Helios - History and Anatomy of a successful in-House Enterprise High-Throughput Screening and Profiling Data Analysis System

Hanspeter Gubler[1], Nicholas Clare[1], Laurent Galafassi[1], Uwe Geissler[1,2], Michel Girod[1], Guy Herr[1]

[1] Novartis Institutes for BioMedical Research, NIBR Informatics Department, CH-4002 Basel, Switzerland

[2] Current affiliation: Cognizant Business Consulting, CH-8048 Zurich, Switzerland

Supplemental Information

Table of Contents

*) incl. correction of a typesetting error for the initial value estimation expression for Hill slope parameter in original Supplemental Information document

1 - Types of Calculation Modules available in the Helios Application

A very high percentage of screening and compound profiling assays can be fully analyzed with just these first 4 types of calculation modules:

1. **Data transformation** – simple arithmetic operations on one or multiple raw (or intermediate) input data, transformation functions (e.g. log-transformation)
2. **Data normalization** – control-based or sample-based %-effect determination, *Z*- or other (robust) score value determination[1] (plate wise or plate batch-wise). Multiple normalization models for data scaling between neutral and active control averages, or with neutral controls alone (e.g. 'fold change' calculations) are available. Variance-based scaling to obtain score values.
3. **Plate response data correction** – multiple methods are available for plate-wise or plate-batch wise (spatial and temporal) pattern detection and correction[1,11,12,13] with visualizations and numerical outputs of systematic pattern and corrected values. The most frequently used plate data correction methods are robust local regression (plate-wise) and principal pattern modeling (plate-batch-wise): The pattern models for describing responses at row and column coordinates r and c can be represented as
   - $\hat{y}(r,c) = RobustLocalRegression(y(r_i, c_j), h)$, where $\hat{y}$ is the 'smoothed' response and where robustness (outlier resilience) is obtained by prior weighting of response data based on their distribution and deviation from the median of the of the neutral controls $\tilde{y}_{NC}$ on a particular plate, iterated reweighting of the residuals (using Tukey's biweight[22]) of the local regression surface using bandwidth *h* for every row and column position r_i and c_j and refitting until $\hat{y}$ has converged.[11] This procedure has in practice a very high robustness breakdown margin, often aided by the prior weighting step which provides good starting values for the *initial* local regression surface.
   - $\hat{y}_k(r_i,c_j) = SVD_{k,t}^{-1}(SVD(y_k(r_i,c_j), k=1..n)$, where the model response $\hat{y}$ for each plate *k* in a batch of *n* plates is based on the determination of the principal patterns across all plates based on singular value decomposition of the normalized response data matrix and retaining only the *t* leading components in the back-calculation (symbolically denoted as SVD_t^{-1}).[1,11] Variants of this procedure can additionally employ outlier resistant representations of the principal patterns (e.g. again by modeling the patterns

as derived by SVD as a smooth local regression surface if this is merited, i.e. if a particular systematic error is not confined to a single, or a few wells only), or plate data going into the SVD step can be pre-filtered for outliers, e.g. based on a simple and fast median polish preprocessing step. All these variants are available in the Helios application as separate modules and users need to assess the merits of the various approaches by reviewing outcomes aided by summary statistics of corrected data. This family of methods needs *at least* 10 plates to enter the analysis, otherwise there is a danger that the model will pick up too much noise and/or represent actually active wells as a 'pattern to be corrected away'. An additional easy visual cue to detect pattern-model 'overfitting' is the appearance of seemingly random 'noise' and 'speckles' in the pattern heatmaps which are similar to the 'pattern' of the (likely) active wells in the uncorrected data set. When dealing with a small number of plates *n* this can already happen when considering the leading principal pattern components below the limit *t*, whereas in a stable model the noise components will only appear later in the pattern series.

- *Median-polish*[12,22] based pattern modeling approaches are not often used by our scientists because modeling and correction artifacts occur relatively frequently with these methods, mainly due to the fact that some of the row-wise and column-wise model terms can have disturbing effects on the model values elsewhere on the plates due to possible long-range spatial correlations.
- Given the model response values $\hat{y}$ and measured values y at each well location the 'corrected sample responses' r can be obtained through different *data correction approaches* (additive: $r_a = y - \hat{y}$, multiplicative: $r_m = y$ median($\hat{y}$) / $\hat{y}$, score calculation: $R\text{-}score = r_a$ / mad(r_a)).[1,11,12,13]

4. **Dose-response curve fitting** – with various curve fit visualizations (single curves, tiled curves, curve overlays), curve clustering to easily identify similar curve shapes and classes in a plate group working set, as well as a curve classification engine, using freely definable classification rules, etc. See **sections 2 and 3** of this **Supplemental Information** document for more details.

Auxiliary data reduction and result derivation steps can be applied when necessary using these modules:

5. **Calibration-curve fitting** and concentration lookup for all wells of a plate or the full plate group (standard curve fitting and analyte quantitation by interpolation, includes determination of limits of quantitation for both accuracy and precision of standard curve recovery values)
6. **Custom transformations**, free mathematical expression entry and execution by using the available catalog of basic mathematical and statistical functions, arithmetic and logical operations with up to 5 different input connections and a single output. The custom transformation module supports array arithmetic which allows to write expressions in an intuitive high level fashion, e.g. as 'X1/(X1 + X2)', where X1 and X2 are the module input datasets (plate readouts) and the operations are automatically executed for all wells. Statistical functions can be applied to subsets of wells, either particular well types or enumerated wells, e.g. as in 'mean(X1, NC)' for the mean of the neutral controls on a plate, or 'mean(X1, myWellSubset)' for a set of wells which are identified and named in a graphical layout display. The expression interpreter is based on the JEP parsing software package (Singular Systems, Edmonton, Alberta, Canada).
7. **Well-level masking modules**, e.g. to allow automated knock-out of values from plate wells which have experienced a liquid handling problem (e.g. by channeling of acoustic dispensing logging information as a 'readout' into the analysis), or where other measured quantities lie outside some define tolerance range (e.g. when total FRET intensity measurements are 'too small' (due to signal quenching) or 'too large' (due to auto-fluorescence).
8. **Automatic plate-level flagging (annotation) or masking (knock-out)** can be accomplished by setting suitable thresholds on plate-level QC summary statistics (e.g. a Z' factor) at each module output. So for this purpose no special calculation module is required.
9. Linear- and first order **kinetic time-course fitting** – i.e. linear and exponential time-course fitting with possibility to use the fitted reaction rate parameter *k* for each well as 'readout' for the downstream modules. These can include transformation, normalization, pattern correction and dose-response curve fitting, among others. The kinetic modules have auxiliary outputs for the standard error of the rate parameter and the coefficient of determination R^2.

All calculation modules possess a number of preset default parameters which are chosen to be appropriate for the majority of use cases (e.g. most frequent type of normalization expression, choice of median of the controls on each plate as average for normalization, choice of 'inhibition' experiments to indicate 'direction' of the signals, default parameters for outlier removal in the control sample area, etc.) and this allows for very quick assembly and definition of new, or alternate, ways of result calculations, also for simple exploratory purposes.

2 - Curve Fitting Progression Steps and Algorithmic Aspects

Each individual concentration response curve is analyzed by going through several steps:

1. First, a prior nonparametric fit to data $\hat{y}(x) = f_{initial}^{NP}(x_i, y_i)$ is done in order to determine the key curve characteristics and to already generate a good set of starting parameter estimates for a possible follow-up Hill-curve model fit. Having good starting values is especially important when employing outlier resistant parametric regression methods in a later step. $f_{initial}^{NP}$ is a chosen to be a spline approximation function.[21] And the *initial parameters* for a later parametric 4PL model fit (see step 3) are derived as e.g. $A_{0,init} = \hat{y}(x: x = \min(x))$, $A_{inf,init} = \hat{y}(x: x = \max(x))$, $AC_{50,init} = \{x \mid \hat{y}(x) = (A_{0,init} + A_{inf,init}) / 2\}$, and the Hill slope α as $\alpha = 4 \cdot (df/dx)_{x=AC50,init} / (ln(10) \cdot (A_{inf,init} - A_{0,init}))$
2. If it is found that a sensible Hill-curve model fit will most likely not be possible because only insignificant effect variations across the concentration series are present, then we only determine the median effect level **(**giving a possible indication of normalization problems or a small residual effect if it is significantly different from 0), but we do not attempt to generate a more complex model fit: $\hat{y}(x) = median(y_i)$
3. If a significant response variation is detected, then we attempt a parametric (sigmoid) model fit: A 4-parameter logistic (4PL), or Hill-curve fit, $\hat{y}(x) = A_{inf} + (A_0 - A_{inf})/(1 + 10^{\alpha(log_{10}x - log_{10}AC_{50})})$ with both the default outlier resistant (IRLS)[17] and classical non-linear regression optimization methods available. In this model A_0 and A_{inf} are the low and high concentration plateau values, $AC_{50\ is}$ the inflection point concentration and α the Hill slope parameter. We note that fitting of the 4PL Hill-curve model is often only a convenient (and frequently used) method to provide a phenomenological description for concentration-response relationships, even if the simple mechanistic assumptions of the Hill theory are not fulfilled in practice, especially for complex cell-based assay systems. The IRLS procedure which is used for outlier resistant fitting is based on using Tukey's biweights[22] which are recalculated on each iteration, and serve to down-weight the influence of outlier data points. Residual values between data and fit are denoted by r_i and the weights $w_B(r_i)$ are calculated as $w_B(r_i) = (1 - (r_i / k)^2)^2$ for $|r_i| \le k$ and $w_B(r_i) = 0$ otherwise. The scaling value k is calculated as $k = 4.685\ mad(r_i)$, or $k = 4.685\ max(mad(r_i), err_{min})$ with a user-settable minimal error err_{min}

to guarantee numerical stability for cases with fully saturated response values (constant for the majority of data) response values. The nonlinear regression optimizer is using the Marquardt-Levenberg optimization method[23] with scaled parameter values. In cases where the modified Hessian matrix becomes close to singular the Moore-Penrose generalized inverse is used to obtain a balanced parameter solution. Subsequent covariance matrix analysis[24] will then most likely reveal high correlations and/or high parameter dependency value(s) and raise appropriate warning flags and corresponding information to the user. These mechanisms together with the 'model selection' described in this section will almost always lead to a sensibly fitted model *curve,* albeit sometimes with 'unstable' result parameters which are all appropriately flagged. Very rare complete fitting failures for rogue data sets are of course also flagged. Such automatic curve 'annotations' together with suitable statistics and visualizations (see next section of supplementary information) allow result curation for large numbers of curves in an effective manner.

4. If the fit quality (residual standard error of fitted curve) is above some preset limit, or if we already have determined through the initial nonparametric regression step (1.) that the 'response-shape' is very likely non-sigmoidal, then we perform a final nonparametric curve fit $\hat{y}(x) = f_{final}^{NP}(x_i, y_i)$, with using user selectable models: a) local polynomial regression (default), b) smoothing spline, or c) interpolation spline approaches. For example we can postulate a bell-shaped dose-response if we detect that $\hat{y}$(x: highest 2 or 3 concentrations) << max($\hat{y}(x)$), among other criteria. So, specifically for pronounced bell-shaped curves, where we presently do not attempt a representation with a parametric model and also for short dose series with 2 to 4 concentration points only we choose a purely phenomenological 'model' and derive surrogate 'parameter' values in place of the Hill parameters. For example, we approximate an AC_{50} value (IC_{50} or EC_{50}, as the case may be) by numerically determining the concentration(s) where the fitted curve values intersect with a suitably defined effect level, either with the 50% effect (absolute AC_{50})[18], or with the $(A_{min}+A_{max})/2$ effect as a fill-in value for the relative (inflection point) AC_{50}. A_{min} and A_{max} are defined as min($\hat{y}(x)$) and max($\hat{y}(x)$), respectively.
5. If the final fitted curve values are such that the observed differences between $\hat{y}(x: x = \max(x))$ and $\hat{y}(x: x = \min(x))$, or between max($\hat{y}(x)$) and min($\hat{y}(x)$) are

below a given limit (a user-set limit below which a measured assay effect difference across the measured concentration range becomes irrelevant, by default set to 30% in the Helios application), then the (constant) median effect model $\hat{y}(x) = median(y_i)$ is returned, as in step 2.

In order to assess the 'validity' of a parametric fit we perform a post-regression covariance analysis as described by Curtis[24] and provide diagnostics information to the user. This is in practice very important, as otherwise ill-conditioned and almost meaningless parameter values, (due to high inter-parameter correlations) could unknowingly end up in the result data. High parameter dependency values and the model parameters responsible for them are flagged and automatically annotated ('validity' and 'diagnostic' results, see next section). Parameter constraints can then interactively be set by the user, or the system can automatically enter an auto-constraining fitting round where parameters are only constrained to a user-defined preset interval or to a fixed value if such ill-conditioning is actually detected. In practice this is most often due to the A_{inf} (high concentration plateau value) and/or Hill slope model parameter α when curves are 'incomplete', i.e. do not reach the maximal effect and/or if the transition region appears to be 'steep' because only few, if any, experimental data points are located around the AC_{50} value. Besides the model fit parameters the curves are also always described by more 'descriptive' model-independent values, as e.g. the 'absolute AC_{50}' ($absAC_{50}$)[18] which is simply the intersection concentration of the fitted curve with the 50% effect line, among many other similar quantities, e.g. the A_{min} and A_{max} parameters mentioned earlier.

3 - Curve Quality Assessment, Curve Review and Curation

The following curve fit parameters and fit quality parameters which are created for every curve can be used to assess the results of the Helios dose-response curve fit procedure. Dose-response analysis spreadsheet columns can be easily sorted or filtered by any of these quantities to allow the user to focus on a whole subset of curves and interactively set the 'Reviewed' and 'Publish' flags for them. Very frequently a whole group of curves can be dealt with at once.

The possibility to interactively define the set of curves which can be published to warehouses was an important user requirement, but at the same time the ability to assess the quality of the fit results from various angles for a large set of curves (often several thousands in a HTS campaign) needed powerful curve selection and grouping tools. The information created through clustering, rule based classification and the ability to focus on specific aspects of data and fitting quality parameters as shown in the following table support this in a very effective and efficient way.

	statistic	comment
1	fit type	parametric, nonparametric, constant
2	curve direction	increasing, decreasing, constant
3	RSE, residual standard error of fit	
4	MRES, mad(abs res)	approximate robust equivalent of RSE
5	fit quality ratio, FQR = RSE/MRES	signals likely presence of outliers if FQR > 1.5 - 2
6	robust fit improvement, RFI= RSE(unweighted data)/RSE(weighted data)	signals likely presence of outliers if RFI > 1.5 - 2
7	coefficient of determination R^2	
8	# of masked data points	
9	# of flagged data points	
10	sum(weights)	Equals n (number of data points) if performing an unweighted fit, will likely be < n when doing a robust fit. Degree of reduction gives an indication on total 'strength' of outliers
11	validity	valid, invalid setting based on findings of post-fitting covariance analysis (parameter dependency values)
12	diagnostic	Standardized description of reason for invalid settings of previous indicator, e.g. ill-determined A_{inf}, AC_{50} or Hill-slope parameters, fit failures, etc.
13	Z', RZ'	plate level quality indicators propagated to curve results, is a median value if curve data originate from multiple assay plates
14	classification	Standardized (user selectable) rule-based curve classification based on Boolean logic/conditions of

		any of the available curve fit parameters and curve-fit quality characteristics
15	cluster	cluster number of curve-shape similarity clustering step. Allows easy identification of 'similar' curves in the data set.
16	standard errors of 4PL fit parameters or their surrogate (replacement) values originating from nonparametric fits.	
17	A_0 and A_{inf} parameter bias	Deviation of 4PL plateau values from the ideally expected values (0%, 100%), allows easy determination of normalization issues, or incomplete curves
18	rAC50 AC_{50} Ratio = AC_{50}(inflection point)/$absAC_{50}$(intersection point)	Deviations from 1 indicate possible normalization issues, incomplete curves, or partial inhibitors/stimulators
19	Delta A, $\|A_{inf} - A_0\|$	Difference of 4PL fit plateau values. Can indicate partial inhibitors or incomplete curves
20	$\|A_{max} - A_{min}\|$	Difference between maximal and minimal value of fitted curve within measured concentration range
21	$C(A_{max})$, $C(A_{min})$	Concentrations where maximal and minimal fitted curve values are reached. For example, for bell-shaped curves C(Amax) will be smaller than Cmax, the maximal concentration of the experiment
22	$C(Data_{max})$, $C(Data_{min})$	Concentrations where maximal and minimal (weighted) mean of the data values are reached. Weights are deduced from the robust IRLS fitting procedure

4 Application Benchmarks, Performance Numbers

Examples of typical application performance numbers are given in the following table:

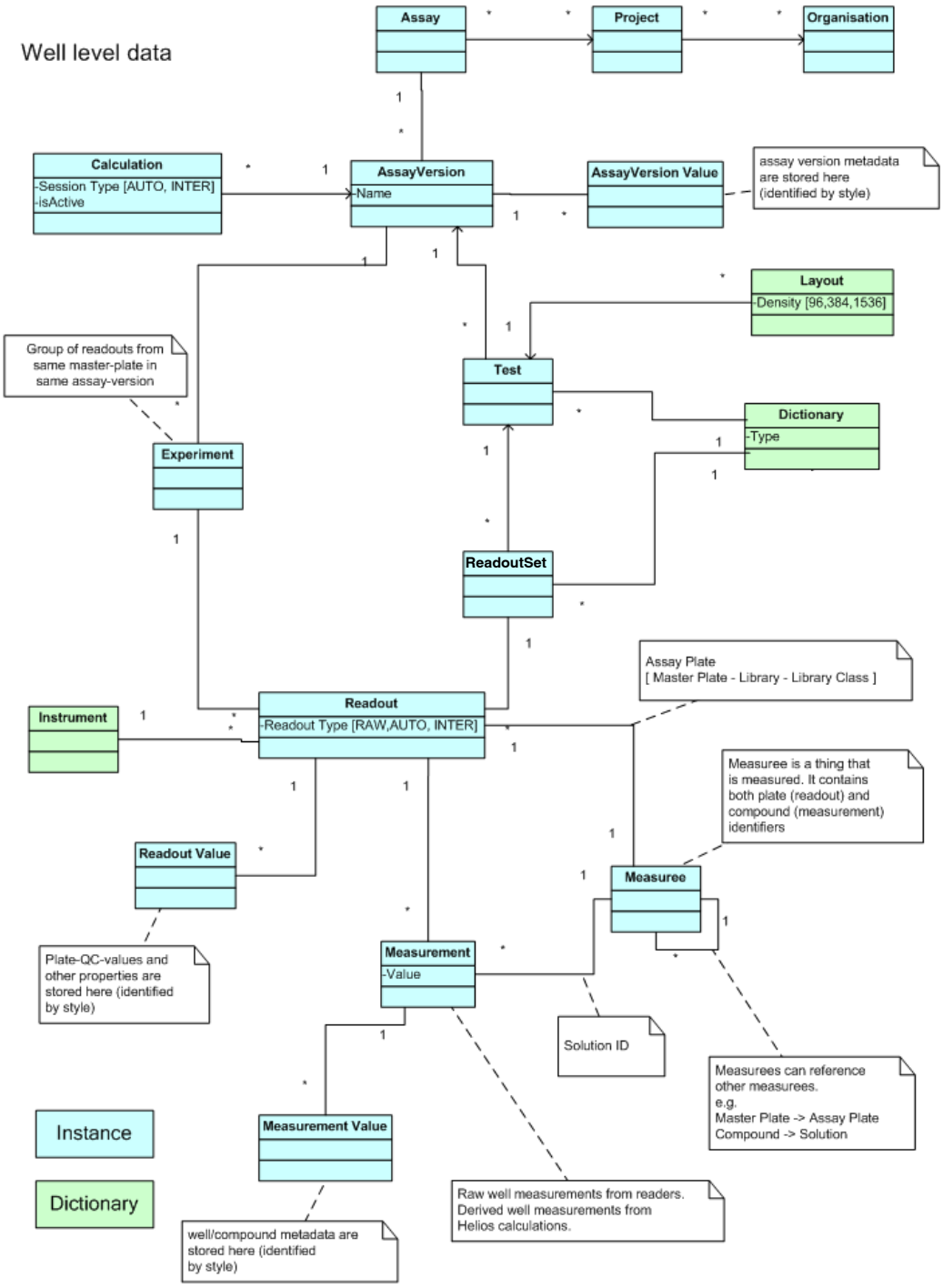
	Operation	Data Set Size	Elapsed Time	Time/Plate or Time/Curve
1	Raw data file loading, (XML parsing, consistency checks, *exp_id*, *m_id* etc. verification against assay configuration and previously loaded data of this assay and experiment, id creation, relational storage of well-level data to Helios database)	1 x 1536 well-plate	5 sec	5 sec
2	Plate group creation (interactive analysis working set), data retrieval from relational database and HDF5 file creation for dose response data	22 x 1536 well dose-response plates	4 sec	0.2 sec
3	Plate group creation (interactive analysis working set), data retrieval from relational database and HDF5 file creation for primary data	891 x 1536 well primary screening plates	25 sec	0.03 sec
4	Plate data analysis (FRET ratio calculation, normalization)	891 x 1536 well-plates	9 sec	0.01 sec
5	Plate data analysis (FRET ratio, normalization, SVD-based principal pattern modeling, data correction)	891 x 1536 well-plates	18 sec	0.02 sec
6	Plate data analysis (FRET ratio, normalization, robust local regression pattern modeling, correction)	22 x 1536 well-plates	28 sec	1.3 sec
7	Dose-response curve fitting, (including model selection steps, rule-based curve classification and storage of result parameters to database)	929 curves	5 sec	0.005 sec
8	Plate heatmap display for a large screen on the client PC	891 x 1536 well-plates (1.37 Mio wells)	17 sec	0.02 sec (50 plates/sec)

Note the relatively high calculation times (~1.3 sec per 1536 well plate) for the iterative robust local regression pattern modeling method. Nonetheless, it is often used in the analyses because of its high faithfulness and robustness for representing ‘smooth’ patterns on individual plates or small numbers of plates, where the fast SVD based *plate-batch* modeling methods will fail to produce sensible pattern representations. The latter needs *at least* 10 plates to enter into the

analysis, otherwise there is a danger that the model will pick up too much noise and/or represent actually active wells as a ‘pattern to be corrected away’, which is definitely not the desired outcome. An easy visual cue to detect pattern-model ‘overfitting’ is the appearance of seemingly random noise and ‘speckles’ in the pattern model heatmaps which are similar to the ‘pattern’ of the (likely) active wells in the uncorrected data set.

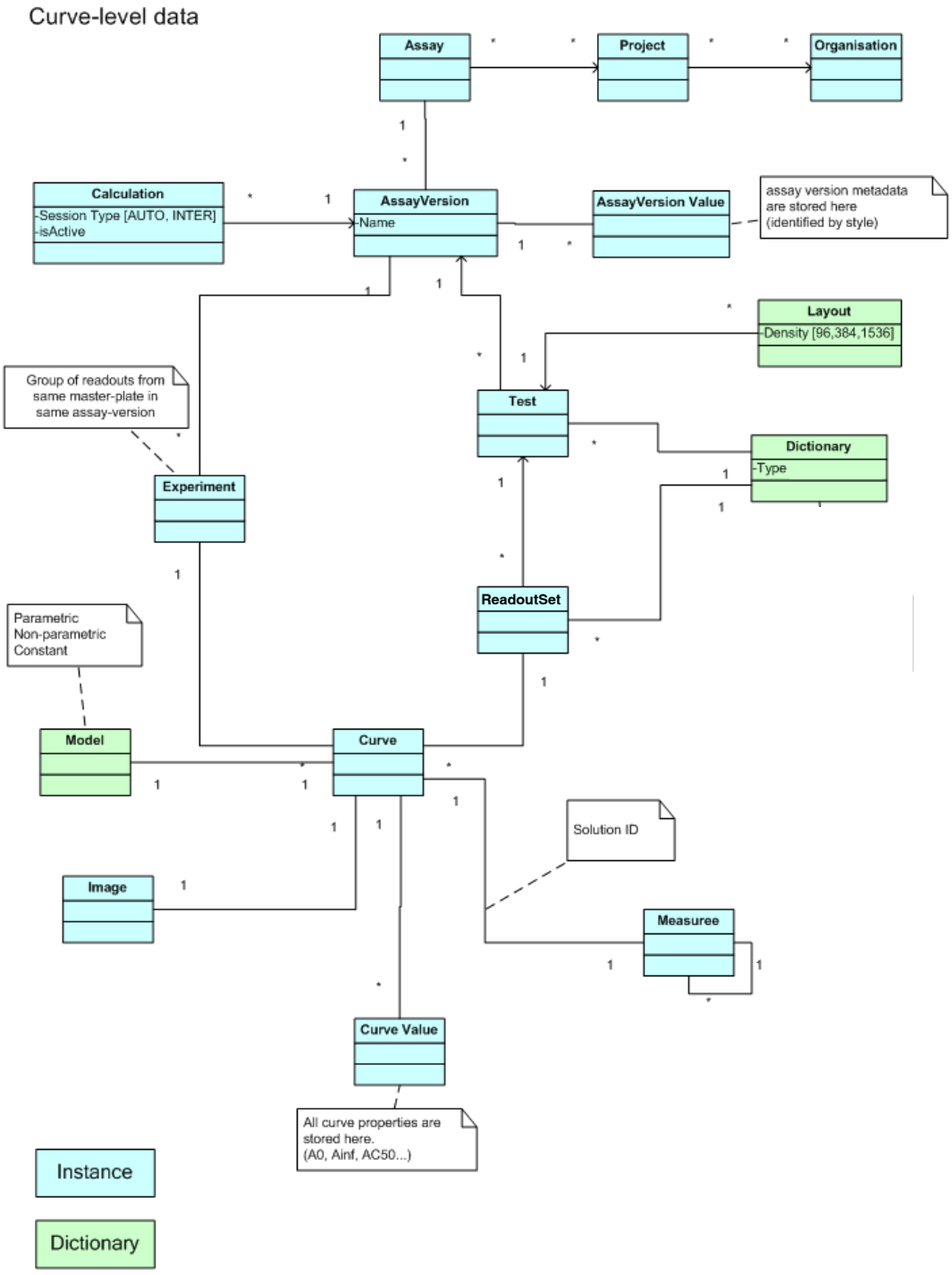
5 – Supplemental Figure S1

High Level Overview of Helios Database Object Schema for Plate Readout Data



6 – Supplemental Figure S2

High Level Overview of Helios Database Object Schema for Curve Data



7 - References, Supplemental References

1. Gubler, H. Methods for Statistical Analysis, Quality Assurance and Management of Primary High-Throughput Screening Data. In *High Throughput Screening in Drug Discovery*; Hüser, J., Ed.; Wiley-VCh: Weinheim, 2006, pp 151-205

2. Sundberg, A. High-throughput and ultra-high-throughput screening: solution- and cell-based approaches. *Curr Opin Biotechnol*. **2000**, *11*, 47-53

3. Wassermann, A.M.; Lounkine, E.; Hoepfner, D.; et al. Dark chemical matter as a promising starting point for drug lead discovery. *Nature Chemical Biology* **2015**, *11*, 958-966

4. Mayr, L.M.; Fuerst, P.: The future of High-Throughput Screening. *Journal of Biomolecular Screening* **2008**, *13*, 443-448

5. Macarron, R.; Banks, M.N.; Bojanic, D.; et al. Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery* **2011**, *10*, 188-195

6. Halford, B., Breakthroughs with bar codes - DNA-encoded libraries help pharma find drug leads. *C&EN* **2017**, *95*, 28–33

7. Bray, M.A.; Singh, S.; Han, H.; et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat Protoc.* **2016**, *11*, 1757-74

8. Snyder, J.; Hong, V. Escaping Flatland: Interactive High-Dimensional Data Analysis in Drug Discovery Using Spark. *Spark Summit East 2016 – Data Science and Engineering at Scale*, New York, **2016**, Feb. 16-18. https://spark-summit.org/east-2016/events/escaping-flatland-interactive-high-dimensional-data-analysis-in-drug-discovery-using-spark/ and presentation slides https://www.slideshare.net/SparkSummit/escaping-flatland-interactive-highdimensional-data-analysis-in-drug-discovery-using-spark-by-josh-snyder-victor-hong-and-laurent-galafassi (accessed August 15, 2017)

9. Landrum, G.; Wrobel, M.; Clare, N. Is one enough? Data warehousing for biomedical research. Basel Life Sciences Week, **2016** https://www.slideshare.net/GregLandrum1/is-one-enough-data-warehousing-for-biomedical-research (accessed August 15, 2017)

10. Inglese, J.; Auld, D.S.; Jadhav, A.; et al. Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc Natl Acad Sci USA* **2006**, *103*, 11473–11478

11. Gubler, H. High Throughput Screening Data Analysis. In *Nonclinical Statistics for Pharmaceutical and Biotechnology Industries*; Zhang, J., Ed.; Springer: Heidelberg, 2016, pp 83-139

12. Brideau, C.; Gunter, B.; Pikounis, B.; et al. Improved Statistical Methods for Hit Selection in High-Throughput Screening. *Journal of Biomolecular Screening* **2003**, *8*, 634-647

13. Kevorkov, D.; Makarenkov, V. Statistical Analysis of Systematic Errors in High-Throughput Screening. *Journal of Biomolecular Screening* **2005**, *10*, 557-567

14. Wu, Z.; Liu, D.; Sui, Y. Quantitative assessment of hit detection and confirmation in single and duplicate high-throughput screenings. *Journal of Biomolecular Screening* **2008**, *13*, 159-167

15. Murie, C.; Barette, C.; Lafanechere, L.; et al. Control-Plate Regression (CPR) Normalization for High-Throughput Screens with Many Active Features. *Journal of Biomolecular Screening* **2014**, *19*, 661-671

16. Varin, T.; Gubler, H.; Parker, C.N.; et al. Compound set enrichment: a novel approach to analysis of primary HTS data. *Journal of Chemical Information and Modeling* **2010**, *50*, 2067-2078

17. Fomenko, I.; Durst, M..; Balaban, D. Robust regression for high throughput drug screening. *Comput Methods Programs Biomed* **2006**, *82*, 31-37

18. Sebaugh, J.L.: Guidelines for accurate EC50/IC50 estimation. *Pharmaceutical Statistics* **2011**, *10*, 128-134

19. Dunn, J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Journal of Cybernetics* **1973**, *3*, 32-57

20. Senisterra, G.; Chau, I.; Vedadi, M. Thermal denaturation assays in chemical biology. Assay and Drug Development Technologies **2012**, *10*, 128-136

21. Ahlberg, J.J..; Nilson, E.N.; Walsh, J.F. Theory of splines and their applications, Acad. Press: New York, 1967

22. Hoaglin, D.C.; Mosteller, F.; Tukey, J.W. Understanding Robust and Exploratory Data Analysis, Wiley, New York, 1982

23. Gill, P.E.; Murray, W. Algorithms for the solution of the nonlinear least-squares problem. SIAM Journal on Numerical Analysis **1978**, *15(5)*, 977–992.

24. Curtis, A.R. Analysis of Covariance after Nonlinear Least Squares Fitting. IMA Journal of Numerical Analysis, **1986**, *6*, 453-461