NON-PARAMETRIC ANALYSIS OF THERMAL PROTEOME PROFILES REVEALS NOVEL DRUG-BINDING PROTEINS

DOROTHEE CHILDS*[14#], KARSTEN BACH*[2,3], HOLGER FRANKEN*[4], SIMON ANDERS[5], NILS KURZAWA[1], MARCUS BANTSCHEFF[4], MIKHAIL SAVITSKI[1] AND WOLFGANG HUBER[1#§]

ABSTRACT. Determining the interaction partners of small molecules in the intact cellular context remains a key challenge in drug research. Thermal proteome profiling (TPP) addresses this need by inferring target engagement from changes in temperature-dependent protein denaturation at a proteome-wide scale. Existing approaches to TPP data analysis have been centered around a single parameter, the melting point. Target engagement is then identified from a compound induced change in this parameter. However, for a substantial number of proteins the melting point shift does not reflect the treatment effect well, or cannot be confidently estimated at all. To overcome these limitations, we present a non-parametric analysis of response curves (NPARC), a functional approach that compares entire curves instead of summary parameters. NPARC projects the data to a space of smooth functions and infers treatment effects by an F-statistic with degrees of freedom estimated from the data. We show that our method outperforms the currently used melting point-centric approach with regard to specificity and sensitivity on five independent datasets. NPARC reliably detected known cancer drug targets for which ligand binding was not reflected by melting point shifts. The proposed method works with arbitrary numbers of replicates, and comparisons can be defined in a flexible manner. We hope that the proposed approach will aid in the detection of novel targets and off-targets for drugs with to date unexplained mechanisms of action or side effects.

KEYWORDS

Drug Discovery, Proteomics, Functional Data Analysis

1. BACKGROUND

Determining the cellular interaction partners of drugs or chemical probes remains a key challenge in cell biology and drug research [1, 2, 3, 4]. Better assays for target engagement, off-target binding, and downstream effects would provide valuable information on drug mode of action, on potential reasons for side effects, and elucidate avenues for drug repurposing.

Thermal proteome profiling (TPP) addresses this need by enabling the detection of small-molecule binding events in living cells on a proteome-wide scale [5, 6]. TPP combines multiplexed quantitative protein mass

[1] EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg
[2] Department of Pharmacology, University of Cambridge, Cambridge CB2 1PD, UK
[3] Cancer Research UK Cambridge Cancer Centre, Cambridge CB2 0RE, UK
[4] Cellzome, GlaxoSmithKline, Meyerhofstraße 1, 69117 Heidelberg
[5] Center for Molecular Biology of University of Heidelberg (ZMBH), Heidelberg, Germany
[#] Correspondence: dorothee.childs@embl.de (D.C.), wolfgang.huber@embl.de (W.H.)
[§] Lead contact: wolfgang.huber@embl.de (W.H.)

spectrometry with the cellular thermal shift assay (CETSA) [7], which identifies binding events as shifts in protein thermostability (see Fig. S1 for a detailed explanation). A typical TPP experiment generates temperature dependent abundance measurements for a large fraction of the cellular proteome. Drug binding events can then be inferred from comparing the melting curves of samples with drug versus vehicle (negative control without drug) treatment.

TPP has been successfully applied to identify previously unknown protein-ligand interactions [5], protein complexes [8], and downstream effects of drug binding in signaling networks [6, 9, 10, 11] in human cells, but has recently also been extended to bacteria [12]. There is an increasing interest in the technology, including both its experimental and data analytic aspects, in order to identify relevant proteins in the data and their interactions in the living system [13, 14, 15, 16].

From the data-analysis perspective, the typical goal of a TPP experiment is to quantify the evidence for changes in thermostability. The present statistical workflow [5, 6, 17, 18] fits sigmoid melting curves to temperature-dependent concentrations for each protein. The fitted curves imply estimates of the melting point (T_m), and this single number summary is used to test for differences between control and treatment conditions (Fig. 1A) [6]. We refer to this published strategy as 'the T_m-based approach'. While this approach is intuitive and the estimated T_m-shifts yield a readily interpretable output, we show that it leads to a needless loss of sensitivity when applied in the high-throughput setting of TPP.

The main reasons for this sensitivity loss are two-fold: First, drug induced effects on thermostability are not always reflected by shifts in the fitted T_m (Fig. 1B-C). Second, some T_m values lie outside of the measured temperature range, which impairs their estimation (Fig. 1D). Both scenarios can result in important targets being missed in the analysis (Fig. 1E).

Furthermore, the z-test used in the T_m-based approach does not incorporate the goodness-of-fit of the sigmoid model into the test statistic. However, such an assessment is important because of the possibility that a protein does not meet the modeling assumptions of sigmoid curve shape in the cellular environment. To reduce the risk of false discoveries related to such proteins, Savitski et al. [5] had employed *a priori* filters on the melting curve shape of each protein prior to computing the test statistic (Table 4). A comparison of the effects of these filters on five published experiments revealed that they discarded between 14-27% of all proteins in each dataset (Fig. 1E), a significant reduction in the 'testable' part of the proteome (Fig. S2).

Here, we overcome these limitations through a nonparametric analysis of response curves (NPARC), which is independent of T_m estimation. NPARC applies concepts from functional data analysis by approximating the data generating function underlying the measurements and by performing hypothesis testing based on the fitted functions instead of individual parameter estimates. This offers a twofold benefit compared to the present methodology. First, we test on differences in curves rather than differences in T_m, thereby using more information directly rather than first lossfully compressing it into a single variable. Second, the model is more robust against deviations from the expected sigmoid shape. Thus, we increase the coverage of the proteome, including biologically and pharmaceutically relevant proteins. Because melting behavior can be modeled without imposing any specific curve shape *a priori*, the model could also in principle be adapted to more complex melting curves [19, 20, 21].

We perform statistical assessment of the improvements in goodness-of-fit by an F-statistic. Other types of test statistics exist in functional data analysis applications and would be possible in this scenario. For example, Maldonado et al. [22] use the sample correlation coefficient between curves and combine it with permutation

tests, Storey et al. use an F-like statistic (omitting scaling by degrees of freedom) [23], and Berck et al. derive a moderated t-statistic in combination with smoothing-spline mixed-effects-models [20]. All these approaches share the problem that the null distribution is not analytically tractable in curve data that exhibits dependent measurements. The mentioned approaches rely on bootstrapping or permutations in order to control the false discovery rate (FDR). However, this is was not feasible for a typical TPP experiment due to the low numbers of biological replicates. Instead, we construct an F-statistic for which we can borrow information across proteins. In particular, we propose to estimate distribution parameters in order to approximate the χ^2-square mixture distributions incorporated into this statistic.

Using the five published datasets introduced in Fig. 1E and Table 1, we show that NPARC improves specificity and sensitivity compared to the established parameter-centric approach. Among these datasets, three experiments were conducted in intact cells treated by the cancer drugs panobinostat [6] or dasatinib in different concentrations [5]. Furthermore, two experiments were conducted in which cell lysates were treated with the high-affinity, ATP-competitive pan-kinase inhibitor staurosporine [5], or with the cellular metabolite ATP [9] (see Table 1 for details).

2. Results

2.1. **Non-parametric analysis of response curves.**

Our approach is based on fitting two competing models to the data, a null model and an alternative model. The null model states that the relative concentration of a protein at increasing temperatures t is explained by a single smooth function $\mu(t)$ irrespective of treatment condition and replicate (Fig. 2A). The deviations between data and the fitted model are referred to as residuals. The function $\mu(t)$ is fitted to the data in such a way that the sum of squared residuals (RSS) is minimized. The alternative model replaces this common function by condition-specific functions: $\mu_T(t)$ for the treatment group, and $\mu_V(t)$ for the vehicle group (Fig. 2B). It is fitted by minimizing the sum of squared residuals of the data in each group. The approach could be generalized to more complex experimental designs, as is often done in analysis of variance (ANOVA) or linear regression.

2.1.1. *Hypothesis testing in a functional framework.*

In order to detect proteins with different melting curve shapes between treatment conditions, we construct an F-statistic that enables testing for significant improvements in goodness-of-fit of the alternative model relative to the null model (Fig. 2C). Conceptually, the F-statistic quantifies the relative reduction in residuals from null to alternative model. By definition, F is always positive, but for proteins not affected by the treatment, we expect values of F close to zero, whereas higher values indicate that thermostability was affected by the drug treatment. For example, as shown in Fig. 2B, serine/threonine protein kinase 4 (STK4) is strongly shifted by staurosporine treatment. Its F-statistic is among the highest ones in the dataset, as highlighted by the arrow in Fig. 2C. The F-statistic serves as base for computation of p-values and control of the false discovery rate (FDR) [24].

2.1.2. *Choice of mean function.*

The mean functions $\mu(t)$, $\mu_T(t)$ and $\mu_V(t)$ need to be chosen from a space of smooth functions. There are various options for choosing these functions which can be conceptually divided into two categories: (1) nonparametric smooth functions that allow flexible shapes, or (2) nonlinear models that impose shape constraints based on prior knowledge. Under idealizing assumptions often made in protein

thermodynamics, the thermal degradation of a protein follows a sigmoid trend [5], and we thus parametrize the mean functions by the three-parameter equation

$$f(t) = \frac{1 - f_\infty}{1 + e^{-(\frac{a}{t} - b)}} + f_\infty. \tag{1}$$

2.1.3. *Estimating the null distribution.* In order to compute p-values, we need to compare the observed test statistics to their expected distribution under the null hypothesis (H_0). If we could assume independent and identically normally distributed residuals, this distribution would be an F-distribution whose degrees of freedom (DOF) only depend on the numbers of measurements and model parameters. For curve data with multiple measurements per protein, the test statistics still approximate an F-distribution under H_0, but the DOF need to be adjusted because of correlations between residuals. Hence we derived an approach to estimate the DOF for all proteins from the empirical distributions (See Methods for details).

2.2. **Application to cancer drugs.** To assess the capability of NPARC to detect expected drug targets, we analysed a dataset on panobinostat (Table 1). Panobinostat is a broad-spectrum histone deacetylase (HDAC) inhibitor known to interact with HDAC1, HDAC2, HDAC6, HDAC8, HDAC10, and tetratricopeptide repeat protein 38 (TTC38).

Out of 3649 proteins reproducibly quantified across both biological replicates and treatment groups, 15 proteins yielded Benjamini-Hochberg adjusted p-values ≤ 0.01. All five expected HDAC targets were among the top ranked NPARC hits (Fig. 3A-E). Besides known HDACs, the strongest effects on thermostability were observed for TTC38, the histone proteins H2AFV or H2AFZ (the two variants could not be distinguished by mass spectrometry), and zinc finger FYVE domain-containing protein 28 (ZFYVE28) (Fig. 3F-H). All these proteins have been previously identified as direct or indirect targets of panobinostat [6, 11].

In principle, these results coincide with a T_m-based ranking on the same dataset [6]. However, the advantage of NPARC becomes apparent for HDAC1 and HDAC2. For both proteins, not only the thermostability, but also the variance of the measurements appears to be affected by the treatment. This impairs the accuracy of the parameter estimates of μ_T. For this reason the requirements placed on the goodness-of-fit by the T_m-based approachdiscard HDAC1 and HDAC2 from the analysis regardless of their high effect sizes. Instead of relying on a single variable, T_m, derived from these estimates to summarize the treatment effect, the F-test enables weighting model uncertainty against the magnitude of the effect along the whole curve. Furthermore, by including additional covariates in the alternative model, the F-statistic can accommodate more complex experimental design, such as treatments with different drug concentrations. This is illustrated by the identification of known targets of the BCR-ABL inhibitor dasatinib (Figure S4).

2.3. **NPARC enables the detection of treatment effects not reflected by changes at the melting point.** In the analysis of the staurosporine data, we observed several cases where treatment effects were not well reflected by T_m-shifts. To assess the capability of NPARC to detect such proteins, we compared the targets detected by NPARC to those obtained by the T_m-based approach (Fig. 4).

Among the proteins only detected by NPARC, the RanGTP binding tRNA export receptor exportin-T (XPOT) exhibited the smallest T_m-shift followed by two members of the p38 MAPK signaling pathway: Mitogen-activated protein kinase 14 (MAPK14) and MAP kinase-activated protein kinase 2 (MAPKAPK2). Despite a small effect size, the proteins reached statistical significance due to high reproducibility of the observation.

The advantage of NPARC is also exemplified by protein kinase C beta (PRKCB), a member of the protein kinase C (PKC) family and destabilized upon target engagement. Members of the PKC family were the first reported staurosporine targets [25, 4]. The effect is reflected by reproducible changes in curve shape before the melting point (Fig. 1C) which can be successfully detected by NPARC. Interestingly, several other PKC members also display a destabilization prior to the T_{m} (Fig. S5), highlighting again that the T_{m} is an unsuitable summary parameter for the treatment effect in this protein family. In a similar manner, NPARC is able to identify the small but reproducible shift caused by dasatinib treatment on Bruton tyrosine kinase (BTK), an important drug target in B-cell leukemia (Fig. 1B).

2.4. **NPARC enables the detection of targets with missing melting point estimates.** In each dataset analyzed for this work, we observed a subset of proteins for which the T_{m} in one or both treatment conditions was outside of the tested temperature range (Fig. 1E). One example is NAD(P)H quinone dehydrogenase 2 (NQO2), a cytosolic flavoprotein and a common off-target of kinase inhibitors [26, 27, 28]. Previous CETSA studies have described NQO2 to be highly stable [29]. This is in concordance with our data, where we observe denaturation under normal conditions only beginning at 67 °C (Fig. 1D). Staurosporine treatment further stabilizes NQO2 to an extent that it shows no sign of melting in the tested temperature range. Theoretically, one could predict the T_{m} by extrapolating the fitted mean functions $\mu_T(t)$ and $\mu_V(t)$. However, their estimates become increasingly unreliable with increasing temperatures. Instead, the curve-based comparison by NPARC allows the assessment of effects directly observed in the measured temperature range.

2.5. **NPARC exhibits increased sensitivity compared to the T_{m}-based approach .** So far we have illustrated that the proposed NPARC approach allows the assessment of effects that were previously not identifiable. In order to assess whether this improves the overall performance, we compared its ability to detect targets with high specificity and sensitivity to the T_{m}-based approach on the staurosporine data and the ATP data. For both compounds, comprehensive lists of expected targets were available based on Gene Ontology annotation (see Table 1 for details).

In both datasets, NPARC enabled the detection of more known targets for a given number of unexpected hits compared to the replicate-wise z-tests for both datasets (Fig. 5). When applying the filters in Table 3, the performance of the T_{m}-based approach becomes similar to NPARC, particularly in the staurosporine dataset. Yet, NPARC exhibits a higher sensitivity than the T_{m}-based approach while sustaining equal specificity. Additional benefits are its ability to rank protein according to effect strength and that it does not rely on excluding proteins from the test by a separate filtering step.

3. Discussion

Thermal proteome profiling offers the possibility to comprehensively characterize ligand-protein interactions on a proteome-wide scale in living cells. However, the method poses the analytical challenge of how to identify statistically significant shifts in thermostability amongst thousands of measurements. In this work, we show that previous analytical workflows, relying on T_{m} estimation, suffer from needless loss of sensitivity. In particular, we demonstrate that in the high-throughput setting of TPP experiments, T_{m} is an insufficient summary of treatment effects. We illustrate examples of proteins whose observed melting curves are affected by drug treatment without substantially affecting the T_{m}. In addition, we highlight cases with high thermostability for which the T_{m} is located outside of the tested temperature range, impairing confident estimation.

In order to tackle this challenge, we introduce a functional approach to test for treatment effects. NPARC compares two competing models by their goodness-of-fit instead of testing on a single parameter. This enables detection of treatment effects for proteins where a (de-)stabilization cannot be captured by the T_{m}. In addition, it increases the coverage of testable proteins and allows ranking of proteins by p-values without the ad hoc thresholds on which the original workflow relied [5]. Its flexibility allows adaptation to more complex experimental designs by including additional covariates, or by allowing multiple levels (e.g. drug concentrations) per covariate.

The suggested framework is also flexible with regard to the mean function used to represent the melting behavior and can be adapted to the particular biological effect of interest. For nonlinear regression problems, spline-based regression [23, 20] or nonlinear regression are most commonly employed. We choose the latter as it incorporates *a priori* knowledge about the data and is thus able to provide a better representation. For example, sigmoid curves require horizontal asymptotes at plateau regions. These are typically poorly represented by splines, which tend to overfit data in these regions.

In a cellular environment we occasionally observe non-sigmoid melting curves for subsets of proteins. A possible reason is the presence of protein subpopulations with distinct melting curves [13]. For example, the formation of protein complexes, the binding to other molecules, or the localization in cellular compartments can lead to deviations from the idealized sigmoid melting curve expected from the same protein in purified form. Our model currently does not account for such systematic and reproducible shape deviations. This could be adapted in future work by adding a low-parametric systematic modification to the sigmoid mean function.

We approximate an 'average' null distribution for all proteins by estimating the DOF of the F-statistic numerator and denominator from the respective observed χ^2-distributions. In doing so, we assume that the empirical distribution of all test statistics is an F-distribution under H_0 whose parameters are accessible by numerical optimization. Nonetheless, we could show that the current approach offers very good sensitivity and reliable and robust specificity for several example studies, including ones with a very specific and promiscuous binder.

4. Conclusions

We present a functional approach to compare response curves like those arising in TPP experiments. By mapping the data to a space of smooth functions and testing on these functions, it uses information from the whole curve and does not rely on further abstraction of these functions to summary parameters like melting points. This enables the detection of a wider range of treatment effects that leads to changes in curve shape, but not the estimated melting point. By these means, it makes better use of information contained in the data than the established statistical workflow and finds a larger fraction of biologically or pharmacologically relevant target proteins. The method is easy to use and all necessary parameters are derived from the data, removing the need for manual tuning. We are currently finalizing the R-package NPARC, which will soon be released as an open-source Bioconductor package.

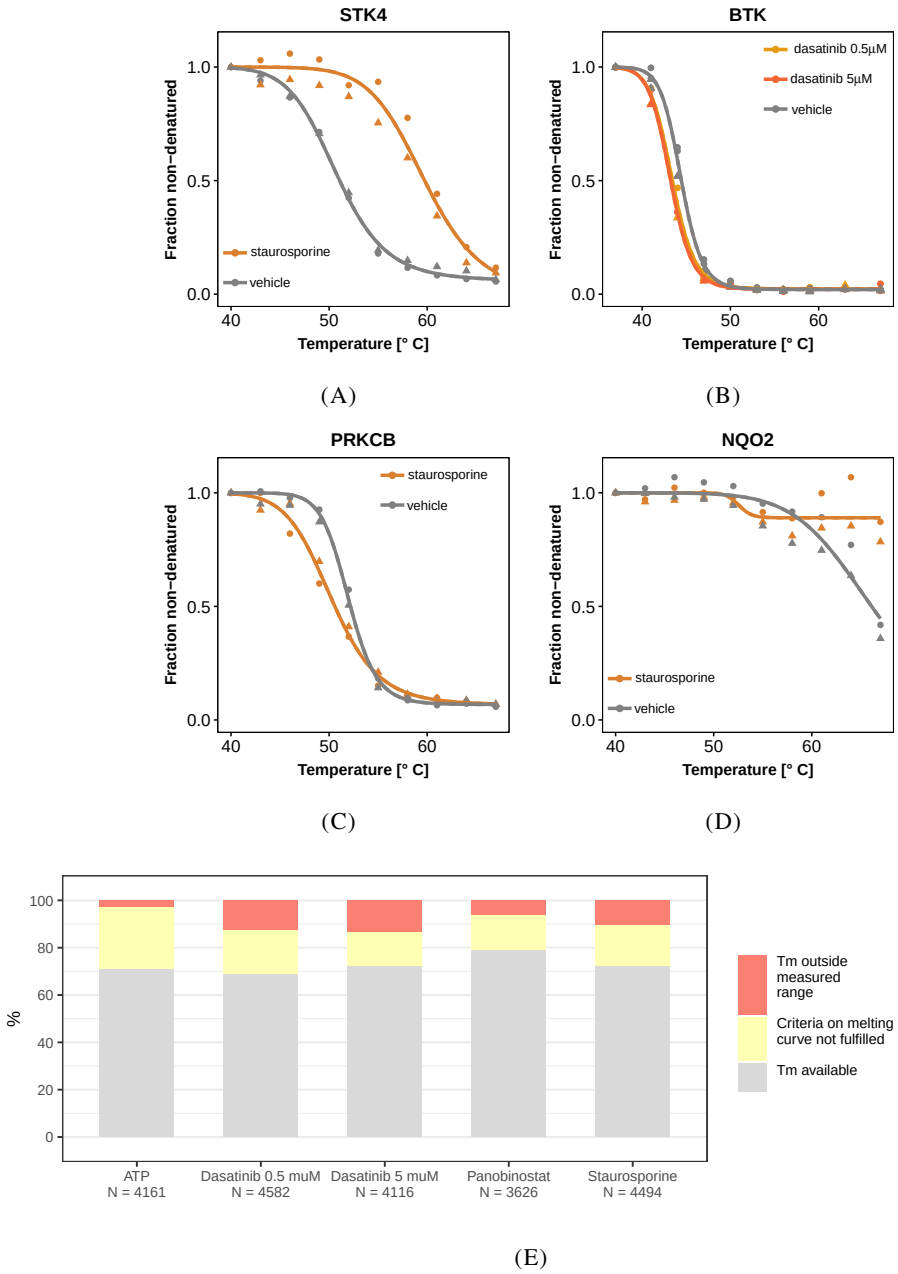


FIGURE 1. Exemplary melting curves. **(A)** Serine/threonine protein kinase 4 (STK4) satisfies the model assumptions of the T_{m}-based approach. **(B)**-**(D)**: Cases in which the T_{m}-based approach fails to detect the treatment effect. **(B)** Bruton tyrosine kinase (BTK) shows a small but highly reproducible T_{m}-shift that does not reach statistical significance when testing on this parameter. **(C)** Protein kinase C beta (PRKCB) is destabilized by staurosporine but the main effect occurs at temperatures lower than the estimated T_{m}. **(D)** NAD(P)H quinone dehydrogenase 2 (NQO2) is stabilized by staurosporine with the T_{m} being shifted outside the measured temperature range. **(E)** In the analysis with the T_{m}-based approach of several data sets (Table 1), a substantial fraction of proteins cannot be analysed due to failure to estimate T_{m}-values or to meet goodness-of-fit criteria (Table 4).

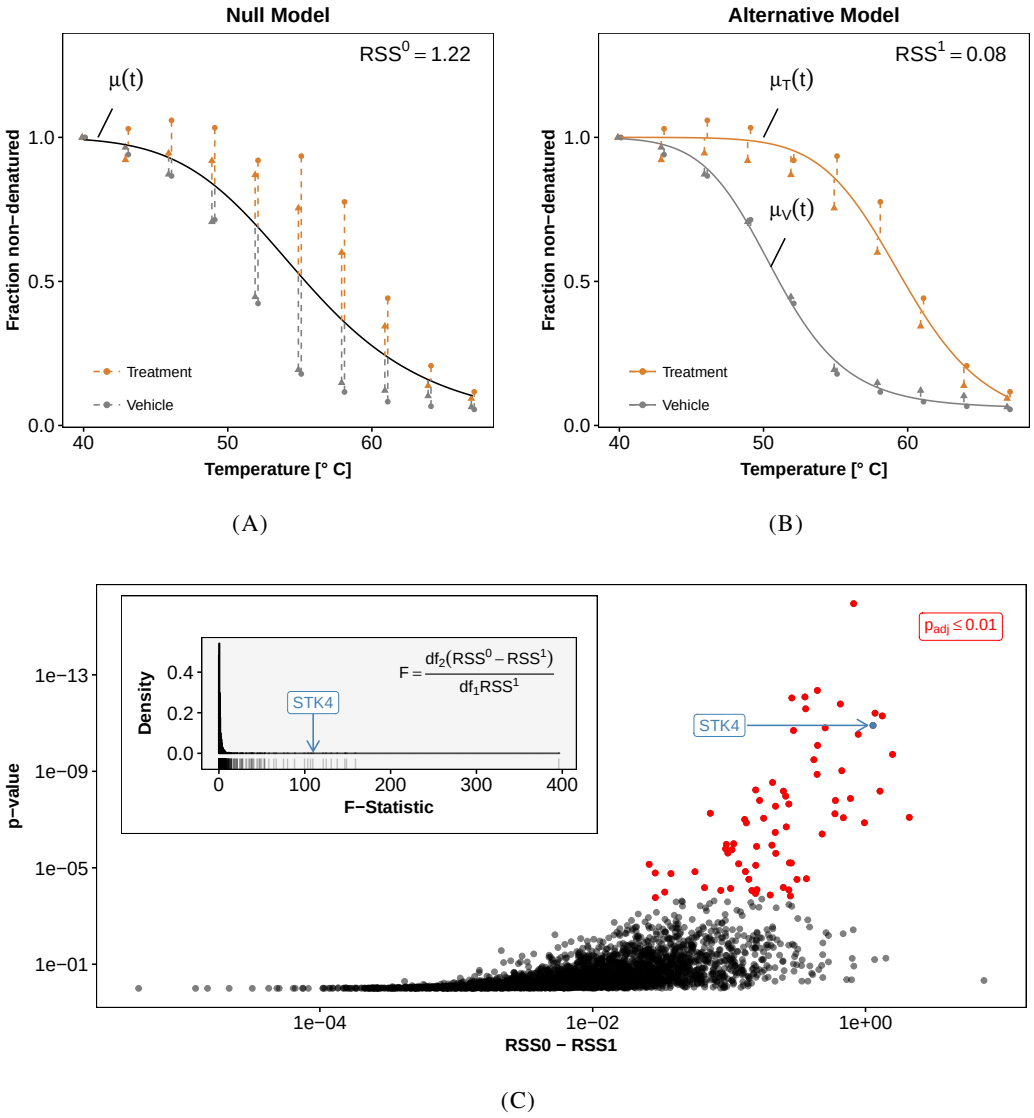


FIGURE 2. Principles of nonparametric analysis of response curves (NPARC), illustrated for the protein STK4 in the staurosporine data set. **(A)** A null model is fitted to all experiments under the assumption of no treatment effects (black line). The goodness-of-fit is quantified by the sum of squared residuals (dashed lines) (RSS^0). **(B)** In the alternative model separate curves are fitted to the treated (orange) and the vehicle group (grey). The sum of squared residuals in the alternative model (RSS^1) is less than or equal to RSS^0. **(C)** The improvement in the goodness-of-fit is reflected by the reduction of the RSS ($\mathrm{RSS}^0 - \mathrm{RSS}^1$) and can be assessed for statistical significance with an F-test. A significant reduction of the RSS, as exemplified by STK4, yields a high F-statistic and a low p-value.

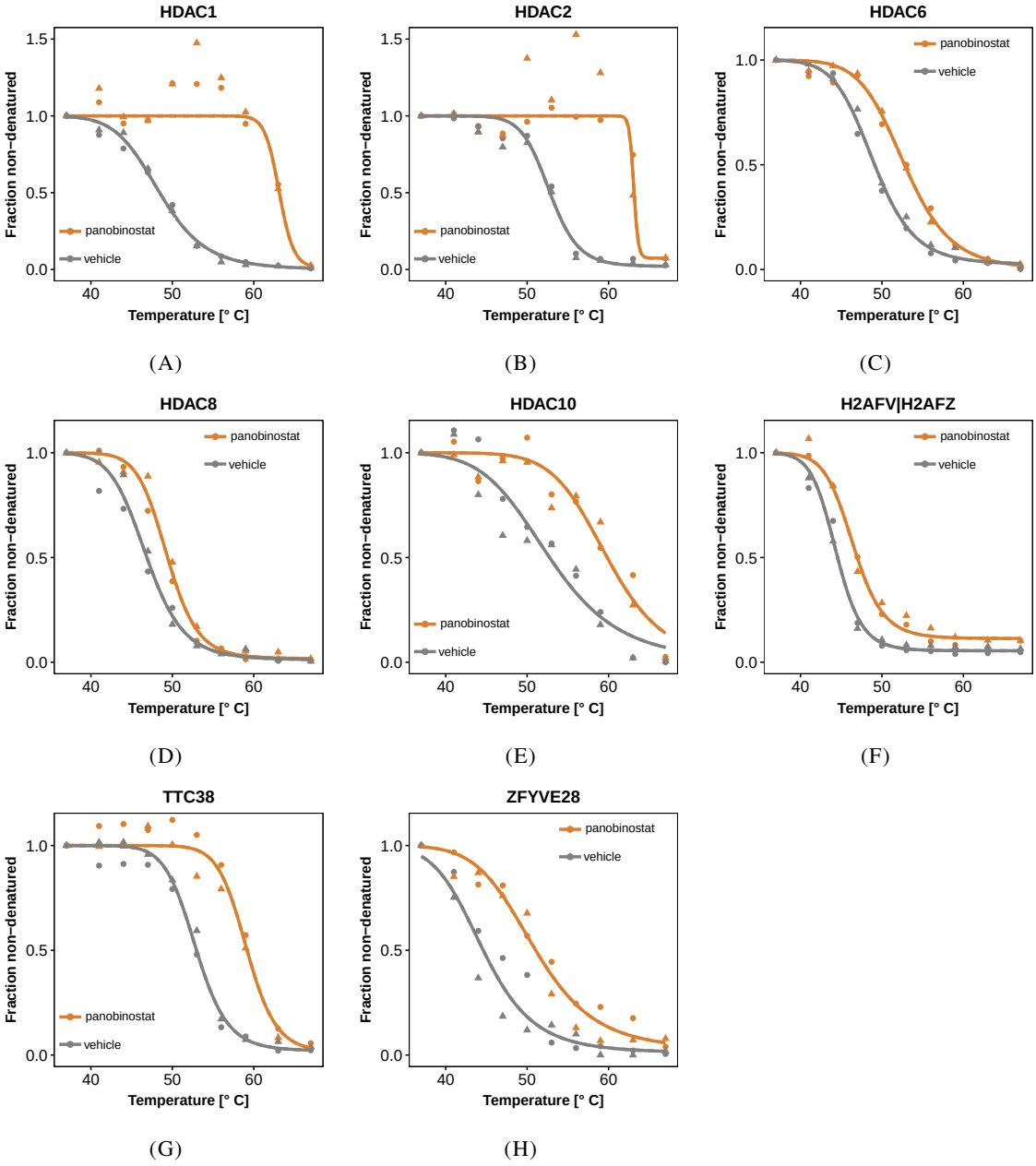


FIGURE 3. NPARC correctly identifies direct and indirect targets of the HDAC inhibitor panobinostat. **(A)** - **(E)**: Melting curves for several HDACs that show significant shifts in their thermostability. As illustrated by HDAC1 and HDAC2, drug treatment can affect the variance of the measurements so that the proteins do not pass the stringent requirements of the T_{m}-based approach. NPARC takes this variance into account when constructing the test statistic and does not require additional filtering steps. **(F)** - **(H)**: Melting curves for known non-HDAC targets that were identified by NPARC. All proteins were detected with Benjamini-Hochberg adjusted p-values ≤ 0.01. Seven more proteins were detected at the same p-value cutoff (Fig. S3).

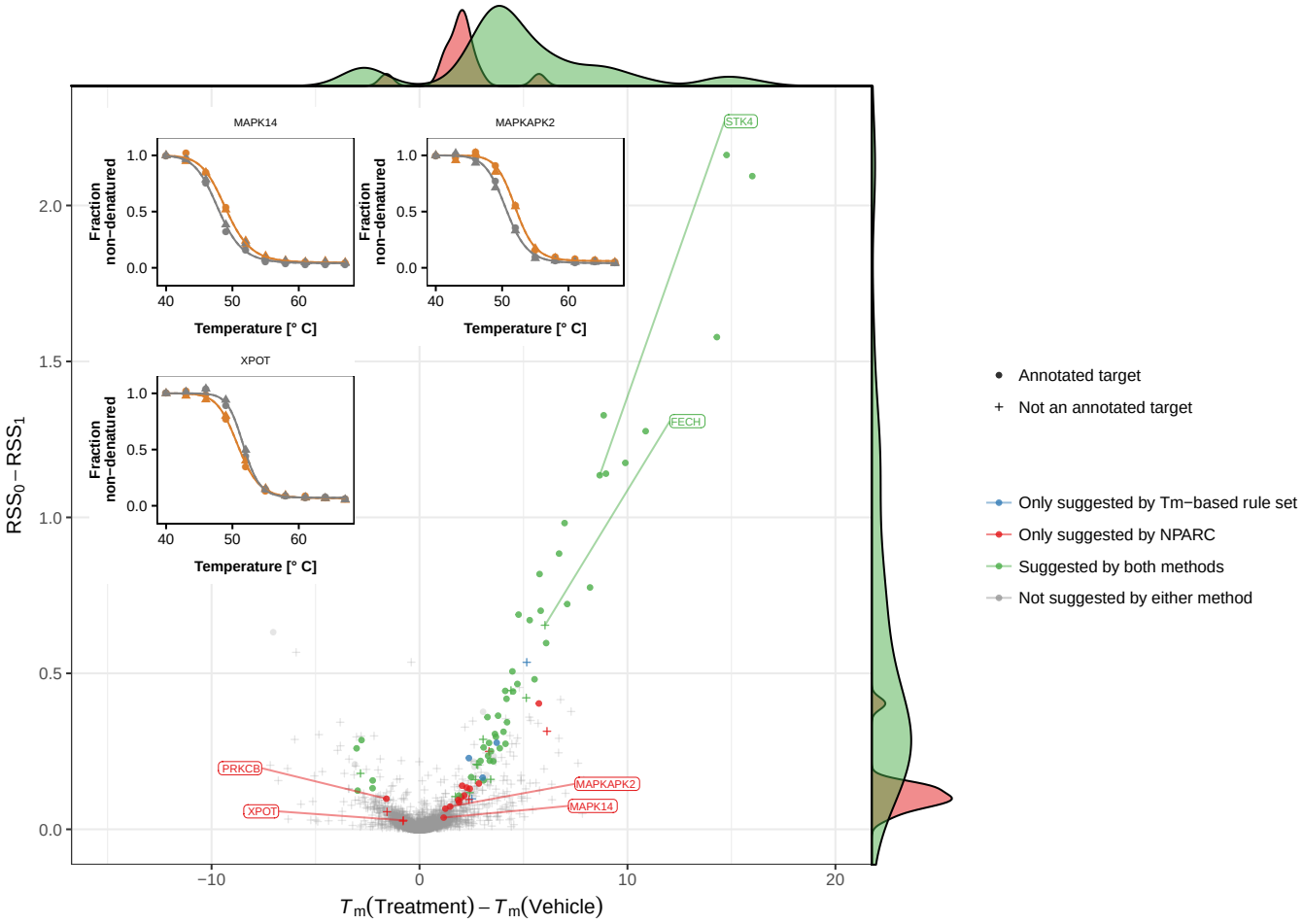
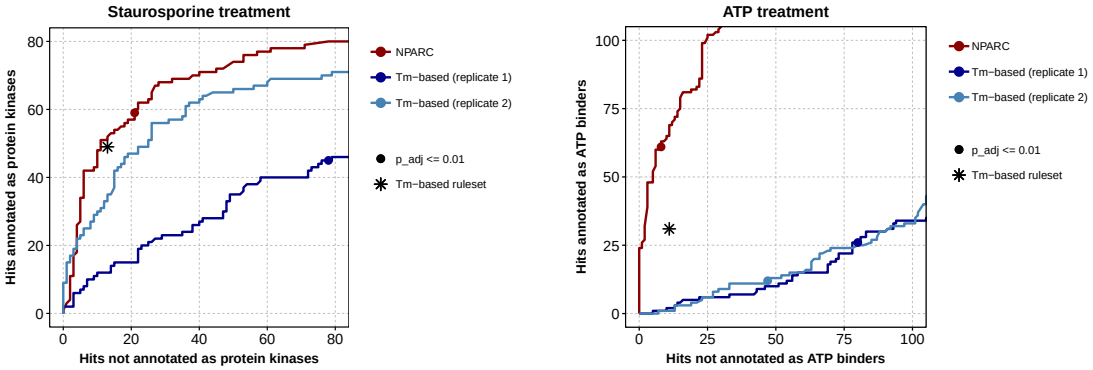


FIGURE 4. NPARC has higher sensitivity to small but reproducible T_{m}-shifts. The plot shows the effect sizes used for significance assessment by NPARC and by the T_{m}-based approach for all tested proteins in the staurosporine dataset. Proteins with Benjamini-Hochberg adjusted p-values ≤ 0.01 are marked in red if they were exclusively found by NPARC, or green if they were also detected by the T_{m}-based approach. NPARC can detect targets with smaller T_{m}-differences as long as the measurements are highly reproducible between replicates. Only proteins for which a T_{m} could be estimated are displayed in the plot.



(A) (B)

FIGURE 5. Pseudo-ROC curves with absolute values on both axes for the staurosporine data **(A)** and the ATP data **(B)**. The curves show the expected hits (y-axis) and unexpected hits (x-axis) for varying p-value cutoffs of NPARC and the z-test on T_{m}, which is the test utilized by the T_{m}-based approach. Asterisks indicate the results obtained by the T_{m}-based approach when applying the filters in Table 3 on the z-test results. The dots indicate a threshold of 0.01 on the Benjamini-Hochberg adjusted p-values. While the T_{m}-based approach is similar to NPARC in the staurosporine dataset, it shows clear benefits on the ATP dataset by enabling the detection of substantially more ATP binders at a given rate of non-ATP binders. All proteins found by NPARC at a Benjamini-Hochberg adjusted p-value ≤ 0.01 are also shown in the Supplementary Figures S6 and S7.

5. METHODS

5.1. **Datasets and preprocessing.** The performance of NPARC was assessed by re-analyzing five published TPP data sets (Table 1). Abundance measurements per protein and temperature were extracted from the supplements of the respective publications [5, 6, 9]. All measurements were scaled per protein and replicate to the abundance at 37°C (the lowest of the ten temperatures at which denaturation was assessed), and subjected to the global normalization procedure described by Savitski et al. [5].

Only proteins reproducibly quantified with at least one unique peptide in the vehicle and compound treated groups of both replicates were included in the analysis (see Table 1 for the final sample sizes).

TABLE 1. Datasets and sample sizes.

Dataset	Treatment	Concentration	Buffer	Cell line	Intact cells or lysate	Proteins	Reference
ATP data	MgATP	2 µM	PBS	K562	Lysate	4177	[9]
Dasatinib 0.5 µM data	Dasatinib	0.5 µM	PBS	K562	Intact Cells	4625	[5]
Dasatinib 5 µM data	Dasatinib	5 µM	PBS	K562	Intact Cells	4154	[5]
Panobinostat data	Panobinostat	1 µM	PBS	K562	Intact Cells	3649	[6]
Staurosporine data	Staurosporine	20 µM	PBS	K562	Lysate	4505	[5]

5.2. **Assignment of expected targets.** In order to obtain objective criteria for the comparison of test sensitivity and specificity, expected targets for the pan-kinase inhibitor staurosporine and ATP were assigned by Gene Ontology (GO) annotation, using the Bioconductor annotation packages 'AnnotationDbi' (version 1.36.2), 'org.Hs.eg.db' (version 3.4.0), and 'GO.db' (version 3.4.0). The resulting numbers of assigned targets are shown in Table 2.

TABLE 2. Expected targets per dataset.

Dataset	GO term	Proteins with GO term in dataset
ATP data	'ATP-binding'	558
Staurosporine data	'protein kinase activity'	187

5.3. **Mathematical model.** We denote with y_{ijk} the relative abundance measurement for protein i, temperature t_j and experiment k. Each experiment can be assigned to a treatment group c (e.g. $c = \mathrm{T}$ for the compound-treated group and $c = \mathrm{V}$ for the vehicle-treated group). Each experiment yields measurements for N temperature points.

We regard the set of measurements $\mathbf{y}_{ik} = [y_{i1k}, y_{i2k}, \dots y_{iNk}]$ as realizations of a stochastic process $\{Y_{ic}\} = \{Y_{ic}(t) : t > 0\}$. This process is specific for each protein and treatment group, and consists of random variables $Y_{ic} : \mathbb{R} \to \mathcal{L}^2$ defined by a treatment-specific smooth mean function $\mu_{ic}(t) : [t_{\text{min}}, t_{\text{max}}] \to \mathbb{R}_{\geq 0}$ and a covariance function $\gamma_{ic}(s, t)$.

The covariance function defines the random fluctuation of each process realization $Y_{ic}(t)$ around the mean function $\mu_{ic}(t)$. The variances $\gamma_{ic}(t, t)$ consist of measurement errors and systematic fluctuations whose extent depends non-monotonically on temperature. This is caused by the fact that the variation is higher at steep areas of the curve than at the plateaus (heteroscedasticity across temperatures).

Motivated by basic protein thermodynamics, we choose Equation (1) for the mean function. It reflects an idealized melting behavior that should well approximate a majority of proteins in the sample [30]. In this

model, $a, b \geq 0$ control the shape of the curve and $f_\infty \in \mathbb{R}_{\geq 0}$ is the value that $f(t)$ approaches for large temperatures. Similar to Storey et al. [23], we assume the intercept to be implicitly the same in both conditions to enable direct comparison of different responses to heat-stress, independent of abundance differences at the lowest temperature. For this purpose, all measurements are scaled by $f_0 = f(t_{\text{min}})$, so that Equation (1) is parametrized by three free parameters a, b, f_∞.

The model could be extended to fit f_0 as an additional free parameter, or by completely substituting Equation (1) by a more flexible model like smoothing splines [20].

5.4. **Test statistic and p-value calculation.**

5.4.1. *Test statistics.* Our goal is to compare two models per protein: The null model states that the relative concentrations of protein i at the j^{th} temperature t_j are characterized by a single mean function irrespective of the treatment condition:

$$\mathbb{E}(y_{ijk}) = \mu_i(t_j). \tag{2}$$

To account for systematic treatment effects, the alternative model replaces the common mean function μ_i with condition-specific mean functions μ_{ic}:

$$\mathbb{E}(y_{ijk}) = \mu_{ic}(t_j). \tag{3}$$

To assess changes in the melting curve of a protein caused by treatment, we compare the variances explained and unexplained by condition-specific modeling for this protein. We estimate these variances from the residual sums of squares of the null model (RSS^0) and the alternative model (RSS^1) and compare them by an F-statistic:

$$\mathrm{F}_i = \frac{\mathrm{d}_{i2}}{\mathrm{d}_{i1}} \cdot \frac{\mathrm{RSS}_i^0 - \mathrm{RSS}_i^1}{\mathrm{RSS}_i^1}. \tag{4}$$

Here, the sums of squared residuals for the null and alternative model are computed by

$$\mathrm{RSS}_\mathrm{i}^0 = \sum_{k=1}^{E} \sum_{j=1}^{N} (y_{ijk} - \hat{\mu}_i(t_j))^2, \tag{5}$$

and

$$\mathrm{RSS}_\mathrm{i}^1 = \sum_{k=1}^{E} \sum_{j=1}^{N} (y_{ijk} - \hat{\mu}_{ic}(t_j))^2 \tag{6}$$

where E is the total number of experiments, i.e. the number of all replicates in all treatment groups.

5.4.2. *Estimating the distribution under H_0.* If we could assume independently and identically normally distributed residuals, the F-statistics would be F-distributed under H_0 with DOF depending only on the numbers of measurements and parameters:

$$F_i \overset{\mathrm{H}_0}{\sim} \mathrm{F}(\mathrm{d}_{i1} = p_1 - p_0, \mathrm{d}_{i2} = E \cdot N - p_1). \tag{7}$$

Here, p_0 is the number of parameters in the null model, and $p_1 > p_0$ is the number of parameters in the alternative model.

In practice, the residuals have different variances at different temperatures (heteroscedasticity) and show dependencies between consecutive temperatures, which reduces the effective DOF and leads to over-optimistic p-values when applying Equation (7). While we can still approximate the empirical distribution of F_i by an F-distribution under H_0, we have to estimate the DOF from the data. If we had a sufficiently large number of replicates for each protein, we could try to estimate the DOF separately per protein by established methods from functional data analysis like the Welch-Satterthwaite approximation [31]. However, this method requires estimation of the covariance function, which is not robust for the small sample sizes and the sparse sampling grid in a typical TPP dataset.

Instead, we make the simplifying assumption that the null distribution can be approximated reasonably well by common DOF across all proteins. To estimate them, we make use of the fact that the F-distribution is a ratio distribution with χ^2-distributed numerator and denominator, so that:

$$\frac{\mathrm{RSS}^0 - \mathrm{RSS}^1}{\mathrm{d}_1} \overset{\mathrm{H}_0}{\sim} \frac{\sigma_0^2}{\mathrm{d}_1} \cdot \chi^2(\mathrm{d}_1), \tag{8}$$

$$\frac{RSS^1}{\mathrm{d}_2} \overset{\mathrm{H}_0}{\sim} \frac{\sigma_0^2}{\mathrm{d}_2} \cdot \chi^2(\mathrm{d}_2), \tag{9}$$

with a global scaling parameter σ_0^2 to reflect the average variance across all proteins.

The fit is performed by numerical optimization using the `MASS::fitdistr` function in `R` on the values of $(\mathrm{RSS}^0 - \mathrm{RSS}^1)$ and RSS^1 after scaling by σ_0. The optimization critically relies on a reasonable estimate of the scaling parameter σ_0. This parameter is estimated from the sample mean and variance of $(\mathrm{RSS}^1 - \mathrm{RSS}^0)$.

We define $M := \mathrm{mean}(\mathrm{RSS}^1 - \mathrm{RSS}^0)$ and $V := \mathrm{Var}(\mathrm{RSS}^1 - \mathrm{RSS}^0)$, the first moment, and the second central moment of the distribution of $(\mathrm{RSS}^1 - \mathrm{RSS}^0)$. We then make use of the fact that the scaled χ^2-distribution can be re-parameterized as a Γ-distribution, and that there is an established relationship between the moments of the Γ-distribution and its shape and scale parameters. We first re-parameterize the scaled χ^2-distribution (8) as a Γ-distribution with shape parameter

$$\alpha = \frac{\mathrm{d}_1}{2} \tag{10}$$

and scale parameter

$$\theta = 2\sigma_0^2. \tag{11}$$

Making use of the relationships between the moments of the Γ-distribution and its shape and scale parameters:

$$M = \alpha \cdot \theta \tag{12}$$

$$V = \alpha \cdot \theta^2, \tag{13}$$

and substituting (10) and (11) into (12) and (13), we obtain expressions of M and V in terms of df_1 and σ_0:

$$M = \sigma_0^2 \cdot \mathrm{d}_1 \tag{14}$$

$$V = 2 \cdot \sigma_0^4 \cdot \mathrm{d}_1\,. \tag{15}$$

Combining (14) and (15), and solving for d_1 yields:

$$\frac{M^2}{V} = \frac{\mathrm{d}_1}{2} \tag{16}$$

$$\Leftrightarrow \mathrm{d}_1 = 2 \cdot \frac{M^2}{V}. \tag{17}$$

This allows to solve for the scaling factor σ_0^2:

$$\sigma_0^2 = \frac{\theta}{2} = \frac{M}{2\,\alpha} = \frac{M}{\mathrm{d}_1} = \frac{M}{2 \cdot \frac{M^2}{V}} = \frac{1}{2}\frac{V}{M}. \tag{18}$$

These equations coincide with those used by Brown's method [32]. Brown proposes an adaptation of Fisher's method for combining multiple p-values to the scenario of correlated tests by estimating χ^2-distribution parameters from the data in a similar manner.

To increase robustness, we estimated M and V by their *D-estimates* [33] (median and median absolute deviation), with possibility for refinement by more unbiased estimators. Values of $(\mathrm{RSS}^1 - \mathrm{RSS}^0)$ within the upper 5%-quantile were excluded from estimation of M and V, but their respective proteins were included again for p-value calculation.

5.4.3. *Calculation of p-values by NPARC.* For each protein, an F-statistic F_i was computed according to Equation (4) with d_1 and d_2 estimated as described above. For each F_i, p-values were calculated from the cumulative F-distribution with the estimated DOF. All p-values were corrected for multiple testing by Benjamini-Hochberg adjustment.

5.5. **Model fitting.** All models were fitted by nonlinear least squares regression using the `nls` function in `R`. For the NPARC analysis, the melting curve model (Eq. (1)) was fitted separately per protein to obtain $\hat{\mu}_i(t)$, or per protein and treatment condition to obtain $\hat{\mu}_{ic}(t)$. To reproduce the results of the T_m-based approach (Table 3), the model fits were repeated per replicate and treatment condition for each protein.

5.6. **Summary of the T_m-based approach.** The results of the T_m-based approach were obtained with the `R` package `TPP` [17]. The package is comprehensively described elsewhere [6] and will only be briefly summarized here. For each curve obtained by the replicate- and condition-wise model fits, T_m was calculated as $T_\mathrm{m} = \frac{a}{b - \ln(\frac{1-f_\infty}{0.5-f_\infty} - 1)}$ so that it fulfilled $f(T_\mathrm{m}) = 0.5$. After melting curve fitting, a filtering step was applied to remove curves with undesirable shape or goodness-of-fit by setting a threshold on the R^2, the slope and the plateau parameter (Table 4). Within each replicate, the difference in T_m of the treatment and control condition (ΔT_m) was computed per protein and converted to z-scores. Robust versions of the z-scores were computed by replacing the mean and standard deviation by the corresponding quantiles of the empirical distributions of ΔT_m. In order to minimize the influences of values with high estimator uncertainty when calculating these quantiles, proteins were binned by the slopes of their curves, and z-scores were calculated separately for each bin as described in [34]. Next, p-values were calculated by comparing the z-scores to the normal distribution. To reach the final decision for each protein, the p-values were combined heuristically across replicates using the rules in Table 3.

TABLE 3. Overview of the criteria suggested in the original TPP analysis workflow to combine z-test p-values across replicates in an experimental design with two biological replicates [5].

Rule number	Rule
1	The Benjamini-Hochberg adjusted z-test p-values fulfill predefined thresholds in each replicate.
2	Both melting point differences are either positive or negative in the two biological replicates.
3	The smallest absolute difference between treatment and vehicle T_{m} is greater than the absolute T_{m} difference between the two vehicle experiments.

TABLE 4. Overview of the ad hoc filters suggested in the original TPP analysis workflow to increase accuracy and precision of the T_{m}-estimators prior to hypothesis testing.

Rule number	Rule
1	Both fitted curves for the vehicle and compound treated condition have an $R^2 > 0.8$.
2	Both vehicle curves have a plateau $f_\infty < 0.3$.
3	In each biological replicate, the steepest slope of the melting curve in the vehicle and treatment group is ≤ -0.06.

6. List of abbreviations

- ANOVA: Analysis of variance
- CETSA: Cellular thermal shift assay
- DOF: Degrees of freedom
- FDR: False discovery rate
- H_0: Null hypothesis
- NPARC: Non-parametric analysis of response curves
- RSS: Sum of squared residuals
- T_m: Melting point
- TPP: Thermal proteome profiling

7. Declarations

7.1. **Ethics approval and consent to participate.** Not applicable

7.2. **Consent for publication.** Not applicable

7.3. **Availability of data and materials.** The TPP-TR experiments based on staurosporine-, panobinostat-, ATP-, and dasatinib treatments are included in the supplementary materials of references [5] and [9]. All results generated from this data are provided in the Supplementary material attached to this work. An R-package will soon be released as an open-source Bioconductor package.

7.4. **Competing interests.** HF, MS and MB are employees or shareholders of GlaxoSmithKline.

7.5. **Funding.** SA is funded by the Deutsche Forschungsgemeinschaft, SFB 1036. KB is funded by a Cambridge Cancer Centre studentship. WH acknowledges funding from the European Commission's H2020 Programme, Collaborative research project SOUND (Grant Agreement no 633974).

7.6. **Authors' contributions.** KB, DC and HF contributed equally. KB, DC, HF and SA conceived the model. KB and DC implemented the model. KB and DC performed the analysis. KB, DC, HF, NK, MS and WH interpreted the results. KB, DC, HF, NK and SA wrote the manuscript. MB, MS and WH supervised the project. All authors read and approved the final manuscript.

8. Figures, tables and additional files

The following Figures and Tables can be found in the Supplementary Material.

TABLE S1. Results of the NPARC approach and the T_m-based approach for all datasets listed in Table 1.

FIGURE S1. Experimental workflow of thermal proteome profiling (TPP) experiments for generation of the datasets listed in Table 1.

FIGURE S2. Effects of QC-filters administered to restrict analysis to 'well-behaved' curves for hypothesis testing.

FIGURE S3. Melting curves of the seven panobinostat hits detected with Benjamini-Hochberg adjusted p-values ≤ 0.01 in addition to the expected targets.

FIGURE S4. The NPARC approach provides the flexibility to accommodate complex experimental designs as illustrated by the treatment with dasatinib at two different concentrations.

FIGURE S5. Further examples of the protein kinase C family for which treatment effects are poorly reflected by shifts in T_{m}.

FIGURE S6. All staurosporine hits of the NPARC approach with Benjamini-Hochberg adjusted F-test p-values ≤ 0.01.

FIGURE S7. All ATP hits of the NPARC approach with Benjamini-Hochberg adjusted F-test p-values ≤ 0.01.

REFERENCES

[1] Kenneth M. Comess, Shaun M. McLoughlin, Jon A. Oyer, Paul L. Richardson, Henning Stöckmann, Anil Vasudevan, and Scott E. Warder. Emerging approaches for the identification of protein targets of small molecules - a practitioners perspective. *Journal of Medicinal Chemistry*, 0(0):null, 0.

[2] Gabriel M Simon, Micah J Niphakis, and Benjamin F Cravatt. Determining target engagement in living systems. *Nature chemical biology*, 9(4):200–5, 2013.

[3] Mark E Bunnage, Eugene L Piatnitski Chekler, and Lyn H Jones. Target validation using chemical probes. *Nature Chemical Biology*, 9(4):195–199, mar 2013.

[4] Cheryl H Arrowsmith, James E Audia, Christopher Austin, Jonathan Baell, Jonathan Bennett, Julian Blagg, Chas Bountra, Paul E Brennan, Peter J Brown, Mark E Bunnage, Carolyn Buser-Doepner, Robert M Campbell, Adrian J Carter, Philip Cohen, Robert a Copeland, Ben Cravatt, Jayme L Dahlin, Dashyant Dhanak, Aled M Edwards, Mathias Frederiksen, Stephen V Frye, Nathanael Gray, Charles E Grimshaw, David Hepworth, Trevor Howe, Kilian V M Huber, Jian Jin, Stefan Knapp, Joanne D Kotz, Ryan G Kruger, Derek Lowe, Mary M Mader, Brian Marsden, Anke Mueller-Fahrnow, Susanne Müller, Ronan C O'Hagan, John P Overington, Dafydd R Owen, Saul H Rosenberg, Ruth Ross, Bryan Roth, Matthieu Schapira, Stuart L Schreiber, Brian Shoichet, Michael Sundström, Giulio Superti-Furga, Jack Taunton, Leticia Toledo-Sherman, Chris Walpole, Michael a Walters, Timothy M Willson, Paul Workman, Robert N Young, and William J Zuercher. The promise and peril of chemical probes. *Nature Chemical Biology*, 11(8):536–541, jul 2015.

[5] Mikhail M Savitski, Friedrich B M Reinhard, Holger Franken, Thilo Werner, Maria Fälth Savitski, Dirk Eberhard, Daniel Martinez Molina, Rozbeh Jafari, Rebecca Bakszt Dovega, Susan Klaeger, Bernhard Kuster, Pär Nordlund, Marcus Bantscheff, and Gerard Drewes. Tracking cancer drugs in living cells by

thermal profiling of the proteome. *Science*, 346(6205):1255784, 2014.

[6] Holger Franken, Toby Mathieson, Dorothee Childs, Gavain M A Sweetman, Thilo Werner, Ina Tögel, Carola Doce, Stephan Gade, Marcus Bantscheff, Gerard Drewes, Friedrich B M Reinhard, Wolfgang Huber, and Mikhail M Savitski. Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nat. Protoc.*, 10(10):1567–93, oct 2015.

[7] Daniel Martinez Molina, Rozbeh Jafari, Marina Ignatushchenko, Takahiro Seki, E Andreas Larsson, Chen Dan, Lekshmy Sreekumar, Yihai Cao, and Pär Nordlund. Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science*, 341(6141):84–7, 2013.

[8] Chris Soon Heng Tan, Ka Diam Go, Xavier Bisteau, Lingyun Dai, Chern Han Yong, Nayana Prabhu, Mert Burak Ozturk, Yan Ting Lim, Lekshmy Sreekumar, Johan Lengqvist, Vinay Tergaonkar, Philipp Kaldis, Radoslaw M. Sobota, and Pär Nordlund. Thermal proximity coaggregation for system-wide profiling of protein complex dynamics in cells. *Science*, 2018.

[9] Friedrich B M Reinhard, Dirk Eberhard, Thilo Werner, Holger Franken, Dorothee Childs, Carola Doce, Maria Fälth Savitski, Wolfgang Huber, Marcus Bantscheff, Mikhail M Savitski, and Gerard Drewes. Thermal proteome profiling monitors ligand interactions with cellular membrane proteins. *Nat. Methods*, 12(12):1129–31, dec 2015.

[10] Kilian V M Huber, Karin M Olek, André C Müller, Chris Soon Heng Tan, Keiryn L Bennett, Jacques Colinge, and Giulio Superti-Furga. Proteome-wide drug and metabolite interaction mapping by thermal-stability profiling. *Nature Methods*, 12(11):1055–1057, sep 2015.

[11] Isabelle Becher, Thilo Werner, Carola Doce, Esther A Zaal, Ina Tögel, Crystal A Khan, Anne Rueger, Marcel Muelbaier, Elsa Salzer, Celia R Berkers, Paul F Fitzpatrick, Marcus Bantscheff, and Mikhail M Savitski. Thermal profiling reveals phenylalanine hydroxylase as an off-target of panobinostat. *Nature Chemical Biology*, (11):908–910, sep 2016.

[12] André Mateus, Jacob Bobonis, Nils Kurzawa, Frank Stein, Dominic Helm, Johannes Hevler, Athanasios Typas, and Mikhail M Savitski. Thermal proteome profiling in bacteria: probing protein state in vivo. *Molecular Systems Biology*, 14(7):e8242, jul 2018.

[13] Isabelle Becher, Amparo Andrés-Pons, Natalie Romanov, Frank Stein, Maike Schramm, Florence Baudin, Dominic Helm, Nils Kurzawa, André Mateus, Marie-Therese Mackmull, et al. Pervasive protein thermal stability variation during the cell cycle. *Cell*, 173(6):1495–1507.e18, may 2018.

[14] Alireza Azimi, Stefano Caramuta, Brinton Seashore-Ludlow, Johan Boström, Jonathan L Robinson, Fredrik Edfors, Rainer Tuominen, Kristel Kemper, Oscar Krijgsman, Daniel S Peeper, Jens Nielsen, Johan Hansson, Suzanne Egyhazi Brage, Mikael Altun, Mathias Uhlen, and Gianluca Maddalo. Targeting CDK2 overcomes melanoma resistance against BRAF and Hsp90 inhibitors. *Molecular Systems Biology*, 14(3):e7858, mar 2018.

[15] Teemu P Miettinen, Julien Peltier, Anetta Härtlova, Marek Gierliński, Valerie M Jansen, Matthias Trost, and Mikael Björklund. Thermal proteome profiling of breast cancer cells reveals proteasomal activation by CDK4/6 inhibitor palbociclib. *The EMBO Journal*, 37(10):e98359, may 2018.

[16] Mikhail M. Savitski, Nico Zinn, Maria Faelth-Savitski, Daniel Poeckel, Stephan Gade, Isabelle Becher, Marcel Muelbaier, Anne J. Wagner, Katrin Strohmer, Thilo Werner, Stephanie Melchert, Massimo Petretich, Anna Rutkowska, Johanna Vappiani, Holger Franken, Michael Steidel, Gavain M. Sweetman,

Omer Gilan, Enid Y.N. Lam, Mark A. Dawson, Rab K. Prinjha, Paola Grandi, Giovanna Bergamini, and Marcus Bantscheff. Multiplexed proteome dynamics profiling reveals mechanisms controlling protein homeostasis. *Cell*, 173(1):260–274.e25, mar 2018.

[17] Dorothee Childs, Nils Kurzawa, Holger Franken, Carola Doce, Mikhail Savitski, and Wolfgang Huber. *TPP: Analyze thermal proteome profiling (TPP) experiments*, 2017. R package version 3.6.0.

[18] Jeremy Volkening. *mstherm: Analyze MS/MS protein melting Data*, 2017. R package version 0.4.7.

[19] J.O. Ramsey and B.W. Silverman. *Functional data analysis*. Springer, New York, NY, 2 edition, 2005.

[20] Maurice Berk, Timothy Ebbels, and Giovanni Montana. A statistical framework for biomarker discovery in metabolomic time course data. *Bioinformatics*, 27(14):1979–1985, jul 2011.

[21] Shahid Ullah and Caroline F Finch. Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 13(1):43, dec 2013.

[22] Yolanda MuÑoz Maldonado, Joan G Staniswalis, Louis N Irwin, and Donna Byers. A similarity analysis of curves. *The Canadian Journal of Statistics*, 30(3):373–381, 2002.

[23] John D Storey, Wenzhong Xiao, Jeffrey T Leek, Ronald G Tompkins, and Ronald W Davis. Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci. USA*, 102(36):12837–42, sep 2005.

[24] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, 57(1):289–300, 1995.

[25] N E Ward and C A O'Brian. Kinetic analysis of protein kinase C inhibition by staurosporine: evidence that inhibition entails inhibitor binding at a conserved region of the catalytic domain but not competition with substrates. *Molecular pharmacology*, 41(2):387–92, feb 1992.

[26] Kevin K. K. Leung and Brian H. Shilton. Quinone reductase 2 is an adventitious target of protein kinase CK2 inhibitors TBBz (TBI) and DMAT. *Biochemistry*, 54(1):47–59, jan 2015.

[27] Jonathan A Winger, Oliver Hantschel, Giulio Superti-Furga, and John Kuriyan. The structure of the leukemia drug imatinib bound to human quinone reductase 2 (NQO2). *BMC Structural Biology*, 9(1):7, 2009.

[28] Marcus Bantscheff, Dirk Eberhard, Yann Abraham, Sonja Bastuck, Markus Boesche, Scott Hobson, Toby Mathieson, Jessica Perrin, Manfred Raida, Christina Rau, Valérie Reader, Gavain M A Sweetman, Andreas Bauer, Tewis Bouwmeester, Carsten Hopf, Ulrich Kruse, Gitte Neubauer, Nigel G. Ramsden, Jens Rick, Bernhard Kuster, and Gerard Drewes. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nature Biotechnology*, 25(9):1035–1044, sep 2007.

[29] Teemu P. Miettinen and Mikael Björklund. NQO2 is a reactive oxygen species generating off-target for acetaminophen. *Molecular Pharmaceutics*, 11(12):4395–4404, dec 2014.

[30] J A Schellman. The thermodynamics of solvent exchange. *Biopolymers*, 34(8):1015–1026, 1994.

[31] Jin-Ting Zhang. *Analysis of variance for functional data*. CRC Press, 2013.

[32] Morton B. Brown. 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31(4):987, dec 1975.

[33] A Marazzi. Bootstrap tests for robust means of asymmetric distributions with unequal shapes. *Computational statistics & data analysis*, 39(4):503–528, 2002.

[34] Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367–1372, 2008.