MDPI

Article

Principled decision-making workflow with hierarchical Bayesian models of high throughput dose-response measurements

Eric J. Ma [1] **and Arkadij Kummer** [2]

[1] ericmajinglong@gmail.com
[2] ETH Zurich, arkadij.kummer@gmail.com
* Correspondence should be sent to Eric J. Ma

Version May 2, 2021 submitted to Entropy

Abstract: We present a case study applying hierarchical Bayesian estimation on high throughput protein melting point data measured across the tree of life. We show that the model is able to impute reasonable melting temperatures even in the face of unreasonably noisy data. Additionally, we demonstrate how to use the variance in melting temperature posterior distribution estimates to enable principled decision-making in common high throughput measurement tasks, and contrast the decision-making workflow against simple maximum-likelihood curve fitting. We conclude with a discussion of the relative merits of each workflow.

Keywords: hierarchical modelling; Bayesian statistics; probabilistic programming; high throughput measurements

1. Introduction

High throughput measurements are a staple of biological measurements. A wealth of literature exists for the statistical analysis of high throughput measurement data [1–3] However, to the best of our knowledge, the application of hierarchical Bayesian models in high throughput assay measurements is not widespread, with only a countably small number of papers leveraging hierarchical Bayesian methods [4,5]. Yet, the advantages of hierarchical models for estimation are well-known. For example, in [6], baseball players' performance estimates are regularized away from extreme values. Players with fewer replicate observations of their fielding statistics had estimates shrunk closer to the population mean, though as more replicate measurements are obtained, the regularization effect diminishes. This property of hierarchical models may be desirable in a high throughput measurement setting, providing a guardrail against being fooled by apparently desirable extreme measurements generated at random, or worse, through systematic error. Because hierarchical Bayesian estimation models require explicit, hand-crafted distributional assumptions, they also helps us avoid canned statistical tests where our data might not necessarily fit the test's assumptions (e.g., the t-test) [7]. We thus see a gap in the *application* of hierarchical Bayesian estimation in high throughput biological assay measurement.

Recently, a global protein 'meltome' was published, in which 41,730 proteins from across 16 species of life had their protein melting points measured in a high throughput fashion, spanning 1,114,710 data points that were released publicly [8]. Protein melting occurs when a protein unfolds under thermal stress; a protein's melting point is thus of interest to biological research. In ecological studies, protein melting points can reveal their host organisms' properties (such as its probable characteristic temperature range). For biomedical applications, having a high melting temperature (i.e., high stability) while maintaining activity is a desirable protein therapeutics property, this property relates to a protein's stability, which can be crucial for preserving a protein therapeutic until it is used.

The characteristics of this data are such that they mimic very closely the kind of measurement data generated in biological screening experiments used for drug hunting and protein engineering. Firstly, it is of the dose-response curve form; the melting temperature, being the mid-point of the curve, is analogous to the IC50 concentration values generated in chemical and enzyme screens. Secondly, it is measured in a high throughput fashion, with extremely high numbers of samples measured, and with every sample (here, a protein from an organism) being treated equally in bulk. Importantly, they only have a single replicate measurement on which inferences, and hence decisions, are to be made. This means that measurements are effectively *sparse* with respect to a sample; even though there may be anywhere from 6-10 measurements of a sample, these are not biological replicate measurements of the same sample. This mirrors very closely what we observe in high throughput screen measurements in drug discovery and high throughput protein engineering, where most samples are measured once, and prioritization decisions have to be made on the basis of these measurements. As such, we reasoned that we could use this public data to illustrate how hierarchical Bayesian models can be used to guide decision-making under uncertainty in a high throughput measurement setting. Where appropriate, we will contrast the Bayesian workflow against the traditionally used "separate curve fitting" with maximum likelihood estimation. [1]

1.1. Base Model Definition

The model provided in [8] for the melting temperature of a protein, as done by their measurement technique, is given by

$$f\left(L, a, b, T\right) = \frac{1-L}{1+e^{-\left(\frac{a}{T}-b\right)}} + L$$

Here, L is the lower bound of the melting curve, which is treated as a random variable to estimate; $\frac{a}{b}$ is the characteristic melting point of a protein, and T is the temperature at which a measurement is taken.

The modifications we made to the model to enable hierarchical modelling are detailed in the Materials and Methods section.

1.2. Data characteristics and summary statistics

The data that was provided by the authors are 1.1 million rows of measurements, which contain:

1. The species and extraction method that a protein was isolated from.
2. The protein ID (a unique identifier encompassing its gene name)
3. The gene from which the protein is expressed
4. The temperature at which an observation was taken
5. The fold change of detected stable protein at that temperature, relative to the level at the lowest measured temperature.

Of the 41,730 proteins that were measured, 11,142 of them had no melting point assigned. This is a direct result of the curve fitting protocol used in the original analysis [8], which included criteria for data quality checks.

---

[1] As we are focusing on the application of hierarchical Bayesian methods to high throughput measurements, we will only be discussing the relevant protein biochemistry in light detail. Any biological assumptions and conclusions we draw are held lightly; specific improvements to the model are discussed only in brief.

2. Validation of model implementation

2.1. Melting curves

To validate whether our model implementation was done correctly or not, we employed a few diagnostics. Firstly, we rank-ordered proteins by their melting temperatures and spot-checked a subset of curves. Proteins that had low posterior uncertainty generally had high quality curves, such as the blue curves in Fig. 1. In this regime, the estimated melting temperatures from separate curve fitting (dotted lines) generally agreed with the hierarchical Bayesian fits (solid line and shaded area), with minimal differences in the estimated melting temperature. As the posterior uncertainty increases, we gradually observe lower quality curves (yellow and red in Fig. 1. In particular, the red curves come from data that failed quality checks in the original analysis method, and hence did not have an assigned melting temperature. For the proteins that did have a melting temperature assigned, the magnitude of the difference between the hierarchical Bayesian estimate and the maximum likelihood estimate co-varies with the magnitude of the posterior distribution standard deviation, but is generally centered around zero (Fig. 2).

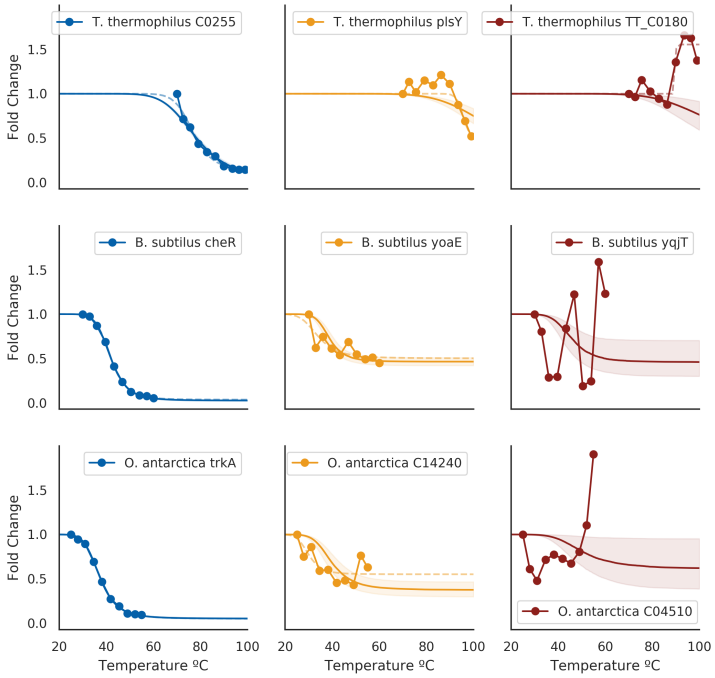

Figure 1. Example estimated melting curves against original measurement data for three species (by row). Blue figures are curves from proteins that had the lowest variance in estimated melting temperatures for each species. Yellow figures are curves from proteins for which melting points are not obvious from the data and did not have an assigned melting temperature, but nonetheless could plausibly be assigned one. Red figures are curves from proteins that exhibited highest variance in estimated melting temperature for each species. Dotted lines indicate separate curve fit using SciPy's curve fitting facilities (Materials and Methods); cases where errors were raised in curve fitting are omitted.

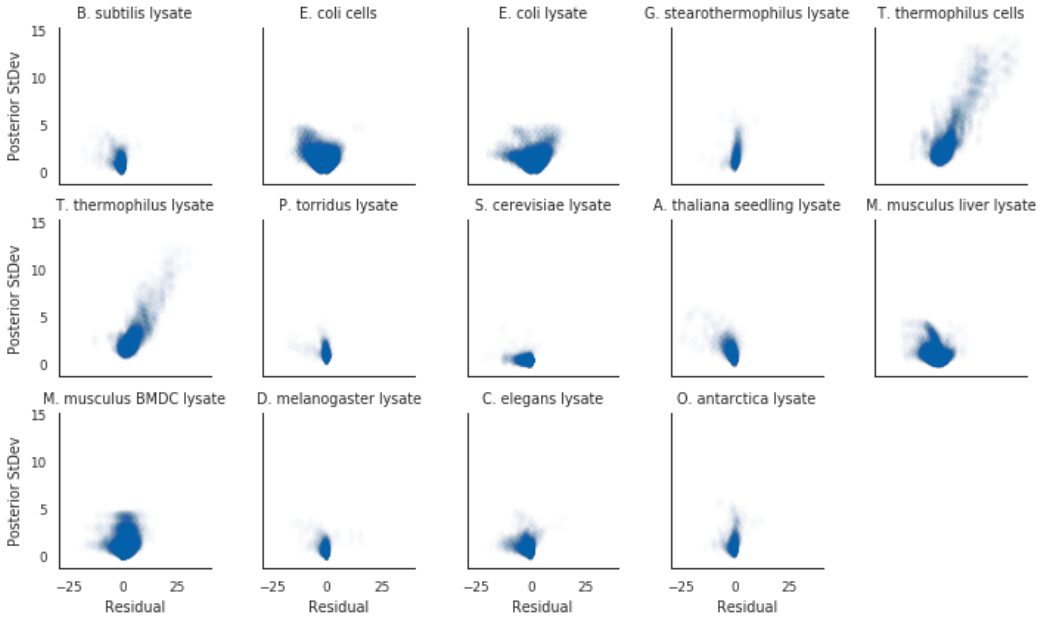

Figure 2. Proteins with greater discrepancies between the two methods for their estimated melting points also had greater uncertainties for their Bayesian estimated melting points. Residual is calculated by taking the Bayesian estimated melting point minus the separate curve estimated melting point.

2.2. Posterior variance

In the meltome paper, proteins that had melting curves that failed quality control checks did not get assigned a melting temperature. In a hierarchical Bayesian setting, we expect that these proteins' curves should give us high uncertainty in their posterior estimates. Indeed, in our model fits, we observe this phenomena. Where our model imputed a protein's melting temperature because of bad curve data (as assessed by the curve quality control checks), the posterior uncertainties for those melting temperatures were much higher than those without imputation (Fig. 3).

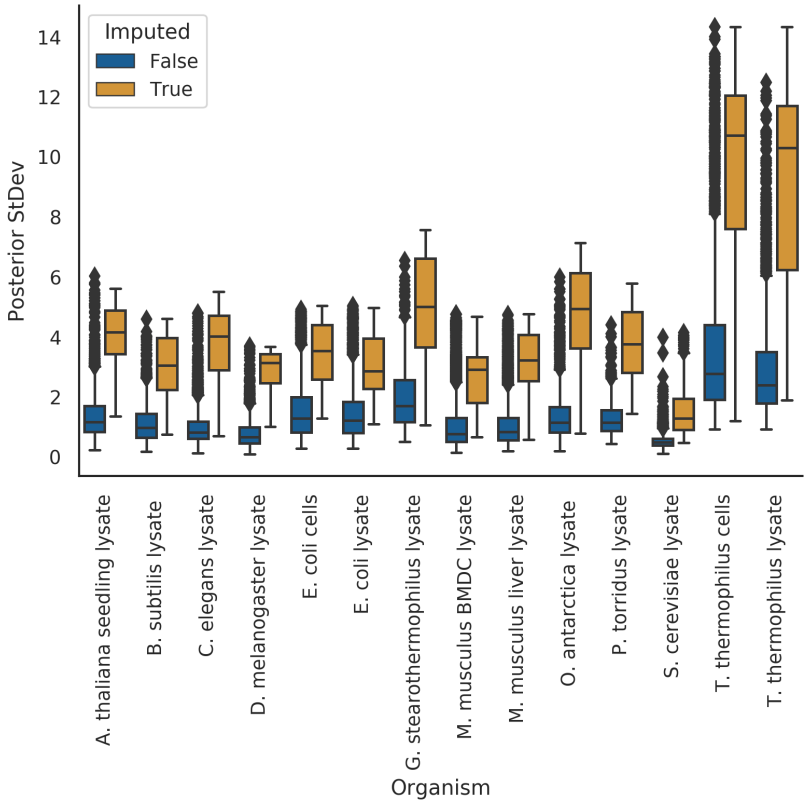

Figure 3. Imputed melting temperatures have higher uncertainty than non-imputed melting temperatures.

3. Principled, Bayesian decision-making in high throughput settings

In this section, we discuss what hierarchical Bayesian modeling enables when used in lieu of separate curve fitting.

In a setting where we measure high throughput dose response data, the goal is to find extreme-valued entities on the desirable side, such as low IC50 molecules for protein binders, or high stability proteins for stability. Some downstream questions that need to be answered generally fall into the categories of:

1. Which samples need *higher quality confirmatory measurements*?
2. Which samples should we *take forward for further investigation* in other measurement modes?

A classic constraint we find ourselves in is that of capacity: it is infeasible to take everything that is desirable. Additionally, it is desirable for us to have a ranking principle that factors in the confidence we have in any particular sample measurement. Leveraging the uncertainty in our estimates gives us a path towards principled decision making.

In the next two sections, we outline how to address the decision-making dilemma in a principled fashion.

3.1. Acquiring informative measurements

One of our goals might be to acquire re-measurements of samples to improve the quality of our data set. To do so, we could rank-order samples by their posterior uncertainty in their measurements. (Fig. 4 (a)) In our example, this would be samples that have the highest posterior variance in melting temperatures; in classic molecular screening settings, this may be samples with highest IC50 measurement uncertainties. Doing so would allow us to improve our data set quality in an iterative fashion, with uncertainty being the guiding principle for re-measurement.

3.2. Confirming optimal measurements

In a vein similar to the acquisition of more informative measurements, we may wish to acquire confirmatory or secondary measurements on samples that have measurement values greater than a particular threshold. An example where this shows up is in enzyme engineering. Our primary assay measurement may be cheap to conduct, but the confirmatory assay may be more expensive. In addition, we may be doing multi-objective optimization, and the secondary properties that we are optimizing for may be similarly expensive to measure. Made concrete, enzyme thermal stability is often a cheap and high-throughput measurement to acquire, while enzyme chiral activity requires more sophisticated instrumentation and is hence more difficult to conduct. In scenarios like this, our desired selection of samples are the ones that have the highest probability of being above some threshold we define *a priori*. (Fig. 4 (b) Here, access to the posterior distribution allows us to calculate the probability of a sample being greater than a particular value and thus rank-order all samples according to this principle.

3.3. Prioritizing samples for further modification

One other goal we might have would be to select samples from the pool of measurements as a baseline for further optimization. This is a classic protein and molecular engineering problem. Given the uncertainties in our measurements, by what principle could we select samples as our baseline?

This problem is similar to the previous section, where the desired outcome is the selection of the *best* new starting point. The difference lies in that in molecular and protein engineering, we are effectively optimizing, or searching, for samples that have *extreme* values. Higher activities, greater binding affinities, or smaller catalysis rates are what we are in search of. Hence, we once again desire samples that have the highest probability of being better than the rest as our new starting point. To calculate this, we would simply compare each sample's posterior estimates and calculate the probability of superiority w.r.t. other samples' estimates (Fig. 4 (c)). By contrast, under a classical setting, picking the sample with the most desirable point estimate would only give us samples that have the highest expected value, which does not help us explore extremities as effectively.

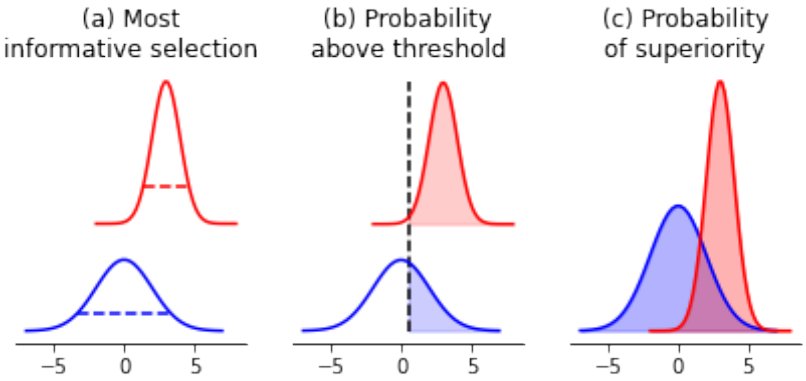

Figure 4. Probabilistic decision-making framework leveraging posterior distributions. (a) In choosing the next most informative re-measurement, we would suggest taking the blue sample because it has the highest uncertainty. (b) In choosing samples for confirmatory measurements that are above a threshold value defined *a priori*, we would suggest taking the red sample because it has the highest probability of being greater than a threshold value. (c) To decide which samples to use as a base for further modification towards extreme values, we would calculate the probability of superiority between all pairs of samples and identify the one that has the highest probability.

Run Name	Mean	StDev	Min	25%	50%	75%	Max
A. thaliana seedling lysate	43.8	1.5	34.6	43.1	43.9	44.5	49.6
B. subtilis lysate	43.7	3.0	36.8	41.9	43.8	44.8	58.4
C. elegans lysate	44.0	3.5	34.2	42.2	44.5	45.4	57.6
D. melanogaster lysate	43.3	2.6	39.2	41.9	42.5	43.7	54.6
E. coli cells	54.1	3.5	45.4	51.8	53.8	55.7	67.1
E. coli lysate	55.2	4.6	45.9	51.8	54.1	57.9	67.3
G. stearothermophilus lysate	81.9	5.7	59.7	77.7	81.3	85.8	97.3
M. musculus BMDC lysate	49.4	2.0	44.0	48.2	49.4	50.7	60.0
M. musculus liver lysate	51.0	2.2	44.4	49.7	51.0	51.8	64.1
O. antarctica lysate	48.8	4.5	36.5	46.0	47.5	51.3	63.7
P. torridus lysate	72.9	3.6	65.2	70.5	72.2	74.5	83.6
S. cerevisiae lysate	47.1	2.3	40.9	45.7	46.8	48.4	55.4
T. thermophilus cells	108.5	8.5	80.7	104.8	110.3	114.5	125.0
T. thermophilus lysate	107.3	8.4	79.2	101.8	109.0	113.4	125.0

Table 1. Summary statistics of the imputed melting temperatures, excluding *D. rerio*.

4. Discussion

4.1. Hierarchical Bayesian methods enable reasonable estimates where separate curve-fitting fails to provide one

The key analysis "safeguards" that the meltome authors used to call a non-melter, and hence assign no melting temperature to a protein, was a combination of a threshold value for the lower-bound and a normalized area under the curve value [8,9]. However, one may ask the question: do we expect proteins to "not melt" under increasing thermal stress? Using a hierarchical Bayesian model, we effectively express a prior in the model structure that all proteins do *eventually* "melt" under thermal stress; proteins with low quality measurements can still be assigned a melting temperature. As shown by our imputation distributions, the imputed values generally fall within the regime of This stands in contrast to the original statistical analysis protocol, in which this model-wise structural assumption cannot be encoded in the model, thus requiring external heuristics as a safeguard for "quality" of data. As for the question of which assumption is more reasonable, the answer would require debate and justification.

In the meltome atlas authors' statistical methods, data that did not fulfill quality control criteria were given a null value for their estimated melting temperatures. By contrast, our use of a hierarchical Bayesian model that baked in biologically-relevant priors enabled us to provide imputed melting points that fell within the general regime of those that *did* have melting points estimated (Table 1), Because we have the posterior distributions inferred, further statistical analysis can leverage this model-based multiple imputation [10]. In doing so, we have managed to preserve hard-won data points without discarding them, even if their quality were questionable, and we have a framework for deciding which to re-measure if we so desired.

A natural caveat to this imputation method is that the imputed values are only as good as the model's assumptions. For example, we do not factor in horizontal gene transfer or prior knowledge of a protein's known function, both of which might affect our estimates of that protein's melting point. We reiterate, as a reminder to the reader, that where the data quality are good, separate curve fitting and hierarchical Bayesian curve fitting will generally match up; one of the value propositions of hierarchical Bayesian curve fitting lies in *principled model-based imputation* where the data quality might *not* permit such a thing.

4.2. Limitations of our model and inferential procedure

We used approximate Bayesian estimation through ADVI [11] because of the large number of data points present. Though we ran it for a long number of steps (Materials and Methods), and could visually inspect for convergence loss, we could still be under-fitting. One other consequence of using

ADVI is that the posterior distributions will have an under-estimated variance [12]. Suppose we are concerned mostly with the expected values of our posterior distributions. In that case, we expect to run into few issues. By contrast, if we are concerned with leveraging the posterior distribution for input design problems [13], such as optimizing melting temperatures, where the extremities of the posterior distributions are important, then the use of ADVI would give us less extreme extremities.

For this particular model, further hierarchical structure could have been imposed on protein measurements that came from both cells and lysate, or from two tissues. For simplicity's sake, we treated these measurements as independent. However, incorporating this knowledge into the structure of the model would be an obvious next step to improve it for this dataset. By extension, similar knowledge may be incorporated into other estimation models for different experiments. However, we do not anticipate that this point interferes with the main point of our article, which describes a decision-making protocol in high throughput experimentation that leverages hierarchical modelling's advantages.

4.3. The promise of hierarchical Bayesian models in high throughput biological measurements

High throughput biological measurements are known to be extremely noisy. Small sample volumes, large numbers of samples, and measurement of non-control samples with single replicates, all contribute to the noise. The corollary is that these we may expect extreme-valued measurements to show up by pure random chance and systematic error. Even though in some settings we may wish to find extreme values, we desire not to be fooled by random and systematic error. With shrinkage and posterior analysis, hierarchical Bayesian modelling provides us with a principled way out.

In our case study, we have provided an example where in curve-fitting scenarios, such as dose response curves, regularization provided by hierarchical Bayesian models can act as a model-based safeguard against random extreme and noisy measurements, allowing us to provide a curve parameter estimate *despite* noisy measurements, while providing posterior uncertainties as a quality control measure. These posterior uncertainties can also help inform downstream experiments, such as deciding whether to re-measure a sample, re-measure groups of samples, or deciding which samples to take forward other measurements. We believe that developing hierarchical models in high throughput biological measurements may better leverage all available data and guide better iterative experimental design.

5. Materials and Methods

5.1. Hierarchical Bayesian Estimation Model

Here, we describe the model structure. A graphical representation of the model is available in Fig. 5

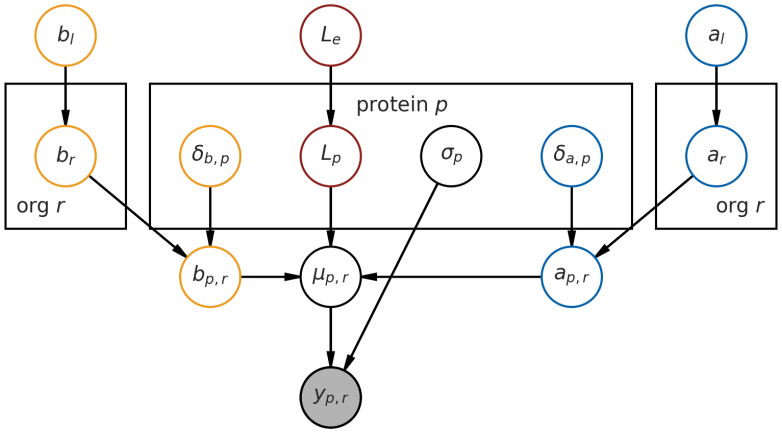

Figure 5. Graphical representation of the hierarchical model.

Our observed data, the fold change y across all proteins, are assumed to be Gaussian-distributed, with the central tendency μ for each observation given by the melting curve function f (which is a function of both temperature and the L, a and b parameters), and errors assumed to be homoskedastic.

$$y \sim \text{Normal}(f(L_{p,r}, a_{p,r}, b_{p,r}, T), \sigma_p)$$

Now, we describe the prior distribution assumptions that lead to each of the parameters.

L, the lower-bound, is given by a hierarchical prior, and is indexed on a per-protein basis, L_p, where $p \in \{0, 1, 2, ...\}$ indexes each protein. It is modeled by an Exponential distribution, with the population rate prior L_e also given by a relatively flat Exponential distribution. (Subscript e indicates *experiment*.) This parameter helps us capture a *global* measurement lower-bound for each experiment, as we know from the experiment description [8] that the experimental conditions primarily influence the lower-bound.

$$L_e \sim \text{Exponential}(\frac{1}{30})$$

$$L_p \sim \text{Exponential}(L_e)$$

Together, a and b give us the melting point $T_m = \frac{a}{b}$, and are indexed on a per-protein basis. However, to model the organism-wide melting temperature, we introduce a per-organism a_r and b_r, which are random variables modeled by a positive-only Normal distribution. (The subscript r is used instead of o for visual clarity, and r is the second letter of "organisms".)

Given the curve equation, a and b have to be positive for the curve to take on a denaturation shape (i.e. decreasing y as temperature increases.) We then model a shift in a and b that is indexed by protein $\delta_{a,p}$ and $\delta_{b,p}$, which shifts the value of a and b from the organism level to give us a per-protein $a_{p,r}$ and $b_{p,r}$.

Additionally, a_r and b_r are given population priors (a_l and b_l). (We use subscript l to denote that these are global priors across the tree of life.) Taken together, this model structure expresses that proteins from one organism most likely shares an underlying melting temperature distribution with other proteins from the same organism.

In mathematical notation, starting with the random variable a:

$$a_l \sim \text{Normal}(500, 1)$$

$$a_r \sim \text{BoundNormal}(a_l, 10)$$

$$\delta_{a,p} \sim \text{Normal}(0, 3)$$

$$a_p = a_r + \delta_{a,p}$$

And now for the random variable b:

$$b_l \sim \text{Normal}(10, 1)$$

$$b_r \sim \text{BoundNormal}(b_l, 1)$$

$$\delta_{b,p} \sim \text{Normal}(0, 1)$$

$$b_p = b_r + \delta_{b,p}$$

Their noise is assumed to be homoskedastic on a per-protein basis, i.e. σ_p, and is modeled with a standard HalfCauchy prior (with $\beta = 1$).

$$\sigma_p \sim \text{HalfCauchy}(1)$$

We caution that this model is bespoke for the meltome dataset; other datasets with different structural assumptions will require a different model.

The model was implemented in PyMC3 [14]. Because of the size of the data set, we used ADVI for 200,000 steps with default settings instead of the default NUTS sampling, yielding an approximation to the posterior distribution from which we drew 2000 samples.

5.2. High Throughput Measurement Data

High throughput measurement data were sourced from the Meltome Atlas' public-facing web server (at http://meltomeatlas.proteomics.wzw.tum.de:5003/) on 15 April 2020.

5.3. Separate curve fitting

Separate curve fitting was performed using the "curve fit" function in the "optimize" submodule of the SciPy library [15].

Funding: This research received no external funding.

Acknowledgments: We would like to acknowledge the developers of PyMC3 for providing a high quality probabilistic programming language in Python.

Conflicts of Interest: The authors are employed by the Novartis Institutes for Biomedical Research, and declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADVI	Automatic Differentiation Variational Inference

References

1. Zhang, J.; Chung, T.; Oldenburg, K. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J Biomol Screen* **1999**, *4*, 67–73.
2. Sui, Y.; Wu, Z. Alternative statistical parameter for high-throughput screening assay quality assessment. *J Biomol Screen* **2007**, *12*, 229–34.
3. Malo, N.; Hanley, J.; Cerquozzi, S.; Pelletier, J.; Nadon, R. Statistical practice in high-throughput screening data analysis. *Nat Biotechnol* **2006**, *24*, 167–75.
4. Wilson, A.; Reif, D.M.; Reich, B.J. Hierarchical dose–response modeling for high-throughput toxicity screening of environmental chemicals. *Biometrics* **2014**, *70*, 237–246.
5. Shterev, I.D.; Dunson, D.B.; Chan, C.; Sempowski, G.D. Bayesian multi-plate high-throughput screening of compounds. *Scientific reports* **2018**, *8*, 1–10.
6. Jensen, S.T.; Shirley, K.E.; Wyner, A.J. Bayesball: A Bayesian hierarchical model for evaluating fielding in major league baseball. *The Annals of Applied Statistics* **2009**, *3*, 491–520. doi:10.1214/08-aoas228.
7. Kruschke, J. Bayesian estimation supersedes the t test. *J Exp Psychol Gen* **2013**, *142*, 573–603.
8. Jarzab, A.; Kurzawa, N.; Hopf, T.; Moerch, M.; Zecha, J.; Leijten, N.; Bian, Y.; Musiol, E.; Maschberger, M.; Stoehr, G.; Becher, I.; Daly, C.; Samaras, P.; Mergner, J.; Spanier, B.; Angelov, A.; Werner, T.; Bantscheff, M.; Wilhelm, M.; Klingenspor, M.; Lemeer, S.; Liebl, W.; Hahne, H.; Savitski, M.; Kuster, B. Meltome atlas-thermal proteome stability across the tree of life. *Nat Methods* **2020**, *17*, 495–503.
9. Savitski, M.M.; Reinhard, F.B.; Franken, H.; Werner, T.; Savitski, M.F.; Eberhard, D.; Molina, D.M.; Jafari, R.; Dovega, R.B.; Klaeger, S.; others. Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **2014**, *346*.
10. Schafer, J.L. Multiple imputation: a primer. *Statistical methods in medical research* **1999**, *8*, 3–15.
11. Kucukelbir, A.; Tran, D.; Ranganath, R.; Gelman, A.; Blei, D.M. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research* **2017**, *18*, 1–45.
12. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* **2017**, *112*, 859–877. doi:10.1080/01621459.2017.1285773.
13. Brookes, D.; Park, H.; Listgarten, J. Conditioning by adaptive sampling for robust design; PMLR: Long Beach, California, USA, 2019; Vol. 97, *Proceedings of Machine Learning Research*, pp. 773–782.
14. Salvatier, J.; Wiecki, T.V.; Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* **2016**, *2*, e55. doi:10.7717/peerj-cs.55.
15. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; others. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **2020**, *17*, 261–272.