A Statistical Thermodynamic Approach to Understanding Protein Stability in a Multiple Cosolvent-Solvent System

Abstract

The human body contains various chemical environments. Each of these chemical environments will contain solvents, cosolvents and proteins. These cosolvents affect the stability of a protein in different ways, but they can be generally categorised into stabilisers and denaturants. Biopharmaceutical drugs, many of which are proteins, need to be kept stable in the body if they are to be of medical use. The aims of this project are twofold: (i) prediction of a chemical's effects on protein thermal stability, and (ii) the development of a routine in R to carry out predictions based on thermal denaturation data. This will be achieved using statistical thermodynamics. The effects of single cosolvent interactions have previously been modelled using the Kirkwood-Buff theory. In this report, the theory is extended to multiple cosolvents, and a tool has been developed to automate the predictive modelling and graphical representation of our additive formula using previously published experimental thermodynamic data involving multiple cosolvents. Our predictive model is used to elucidate the origin of cosolvent additivity effects

Introduction

Biopharmaceutical drugs (biologics) are a class of drugs that are derived from natural sources whose pharmaceutical component is too complex to be produced *en masse* by organic synthesis (Revers and Furczon, 2010; Pham et al., 2019). These can be compiled into 5 different classes that are primarily protein-based, and can have a broad range of uses in genomics and medicine (Lacaná et al., 2007). Biologics are drugs and, as such, they have specific targets in the treatment of myriad afflictions such as anaemia, cystic fibrosis and cancers (Morrow and Felcone, 2004). Biologics also have protein components. One property of a protein is such that it may find itself in many different conformations that may either be useful and functional, or defunct. For the sake of simplicity, we refer to the functional state as the native state, and all others, regardless of their functionality, denatured. For a biologic to function, the native state must be stabilised.

The body comprises many different chemical environments. The stomach's acidic chemical contents differ wildly from those found in tumorigenic tissue, whose components vary with respect to skeletal muscle tissue (Hollander, 1949; Jorgenson, Zhong and Oberley, 2013). Each of these environments contain solvents, solutes and cosolvents that interact with each other dynamically in solution. For example, in the case of the stomach, the chemical environment predominantly comprises water (solvent), pepsin (protein solute) and hydrochloric acid (cosolvent). In order to derive new biologics from currently available ones, they must be proven to be stable in their target's environment. The stability of proteins is an important theme explored throughout this project.

Proteins and their chemical stability provide an effective segue into the purpose of this project: to predict the stability of any given protein in any given chemical environment comprising a solvent and multiple cosolvents. A cosolvent is defined for our purposes as a chemical substance added to a solvent to affect the stability of a protein's native state. Therefore, we can categorise relevant cosolvents as either stabilisers or denaturants of a protein (fig.1). There are thermodynamic equations that can be used to

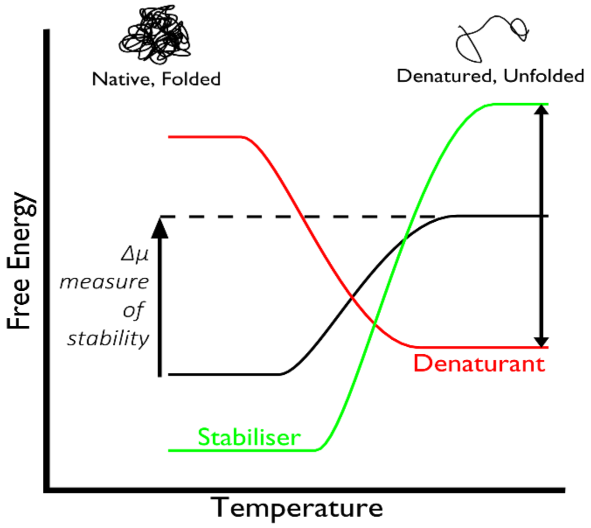

Figure 1 – schematic shows how addition of cosolvents affects the free energy of the native and denatured states of a protein. The difference between these two energy levels (Δμ) can be used as a direct measure of a protein's ability to stay in the native state.

model a cosolvent's effect on a protein's stability – the Wyman-Tanford formula proves useful in many cases (Wyman, 1964; Lee and Timasheff, 1981). However, this formula has several issues which arise from the fact that the Wyman-Tanford formula is purely thermodynamic in nature and cannot describe interactions on a molecular level (Timasheff, 1998). The binding sites assumed by the Wyman-Tanford model cannot account for the exclusion of stabilisers from a protein's surface (Shimizu, 2020). The most prominent and relevant shortcoming is that it will only model systems involving a solute, solvent and one cosolvent (binary systems), as opposed to those with multiple cosolvents (ternary systems) found in the body.

To overcome this, a statistical thermodynamic theory elucidates what effects one might expect multiple cosolvents to have on protein stability, in both solvent in dilution and in larger concentrations (Shimizu, Stenner and Matubayasi, 2017). Thermal denaturation calorimetry can be used to forcefully transition a protein between the native and denatured states. The temperature at which this transition occurs is a useful, indirect measure of a protein's stability.

Since thermal denaturation calorimetry experiments are relatively common (Chiu and Prenner, 2011; Ranjbar and Gill, 2009), there is an abundance of available data that can be used. However, there currently are no dedicated tools for processing this data in a way that is useful for this study. As such, this report will detail the development of a tool that meets two criteria: (i) it must model the additivity of binary systems to predict the stabilising effects that cosolvents in a ternary system will have on a protein; and (ii) produce a graph to illustrate the additivity of multiple cosolvents, and output generated predictive data.

This report sets out to define a series of mathematical equations to model the thermodynamic effects of individual cosolvents on protein stability (see Theory), to elucidate the development of a tool that can use experimental thermodynamic data to illustrate these effects, and to put the information gathered into the context of the development of biologics.

Theory

The Wyman-Tanford formula (eq. 1), developed in 1948 and expanded upon in 1968 (Wyman, 1948; Tanford, 1968), was used to describe how the number of solvent and cosolvent (denoted as species 1 and 2, respectively) molecules affects a fixed protein's (denoted as u^*) stability. The addition of a cosolvent changes the activity, and thus chemical potential, μ, of water, and it is this change in chemical potential that drives the change in protein stability. This can also be expressed as a competition between the number of molecules of solvent and cosolvent (ΔN_{ui}) interacting with the surface of the protein.

$$-\left(\frac{\partial \Delta\mu_u^*}{\partial \mu_1}\right)_{T,P,n_u\to 0} = \Delta N_{u1} - \frac{n_1}{n_2}\Delta N_{u2} \tag{1}$$

This equation alone, however, cannot explain which species (1 or 2) makes a more significant contribution to the conformational transition of a protein. Several context-based workarounds were employed to modify the above equation to fit different models of cosolvent activity, each with their own assumption on the dominance of each species' contribution to the protein's stability. These assumptions assume equal influence (Timasheff, 1998), the dominance of solvent movement in osmotic stress (Parsegian, Rand and Rau, 1995), or dominance of bulky cosolvents in molecular crowding examples (Davis-Searles et al., 2001). Having to modify the equation itself to suit one's specific needs is not ideal, so an all-encompassing alternative must be developed.

The most significant problem with the Wyman-Tanford formula is that it does not give enough information about what interactions are occurring at a molecular level. In the context of hydrotropy, cosolvents – categorised into denaturants and stabilisers – are either included (denaturants) or excluded (stabilisers) from the surface of a protein. The Wyman-Tanford formula fails to recognise that stabilisers are excluded from the protein's surface and limits $\Delta N_{u2} \geq 0$, regardless of the degree of exclusion, which has often led to an artificially large ΔN_{u1}. We need to be able to link thermodynamic measurements of the macroscopic world to interactions between individual molecules in the microscopic world. This can be achieved using statistical mechanics in conjunction with classical thermodynamics. The Kirkwood-Buff theory derives thermodynamic properties of all components of a two-component solution easily and can illustrate which component – solvent or cosolvent – contributes more greatly to changes in protein stability (eqs. 2, 3) (Kirkwood and Buff, 1951; Shimizu, 2011).

KB theory was used by Shimizu (2004) to rederive equation 1. This results in the redefinition of ΔN_{u1} and ΔN_{u2} via molecular distribution functions such that ΔN_{u2} can be reinterpreted to be negative for stabilisers. This accurately illustrates the exclusion of a stabiliser from the surface of a protein. These terms are redefined as Kirkwood-Buff Integrals.

ΔV_i refers to the change in partial molar volume of solvent (1) and solute (u), while ΔG_{ij} refers to the change in G – KB integrals that reflect excess numbers of solvent and cosolvent (2) molecules on a protein's surface – associated with a conformational transition.

$$\Delta G_{u1}^0 = -\Delta V_u^0 \tag{2}$$
$$\Delta G_{u2}^0 = -V_1^0 \Delta v_{u1}^0 - \Delta V_u^0 \tag{3}$$

If $|\Delta G_{u1}^0| > |\Delta G_{u2}^0|$, we can conclude that the change in the number of solvent molecules is a more dominant driving force than that of the number of cosolvent molecules, and vice versa. This gives an insight into the thermodynamics of binary systems of solvent-cosolvent competition. There are two drawbacks to this approach, however: (i) the use of this model requires there be a determined number of specific cosolvent binding sites on a protein (Shimizu and Matubayasi, 2017) and (ii) this model still only accounts for the effects of one dissolved cosolvent. A more general statistical thermodynamic approach is required to combine the effects of two cosolvents.

Our Theory for Additivity

Here, we show (i) how we derived our theory, (ii) what it is useful for and (iii) what restrictions were put in place for them to be useful. This derivation is a continuation from work done by Shimizu and Matubayasi (2017).

Consider our solution of 4 components – a protein, a solvent, and two cosolvents. The protein is fixed in position and is thus considered an external field, essentially giving us 3 components. According to the Gibbs phase rule, with three components and one phase, we are given $F = 3 - 1 + 2 = 4$ degrees of freedom, meaning that there are four independent thermodynamic variables that can modulate the chemical potential of the protein solute in a given state, a (Shimizu and Matubayasi, 2017). These four parameters are namely the chemical potentials of cosolvent 2 and 3 (μ_2, μ_3), temperature, and pressure, which can be used to generate the following generalised Clausius-Clapeyron equation (eq. 4).

$$d\mu_u^{(a)} = \left(\frac{\partial \mu_u^{(a)}}{\partial \mu_2}\right)_{T,P,\mu_3} d\mu_2 + \left(\frac{\partial \mu_u^{(a)}}{\partial \mu_3}\right)_{T,P,\mu_2} d\mu_3 + \left(\frac{\partial \mu_u^{(a)}}{\partial T}\right)_{P,\mu_2,\mu_3} dT + \left(\frac{\partial \mu_u^{(a)}}{\partial P}\right)_{T,\mu_2,\mu_3} dP \qquad (4)$$

The last two partial differentials can be interpreted respectively as $-S_u^{(a)}$ and $V_u^{(a)}$ – the partial molar entropy and volume of solute u in state a. Hence, equation 4 can be rewritten for the native and denatured states thusly:

$$d\mu_u^{(n)} = \left(\frac{\partial \mu_u^{(n)}}{\partial \mu_2}\right)_{T,P,\mu_3} d\mu_2 + \left(\frac{\partial \mu_u^{(n)}}{\partial \mu_3}\right)_{T,P,\mu_2} d\mu_3 - S_u^{(n)} dT + V_u^{(n)} dP \qquad (5)$$

$$d\mu_u^{(d)} = \left(\frac{\partial \mu_u^{(d)}}{\partial \mu_2}\right)_{T,P,\mu_3} d\mu_2 + \left(\frac{\partial \mu_u^{(d)}}{\partial \mu_3}\right)_{T,P,\mu_2} d\mu_3 - S_u^{(d)} dT + V_u^{(d)} dP \qquad (6)$$

These equations now bear semblance to the Gibbs-Duhem equation ($\sum N_i d\mu_i = -SdT + VdP$) which requires that the effects of our 4 independent variables be additive (Dalal, 2018). This additive condition is what will be checked for when analysing data. Subtracting equation 6 from equation 5 gives us equation 7:

$$\Delta\mu_u = \left(\frac{\partial \Delta\mu_u}{\partial \Delta\mu_2}\right)_{T,P,\mu_3} d\mu_2 + \left(\frac{\partial \Delta\mu_u}{\partial \Delta\mu_2}\right)_{T,P,\mu_2} d\mu_3 - \Delta S_{n\to d} dT + \Delta V_{n\to d} dP \qquad (7)$$

At equilibrium and constant pressure (i.e. $\Delta\mu_u = 0$ and $\Delta V_{n\to d} dP = 0$), equation 7 can be rewritten as

$$\left(\frac{\partial \Delta\mu_u}{\partial \mu_2}\right)_{T,P,\mu_3} d\mu_2 + \left(\frac{\partial \Delta\mu_u}{\partial \mu_3}\right)_{T,P,\mu_2} d\mu_3 = \Delta S_{n\to d} dT_m \qquad (8)$$

where T_m is the melting temperature of the protein solute. At equilibrium, $\Delta S_{n\to d} = \frac{\Delta H_{n\to d}}{T_m}$, so we rewrite 8 for finite changes (δ instead of infinitesimal d):

$$\delta T_m = \frac{T_m}{\Delta H_{n\to d}} \left[\left(\frac{\partial \Delta\mu_u}{\partial \Delta\mu_2}\right)_{T,P,\mu_3} \delta\mu_2 + \left(\frac{\partial \Delta\mu_u}{\partial \Delta\mu_2}\right)_{T,P,\mu_2} \delta\mu_3 \right] \qquad (9)$$

This equation is a general-case model that can be used to describe the effects of multiple cosolvents under higher concentration conditions. It is important to note that $\frac{T_m}{\Delta H_{n\to d}}$ is defined at zero cosolvent concentration (ie pure protein-solvent solution). Equation 9 is useful to us in showing exactly how the *chemical potentials* of different cosolvents affect the transition temperature (our indicator of protein stability). Following Raoult's law, we can convert our chemical potential into a more controllable variable – concentration – via the relation $\mu_i = \mu_i^\infty + RT \ln c_i$ (Raoult, 1889). This places a restriction on our experiment such that this must be operated under infinitesimally dilute conditions, and we assume this to be a dilute ideal solution. As such, we can rewrite equation 9 as:

$$\delta T_m = \frac{T_m}{\Delta H_{n\to d}} \left[\left(\frac{\partial \Delta\mu_u}{\partial \Delta c_2}\right)_{T,P,c_{3\to 0}} \delta c_2 + \left(\frac{\partial \Delta\mu_u}{\partial \Delta c_3}\right)_{T,P,c_{2\to 0}} \delta c_3 \right] \qquad (10)$$

The dilute condition requires that the partial derivatives are operated at the $c_i \to 0$ limit, meaning that the two derivatives are independent of one another. Following this train of thought, we can show that the effects of each cosolvent *in isolation* (eqs. 11a and 11b) can be added together to model the overall effect (eq. 11c):

$$\frac{\Delta H_{n\to d}}{T_m}\delta T_{m,2} = \left(\frac{\partial \Delta\mu_u}{\partial \Delta c_2}\right)_{T,P,c_3\to 0}\delta c_2 \qquad (11a)$$

$$\frac{\Delta H_{n\to d}}{T_m}\delta T_{m,3} = \left(\frac{\partial \Delta\mu_u}{\partial \Delta c_3}\right)_{T,P,c_2\to 0}\delta c_3 \qquad (11b)$$

$$\delta T_m = \frac{\Delta H_{n\to d}}{T_m}\left(\delta T_{m,2} + \delta T_{m,3}\right) \qquad (11c)$$

These equations provide us with a method of modelling a multiple-cosolvent environment by combining the stabilising effects of isolated cosolvents. In order to get a clearer overall picture of what is happening at a molecular level, we must relate the information we have gained from equation 11c back to Kirkwood-Buff theory; ΔG_{u2} values will then prove useful in quantifying the interaction between the protein and cosolvents. This value can be determined through the relation:

$$\Delta G_{u2} = \Delta G_{u1} - \frac{\Delta H_m}{RT_m^2}\frac{\delta T_{m,2}}{\delta c_2} \qquad (12)$$

where ΔG_{u1} is assumed to be negligible with regards to $-\frac{\Delta H_{n\to d}}{RT_m^2}\frac{\delta T_m}{\delta c_2}$ (Shimizu, Stenner and Matubayasi, 2017). Shimizu and Matubayasi (2017) show that cosolvents with a positive ΔG_{u2} are classified as denaturants, and those with negative values are stabilising agents. Equation 12 can be rewritten to retrieve ΔG_{u2} easily as:

$$\Delta G_{u2} = -\frac{\Delta H_m \delta T_m}{RT_m^2}\frac{1}{\delta c_2} \qquad (13)$$

Thus, the derivation of a statistical thermodynamic theory has been achieved. We henceforth set out to show the usefulness of our theory using experimental data.

Methods

This section of the report details how data from thermal denaturation calorimetry experiments is collected, formatted, analysed and visualised via a tool we have developed based on the statistical thermodynamic theory set out above.

Differential Scanning Calorimetry (DSC)

Differential Scanning Calorimetry (DSC) is a non-perturbing technique used to study the thermodynamic properties of biomolecules undergoing a transition. Two isobaric chambers are set up: one contains the solution without protein, and the other contains all species involved (u, 1, 2, 3…). The difference in temperature between chambers is measured and the temperature is controllably incremented. At the melting temperature (T_m), 50% of all protein molecules have undergone a transition from the native state. With this transition comes a release of energy into the surrounding solution which is detected by scanners in the DSC. A peak (fig. 2) is drawn which provides us with the T_m, as well as the enthalpy (Chiu and Prenner, 2011; Privalov and Plotnikov, 1989; Freire, 1995).

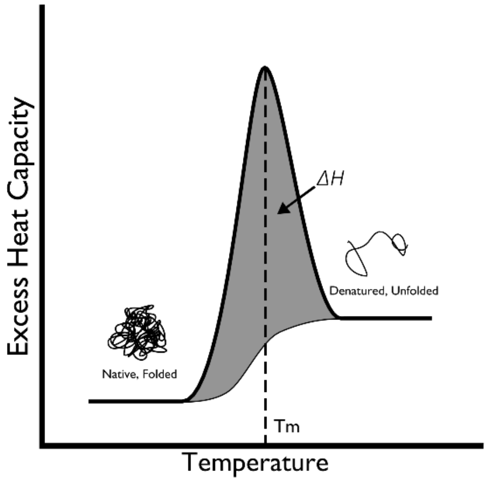

Figure 2 – DSC output graph depicting the heat capacity changes associated with protein denaturation. This difference is derived from the comparison of a solution containing the solute with cosolvents versus without cosolvents. The temperature at the peak of the curve represents the melting temperature, T_m, while the area under the curve represents the enthalpy associated with the transition

Data Collection

Experimental data used throughout this report has been collected from previous publications on the effects different cosolvents have on a protein's stability, both in isolation and in combination. The most important parameters in these papers are the melting temperatures and concentrations of cosolvents involved. Data were extracted from the publications directly where possible – i.e., the numerical data for melting temperatures, enthalpies at zero concentration and cosolvent concentrations. If the data is not presented, it is possible to extract data from a graph via a web app – WebPlotDigitiser (Rohatgi, 2021) – that exports data to a readily-usable comma-separated values (.csv) file. All data that is processed is done so from a .csv file. Since the WebPlotDigitiser web app is only as accurate as its user, some manual modification to the .csv file may be required to line up the X-axis (concentration) values with those used in the publication. Where data is presented in the publication, manual copy-pasting to a .csv file is required.

Data Formatting

Once data is collected into .csv files, headers are required. They need not be in any order, but Table 1 provides an example of how formatting could be done:

SX	SY	DX	DY	CX	CY	*ΔH* at [cosolvent] = 0	Ratio	chem
0	55	0	55	0	55	111	2	Urea
1	56	2	52	1	54			TMAO
2	57	4	49	2	55			

Table 1 – example data formatted as required by tool. Each header is defined as: SX – stabiliser concentration; SY – stabiliser melting temperature; DX – denaturant concentration; DY – denaturant melting temperature; CX – experimental combination concentration (from primary x-axis); CY – combination melting temperature; ΔH – enthalpy measured at 0 cosolvent concentration; Ratio – ratio between the two cosolvent species' concentrations; chem – the cosolvents in question (required by aesthetics for graphing)

Data Analysis

The data that is gathered now needs to be processed. For this we have developed a tool; the criteria that this needs to meet are as follows:

1. Determines how well our model (equation 11c) fits experimental data using statistical analyses
2. Predicts combination data for publications in which there are no combined effects documented
3. Calculates ΔG_{u2} value for each cosolvent
4. Outputs analyses in both graphical and numerical formats (image and .csv)

The code, developed using R (R Core Team, 2021), is aided by simple pseudocode (Appendix A) for the purposes of reproducibility. The programming language was chosen for its statistical analysis capability, as well as having a dedicated graphing module that is simple to manipulate, even for those unfamiliar with R. The processing pipeline for the data is represented graphically in fig. 3.

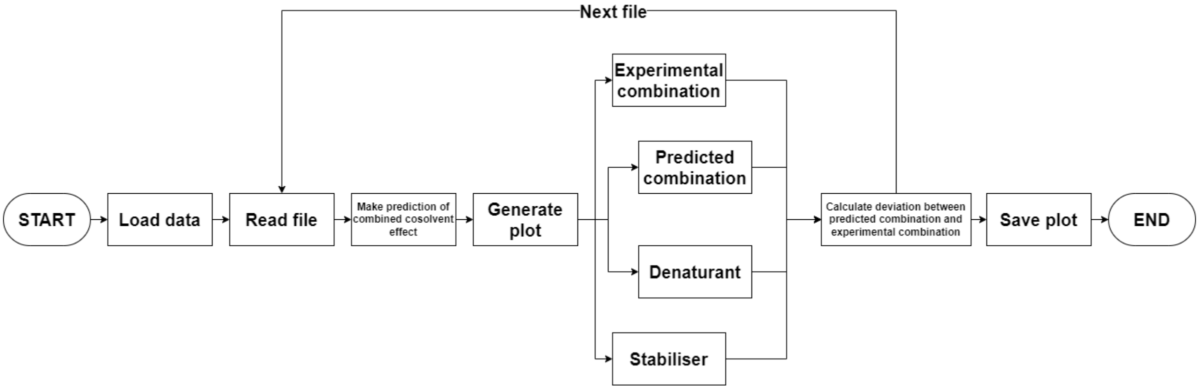
Next file
START
Load data
Read file
Make prediction of combined cosolvent effect
Generate plot
Experimental combination
Predicted combination
Denaturant
Stabiliser
Calculate deviation between predicted combination and experimental combination
Save plot
END

Figure 3 – Flow Diagram. Shows the flow of information of thermal denaturation data. The program is a loop whereby data is loaded, read and processed. This process involves statistical analysis of the data from a paper and modelling the data to show how well our model (equation 11c) fits the data. From there, a prediction of cosolvent additivity is made for publications with no combined data.

The tool was developed using an iterative, agile workflow using pseudocode as a guide. Three R packages have been used to draw plots (Wickham, 2016; Wickham and Seidel, 2022) and to monitor the runtime efficiency of our code (Kusnierczyk, 2012). The structure of the code is simple: a for loop is used to cycle through .csv files that a user has in their directory. For each file, T_m and enthalpy change at zero cosolvent concentration (ΔH_m) values are extracted, then tested against a linear model for stabilisers, denaturants, and combinatory data. This linear model accounts for equations 11a, 11b and 11c. The resulting data is then plotted onto a simple graph of T_m against cosolvent concentration. The initial T_m and ΔH_m are used to calculate KB integrals for all cosolvents individually.

Results

A lack of currently available tools for processing thermodynamic data to output cosolvent effects on protein stability provides an opportunity to develop our own tool. This tool takes in T_m and ΔH_m values across a range of concentrations, and outputs three things: (i) a graph showing the relationship between cosolvent concentration and T_m, as well as a modelled prediction of what the combined cosolvent effect looks like; (ii) R^2 values that show how well our models (equations 11a, b and c) fit the experimental data; and (iii) Kirkwood-Buff integrals associated with the protein conformation transition (ΔG_{ui} where i is any cosolvent species).

Graphs

Here we demonstrate how the relationships between cosolvent concentrations and a protein's stability can be illustrated using graphs. As an initial step in development, we required publications with data regarding the isolated effects of a cosolvent on T_m, as well as in combination. This sets the groundwork for how we can move forward in analysing more data.

Following the workflow set out in figure 3, data is searched for in the user's directory, set by the user, then loaded into R as a data frame. Each data set provides cosolvent concentration and T_m as the X and Y aesthetics of the graph, and the ggplot2 package handles plotting, resulting in figure 4. A code snippet used to generate this is found in Appendix B.

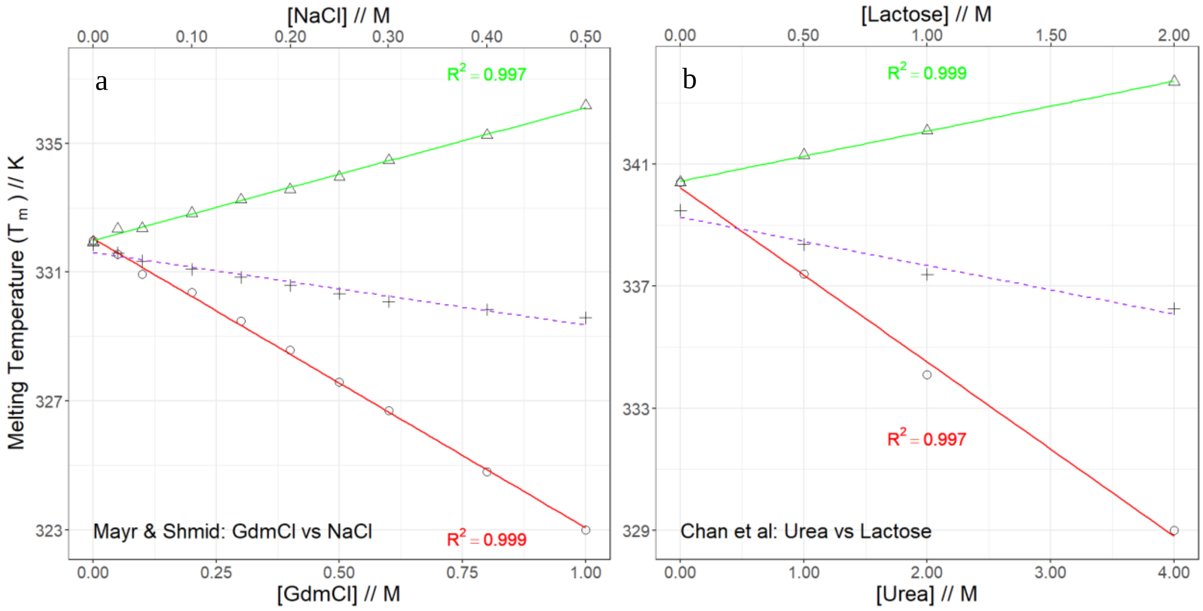

Figure 4 – Using data from (a) Mayr and Schmid (1993) and (b) Chan et al (1996). Data is plotted and colour-coded the same way as in figure 4, except for the lack of a black experimental combination line. Plots use the species: u = (a) Ribonuclease T1, (b) rhDNase; 1 = water; 2 = (a) Guanidinium chloride, (b) urea; 3 = (a) sodium chloride, (b) lactose.

When plotting on two different x-axes such that the concentration of urea is twice that of TMAO in both cases, leading to a lack of visual clarity. This is fixed using a secondary axis scaled by the ratio between denaturant and stabiliser, hence the requirement for manual addition into the .csv file.

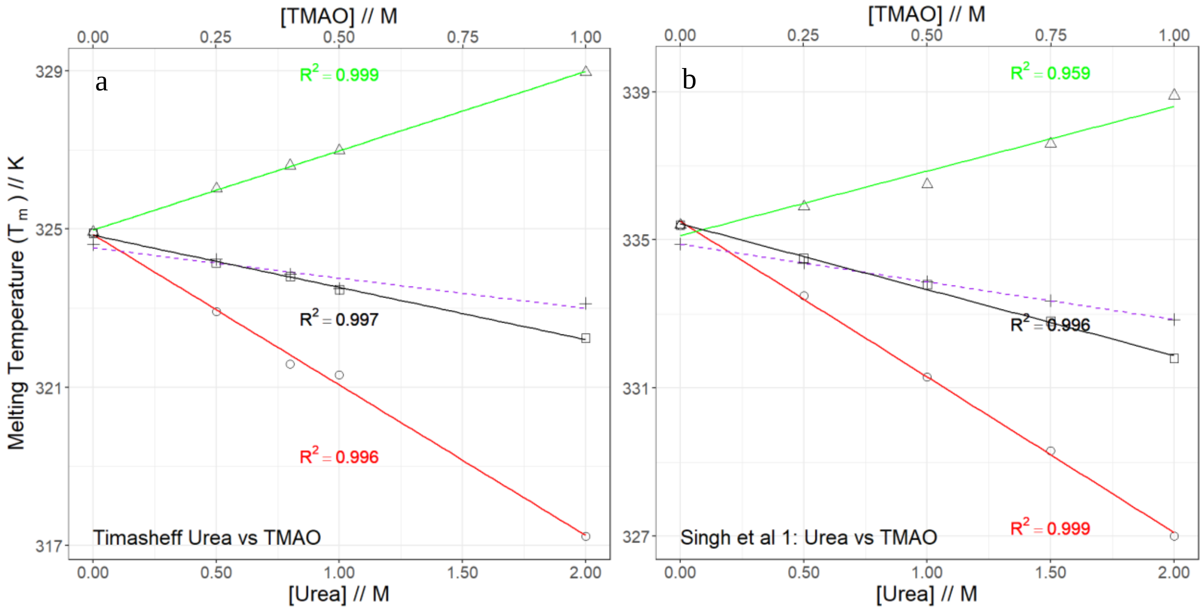

Figure 5 – Using data from Lin & Timasheff (1994) and Singh et al. (2007), data is plotted [cosolvent] vs T_m. R^2 values show how well our explanatory variable describes variation in the response variable. The plots use the species: u = (a) RNase T1, (b) RNase A; 1 = water; 2 = urea; 3 = trimethylamine N-oxide (TMAO). Colours and symbols: green w/ triangle – stabiliser effect; red w/ circle – denaturant effect; purple (dashed) w/ cross – predicted combination; black w/ square – experimental combination.

Once it was evident that we could use our model for predicting additivity, the program was modified to detect whether an experimental combination was present in the data. The graphs are then changed accordingly to show the two isolated cosolvents' data, and the *predicted* combination (fig. 5).

In contrast to figure 4, there are instances in which the experimental combination does not fit our model well – i.e., it has a low R^2 value. Example data from Singh et al. (2007) using a different protein solute (α-lactalbumin) is shown in figure 6.

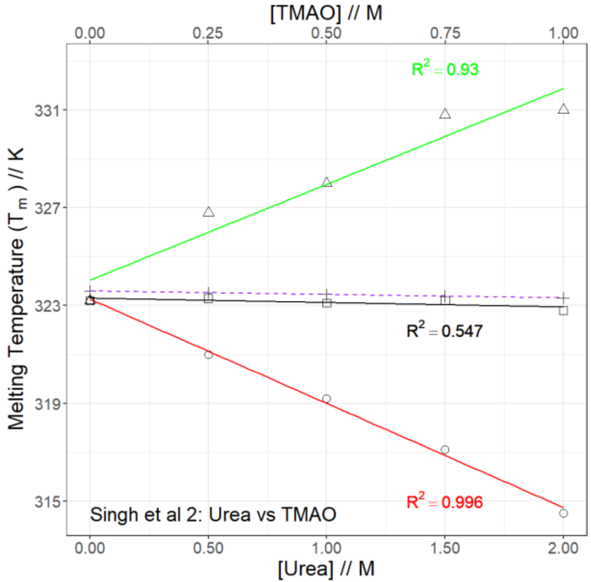

Figure 6 – Example of data (from Singh et al. (2007)) that does not fit our additive model well. The experimental combination line (black) shows strong linearity but has a relatively low R^2 value. Symbols and colours are as in figure 4.

Statistical and Thermodynamic Analysis

The program, in carrying out linear regression as seen in figures 4 and 5 and produces R^2 values for the linear fit. These evaluate how well our explanatory variable (concentration of cosolvent) describes variation in our response variable (T_m). This is essentially an indicator of how well our model fits the data. Equations 11a, b and c are linear in nature and, as such, can be represented by a standard linear regression model which returns an R^2 value. R^2 values range from 0 to 1 – 0 meaning the explanatory variable does not account for any variation in the response, and 1 meaning all variation is explained by the explanatory variable (Steel and Torrie, 1960; Nagelkerke, 1991).

The KB integrals for each isolated cosolvent can be calculated using equation 13. This requires the initial ΔH_m at 0 cosolvent concentration, as well as the gas constant, R. Output data from Singh et al. (2007) and Chan, Au-Yeung and Gonda (1996) containing R^2 values for stabiliser, denaturant and experimental combination, as well as ΔG_{u2} values for each cosolvent can be found in table 2.

Singh et al (2007)					**Chan et al (1996)**			
TMAO $\boldsymbol{R^2}$	Urea R^2	Combination R^2	ΔG_{u2} Urea	ΔG_{u2} TMAO	Lactose R^2	Urea R^2	ΔG_{u2} Urea	ΔG_{u2} Lactose
0.959	0.999	0.996	13.54	-2.82	0.999	0.999	54.2	-2.35

Table 2 – Output data from Singh et al (2007) and Chan et al (1996). These are examples of high R^2 values showing a good fit with our linear models. ΔG_{u2} values highlight the relative effect that each cosolvent has on protein stability

Important Features of R Script

The R script used in our tool is primarily a `for` loop with some preliminary setups. Here we break down how input data is processed into our outputs, and how computationally efficient each section of the script is (Table 3).

Linear regression was used for two purposes: (i) to extrapolate experimental data for isolated cosolvents ready for combination and graphing, and (ii) to extract R^2 values. This regression is represented by lines 38 to 64. Using `rbenchmark,` we see that for a dataset of 1000 publications, regression takes a total of 4.05 seconds. This gives us both the extrapolated data, as well as the R^2 values associated

KB integrals were determined using an `if` statement that runs only if a ΔH_m value is present in the data. Lines 66 to 82 use equation 13 to calculate a ΔG_{u2} value for each cosolvent in isolation. For a dataset of 1000 publications, this part of the script takes a total of 3.65 seconds.

Graphing is a vital part of any scientific report with large datasets. This carries the bulk of the processing power since it requires linear regression models for each line fitted on the graph. For 1000 publications, lines 93 to 160 take 413 seconds to complete.

Code Region	**Time elapsed for 100 publications // s**	**Time elapsed for 1000 publications // s**
Linear Regression (Lines 38 to 64)	0.41	4.05
KB Integrals (Lines 66 to 82)	0.36	3.65
Graphing (Lines 93 to 160)	41.0	413
Whole Script	41.4	429

Table 3 – benchmark times for 100 and 1000 replications of each of the functions of the R script.

Discussion

Results provided in this report show that a statistical thermodynamic approach to computationally modelling the effects that multiple cosolvents have on a protein's stability has been achieved. Each criterion that this project set out to meet will be discussed in parts, evidenced by the results.

Derivation of Statistical Thermodynamic Theory

As set out in the Theory section of this report, our mathematical model clearly shows how the limitations of previous work – inability to manage multiple cosolvents and lack of thermodynamic insight – can be overcome. Equations 11a, 11b, and 11c can be used with ease to describe the stabilising or destabilising effects of a cosolvent. Equation 11c can be modified simply to accommodate more cosolvents in the mixture; this can be accomplished due to the additive nature of the Gibbs-Duhem equation. To give example, equation 11c can be modified as such:

$$\delta T_m = \frac{\Delta H_{n\to d}}{T_m} \sum \delta T_{m,i} \tag{14}$$

Where $\delta T_{m,i}$ is the cosolvent's isolated stabilising effect on a protein. If we expanded our library of cosolvent data, we could create models of much larger environments more closely resembling that of a biopharmaceutical drug's potential target site.

This is a powerful equation on its own to show the overall effect on a protein's stability, but we go a step further to analyse the KB integrals associated with each interaction. Using equation 13, we obtain thermodynamic information about the interactions between a cosolvent and the protein. This tells us the relative strength of a cosolvent's stabilising effects, as well as whether this cosolvent is a stabilising agent (negative ΔG_{u2}) or denaturant (positive ΔG_{u2}) under the experimental conditions.

Data Analysis (tool)

The tool that has been produced in this project has proven to be valuable in utilising the Theory derived. It meets all of the success criteria set out in the Results section to varying degrees, each of which is discussed hereafter.

1. Determines how well our model (equation 11c) fits experimental data using statistical analyses
2. Predicts combination data for publications in which there are no combined effects documented
3. Calculates ΔG_{u2} value for each cosolvent
4. Outputs analyses in both graphical and numerical formats (image and .csv)

Equation 11c was used in data analysis to show the additivity of cosolvent effects. The linear regression performed on the isolated and combined cosolvent effects yielded a model that allows us to make predictions of T_m values at different concentrations of cosolvent. The model also tells us how closely the experimental data, when added together, still fits with linearity using R^2 values. The meaning of the R^2 value is up for interpretation depending widely on the scientific field it is used in (Cohen, 1992; Moksony and Heged, 1990). In statistical thermodynamics and related fields, an R^2 value above 0.7 is regarded as passable. Figure 4 is used to show how the experimental combination (black line) fits with our test for linearity as shown by the R^2 close to 1. The actual statistical significance of R^2 is up for debate since, while it does show how well an explanatory variable explains variance in a response variable, it is not always a complete descriptor of the quality of a model. In order to give a more accurate meaning to this value, more rigorous analysis of a linear regression using prediction intervals may become more relevant in future studies (Olive, 2007).
Our tool found a publication with an anomalous R^2 value (0.547) representing its experimental combination data (Figure 6, Singh et al., 2007). This result does not suffice to show that there is a statistically viable relationship between the concentration of cosolvents and the change in melting temperature. There are two possible explanations for this: (i) this supports the case that, from a general standpoint, the R^2 value is not thorough as an indicator of a high-quality predictive model (which would warrant the use of prediction interval analysis), or (ii) that equation 11c does not explain the variance in the melting temperature as seen experimentally. In terms of the cosolvents in use, the experimental procedure used in this publication is identical to that of those found in figure 4b, with the exception being that the protein solute changes from RNase A to α-Lactalbumin. This may help to highlight differences in the ways these two proteins interact with cosolvents, perhaps by inducing interactions between cosolvents that our theory cannot yet be used to model. Such an interaction occurring between urea and TMAO (and other related methylamines) may be similar in mechanism to that of α-chymotrypsin whereby preferential hydration excludes TMAO from the surface of the protein which, via electrostatic interactions between an electropositive carbon in TMAO and the electronegative oxygen in urea, pulls the denaturant away from the protein's vicinity (Venkatesu, Lee and Lin, 2009).

To speak critically of our tool, one cannot overlook its computational performance. Referring to Table 3 we see that, on runtime, the script is very efficient at analysing the data (linear regression and generation of KB integrals). In comparison, graphing our results using the `ggplot2` package is much slower – approximately 96.3% of runtime is taken up by producing and displaying graphs. For the purposes of the investigation of the effects of cosolvents on protein stability, graphing is not a vital necessity except in presentation of key data. It may be possible to keep the graphical representation of our data as an optional part of the tool. This will allow users to analyse more data, and they will have the option to graphically represent the data that is most relevant to their findings.

Further Study

It may be possible in future to extend our Theory to illustrate cosolvent-cosolvent interactivity if we are to remove the dilution restriction – the assumption that our cosolvents do not significantly interact with each other. This requires that we can obtain information about the chemical potential of a cosolvent at near-limit dilution (μ^{∞}) to relate concentration back to the chemical potential of a cosolvent in larger concentrations via Raoult's Law ($\mu_i = \mu_i^{\infty} + RT \ln c_i$). This brings us back to equation 9. Since thermal denaturation experiments cannot be used to obtain the μ_i^{∞} value, an alternative approach – either experimental or theoretical – will be required to show how the chemical potentials of each cosolvent are affected by one another.

Future work should also include adjustments to the tool that has been developed. Two vital features that should be updated or included are (i) capability for more cosolvents than two and (ii) a user interface that can show all graphical outputs in an interactive manner.

The first is significantly more important within the scope of our work since it would allow us to model systems with more chemical species, much closer in likeness to an *in vivo* chemical environment. Criteria for the analysis of thermal denaturation data are as follows:

- Individual cosolvent data and, if available, experimental combination data are analysed, yielding R^2 values and ΔG_{u2} values where appropriate.
- Predicted values for a combination are calculated using equation 14, rather than equation 11c (to accommodate the presence of potentially more than 2 cosolvents).
- Should be able to differentiate between data containing different numbers of cosolvent species.

The second feature – a user interface – is required to make the tool more user-friendly, as well as accessible to a wider range of people. One such way of doing this is to help manage large numbers of output graphs using the `trelliscopejs` R package (Hafen and Schloerke,2021). This can be used with efficiency to collate and display all the output graphs in an interactive format, and it is designed to work in conjunction with the ggplot2 workflow. An example of this package in use is found at http://hafen.github.io/trelliscopejs-demo/gapminder_plotly/. This package also allows for effective presentation and communication of results as trelliscopes can be output as an HTML element for embedding in webpages or emails.

A final addendum to the note of further study is to rederive our theory for application using pressure denaturation as opposed to thermal denaturation. A protein's conformational stability is also afflicted under high pressures (Bridgman, 2013; Huang et al., 2016). Using the framework set out by Shimizu and Smith (2017) and the Theory set out in this report, modifying equation 8 for constant temperature, it is possible to link denaturation pressure and cosolvent concentration to a protein's stability. An effective technique for probing pressure dependence is pressure-jump, which increases or decreases pressure on a sample and measures the resulting volume change (Takahashi and Alberty, 1969; Nölting, 2005). The resulting theoretical model to replace equations 11a, 11b and 14 may have the following form:

$$\Delta \mathrm{V} \delta P_{m,2} = \left(\frac{\partial \Delta \mu_u}{\partial \Delta c_2}\right)_{T,P,c_3 \to 0} \delta c_2 \tag{15a}$$

$$\Delta V \delta P_{m,3} = \left(\frac{\partial \Delta \mu_u}{\partial \Delta c_3}\right)_{T,P,c_2 \to 0} \delta c_3 \tag{15b}$$

$$\delta P_m = \Delta \mathrm{V_m} \sum \delta P_{m,i} \tag{16}$$

This kind of investigation may become useful in the discovery of biologics depending on the environment of the drug’s target site. Various regions in the human body may be prone to frequent, drastic changes in pressure such as blood vessels in and around the heart. For biologic delivery to these regions, it may be necessary to keep a protein stable with respect to the surrounding pressures.

Conclusions

In this report, we show that the thermal stability of a protein is affected by the presence of chemical cosolvents in a way that can be modelled. Current knowledge of cosolvent-mediated thermal stability of proteins was limited to systems involving only one cosolvent, and it is this knowledge that needed to be built upon to begin to understand the complexities of larger systems. These larger systems should then act as microcosmic models to represent the more complex systems found in nature. We have shown in this report the derivation of a theory rooted in statistical thermodynamics that can be used to manipulate experimental thermal denaturation data to provide an insight into the additivity of cosolvent effects. The theory set out in this report is supported by the development of a new tool. It is shown, with example, that this tool can execute statistical analyses on thermodynamic data with relative ease, as well as produce publication-quality graphs to show the additivity of cosolvent effects. What remains to subsequently be seen are the following: (i) whether pressure denaturation can be analysed using the theory set out in this paper, and (ii) updates to, and publishing of, the tool developed.

The findings of this report can be used not only in the field of biologics – i.e., their discovery, development and administration – but also in understanding food chemistry(Shimizu, Stenner and Matubayasi, 2017), deep sea marine biochemistry (Shimizu and Smith, 2017) and cancer treatment (Morrow and Felcone, 2004).

References

Bridgman, P. W. (2013). THE COAGULATION OF ALBUMEN BY PRESSURE. In: *THE COAGULATION OF ALBUMEN BY PRESSURE*. Harvard University Press. pp.735–736. [Online]. Available at: doi:10.4159/harvard.9780674287808.c8 [Accessed 17 April 2022].

Chan, H.-K., Au-Yeung, K.-L. and Gonda, I. (1996). Effects of Additives on Heat Denaturation of rhDNase in Solutions. *Pharmaceutical Research*, 13 (5), pp.756–761. [Online]. Available at: doi:10.1023/A:1016007818575.

Chiu, M. H. and Prenner, E. J. (2011). Differential scanning calorimetry: An invaluable tool for a detailed thermodynamic characterization of macromolecules and their interactions. *Journal of Pharmacy and Bioallied Sciences*, 3 (1), pp.39–59. [Online]. Available at: doi:10.4103/0975-7406.76463.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, (112(1)), pp.155–159.

Dalal, M. (2018). A Textbook of Physical Chemistry - Volume 1. In: *A Textbook of Physical Chemistry*. 1. Dalal Institute. pp.108–110. [Online]. Available at: https://www.dalalinstitute.com/books/a-textbook-of-physical-chemistry-volume-1/.

Davis-Searles, P. R. et al. (2001). Interpreting the Effects of Small Uncharged Solutes on Protein-Folding Equilibria. *Annual Review of Biophysics and Biomolecular Structure*, 30 (1), pp.271–306. [Online]. Available at: doi:10.1146/annurev.biophys.30.1.271.

Freire, E. (1995). Differential Scanning Calorimetry. In: Shirley, B. A. (Ed). *Protein Stability and Folding: Theory and Practice*. Methods in Molecular Biology^{TM}. Totowa, NJ: Humana Press. pp.191–218. [Online]. Available at: doi:10.1385/0-89603-301-5:191 [Accessed 16 April 2022].

Hollander, F. (1949). The composition and mechanism of formation of gastric acid secretion. *Science (Washington)*, 110, pp.57–63.

Huang et al. (2016). A molecular perspective on the limits of life: Enzymes under pressure. *Condensed Matter Physics*, 19 (2), p.22801. [Online]. Available at: doi:10.5488/CMP.19.22801.

Jorgenson, T. C., Zhong, W. and Oberley, T. D. (2013). Redox Imbalance and Biochemical Changes in Cancer. *Cancer research*, 73 (20), pp.6118–6123. [Online]. Available at: doi:10.1158/0008-5472.CAN-13-1117.

Kirkwood, J. G. and Buff, F. P. (1951). The Statistical Mechanical Theory of Solutions. I. *The Journal of Chemical Physics*, 19 (6), pp.774–777. [Online]. Available at: doi:10.1063/1.1748352.

Kusnierczyk, W. (2012). *rbenchmark: Benchmarking routine for R*. [Online]. Available at: http://rbenchmark.googlecode.com/.

Lacaná, E. et al. (2007). The Emerging Role of Pharmacogenomics in Biologics. *Clinical Pharmacology & Therapeutics*, 82 (4), pp.466–471. [Online]. Available at: doi:10.1038/sj.clpt.6100334.

Lee, J. C. and Timasheff, S. N. (1981). The stabilization of proteins by sucrose. *Journal of Biological Chemistry*, 256 (14), pp.7193–7201. [Online]. Available at: doi:10.1016/S0021-9258(19)68947-7.

Moksony, F. and Heged, R. (1990). Small is beautiful. The use and interpretation of R2 in social research. *Szociológiai Szemle, Special issue*, pp.130–138.

MORROW, T. and Felcone, L. H. (2004). Defining the difference: What Makes Biologics Unique. *Biotechnology Healthcare*, 1 (4), pp.24–29.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78 (3), pp.691–692.

Nölting, B. (2005). *Protein Folding Kinetics: Biophysical Methods*. Springer Science & Business Media.

Olive, D. J. (2007). Prediction intervals for regression models. *Computational Statistics & Data Analysis*, 51 (6), pp.3115–3122. [Online]. Available at: doi:10.1016/j.csda.2006.02.006.

Parsegian, V. A., Rand, R. P. and Rau, D. C. (1995). [3] Macromolecules and water: Probing with osmotic stress. In: *Methods in Enzymology*. Energetics of Biological Macromolecules. 259. Academic Press. pp.43–94. [Online]. Available at: doi:10.1016/0076-6879(95)59039-0 [Accessed 25 March 2022].

Pham, J. V. et al. (2019). A Review of the Microbial Production of Bioactive Natural Products and Biologics. *Frontiers in Microbiology*, 10. [Online]. Available at: https://www.frontiersin.org/article/10.3389/fmicb.2019.01404 [Accessed 24 March 2022].

Privalov, P. L. and Plotnikov, V. V. (1989). Three generations of scanning microcalorimeters for liquids. *Thermochimica Acta*, 139, pp.257–277. [Online]. Available at: doi:10.1016/0040-6031(89)87027-3.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. [Online]. Available at: https://www.R-project.org/.

Ranjbar, B. and Gill, P. (2009). Circular Dichroism Techniques: Biomolecular and Nanostructural Analyses- A Review. *Chemical Biology & Drug Design*, 74 (2), pp.101–120. [Online]. Available at: doi:10.1111/j.1747-0285.2009.00847.x.

Raoult, F.-M. (1889). Recherches expérimentales sur les tensions de vapeur des dissolutions. *Journal de Physique Théorique et Appliquée*, 8 (1), pp.5–20. [Online]. Available at: doi:10.1051/jphystap:0188900800500.

Revers, L. and Furczon, E. (2010). An Introduction to Biologics and Biosimilars. Part II: Subsequent Entry Biologics: Biosame or Biodifferent? *Canadian Pharmacists Journal / Revue des Pharmaciens du Canada*, 143 (4), pp.184–191. [Online]. Available at: doi:10.3821/1913-701X-143.4.184.

Rohatgi, A. (2021). *WebPlotDigitiser*. [Online]. Available at: https://automeris.io/WebPlotDigitizer.

Shimizu, S. (2004). Estimating hydration changes upon biomolecular reactions from osmotic stress, high pressure, and preferential hydration experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (5), pp.1195–1199. [Online]. Available at: doi:10.1073/pnas.0305836101.

Shimizu, S. (2011). Molecular origin of the cosolvent-induced changes in the thermal stability of proteins. *Chemical Physics Letters*, 514 (1), pp.156–158. [Online]. Available at: doi:10.1016/j.cplett.2011.08.038.

Shimizu, S. (2020). Formulating rationally via statistical thermodynamics. *Current Opinion in Colloid & Interface Science*, 48, pp.53–64. [Online]. Available at: doi:10.1016/j.cocis.2020.03.008.

Shimizu, S. and Matubayasi, N. (2017). Unifying hydrotropy under Gibbs phase rule. *Physical Chemistry Chemical Physics*, 19 (35), pp.23597–23605. [Online]. Available at: doi:10.1039/C7CP02132A.

Shimizu, S. and Smith, P. E. (2017). How Osmolytes Counteract Pressure Denaturation on a Molecular Scale. *ChemPhysChem*, 18 (16), pp.2243–2249. [Online]. Available at: doi:10.1002/cphc.201700503.

Shimizu, S., Stenner, R. and Matubayasi, N. (2017). Gastrophysics: Statistical thermodynamics of biomolecular denaturation and gelation from the Kirkwood-Buff theory towards the understanding of tofu. *Food Hydrocolloids*, 62, pp.128–139. [Online]. Available at: doi:10.1016/j.foodhyd.2016.07.022.

Singh, L. R. et al. (2007). Testing the paradigm that the denaturing effect of urea on protein stability is offset by methylamines at the physiological concentration ratio of 2:1 (urea:methylamines). *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1774 (12), pp.1555–1562. [Online]. Available at: doi:10.1016/j.bbapap.2007.09.006.

Steel, R. G. D. and Torrie, J. H. (1960). *Principles and procedures of statistics: with special reference to the biological sciences*. New York.

Takahashi, M. T. and Alberty, R. A. (1969). [2] The pressure-jump method. In: Kustin, K. (Ed). *Methods in Enzymology*. Fast Reactions. 16. Academic Press. pp.31–55. [Online]. Available at: doi:10.1016/S0076-6879(69)16005-X [Accessed 17 April 2022].

Tanford, C. (1968). Protein Denaturation. In: Anfinsen, C. B. et al. (Eds). *Advances in Protein Chemistry*. 23. Academic Press. pp.121–282. [Online]. Available at: doi:10.1016/S0065-3233(08)60401-5 [Accessed 25 March 2022].

Timasheff, S. N. (1998). In disperse solution, “osmotic stress” is a restricted case of preferential interactions. *Proceedings of the National Academy of Sciences,* 95 (13), pp.7363–7367. [Online]. Available at: doi:10.1073/pnas.95.13.7363.

Venkatesu, P., Lee, M.-J. and Lin, H. (2009). Osmolyte Counteracts Urea-Induced Denaturation of α-Chymotrypsin. *The Journal of Physical Chemistry B*, 113 (15), pp.5327–5338. [Online]. Available at: doi:10.1021/jp8113013.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. [Online]. Available at: https://ggplot2.tidyverse.org.

Wickham, H. and Seidel, D. (2022). *scales: Scale Functions for Visualization*. [Online]. Available at: https://scales.r-lib.org.

Wyman, J. (1948). Heme Proteins. In: Anson, M. L. and Edsall, J. T. (Eds). *Advances in Protein Chemistry*. 4. Academic Press. pp.407–531. [Online]. Available at: doi:10.1016/S0065-3233(08)60011-X [Accessed 23 March 2022].

Wyman, J. (1964). Linked Functions and Reciprocal Effects in Hemoglobin: A Second Look. In: Anfinsen, C. B. et al. (Eds). *Advances in Protein Chemistry*. 19. Academic Press. pp.223–286. [Online]. Available at: doi:10.1016/S0065-3233(08)60190-4 [Accessed 24 March 2022].