

The Thermodynamic Basis for Binding Affinity Approximation in OSPREY

Henry Childs
Donald Lab

Fall 2024

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 1 |
| 2 | Deriving the Equilibrium Binding Constant | 1 |
| 2.1 | Derivation from Chemical Potential | 1 |
| 2.2 | Derivation from Number of Particles | 2 |
| 2.3 | K is Dimensionless | 2 |
| 3 | Types of Equilibrium | 2 |
| 3.1 | Thermal Equilibrium | 3 |
| 3.1.1 | Entropy | 3 |
| 3.1.2 | Temperature | 3 |
| 3.2 | Mechanical Equilibrium | 3 |
| 3.3 | Diffusive Equilibrium | 4 |
| 4 | Derivation of the Central Thermodynamic Identity | 5 |
| 4.1 | Fixed N | 5 |
| 4.2 | Variable N | 6 |
| 5 | Helmholtz Free Energy (F) | 6 |
| 5.1 | Creating a System | 6 |
| 5.2 | Measuring Small Changes to a System | 7 |
| 6 | Boltzmann Statistics | 7 |
| 6.1 | The Fundamental Assumption of Statistical Mechanics | 7 |
| 6.2 | The Boltzmann Factor | 8 |
| 6.3 | The Partition Function (Z) | 9 |
| 6.4 | Composite Systems with Indistinguishable Particles | 9 |
| 6.5 | Using Partition Functions with Helmholtz | 10 |
| 6.6 | Using Partition Functions with Chemical Potential | 11 |
| 7 | Deriving K^* | 11 |
| 7.1 | Derivation | 11 |
| 7.2 | Components of Approximation | 12 |

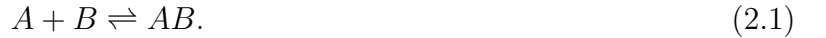
1 Introduction

The rapid and accurate prediction of non-covalent interaction between an *in silico* design and a biological target is a task of central importance in computational protein design. Expanding on the Support Information of (Lilien et al., 2005), this document serves to both demonstrate the logic and derive the biophysical basis for OSPREY’s K^* algorithm (reviewed in Donald, 2011; Gainza et al., 2012; Georgiev et al., 2008; Lilien et al., 2005). Given total partition functions, K^* can be shown to exactly calculate the binding constant of a biophysical model. The thermodynamic concepts reviewed in this article are mainly derived from (Reif, 1965 and Schroeder, 2021).

2 Deriving the Equilibrium Binding Constant

2.1 Derivation from Chemical Potential

We will now consider the derivation of K (the equilibrium binding constant) from the chemical potential, as reviewed in (Gilson et al., 1997). Let us consider the non-covalent interaction of two proteins, A and B , to form the protein:protein complex AB in an aqueous solution:



For the reaction defined above, the equilibrium condition for chemical potential (μ , discussed later in Section 3.3) is:

$$\mu_A + \mu_B = \mu_{AB}. \quad (2.2)$$

For now, we will define the chemical potential for species i in solution as:

$$\mu_i = \mu_i^\circ + RT \ln \frac{\gamma_i C_i}{C^\circ}. \quad (2.3)$$

Where R and T are the gas constant and absolute temperature, respectively. The standard chemical potential, μ_i° , is the chemical potential of species i in the standard state. We represent the concentration of species i as C_i , with the standard concentration C° (typically 1 M) in the same units. For brevity, some texts may exclude these concentrations in favor of mole ratio. To account for deviations for ideal gas behavior, the activity coefficient γ_i is included. An ideal gas will have a γ_i of unity, while deviations of ideal behavior for species i are accounted for with $\gamma_i < 1$ or $\gamma_i > 1$.

Consider a standard state where A , B , and AB are at standard concentration in solvent, but do not interact with other A , B , or AB molecules. For this standard state at equilibrium, we can calculate the standard free energy (discussed later in Section 5) for the forward reaction (Equation (2.1)) using Equations (2.2) and (2.3):

$$\begin{aligned} \Delta G_{AB}^\circ &\equiv \mu_{AB}^\circ - \mu_A^\circ - \mu_B^\circ \\ &= \left(\mu_{AB} - RT \ln \frac{\gamma_{AB} C_{AB}}{C^\circ} \right) - \left(\mu_A - RT \ln \frac{\gamma_A C_A}{C^\circ} \right) - \left(\mu_B - RT \ln \frac{\gamma_B C_B}{C^\circ} \right) \\ &= -RT \ln \left(\frac{\gamma_{AB} C_{AB} C^\circ}{\gamma_A \gamma_B C_A C_B} \right) = -RT \ln \frac{C_{AB} C^\circ}{C_A C_B} = -RT \ln K_{AB} \end{aligned} \quad (2.4)$$

where K_{AB} is the binding constant. In this standard state, γ_i approaches 1 as C_i approaches zero. If we assume a low concentration state, $\gamma_i = 1$ is a sufficient approximation.

2.2 Derivation from Number of Particles

For systems at constant volume, K can be calculated directly from the number of particles per species (N). Consider the relation between the number of particles (or molecules), number of moles (n), and Avogadro's number (A_N):

$$N = nA_N. \quad (2.5)$$

From Equations (2.4) and (2.5):

$$\begin{aligned} K_{AB} &= \frac{C_{AB}C^\circ}{C_A C_B} \\ &= \frac{\frac{n_{AB}}{V} \frac{n^\circ}{V}}{\frac{n_A}{V} \frac{n_B}{V}} \\ &= \frac{n_{AB}n^\circ}{n_A n_B} * \frac{(A_N)^2}{(A_N)^2} \\ &= \frac{N_{AB}N^\circ}{N_A N_B} \\ &= \frac{N_{AB}}{N_A N_B}. \end{aligned} \quad (2.6)$$

In standard aqueous solutions, N° is 1 (because C° is 1 M). This property of constant volume systems generalizes to reaction with different ratios of reactants to products.

2.3 K is Dimensionless

Despite traditional notation, K is a dimensionless value. The standard concentration C° is not always explicitly written, and is often implied. For the reaction defined by equation (2.1), the units are M^2 divided by M^2 . For reactions with different ratios of reactants to products, K remains dimensionless. For example, in a reaction with two products



the standard concentration is eliminated in the numerator and denominator:

$$\begin{aligned} \Delta G_{C,D}^\circ &\equiv \mu_C^\circ + \mu_D^\circ - \mu_A^\circ - \mu_B^\circ \\ &= -RT \ln \left(\frac{\gamma_C \gamma_D C_C C_D}{\gamma_A \gamma_B C_A C_B} \right) = -RT \ln \frac{C_C C_D}{C_A C_B} = -RT \ln K_{C,D}. \end{aligned} \quad (2.8)$$

Once again, K is dimensionless.

3 Types of Equilibrium

We stand to gain great insight into chemical systems by studying how two interacting systems change when certain parameters are held constant while other are permitted to evolve. In this section, we will investigate three type of equilibrium from the perspective of energy, temperature, volume, and number of particles. By defining systems that hold some of these conditions constant, we can establish special cases of equilibrium that are helpful for studying chemical properties.

The following table serves as a guide for different types of equilibrium:

| Type of Eq. | Exchanged |
|-------------|-----------|
| Thermal | Energy |
| Mechanical | Volume |
| Diffusive | Particles |

Table 1: Different Types of Equilibrium. Reproduced from (Schroeder, 2021).

3.1 Thermal Equilibrium

Two objects are in thermal equilibrium if they are at the same temperature and their entropy is at its maximum value. Because defining thermal equilibrium requires an understanding of both entropy and temperature, herein we briefly review entropy (from the perspective of multiplicity and the second law) and derive the formula for temperature.

3.1.1 Entropy

The number of microstates for a given macrostate is termed multiplicity (Ω). The second law of thermodynamics states that at a fixed temperature, the entropy (S) is maximized. In other words, a large system in thermal equilibrium will be found in the macrostate with the greatest multiplicity (i.e., greatest entropy). We can then intuitively define entropy as:

$$S = k \ln \Omega. \quad (3.1)$$

We multiply by the Boltzmann factor, k (J/K), so S has units of energy. Simply put, entropy is a measure of the number of configurations of a system. We can describe thermal equilibrium, in part, as the state of a system that maximizes S .

3.1.2 Temperature

For objects A and B at thermal equilibrium, entropy has reached its maximum value. Therefore, we can represent the change in entropy relative to the object's internal energy, U_A , as:

$$\frac{\partial S_A}{\partial U_A} = 0. \quad (3.2)$$

Note that this relation has units of inverse Kelvin. We can therefore write the equation for temperature (holding volume and number of particles constant) as:

$$\frac{1}{T} = \left(\frac{\partial S}{\partial U} \right)_{N,V}. \quad (3.3)$$

Therefore, a fundamental way to define thermal equilibrium is that if two objects are at equilibrium, then the following relation is true:

$$\frac{\partial S_A}{\partial U_A} = \frac{\partial S_B}{\partial U_B}. \quad (3.4)$$

3.2 Mechanical Equilibrium

Consider system A and system B separated by a barrier with each able to exchange energy and volume with one another (in other words, they are separated by a non-insulated, moving barrier). S_{total} (the

total entropy) is therefore a function of two values: U_A (internal energy of A) and V_A (volume of A), as visualized in Figure 1 .

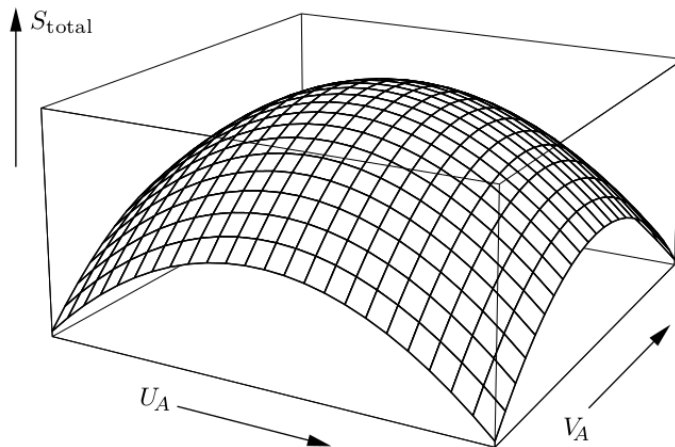


Figure 1: Graph of entropy (S) vs internal energy (U) and volume (V). This figure was borrowed directly from Chapter 3 of (Schroeder, 2021).

At equilibrium, S_{total} is maximized and the partial derivatives go to 0 in both directions (visualized in Figure 1). Because we've already reviewed changes in entropy relative to internal energy (Equation (3.3)), let's review the following equilibrium condition:

$$\frac{\partial S_{total}}{\partial V_A} = 0. \quad (3.5)$$

Because any volume lost in one system is gained by the other, $dV_A = -dV_B$. We can then rewrite Equation (3.5) in terms of each system's entropy. Therefore, a fundamental way to define mechanical equilibrium is that if two systems are at equilibrium, the following relation is true:

$$\frac{\partial S_A}{\partial V_A} = \frac{\partial S_B}{\partial V_B}. \quad (3.6)$$

Entropy has units J/K and volume has units m^3 , so the relation in Equation (3.6) has units ($N/m^2/K$). Because we've assumed a system in thermal equilibrium (energy is allowed to be exchanged), we can multiply the relation show in Equation (3.6) by temperature to obtain an expression for pressure:

$$P = T \left(\frac{\partial S}{\partial V} \right)_{U,N}. \quad (3.7)$$

3.3 Diffusive Equilibrium

Now consider two systems, both comprised of the same species (e.g., water) that are free to exchange both energy and particles (in other words, they are separated by a fixed, porous barrier). Assuming the total number of particles and total energy are fixed, we can express entropy as a function of these constants. We've already reviewed changes in S with respect to total energy (see Equation (3.3)), so let's investigate the relation:

$$\left(\frac{\partial S_{total}}{\partial N_A}\right)_{U_A, V_A} = 0. \quad (3.8)$$

Following the same logic as Equation (3.6), we can express this condition for each system. Because the systems are in thermal equilibrium, we can multiply through by $-T$ (the negative sign is convention). Therefore, a fundamental way to describe diffusive equilibrium is by the following relation:

$$-T \frac{\partial S_A}{\partial N_A} = -T \frac{\partial S_B}{\partial N_B}. \quad (3.9)$$

The terms in this relation are defined to be chemical potential (μ). Therefore an alternative definition of μ (see Equation (2.3)) for small changes to a system is

$$\mu \equiv -T \left(\frac{\partial S}{\partial N}\right)_{U, V}. \quad (3.10)$$

This quantity is conserved in diffusive equilibrium (refer back to Equation (2.2)).

4 Derivation of the Central Thermodynamic Identity

Although we derived thermodynamic properties for systems in specific equilibrium (for example, pressure from a system in mechanical equilibrium), we can apply these properties to diverse system conditions. Let's now unify this information into a single expression so we can quickly derive expressions for select properties at different system conditions.

4.1 Fixed N

Consider a process where a system undergoes change in both energy and volume, but at a constant number of particles (i.e., mechanical equilibrium). We would like to measure the total change in entropy. For simplicity, let's consider this process as two different steps:

- Process 1 (ΔS_1): ΔV and constant U
- Process 2 (ΔS_2): ΔU and constant V

We can then write the entropy of the system as the sum of these two process (recall that entropy is merely a description of system states; see Section 3.1.1).

$$\begin{aligned} \Delta S &= \Delta S_1 + \Delta S_2 \\ &= \left(\frac{\Delta S}{\Delta U}\right)_V \Delta U + \left(\frac{\Delta S}{\Delta V}\right)_U \Delta V. \end{aligned} \quad (4.1)$$

If these changes are small, we can represent this change as partial derivatives and evaluate the expressions using Equations (3.3) and (3.7):

$$\begin{aligned} dS &= \left(\frac{\partial S}{\partial U}\right)_V dU + \left(\frac{\partial S}{\partial V}\right)_U dV \\ &= \frac{1}{T} dU + \frac{P}{T} dV. \end{aligned} \quad (4.2)$$

Rewriting this in terms of energy, we obtain the thermodynamic identity for an equilibrium system with a fixed number of particles:

$$dU = TdS - PdV. \quad (4.3)$$

4.2 Variable N

Let us now extend this identity to processes where the number of particles changes. We need only extend the logic of the previous section; we can describe the system changes as three separate processes:

- Process 1 (ΔS_1): ΔV and constant U, N
- Process 2 (ΔS_2): ΔU and constant V, N
- Process 3 (ΔS_3): ΔN and constant U, V

Similar to Equation (4.2), the total entropy can be expressed using Equation (3.10) to evaluate Process 3:

$$\begin{aligned} dS &= \left(\frac{\partial S}{\partial U}\right)_{V,N} dU + \left(\frac{\partial S}{\partial V}\right)_{U,N} dV + \left(\frac{\partial S}{\partial N}\right)_{U,V} dN \\ &= \frac{1}{T}dU + \frac{P}{T}dV - \frac{\mu}{T}dN. \end{aligned} \quad (4.4)$$

We can therefore measure the change in entropy for a system with variable N as:

$$dS = \frac{1}{T}(dU + PdV - \mu dN). \quad (4.5)$$

Rewriting in terms of energy, we obtain the thermodynamic identity for an equilibrium system with a variable number of particles:

$$dU = TdS - PdV + \mu dN. \quad (4.6)$$

This is a very useful formula in statistical thermodynamics. Given system constraints (e.g., fixed U and N) we can easily rewrite Equation (4.6) to calculate desired properties.

5 Helmholtz Free Energy (F)

5.1 Creating a System

If we were to wave a magic wand and conjure a protein from nothing, the minimum energy cost of this fictional process would be internal energy of the protein (U). However, what if this phenomena occurred in an environment of constant temperature? If so, we could reduce our energy cost from the energy that flows spontaneously from the environment into our protein ($-TS$). This energy is called Helmholtz free energy (F):

$$F = U - TS. \quad (5.1)$$

For the rest of this article, we will be discussing energy in terms of Helmholtz free energy (F), not the traditional Gibbs (G) which also accounts for the atmospheric energy input required to make

space for our protein (+ PV). However, because computational ensembles are generally modeled without consideration for volume or pressure, these energies are equivalent. Based on the equilibrium conditions for K^* (described in Section 7), the following relation is established:

$$\begin{aligned} G &= U - TS + PV \\ G &= F + PV \\ dG &= dF + PdV + VdP \\ dG &= dF. \end{aligned} \tag{5.2}$$

5.2 Measuring Small Changes to a System

The definition of Helmholtz free energy can be extended to account for small changes to a system (not just creation). From Equation (5.1), we can use the chain rule to state:

$$dF = dU - TdS - SdT. \tag{5.3}$$

Plugging in our handy thermodynamic identity (Equation (4.6)) for dU , we can obtain the thermodynamic identity for F :

$$dF = SdT - PdV + \mu dN. \tag{5.4}$$

By holding certain system parameters constant (see Section 3) in Equation (5.4), we can compute desired thermodynamic properties for diverse chemical systems.

6 Boltzmann Statistics

Using our knowledge from previous sections on evaluating equilibrium systems, let us now review how to use a very powerful statistical method for determining the probability of finding a system in a given state.

6.1 The Fundamental Assumption of Statistical Mechanics

In Section 3.1.1 we noted that there are many microstates for a corresponding macrostate, and that a system in thermal equilibrium will be found in the system macrostate with the greatest number of microstates (maximizing entropy). Motivated by the ergodic hypothesis, the *equal a priori probability postulate* (commonly referred to as the fundamental assumption) states:

For an isolated system in thermal equilibrium, all accessible microstates are equally probable.

Note that this is an assumption- and therefore may be false. Although we would expect all microstates to be occupied given sufficient time, in reality the number of microstates for a given macrostate may be so large that only a small subset will be occupied in a reasonable time frame. Instead, what we assume is that the microstates that do occur are a representative sample. Assuming all microstates are equally probable allows us to conclude that select macrostates are more probable than others.

Given an *ensemble* (the collection of all macrostates, each with corresponding microstates, for a system) constructed under fixed thermodynamic properties, we can determine the probability distribution for a system. The evaluation of this distribution is very valuable, and can be done efficiently using Boltzmann statistics.

6.2 The Boltzmann Factor

Consider the following system, containing one particle, in contact with a large reservoir containing many particles:

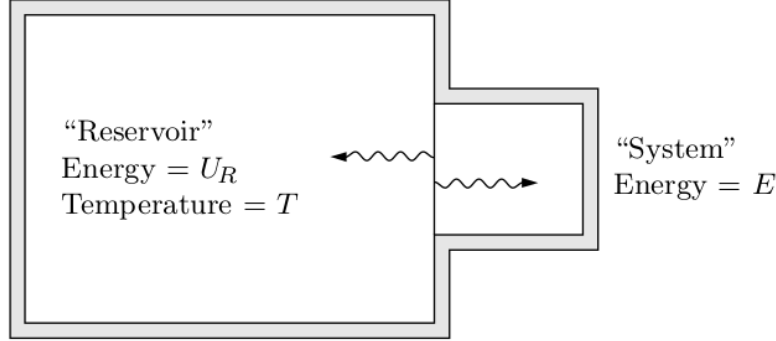


Figure 2: A system containing one particle in contact with a larger reservoir with many particles at a fixed temperature. This figure was borrowed directly from Chapter 6 of (Schroeder, 2021)

For the system illustrated in Figure 2, let us simplify the task of finding the probability of a microstate into investigating the ratio of probabilities for two microstates. We will term these states s_1 and s_2 , where $E(s_2) > E(s_1)$. Although the single-particle system and reservoir alone are not isolated, if we consider the reservoir and system together as an isolated system, we can apply the fundamental assumption (Section 6.1) to state that we are equally likely to find this isolated system in any of its accessible microstates.

Let us define $\Omega_R(s)$ to be the multiplicity of the reservoir when the single particle is in state s . Because s_1 has a lower energy than s_2 , $\Omega_R(s_1) > \Omega_R(s_2)$. Intuitively, we can understand this relationship from the perspective of our fixed total energy: if our single particle is at a lower energy, then more energy is available to the reservoir. Therefore, the probability of finding the single particle in a given microstate is proportional to the number of microstates in the reservoir:

$$\frac{\mathcal{P}(s_2)}{\mathcal{P}(s_1)} = \frac{\Omega_R(s_2)}{\Omega_R(s_1)}. \quad (6.1)$$

Rewriting using the formula for entropy (Equation (3.1)):

$$\frac{\mathcal{P}(s_2)}{\mathcal{P}(s_1)} = e^{(S_R(s_2) - S_R(s_1))/k}. \quad (6.2)$$

For this single particle, $dN = 0$ and dV is negligible compared to dU . Therefore, we can rewrite Equation (4.5) for particle states s_1 and s_2 :

$$S_R(s_2) - S_R(s_1) = \frac{1}{T}(U_R(s_2) - U_R(s_1)) = -\frac{1}{T}(E(s_2) - E(s_1)) \quad (6.3)$$

where $E(s)$ is the energy of the particle in state s . Note that any energy lost by the reservoir is gained by the particle (and vice versa), so we must multiply the last term in Equation (6.3) by -1 to maintain the correct relation between the internal energy of the reservoir and the energy of the particle. Plugging Equation (6.3) into Equation (6.2), we can establish the relationship between the ratio of probabilities and the ratio of exponential factors:

$$\frac{\mathcal{P}(s_2)}{\mathcal{P}(s_1)} = \frac{e^{-E(s_2)/kT}}{e^{-E(s_1)/kT}}. \quad (6.4)$$

Each of the exponential factors in Equation (6.4) are called the Boltzmann factor. These factors are a function of energy and temperature, and correlate to the ratio of probabilities. However, it is not true to state that the probability of a given microstate is equal to the corresponding Boltzmann factor. Establishing the relationship between the probability of a single microstate and the energy and temperature of that state requires the partition function.

6.3 The Partition Function (Z)

Let us rewrite Equation (6.4):

$$\frac{\mathcal{P}(s_2)}{e^{-E(s_2)/kT}} = \frac{\mathcal{P}(s_1)}{e^{-E(s_1)/kT}}. \quad (6.5)$$

We have isolated terms corresponding to s_2 to the left side of the equation, and terms corresponding to s_1 to the right side. If both sides are equal, then both must be equal to a constant of proportionality that is true for all states in the system. This constant is $1/Z$, where Z is called the partition function. If we insert Z into a Boltzmann factor, we convert it to a probability:

$$\mathcal{P}(s) = \frac{1}{Z} e^{-E(s)/kT}. \quad (6.6)$$

Some texts may refer to Equation (6.6) as the Boltzmann or canonical distribution. But what if we want to calculate Z ? We simply recall that the sum of probabilities for all states is unity:

$$1 = \sum_s \mathcal{P}(s). \quad (6.7)$$

Plugging Equation (6.6) into Equation (6.7) and solving for the partition function:

$$Z = \sum_s e^{-E(s)/kT} \quad (6.8)$$

we find the sum of all Boltzmann factors. In fact, the traditional term for the partition function, Z , is German for "*Zustandsumme*", which means "sum over states."

For brevity, we will adopt the convention where the fraction $1/kT$ is represented as β . We will herein use the partition function in the form

$$Z = \sum_s e^{-\beta E(s)}. \quad (6.9)$$

6.4 Composite Systems with Indistinguishable Particles

A system is *distinguishable* if particles can be identified by position or intrinsic property (e.g., fixed atoms in a crystal). Further, we often assume a system contains non or weakly-interacting particles. This allows us to represent the total energy of system as the sum of state energies. For example, for a system of two non-interacting particles, the energy is calculated:

$$E_{total} = E_1 + E_2. \quad (6.10)$$

Extending this logic, we can represent the total Z (Equation (6.9)) of a distinguishable two-particle system as:

$$\begin{aligned} Z_{total} &= \sum_s e^{-\beta(E_1(s)+E_2(s))} \\ &= Z_1 Z_2 \end{aligned} \tag{6.11}$$

because the set of all possible states is the set of all pairs of (s_1, s_2) . However, what if we are unable to distinguish particles? A system is *indistinguishable* if interchanging the position of two particles leaves the system in the same state. Equation (6.11) would double-count states, so a better representation of the total partition function for a two-particle system is:

$$Z_{total} = \frac{1}{2} Z_1 Z_2. \tag{6.12}$$

Generalizing Equation (6.12) to a low-density system, the total partition function for a system of indistinguishable particles is

$$Z_{total} = \frac{Z_1^N}{N!} \tag{6.13}$$

where Z_1 is the partition function for any one particle in the system.

6.5 Using Partition Functions with Helmholtz

We previously reviewed entropy and Helmholtz energy individually (Sections 3.1.1 and 5, respectively), but combining these concepts can be very useful.

For a system in equilibrium with a reservoir (see conditions defined in Section 6.2), a value analogous to multiplicity (Ω , see Equation (3.1)) is the partition function (Z , see Equation (6.9)). Ω is equal to the number of microstates at constant energy, while Z is equal to the weighted number of microstates at constant temperature. Intuitively, we can express Helmholtz energy (F , see Equation (5.1)) in terms of the partition function:

$$F = -kT \ln Z. \tag{6.14}$$

Note that we've included a negative sign in Equation (6.14) because at thermal equilibrium entropy is maximized (refer to Section 3), so F will decrease (see Equation (5.1)). Therefore, we expect $-F$ to increase. See Chapter 6.5 of (Schroeder, 2021) for a full derivation.

For a system of indistinguishable, non-interacting particles (Equation (6.13)), we can derive the expression for energy in terms of N and Z (Equation (6.14)) using Stirling's approximation:

$$\begin{aligned} F &= -kT \ln \left(\frac{Z_1^N}{N!} \right) \\ &= -kT [N \ln(Z_1) - \ln(N!)] \\ &= -kT [N \ln(Z_1) - N \ln N + N] \\ &= -NkT \ln \left(\frac{Z_1}{N} \right) + 1. \end{aligned} \tag{6.15}$$

Therefore, we can express the free energy of this system given the number of particles and the partition function for a single particle.

6.6 Using Partition Functions with Chemical Potential

Rewriting Equation (5.4) for a system at constant volume and temperature, we can calculate chemical potential as the partial derivative of F with respect to the number of particles:

$$\begin{aligned} dF &= SdT - PdV + \mu dN \\ \mu &= \left(\frac{\partial F}{\partial N} \right)_{T,V}. \end{aligned} \tag{6.16}$$

We can therefore take the partial derivative of Equation (6.15) to derive an equation for the chemical potential for a system of non-interacting, indistinguishable particles:

$$\begin{aligned} \left(\frac{\partial F}{\partial N} \right)_{T,V} &= -NkT \ln(Z_1) + NkT \ln(N) - NkT \\ &= -kT \ln(Z_1) + kT \ln(N) + \frac{NkT}{N} - kT \\ \mu &= -kT \ln \left(\frac{Z_1}{N} \right). \end{aligned} \tag{6.17}$$

Generalizing μ for an arbitrary species i :

$$\mu_i = -kT \ln \left(\frac{Z_i}{N_i} \right). \tag{6.18}$$

We can now calculate the chemical potential for a system of non-interacting, indistinguishable particles (compare to the general definition given in Equation (2.3)).

7 Deriving K^*

Now that we have established sufficient background on representing chemical systems using statistical thermodynamics, let us derive the provable approximation to K_a , the K^* score.

7.1 Derivation

Plugging Equation (6.18) into Equation (2.2), we can write the equilibrium condition for the non-covalent interaction of two proteins (refer to the reaction defined in Equation (2.1)) as:

$$-kT \ln \left(\frac{Z_A}{N_A} \right) - kT \ln \left(\frac{Z_B}{N_B} \right) = -kT \ln \left(\frac{Z_{AB}}{N_{AB}} \right). \tag{7.1}$$

Eliminating the constants and log establishes the relationship:

$$\frac{Z_{AB}}{Z_A Z_B} = \frac{N_{AB}}{N_A N_B}. \tag{7.2}$$

As demonstrated in Equation (2.6), we can rewrite the ratio of particles for constant-volume systems as the ratio of concentrations for each species. Therefore, the theoretical basis for calculating K from partition functions is:

$$K_a = \frac{[AB]}{[A][B]} = \frac{Z_{AB}}{Z_A Z_B}. \tag{7.3}$$

If we define the K^* score as the same ratio of partition functions, then, given exact partition functions:

$$K^* = K_a. \quad (7.4)$$

7.2 Components of Approximation

In practice, computing exact partition functions is not possible, and a number of simplifications must be made to our chemical system model to efficiently search. K^* systematically evaluates a sub-ensemble that approximates the true partition function. For an arbitrary state i , we define Q_i as the approximation to Z_i :

$$Q_i = \sum_s \exp\left(\frac{-E(s)}{RT}\right), \quad Q_i \approx Z_i. \quad (7.5)$$

Therefore, as K^* approximates the true partition function, it approximates the binding constant:

$$K^* = \frac{Q_{AB}}{Q_A Q_B} \approx K_a. \quad (7.6)$$

For a complete review of these simplifications and their respective assumptions, refer to the following seminar: <https://www.youtube.com/watch?v=MUU1d4NhOo>.

References

- Donald, B. R. (2011). *Algorithms in structural molecular biology*. MIT Press.
- Gainza, P., Roberts, K. E., & Donald, B. R. (2012). Protein design using continuous rotamers. *PLoS Computational Biology*, 8. <https://doi.org/10.1371/journal.pcbi.1002335>
- Georgiev, I., Lilien, R. H., & Donald, B. R. (2008). The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *Journal of Computational Chemistry*, 29, 1527–1542. <https://doi.org/10.1002/jcc.20909>
- Gilson, M. K., Given, J. A., Bush, B. L., & McCammon, J. A. (1997). The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophysical Journal*, 1047–1069. [https://doi.org/10.1016/S0006-3495\(97\)78756-3](https://doi.org/10.1016/S0006-3495(97)78756-3)
- Lilien, R. H., Stevens, B. W., Anderson, A. C., & Donald, B. R. (2005). A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. *Journal of Computational Biology*, 12, 740–761. <https://doi.org/10.1089/cmb.2005.12.740>
- Reif, F. (1965). *Fundamental of statistical and thermal physics*. McGraw-Hill.
- Schroeder, D. V. (2021). *An introduction to thermal physics*. Oxford University Press.