Received: 18 November 2024 | Revised: 23 January 2025 | Accepted: 25 January 2025

DOI: 10.1002/pro.70055

TOOLS FOR PROTEIN SCIENCE

THE PROTEIN SOCIETY WILEY

ShiftScan: A tool for rapid analysis of high-throughput differential scanning fluorimetry data and compound prioritization

Samantha C. Waterworth[1] | **Shilpa R. Shenoy**[1] | **Nirmala D. Sharma**[1] | **Chris Wolcott**[1,2] | **Duncan E. Donohue**[3] | **Barry R. O'Keefe**[1,4] | **John A. Beutler**[1]

[1]Molecular Targets Program, Center for Cancer Research, National Cancer Institute, Frederick, Maryland, USA

[2]Advanced Biomedical Computational Science, Frederick National Laboratory for Cancer Research, Frederick, Maryland, USA

[3]Statistics Department, Data Management Services Inc., Frederick National Laboratory for Cancer Research Sponsored by the National Cancer Institute, Frederick, Maryland, USA

[4]Natural Products Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Frederick, Maryland, USA

Correspondence
John A. Beutler and Barry R. O'Keefe, Molecular Targets Program, Center for Cancer Research, National Cancer Institute, Frederick, MD 21702, USA.
Email: beutlerj@nih.gov; okeefeba@mail.nih.gov

Funding information
National Cancer Institute, Grant/Award Numbers: 75N91019D00024, ZIA BC 011469, ZIA BC 011471

Review Editor: Nir Ben-Tal

Abstract

Differential scanning fluorimetry (DSF) can be an effective high-throughput screening assay in drug discovery for detecting protein-compound interactions that stabilize or destabilize macromolecules. Due to the magnitude and quality of the data produced by this biophysical assay, analyzing and prioritizing compounds from large-scale DSF data sets has proven challenging to the research community. Here, we present ShiftScan—a powerful, stand-alone tool designed for the rapid analysis of DSF data and compound prioritization based on thermal transition patterns. ShiftScan accurately and quickly predicts melting temperatures (Tm values) from both canonical and non-canonical transition patterns, efficiently filtering out spurious data to minimize false positives. We report on the use of this tool for data analysis of screens involving both pure compound and natural product fraction libraries and provide the software to the screening community to aid in the discovery of molecularly-targeted compounds. Instructions for installation and usage of ShiftScan can be found at our GitHub repository: https://github.com/samche42/ShiftScan.

KEYWORDS
biological software, curve fitting, differential scanning fluorimetry, drug discovery, high throughput analysis, thermal shift assay

1 | INTRODUCTION

Differential scanning fluorimetry (DSF), a type of thermal shift assay (TSA), can be used to monitor temperature-induced unfolding events of either protein (Pantoliano et al., 2001) or polynucleotide macromolecules (Sztuba-Solinska et al., 2014). In a typical assay for changes in protein stability, a protein of interest is incubated with a fluorescent dye, most commonly Sypro Orange™, in a multi-well plate and heated in a real-time polymerase chain reaction (RT-PCR) instrument (Pantoliano et al., 2001). The plate is uniformly heated over a temperature gradient, and as the protein unfolds, the hydrophobic residues of the inner regions of the protein structure become exposed and bind the Sypro Orange™ dye, increasing the dye's fluorescence intensity. Fluorescence measurements are routinely taken >250 times per well over the course of an experiment. If the protein unfolds in a two-state (folded to unfolded) manner, the change in fluorescence with respect to temperature typically follows a sigmoidal pattern, and the mid-point of the rise in the

curve can be extrapolated as the protein's melting temperature (Tm). This melting temperature can be monitored in the presence of small molecule libraries; if the binding of a molecule causes a structural perturbation in the protein, this is often reflected in a shift in the Tm of the protein (Pantoliano et al., 2001). A shift in Tm can occur in either direction corresponding to either a stabilization (higher Tm) or a destabilization (lower Tm) of the protein's tertiary structure upon ligand binding.

DSF is used in numerous applications (Wu et al., 2024), including as a primary screening platform for detecting novel macromolecule-compound interactions (Ciulli, 2013; Dai et al., 2015; Gao et al., 2020; Matarlo et al., 2019; Pantoliano et al., 2001). In these DSF screens, compounds (or mixture of compounds from fractionated natural product extracts) can be assayed against a protein of interest in a high throughput manner, the samples tested in a 96-well or 384-well plate. The raw fluorescence readings upon heating are then measured, and the extrapolated melting temperatures (Tm values) of the sample wells relative to that of controls determine which samples may be worth further consideration. DSF is a cost-effective, sensitive assay in early-stage drug discovery (Gao et al., 2020) that can be performed in high-throughput manner on widely available RT-PCR instruments. While the ideal output of a DSF assay is a series of sigmoidal curves that can be easily compared to one another, there are often non-canonical transition patterns that occur due to interference by a compound/extract's intrinsic fluorescence (Wu et al., 2023). Sub-optimal experimental conditions (Gao et al., 2020; Wu et al., 2023, 2024) and non-two-state protein unfolding events due to sample effects can also produce non-canonical transition curves, making the entire process of Tm extrapolation and compound prioritization more challenging.

In the context of a 384-well plate setup, with a 260-step temperature increase, a single plate generates approximately 100,000 data points. This means that in a high throughput setting, datasets can rapidly produce millions of data points for analysis. Various tools are available for processing and inspecting the data, including DSFWorld (Wu et al., 2024), SimpleDSFViewer (Sun et al., 2020), the DSF workflow in KNIME (Samuel et al., 2021), HTSDSF Explorer (Martin-Malpartida et al., 2022), among others (Wu et al., 2023). DSFWorld is a recently developed and useful tool that exploits four robust mathematical models for fitting canonical and non-canonical DSF curve data. However, to our knowledge, neither DSFWorld (Wu et al., 2024) nor SimpleDSFViewer (Sun et al., 2020) process data from high-throughput screening campaigns. The KNIME workflow (Samuel et al., 2021) is powerful but relatively slow, processing one 384-well plate at a time, and requiring heavy user interaction/input in at least 11 different points and visual inspection of individual melting curves for "well-behaved" data (Samuel et al., 2021); this is not a feasible option for high-throughput initiatives. Similarly, this approach excludes curves which follow non-canonical transition patterns which may be of interest to researchers. HTSDSF Explorer (Martin-Malpartida et al., 2022), processes DSF data in a high-throughput manner but aims to provide preliminary binding constants from concentration response assays rather than prioritization of compounds at a single concentration (as is often screened in preliminary HTS campaigns).

A recent investigation into strategies for compound prioritization found DSF to be an optimal first step in identifying hits before confirmation with either surface plasmon resonance (SPR) or temperature-related intensity change (TRIC) assays (Fotsch et al., 2024). Similarly, major pharmaceutical companies such as AstraZeneca have recently reported the automation of the DSF assay for the high-throughput screening of approximately 100,000 compounds (Hansel et al., 2023). At the National Cancer Institute (NCI), we have developed a workflow for screening large libraries of pure compounds and pre-fractionated natural product extracts (Grkovic et al., 2020; Thornburg et al., 2018) against protein targets of interest using DSF. The challenge has been the magnitude of the resultant data generated (i.e., for every hundred 384-well plates assayed across a 260-step temperature gradient (35,200 test samples) ~10 million data points are produced). Along with the size of the data set, additional challenges, such as a variety of non-canonical transition patterns and otherwise noisy data, often result from the assaying of natural product mixtures.

To address these challenges, we developed ShiftScan, a standalone tool that can be run locally or on large computing clusters, with the choice of RAM- or disk-intensive modes to suit different system capabilities. Along with the main processing algorithm, we have developed a companion visualization tool for the rapid identification of hits as defined by a user's criteria. ShiftScan is also available as a Google Colab notebook and a stand-alone GUI application, although the increased user-friendliness in these implementations comes at the cost of processing speed. ShiftScan processes data in a plate-wise fashion and can analyze data from sample sets assaying different proteins of interest simultaneously. We hope that ShiftScan will help provide a valuable tool to the scientific community in the pursuit of novel therapeutics.

2 | RESULTS

2.1 | Algorithm development

ShiftScan is written in Python3 and consists of two parts: (1) the main data processing algorithm and (2) the companion visualization tool developed as a Plotly Dash application (https://dash.plotly.com/). Detailed instructions for installation and usage are provided in our GitHub repository: https://github.com/

samche42/ShiftScan. ShiftScan has been tested and can be used in Mac, Linux, or Windows environments. Users can customize parameters such as the maximum permissible *z*-scores for control well amplitudes and melting temperatures, the threshold for control failures before a plate is flagged as a failure, the amplitude range for experimental curves relative to controls, data normalization options, the smoothing factor used, and control column assignments. Currently, the algorithm supports raw input files from Roche LightCycler 480 II and Bio-Rad (Opticon Monitor) instruments. The import functions are modular, and we encourage users to send additional examples of input data formats for us to incorporate. The algorithm supports plates of any size, with defaults optimized for 384-well plates. Each parameter has a default setting, detailed in the following sections, should the user opt to defer to these.

2.1.1 | Initial processing of all wells

Raw data is read in from a user-provided input folder and concatenated, where the file name is used as a primary key to link the data with user-provided metadata. The data is then normalized per plate (unless the user has specified otherwise) to values between 0 and 1. Data is processed for each individual well. If a well is not assigned as a control or lacks experimental metadata, it is considered blank and skipped. For all wells, the data is smoothed (Figure 1a), 'cleaned,' and subsequently 'split' to isolate the sigmoidal regions within (Figure 1b). This is achieved by splitting the respective curves at local minima and maxima, then assessing the average gradient and the shape of the first and second derivatives of the individual slices. We would expect the data to have a positive average gradient, and for the first derivative from which the data is derived to include a peak. The ratio of positive and negative values for the second derivative gradients is calculated to identify legitimate peaks in the first derivative. Here, we defined an unacceptable ratio of positive and negative values as 90:10 or more, for example, if the positive to negative ratio of second derivative values is 93:7, the first derivative likely does not have a legitimate peak and therefore does not represent sigmoidal data. Each sigmoidal region is then fitted against a Boltzmann sigmoidal distribution model (Figure 1c), mathematically defined as follows:

$$y = \frac{D - A}{1 + e^{-B(x-C)}}$$

where A and D are the minimum and maximum asymptotes, respectively, B is the slope of the curve at C, the point of inflection. The curve is model-fitted using the 'optimize.curve_fit' function from SciPy (Virtanen et al., 2020). The optimization method is initially set to 'lm' (Levenberg–Marquardt), and if this fails, optimization is reattempted with the 'dogbox' (dogleg algorithm with rectangular trust regions) method. The former is the traditional option for modeling of nonlinear data, while the latter is recommended for more complex or noisier data (An et al., 2023). The Boltzmann melting temperature (Tm), mean squared error (MSE), residual sum of squares (RSS), and any errors resulting from modeling failures are reported for each well. The final melting temperature(s) for each well is determined from the inflection point of the smoothed sigmoidal curve(s) data, and the difference of this measurement from the predicted Tm from the Boltzmann fit is documented. If the user only requires Tm value estimates without comparing control and experimental values, they can use the—only_tm flag. This option stops the pipeline early, generating only two files, Only_Tm_values.txt and Only_Tm_curves.txt as the final output. Note that an augmented version of the companion visualization tool is available but has limited functionality given the truncated output.

2.1.2 | Quality control of data

All data corresponding to wells that are designated as control wells by the user are subset out. The *z*-scores for control well melting temperatures (Tm) are calculated per plate, and any control wells that have a Tm *z*-score beyond the user-specified cutoff (default is a maximum of 1.5 standard deviations) are marked as having failed. Similarly, *z*-scores for the amplitudes of the control curves are calculated and curves with amplitude *z*-scores greater than the defined cutoff (default is a maximum of 2) are marked as having failed. These cutoffs were chosen following assessment of the number of failed control wells generated using different cutoff combinations for melting temperature and amplitude *z*-scores across 500 assay plates, including 242 plates with 'Protein A' and 258 plates with 'Protein B' assayed against various compound and extract libraries (Figure S1). Based on data from two assayed proteins across 500 plates (384-well), we found that an amplitude *z*-score cutoff below 2 was overly stringent, resulting in a high rate of control well failures (Figure S1). Notably, combining this with a Tm *z*-score cutoff of 1.5 reduced the effect, leading to a plateau in control well failure rates (Figure S1). To balance accuracy with conservatism, we selected default cutoffs of 1.5 for Tm *z*-scores and 2 for amplitude *z*-scores in control wells. The average control melting temperature and amplitude per plate are calculated only using wells that pass these quality checks. Outlier removal from the control data helps ensure that the calculated control means are accurate and unbiased, reducing the risk of skewed results that could otherwise lead to erroneous hit assignments in downstream analysis.

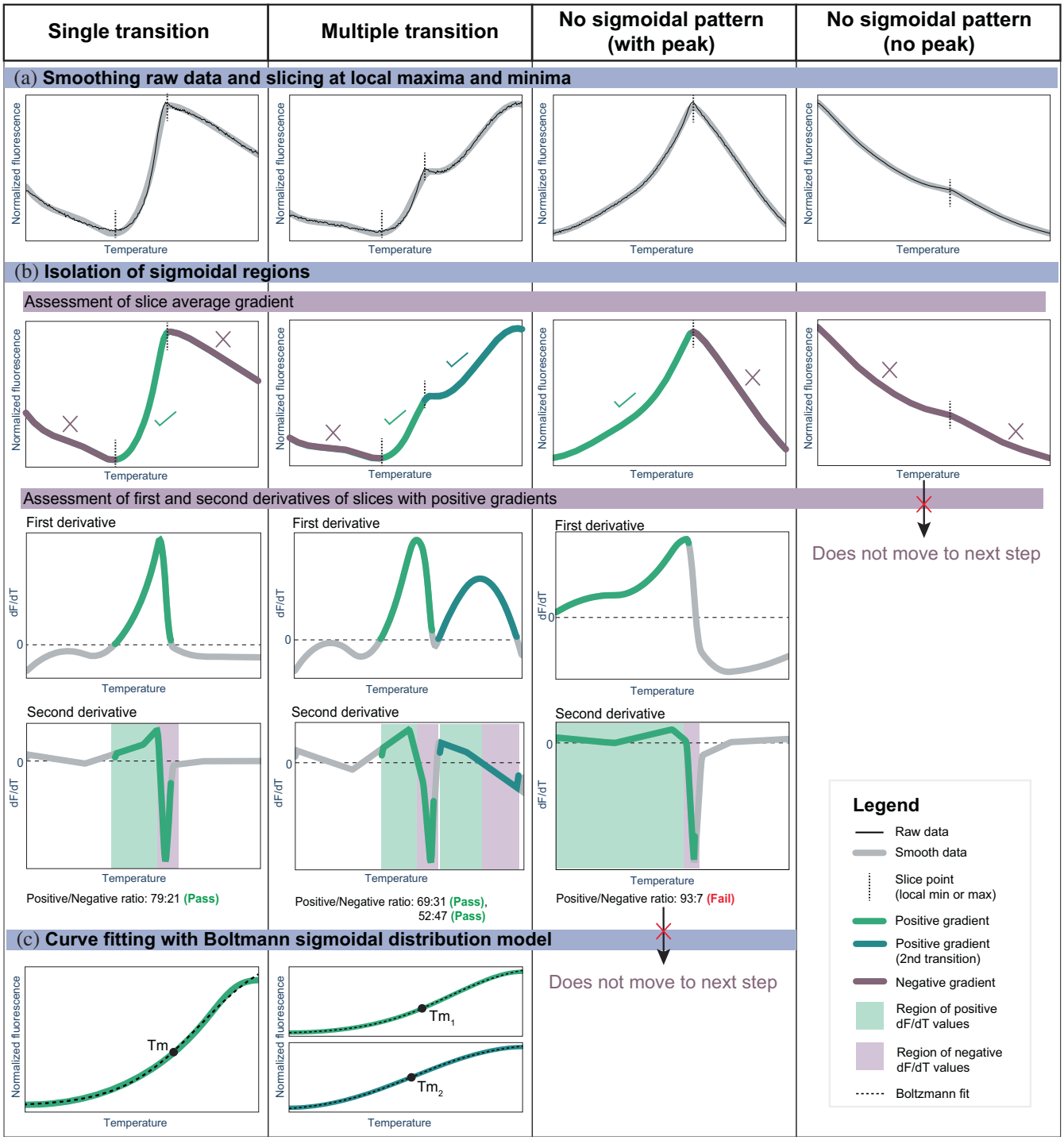

FIGURE 1 Processing of data per well is performed in several key steps: (a) curve data for each well is smoothed and sliced at local minima and maxima, (b) sigmoidal regions are isolated through assessment of the average gradient and shape of the first and second derivatives of the respective data. Finally, the data is (c) fitted against a Boltzmann sigmoidal distribution model.

Next, the experimental wells are assessed relative to the remaining controls per plate. In the case of non-canonical curves, where two or more transitions were detected, each sigmoidal 'subcurve' is treated independently of other subcurves detected for that well, and each is reported as a 'Subplot' for the well in question. The relative amplitude of each experimental (sub)curve is calculated relative to the average control amplitude for the respective plate. The default behavior for the algorithm is to fail any experimental (sub)curves where the relative amplitude is 6 times greater OR a fifth of the average control amplitude, for example, if the (sub) curve amplitude of an experimental well is 0.19 relative to the average control amplitude of that plate, this (sub)

curve is marked as a failure. These defaults were chosen following parameters presented in a published study (Reidenbach et al., 2020). These defaults may not be ideal for all datasets and therefore the user can modify the default cutoff parameters to suit their specific purposes or protocols.

2.1.3 | Preparation of data for 'hit' identification

The primary aim of this tool is to enable users to quickly identify candidate compounds or extracts for further evaluation, as an initial step in the discovery process. Usually, these candidates, often called 'hits,' are identified by determining which experimental Tm values fall outside *N* standard deviations from the average control Tm value (Douse et al., 2015; Hanley, 2019; Malo et al., 2006; Martin-Malpartida et al., 2024; Zhang et al., 1999), that is, determination of which compounds have resulted in statistically significant shifts in the Tm of the protein of interest. An alternative method for hit identification is to detect compounds that shift the native protein Tm by a fixed temperature range (e.g., more than 2°C) (Dai et al., 2015; Douse et al., 2015; Scholle et al., 2021). To this end, the metrics of the number of degrees from the control mean and the number of standard deviations from the control mean are calculated for each experimental well. The option to identify hits via this methodology is available to the user in the companion ShiftScan Viewer tool.

These strategies are predicated on the assumption that the Tm values of the controls follow an approximate normal distribution (Coma et al., 2009; Goktug et al., 2013). However, we noted that this assumption may not always be true in practice. Using the Shapiro–Wilk test for normality from scipy (Shapiro & Wilk, 1965; Virtanen et al., 2020) ($\alpha = 0.05$), a survey of the same 500 assay plates used previously showed that only 88 plates (17.6%) had control Tm values that followed a normal distribution (i.e., 57 out of 242 for Protein A, and 31 out of 258 for Protein B) (Dataset S1). Wells or subcurves that failed were not included in these calculations. Similarly, the average range of control Tm values for Protein A and Protein B were 3.3°C (∓2.9°C) and 5.2°C (∓7.9°C), respectively (Table S1). Such a wide range of Tm values observed in the controls makes establishing a reliable, robust standardized cutoff for hit determination problematic. We could not find a sufficient number of additional, publicly available datasets to determine if this was a phenomenon exclusive to our assays. We therefore erred on the side of caution which led us to explore alternative strategies for hit identification when control Tm values do not follow a normal distribution: First *z*-scores are calculated for all Tm values per plate. If the Tm *z*-score of an experimental well falls outside the range of control Tm *z*-scores for a given plate, it is flagged as a potential hit. The user can then adjust this range (e.g., Hit Tm = control range ± *N*% of the control range) to apply more stringent hit criteria. The data processing stage ends with the identification of control Tm *z*-score maxima and minima per plate, and this third hit identification option is available in the companion ShiftScan Viewer tool.

Finally, four output files are generated:

A. 'Final_results.txt': A table that details data for each subcurve isolated from a well such as the final Tm, modeled Tm, the difference between these two values, the subcurve amplitude, the well *z*-score as well as all other data such as parent plate, well, and compound (if experimental).
B. 'Final_curves.txt': Co-ordinates for original curves and subcurves identified per well per plate.
C. 'Plate_report.txt': A summary of how many control wells failed per plate, where they are situated, and the errors in the event of failure. By default, any plate that has 8 or more failed control wells is marked as a failed plate, but the user may opt to modify this parameter.
D. 'Potential_problems.txt': A small summary of wells that repeatedly fail across plates. We found this useful to identify potential automated pipetting issues.

These files can be manipulated directly by the user for their own statistical analyses and hit identification, or as input for the companion visualization and hit identification tool; ShiftScan Viewer.

2.1.4 | Data visualization with ShiftScan viewer

Processed data can be viewed and manipulated by the user for the identification of hits. Upon initializing ShiftScan Viewer, the user is greeted with four tab options. The first is the 'Plate overview' (Figure 2a) which includes a visualization of the plate report file that details control wells that failed and why, and the potential problem file which identifies wells that repeatedly fail in 3 or more plates. The second tab, 'Control overview' (Figure 2b), includes three graphs; the first shows the distribution of all control Tm values per plate, the second shows the distribution of control Tm values that passed quality control, and the third provides a view of the control curves per plate and whether they passed or failed the quality control criteria. The third tab, 'Melting temp overview' (Figure 2c), displays the data in a 384-plate format. The wells can be colored by the Tm or *z*-score of each well. The greatest absolute value is displayed when two or more subcurves are present in a well. The user can hover over a well to access additional metadata or click on a well to produce the corresponding curve.

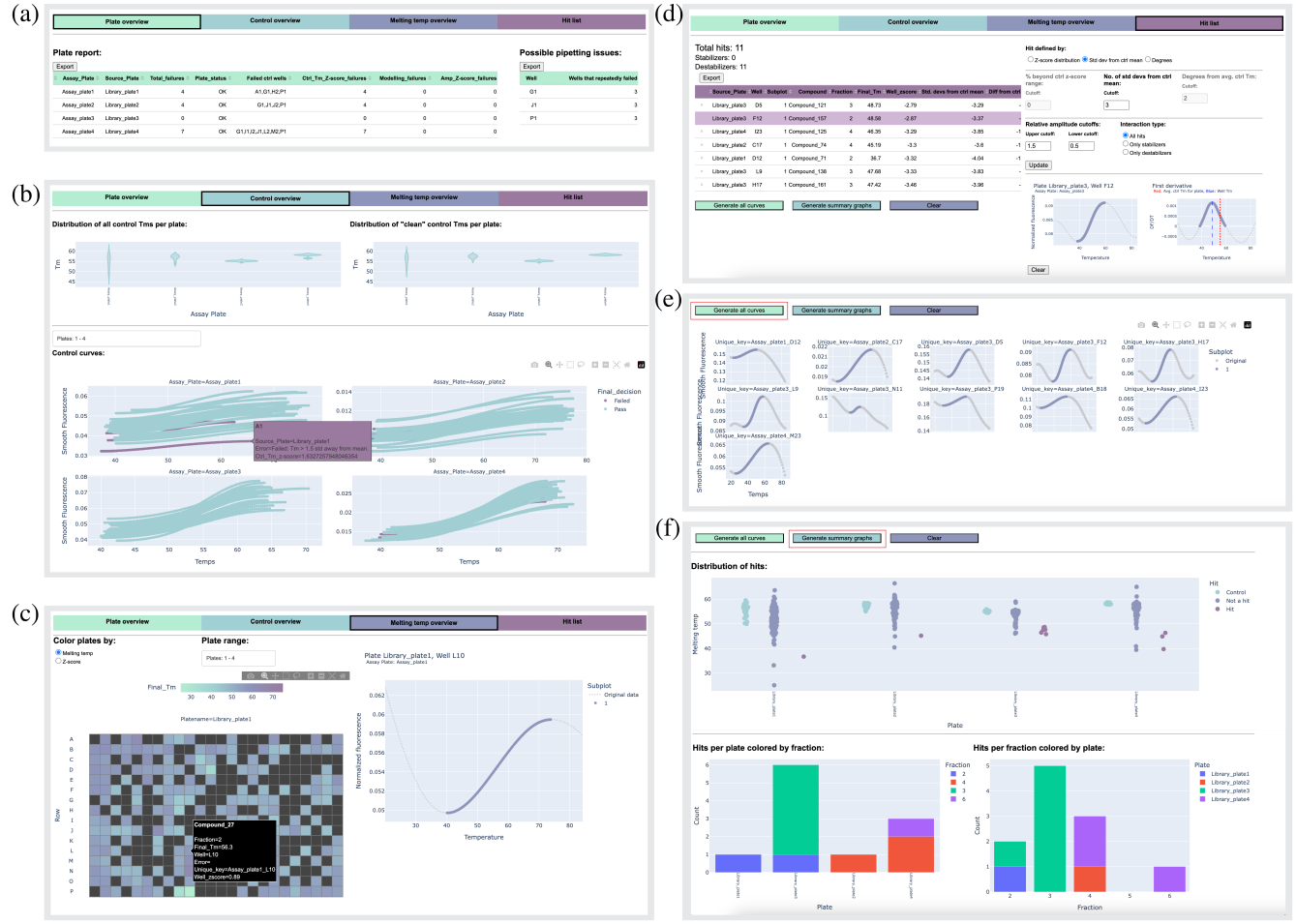

FIGURE 2 The ShiftScan Viewer companion tool offers comprehensive data exploration features. (a) The 'Plate overview' tab summarizes control well failures for each plate and lists wells that have failed in three or more plates. (b) The 'Control overview' tab visually presents control Tm value distributions before and after quality control and displays the individual control curves per plate, colored according to whether they passed or failed quality control. (c) The 'Melting temp overview' tab allows users to view data in a plate format (wells can be colored by Tm or *z*-score), providing additional information and graphics by hovering over or clicking on a well. (d) The 'Hit list' tab enables users to choose and adjust hit definition parameters, where clicking on a row produces two graphs: the original and first derivative curves with associated subcurves. (e) The 'Generate all curves' button creates a matrix of all curves currently in the hit list. (f) The 'Generate summary graphs' button produces a distribution plot of hits, controls, and non-hit experimental Tm values, and two bar plots breaking down the hits by plate and fraction.

In the fourth tab, 'Hit list' (Figure 2d), the user is presented with a table of hits as defined by default parameters on the left of the screen. Counts of the total number of hits, the number of stabilizing hits (i.e., the Tm is increased in the presence of the compound/fraction), and the number of destabilizing hits (i.e., the Tm is decreased in the presence of the compound/fraction) are displayed above the table. To the left are a variety of parameters that the user may change which will alter the final number of hits. These include the choices for how to define hits (the newly developed *z*-score method described here, *N* standard deviations away from the control mean Tm, or *N* degrees away from the control mean Tm), the acceptable relative amplitude of the experimental curve (relative to the average control amplitude), and whether they would like to see all hits, only stabilizers, or only destabilizers. Clicking on a row in the table will trigger a pop-up to the right, displaying two graphs. The first graph shows the original data as a dotted gray line with the isolated sigmoidal subcurve(s) in color (termed "Subplots" here) The second graph presents the first derivative of the data, where subcurve Tm value(s) are marked by a blue line and the average control Tm for the plate is marked by a red line. Additionally, a user may opt to delete a predicted hit from the list by clicking the small gray cross to the left of the row. Finally, the user may generate all curves associated with hits by clicking the 'Generate all curves' button (Figure 2e), or a summary of the hits which includes their distribution and breakdown by plate and fraction combination (Figure 2f). Following the user's modifications the final list of hits can be exported and downloaded to their local computer by clicking the "Export" button above the table of hits. Similarly, images of all figures can be exported in Portable Network Graphics (PNG) format by clicking the

small camera icon that appears on the top right-hand side of the figure when a user hovers over the figure in question.

2.1.5 | Disk-intensive mode

The default mode of ShiftScan is highly efficient for systems with ample RAM, making it ideal for processing hundreds of plates quickly. For systems with limited memory, ShiftScan can be run in a disk-intensive mode, which writes intermediate files to disk to conserve RAM. This mode is slower and will use a larger portion of disk space but ensures equally accurate results without overloading the user's system's memory. In this mode, each plate is loaded and processed as described above, however, intermediate output is written to appropriate files. Intermediate files are deleted upon processing completion leaving only the four output files as generated by the default RAM-intensive mode. Differences in performance between the two modes are detailed in Section 2.3.

2.2 | Installation and basic usage

Detailed instructions for installation and usage can be found in the ShiftScan GitHub repository: https://github.com/samche42/ShiftScan. Briefly, the repository files can be cloned or copied to the user's local machine, and a conda environment can be created and initiated containing all dependencies necessary for ShiftScan to run. At a minimum, the user will be required to specify the folder in which the input files can be found, the associated metadata, and where to place the resultant output files. There are several additional parameters that the user can adjust, such as the location of controls, quality control cut-offs, and the desired smoothing factor. Default values are provided for all parameters and are detailed in the GitHub repository. Demonstration data is included in the GitHub repository. This is to help users adhere to the necessary data formatting guidelines and ensure that the algorithm and visualizations function as intended.

In addition to this command-line implementation of the algorithm, we have made two additional implementations available for those less comfortable with the command-line interface: ShiftScan and the associated ShiftScan Viewer are also available in the form of a Google Colab notebook (Link) and a graphical user interface (GUI)-based application (Link). It is strongly encouraged that users refer to the Github repository for usage instructions. The application is the most user-friendly but is significantly slower than the command line version and is currently only available for MacOS. The Google Colab notebook serves as an intermediate solution, simplifying the command-line experience. Users are encouraged to choose the implementation that best aligns with their preferences and technical expertise.

2.3 | SHIFTSCAN performance

Processing high-throughput data from DSF assays poses significant challenges due to the large size of datasets and the time required to transform raw data into meaningful results. To address this, we implemented a multiprocessing feature in our algorithm. Further, the algorithm is available in two modes: RAM-intensive (default) and disk-intensive.

We evaluated processing time and RAM usage across 4–16 CPUs for datasets comprising 20–200 plates for both modes. The RAM-intensive mode was tested on an exclusive node on a high-performance Linux-based computing cluster owing to the RAM requirements (∼20GB RAM for 200 plates). The disk-intensive mode was tested on a personal MacBook Pro laptop with 16 GB RAM and 8 cores available. Increasing the number of CPUs improves overall processing time in both modes (Figure 3a,b), however, we did note an exponential decay in the average processing time per plate in both modes (Figure 3a). This is likely due to a combination of unavoidable serial steps in the data processing (i.e., processing that cannot be parallelized) and the overhead required to manage the increasing number of processes, that is, the more processes there are, the more communication and coordination between processors is required, which can limit the efficiency of the individual additional processors. The user is therefore advised to consider the balance between the number of CPUs deployed and the trade-off in efficiency relative to the size of their dataset.

As expected, the default RAM-intensive mode is significantly faster than the disk-intensive mode, processing each plate in an average of 3.5 s—3.4 times faster than the 12 s per plate in disk-intensive mode (Figure 3a,b). Memory usage per plate in the disk-intensive mode (0.01 GB ∓ 0.007) is approximately 10 times lower than the RAM-intensive mode (0.1 GB ∓ 0.002) (Figure 3c). In the default RAM-intensive mode, memory requirements scale linearly with the number of input files, averaging 0.1 GB per file (Figure 3d). For example, processing 160 data files from 384-well plates would require about 16 GB of RAM. We recommend adding approximately 20% to the estimated memory usage to account for system overhead. The disk-intensive mode, designed for users with limited computational resources, performs well, requiring only an average of 0.83 GB (∓ 0.28 GB) of memory regardless of the number of input files or CPUs deployed (Figure 3d).

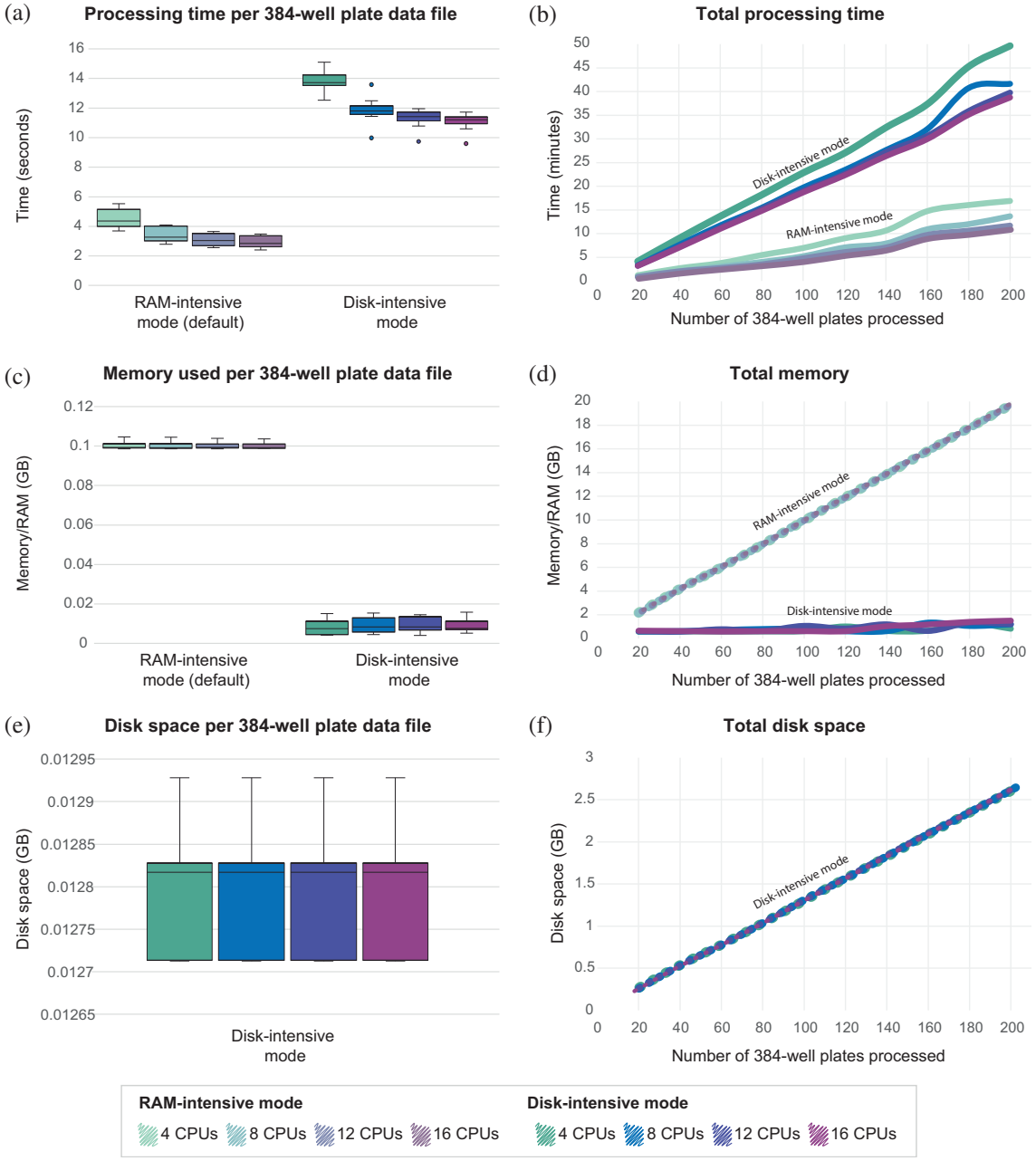

FIGURE 3 Performance of the RAM-intensive and disk-intensive modes available in ShiftScan. The parameter sweep of both modes was performed using datasets of 20–200 data files (increasing in increments of 20 files), across 4, 8, 12, and 16 CPUs. Each data file was taken from a 384-well plate with an average of 256 temperature points. (a) Distribution of processing times per data file. (b) Total processing times for each dataset. (c) Distribution of memory (RAM) usage per data file. (d) Total memory (RAM) used for each dataset. (e) Distribution of disk space usage for each dataset in the disk-intensive mode. (f) Total peak disk space usage for each dataset in the disk-intensive mode.

Finally, the disk-intensive mode relies on intermediate files being stored on the hard drive. This mode requires an average of 0.013 GB (12.67 MB) of disk space per data file from a 384-well plate (Figure 3e). The disk space scales linearly with the number of files in the input dataset (Figure 3f), for example, processing data from 200 plates in this mode will require approximately 2.6 GB of available disk space for intermediate files.

2.4 | SHIFTSCAN benchmarking

We considered DSFWorld to be the state-of-the-art tool due to its ability to assess curves following non-canonical transition patterns and used it as a benchmark to measure the accuracy of ShiftScan. To achieve this, we ran DSFWorld using the example data and layout provided on its website (https://gestwickilab.shinyapps.io/dsfworld/). We selected all four available models for

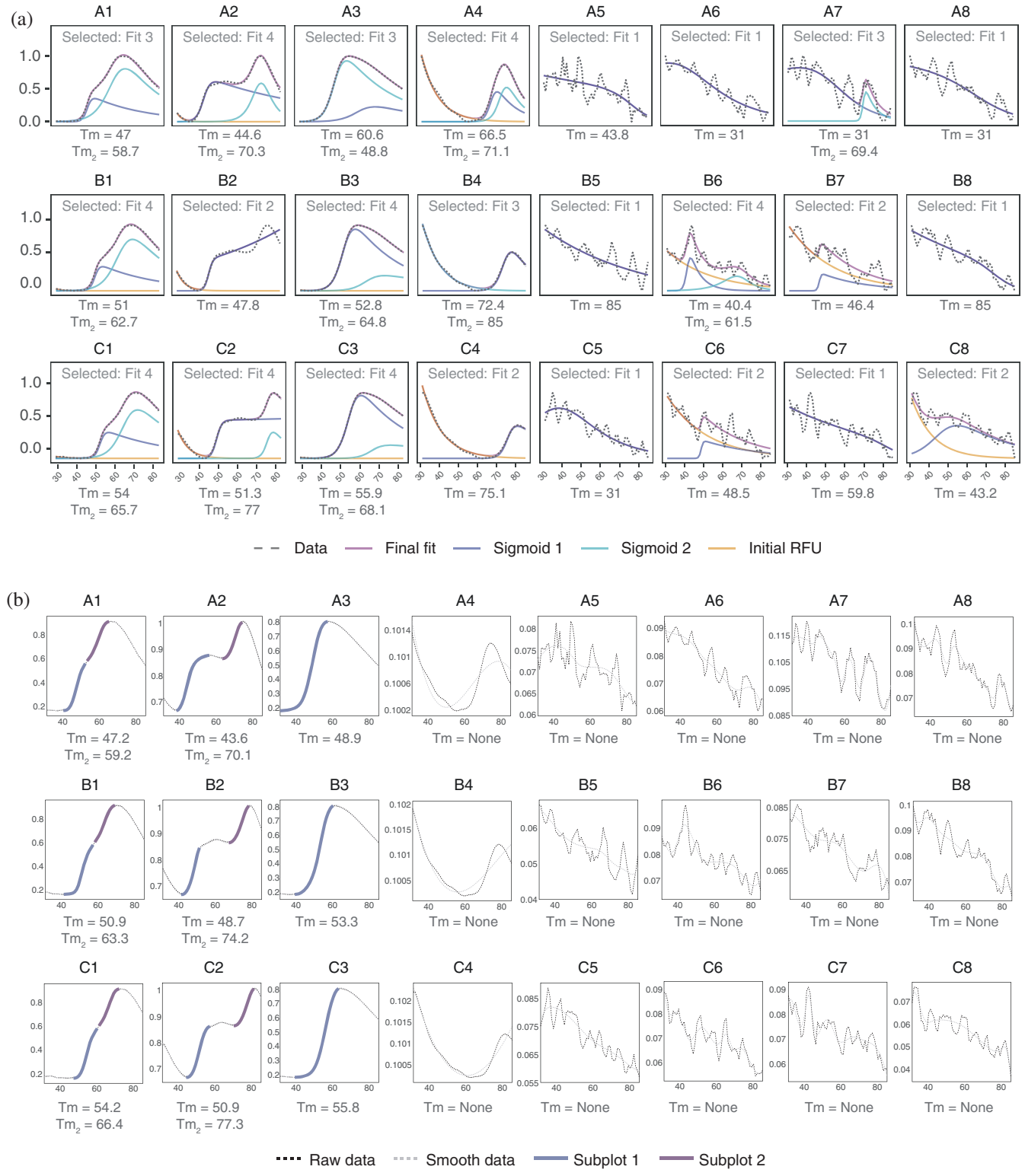

FIGURE 4 Assessment of ShiftScan performance relative to the DSFWorld online tool, using example data supplied by DSFWorld. All plots show normalized fluorescence values on the *y*-axes and temperatures on the *x*-axes. Predicted Tm values for detected transitions are below each graph. (a) Curve fitting and Tm calling using all four models available in DSFWorld. Plots were produced in DSFWorld and exported. Best fit predicted Tm values are provided for each well dataset. The different models are shown by default in DSFWorld as indicated in the legend. Raw data is shown with a dotted line and all other models are shown with colored lines. Default colors as plotted in DSFWorld were altered for easier reference here. (b) Curve fitting and Tm value calling in ShiftScan. Different transitions ("Subplots") are indicated with blue and purple colors for each dataset. Raw and smoothed data are shown for each dataset with dotted lines.

sigmoidal fitting, instructed the algorithm to choose the best fit for each dataset, and exported the plotted results (Figure 4a). After exporting all predicted melting temperature values (Dataset S2), we retained those identified as the best fit for each well (Table 1, Dataset S2). We then analyzed the same data with ShiftScan, using column 3 as the 'controls' since these represented canonical transitions.

ShiftScan successfully found double transitions in the six wells in columns 1 and 2, and a single transition in column 3 (Figure 4b). There was an average difference of 0.15°C in melting temperatures for each transition in columns 1–3 as predicted by the two tools (Table 1). A two-tailed, paired T-test revealed that the predicted Tm values from the two tools for these transitions were not statistically significantly different ($p = 0.28$, $\alpha = 0.05$). The data in columns 5–8 are examples of noisy and inconclusive data. ShiftScan successfully failed these wells: The smoothed data was either recognized as following a negative trend or, if any sigmoidal regions were detected, they were removed due to the small relative amplitudes of these sigmoid curves in comparison to the "controls" (Dataset S2). Finally, ShiftScan successfully detected the sigmoidal regions in data from column 4, however, while these sigmoidal regions failed due to small relative amplitudes, visual inspection revealed that smoothing of the data was not optimal. To assess whether the small amplitudes of the curves (as a result of normalization across all datasets) were problematic, we processed the data for these three specific wells (A4, B4, and C4) using only the Tm calling portion of the ShiftScan algorithm, that is, no comparison is made with control wells (please see Jupyter notebook for full detailed code: https://zenodo.org/records/13838488, DOI: 10.5281/zenodo.13838487). In all three cases, the sigmoidal regions were adequately smoothed, and the predicted Tm values matched those obtained with the DSFWorld algorithm (Figure S2). These results suggest that the issue with fitting the data may be due to inappropriate normalization across datasets that should not be analyzed together as a single plate, rather than a limitation of the algorithm in identifying the sigmoidal region. Finally, DSFWorld predicted double transitions (i.e., two Tm values) for wells A3, A4, B3, B4, and C3, as highlighted in Table 1. Although the source of these

TABLE 1 Comparison of melting temperatures predicted by DSFWorld and ShiftScan using example data taken from the DSFWorld website.

		DSFWorld results			ShiftScan results		Difference?	
Well	Condition	Best fit	Tm1	Tm2	Tm1	Tm2	Tm1	Tm2
A1	Compound1_0uM_Protein1	Fit 3	47	58.7	47.2	59.2	−0.19	−0.51
A2	Compound2_0uM_Protein2	Fit 4	44.6	70.3	43.6	70.1	1.03	0.22
A3	Compound3_0uM_Protein3	Fit 3	60.6	48.8	48.9		−0.10	
A4	Compound4_0uM_Protein4	Fit 4	66.5	71.1	Failed			
A5	Compound1_0uM_Buffer	Fit 1	43.8	N/A	Failed			
A6	Compound2_0uM_Buffer	Fit 1	31	N/A	Failed			
A7	Compound3_0uM_Buffer	Fit 3	31	69.4	Failed			
A8	Compound4_0uM_Buffer	Fit 1	31	N/A	Failed			
B1	Compound1_12.5uM_Protein1	Fit 4	51	62.7	50.9	63.3	0.11	−0.55
B2	Compound2_12.5uM_Protein2	Fit 2	47.8	N/A	48.7	74.2	−0.92	
B3	Compound3_12.5uM_Protein3	Fit 4	52.8	64.8	53.3		−0.47	
B4	Compound4_12.5uM_Protein4	Fit 3	72.4	85	Failed			
B5	Compound1_12.5uM_Buffer	Fit 1	85	N/A	Failed			
B6	Compound2_12.5uM_Buffer	Fit 4	40.4	61.5	Failed			
B7	Compound3_12.5uM_Buffer	Fit 2	46.4	N/A	Failed			
B8	Compound4_12.5uM_Buffer	Fit 1	85	N/A	Failed			
C1	Compound1_25uM_Protein1	Fit 4	54	65.7	54.2	66.4	−0.18	−0.74
C2	Compound2_25uM_Protein2	Fit 4	51.3	77	50.9	77.3	0.37	−0.25
C3	Compound3_25uM_Protein3	Fit 4	55.9	68.1	55.8		0.08	
C4	Compound4_25uM_Protein4	Fit 2	75.2	N/A	Failed			
C5	Compound1_25uM_Buffer	Fit 1	31	N/A	Failed			
C6	Compound2_25uM_Buffer	Fit 2	48.5	N/A	Failed			
C7	Compound3_25uM_Buffer	Fit 1	59.8	N/A	Failed			
C8	Compound4_25uM_Buffer	Fit 2	43.2	N/A	Failed			

Note: Cells highlighted in red indicate second Tm values reported for transition patterns that appear to follow a canonical trend.

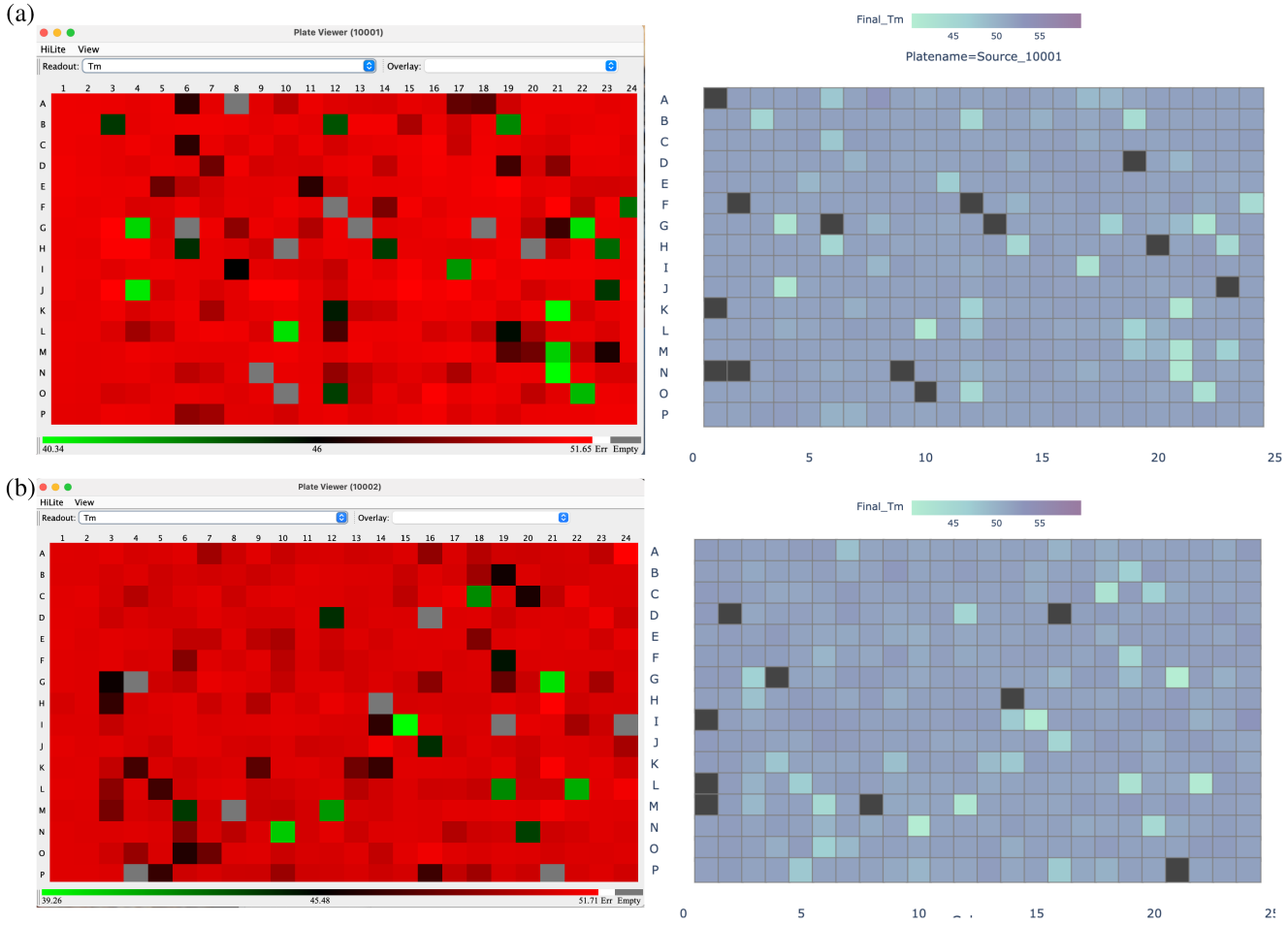

FIGURE 5 Comparison of Tm values in plates (a) 10,001 and (b) 10,002, as taken from the KNIME public dataset as assessed in the KNIME workflow (left) and ShiftScan (right). Failed wells are indicated in light gray for the KNIME workflow and in black for ShiftScan. Continuous color scales for Tm value estimates are provided for each tool's output.

double transitions is unclear, since the modeling appears optimal (Figure 4a), we assumed that users would need to filter out such values through visual inspection of their data. However, visual inspection is impractical for high-throughput data. In contrast, ShiftScan predicted only a single transition with an associated Tm value for wells A3, B3, and C3, which aligns with our expectations and negates the need for user inspection of all data.

Finally, we assessed a dataset of two 384-well plates sourced from the KNIME workflow (Samuel et al., 2021) (https://github.com/loicsamuel/knime-tsa-analysis/tree/main) using ShiftScan and the KNIME workflow. Repeated attempts were made to analyze the same, appropriately formatted, data with DSFWorld but each attempt resulted in either an "Ajax error" or being disconnected from the server. The developers were contacted about the issue, but we had received no response at the time of submission.

Using default parameters, ShiftScan successfully processed both plates in approximately 2 min, and the KNIME workflow (using parameters as set out in the tutorial) successfully processed the two plates in approximately 40 min. A comparison of the predicted Tm values for non-failed wells from both plates showed an average difference of 1.05°C between the two tools (Dataset S3). A two-tailed, paired *T*-test revealed that the predicted Tm values from the two tools for these transitions were statistically significantly different ($p = 4.62^{-28}$, $\alpha = 0.05$). A correlation analysis of Tm values generated by the two tools reveals a strong, positive linear relationship (R^2 values of 0.979 and 0.967 for the two plates, Figure S3a), suggesting that the two tools produce highly similar trends in Tm estimates. However, ShiftScan consistently yields slightly higher Tm values (Figure S3b), indicating a systematic offset rather than random variation. This difference in Tm value estimates may be a result of the order of operations by which these two tools differ in the data processing, that is, the KNIME workflow smooths and then normalizes the data, whereas ShiftScan first normalizes the data and then applies the smoothing function. As hit identification is a relative function of well Tm values, and the patterns are consistent between the

two approaches (Figure 5), we believe that this difference is negligible.

3 | CONCLUSIONS

Prioritizing active compounds is a crucial step in drug discovery. DSF offers a rapid, scalable method for primary screening. However, traditional data analysis methods are not adequate for the scale of data from high-throughput DSF. ShiftScan addresses this gap by quickly and accurately predicting melting temperatures (Tm) from thousands of DSF curves, accommodating both canonical and non-canonical melting curves. Its built-in quality checks reduce false positives compared to conventional modeling, thereby enhancing reliability. The companion tool, ShiftScan Viewer, further empowers researchers to visually inspect and refine their data, making ShiftScan a valuable asset for drug discovery scientists.

AUTHOR CONTRIBUTIONS

Samantha C. Waterworth: Conceptualization; investigation; writing – original draft; methodology; validation; visualization; writing – review and editing; software; formal analysis; data curation. **Shilpa R. Shenoy:** Conceptualization; methodology; writing – review and editing; data curation. **Nirmala D. Sharma:** Data curation; writing – review and editing; methodology. **Chris Wolcott:** Conceptualization; data curation; formal analysis; investigation; software; writing – review and editing. **Duncan E. Donohue:** Conceptualization; investigation; writing – review and editing; software; formal analysis; data curation. **Barry R. O'Keefe:** Conceptualization; project administration; resources; supervision; writing – review and editing; funding acquisition. **John A. Beutler:** Project administration; supervision; resources; writing – review and editing; funding acquisition.

ACKNOWLEDGMENTS

This research was supported in whole or in part with Federal Funds from National Cancer Institute, Center for Cancer Research, National Institutes of Health, Department of Health and Human Services, Intramural Program, under project number ZIA BC 011471 (B.R.O.) Cell-free assay technologies for the identification of active compounds, and ZIA BC 011469 (J.A.B.) This work was also supported by the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N91019D00024.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All code and example data are available from the ShiftScan GitHub repository: https://github.com/samche42/ShiftScan. All additional code and data used for benchmarking can be found in the Zenodo repository, DOI: 10.5281/zenodo.13838487.

ORCID

Samantha C. Waterworth https://orcid.org/0000-0001-6436-0142

REFERENCES

An L, Chen L, Hao X. Indoor fire detection algorithm based on second-order exponential smoothing and information fusion. Information. 2023;14:258.

Ciulli A. Biophysical screening for the discovery of small-molecule ligands. Methods Mol Biol. 2013;1008:357–88.

Coma I, Herranz J, Martin J. Statistics and decision making in high-throughput screening. In: Janzen WP, Bernasconi P, editors. High throughput screening: methods and protocols. 2nd ed. Totowa, NJ: Humana Press; 2009. p. 69–106.

Dai R, Geders TW, Liu F, Park SW, Schnappinger D, Aldrich CC, et al. Fragment-based exploration of binding site flexibility in Mycobacterium tuberculosis BioA. J Med Chem. 2015;58:5208–17.

Douse CH, Vrielink N, Wenlin Z, Cota E, Tate EW. Targeting a dynamic protein-protein interaction: fragment screening against the malaria myosin A motor complex. ChemMedChem. 2015;10:134–43.

Fotsch C, Basu D, Case R, Chen Q, Koneru PC, Lo M-C, et al. Creating a more strategic small molecule biophysical hit characterization workflow. SLAS Discov. 2024;29:100159.

Gao K, Oerlemans R, Groves MR. Theory and applications of differential scanning fluorimetry in early-stage drug discovery. Biophys Rev. 2020;12:85–104.

Goktug A, Chai SC, Chen T. Data analysis approaches in high throughput screening. In: El-Shemy HA, editor. Drug discovery. London: InTech; 2013.

Grkovic T, Akee RK, Thornburg CC, Trinh SK, Britt JR, Harris MJ, et al. National Cancer Institute (NCI) program for natural products discovery: rapid isolation and identification of biologically active natural products from the NCI prefractionated library. ACS Chem Biol. 2020;15:1104–14.

Hanley QS. The distribution of standard deviations applied to high throughput screening. Sci Rep. 2019;9:1268.

Hansel CS, Lanne A, Rowlands H, Shaw J, Collier MJ, Plant H. High-throughput differential scanning fluorimetry (DSF) and cellular thermal shift assays (CETSA): shifting from manual to automated screening. SLAS Technol. 2023;28:411–5.

Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R. Statistical practice in high-throughput screening data analysis. Nat Biotechnol. 2006;24:167–75.

Martin-Malpartida P, Hausvik E, Underhaug J, Torner C, Martinez A, Macias MJ. HTSDSF explorer, a novel tool to analyze high-throughput DSF screenings. J Mol Biol. 2022;434:167372.

Martin-Malpartida P, Torner C, Martinez A, Macias MJ. TPPU_DSF: a web application to calculate thermodynamic parameters using DSF data. J Mol Biol. 2024;436(17):168519. https://doi.org/10.1016/j.jmb.2024.168519

Matarlo JS, Krumpe LRH, Heinz WF, Oh D, Shenoy SR, Thomas CL, et al. The natural product butylcycloheptyl prodiginine binds pre-miR-21, inhibits dicer-mediated processing of pre-miR-21, and blocks cellular proliferation. Cell Chem Biol. 2019;26:1133–1142.e4.

Pantoliano MW, Petrella EC, Kwasnoski JD, Lobanov VS, Myslik J, Graf E, et al. High-density miniaturized thermal shift assays as a general strategy for drug discovery. J Biomol Screen. 2001;6:429–40.

Reidenbach AG, Mesleh MF, Casalena D, Vallabh SM, Dahlin JL, Leed AJ, et al. Multimodal small-molecule screening for human prion protein binders. J Biol Chem. 2020;295:13516–31.

Samuel ELG, Holmes SL, Young DW. Processing binding data using an open-source workflow. J Chem. 2021;13:99.
Scholle MD, McLaughlin D, Gurard-Levin ZA. High-throughput affinity selection mass spectrometry using SAMDI-MS to identify small-molecule binders of the human rhinovirus 3C protease. SLAS Discov. 2021;26:974–83.
Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52:591–611.
Sun C, Li Y, Yates EA, Fernig DG. SimpleDSFviewer: a tool to analyze and view differential scanning fluorimetry data for characterizing protein thermal stability and interactions. Protein Sci. 2020;29:19–27.
Sztuba-Solinska J, Shenoy SR, Gareiss P, Krumpe LRH, Le Grice SFJ, O'Keefe BR, et al. Identification of biologically active, HIV TAR RNA-binding small molecules using small molecule microarrays. J Am Chem Soc. 2014;136:8402–10.
Thornburg CC, Britt JR, Evans JR, Akee RK, Whitt JA, Trinh SK, et al. NCI program for natural product discovery: a publicly-accessible library of natural product fractions for high-throughput screening. ACS Chem Biol. 2018;13:2484–97.
Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods. 2020;17:261–72.
Wu T, Gale-Day ZJ, Gestwicki JE. DSFworld: a flexible and precise tool to analyze differential scanning fluorimetry data. Protein Sci. 2024;33:e5022.
Wu T, Hornsby M, Zhu L, Yu JC, Shokat KM, Gestwicki JE. Protocol for performing and optimizing differential scanning fluorimetry experiments. STAR Protoc. 2023;4:102688.
Zhang JH, Chung TD, Oldenburg KR. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. J Biomol Screen. 1999;4:67–73.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Waterworth SC, Shenoy SR, Sharma ND, Wolcott C, Donohue DE, O'Keefe BR, et al. ShiftScan: A tool for rapid analysis of high-throughput differential scanning fluorimetry data and compound prioritization. Protein Science. 2025;34(3): e70055. https://doi.org/10.1002/pro.70055