\langle\delta(q_1-s)\rangle}\\ &= -\frac{\beta}{N!\lambda^{3N}Q(N,V,T)}\frac{\int\ d^{3N}q\ \left[\frac{\partial\tilde{U}}{\partial q_1}-kT\frac{\partial}{\partial q_1}\ln J(q)\right]e^{-\beta\left(\tilde{U}(q)-kT\ln J(q)\right)}\delta(q_1-s)}{\langle\delta(q_1-s)\rangle}\\ &= -\beta\left\langle\left[\frac{\partial\tilde{U}}{\partial q_1}-kT\frac{\partial}{\partial q_1}\ln J(q)\right]\right\rangle_s^{\rm cond}\end{aligned} \tag{43}$$

Therefore, the free energy profile becomes

$$A(q) = A(s_i) + \int_{s_i}^{q} ds\ \left\langle\left[\frac{\partial\tilde{U}}{\partial q_1}-kT\frac{\partial}{\partial q_1}\ln J(q)\right]\right\rangle_s^{\rm cond} \tag{44}$$

Again, the derivative of $\tilde{U}$, the transformed potential, can be computed form the untransformed potential via the chain rule

$$\frac{\partial\tilde{U}}{\partial q_1} = \sum_{i=1}^{N}\frac{\partial U}{\partial\mathbf{r}_i}\cdot\frac{\partial\mathbf{r}_i}{\partial q_1} \tag{45}$$

Eqn. (44) is useful for simple reaction coordinates in which the full transformation to generalized coordinates is known. We will see shortly how the expression for $A(q)$ can be further simplified in a way that does not require knowledge of the transformation at all. First, however, we must tackle the problem alluded to earlier of computing the conditional ensemble averages from the constrained dynamics employed by the blue moon ensemble method.