

Supplementary information: Supplementary Table 1 and
Supplementary Figures 1 - 9

**A statistical framework for high-content phenotypic profiling using
cellular feature distributions**

Yanthe E. Pearson¹, Stephan Kremb¹, Glenn L. Butterfoss¹, Xin Xie¹, Hala
Fahs¹, Kristin C. Gunsalus^{1,2*}

¹Center for Genomics and Systems Biology, New York University Abu
Dhabi, P. O. Box 129188, Abu Dhabi, United Arab Emirates

²Department of Biology and Center for Genomics and Systems Biology,
New York University, New York, NY 10003, USA

* Corresponding author: Kristin C. Gunsalus
Email: kcg1@nyu.edu

Supplementary Table 1: R scripts and data files underlying each manuscript figure.

Supplementary Figure 1: Adjustment of positional effects and data standardization across different plates.

Supplementary Figure 2: Statistical distance measurement among replicates.

Supplementary Figure 3: Full feature cytological profile.

Supplementary Figure 4: Full feature EMD fingerprints of individual control samples.

Supplementary Figure 5: Summary of phenotypes for 65 compounds.

Supplementary Figure 6: Divergent phenotypes.

Supplementary Figure 7: Feature reduction.

Supplementary Figure 8: Comparison of cytological profiles.

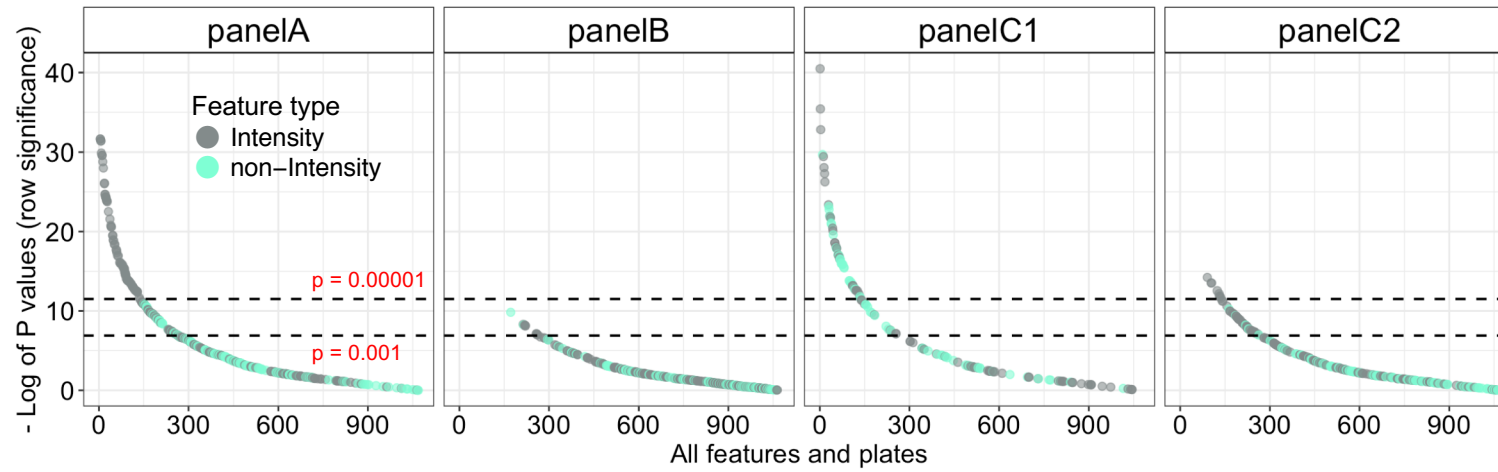
Supplementary Figure 9: Radial plot fingerprints of 65 compounds.

Supplementary Table 1: R scripts and data files underlying each manuscript figure.

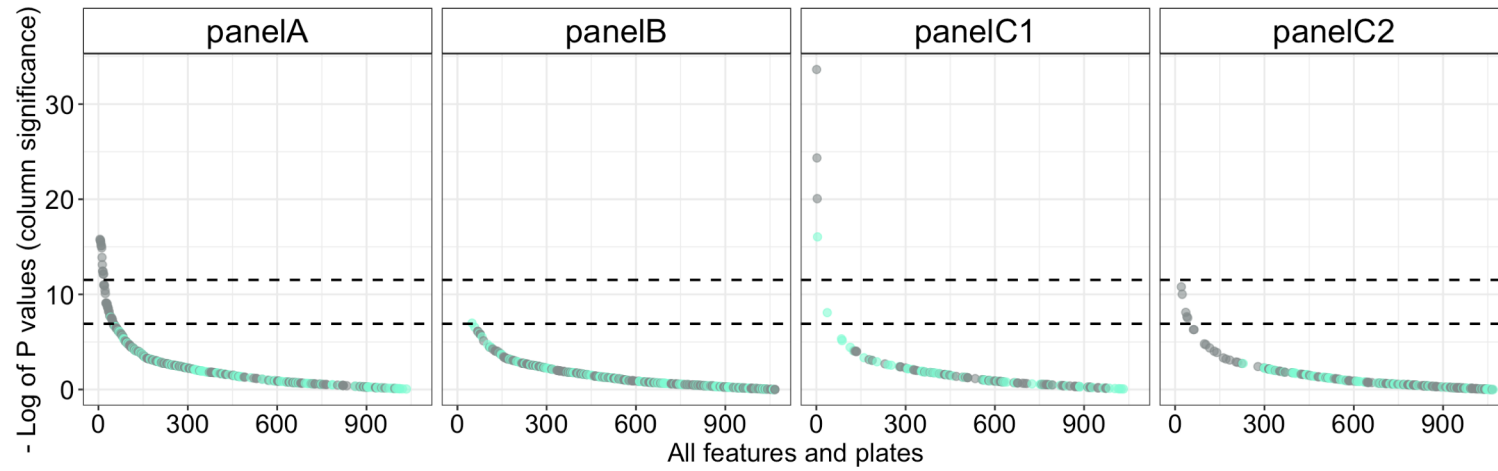
Manuscript Figure	Description	R script	Data files
Figure 2a	Heatmap of compound similarity based on Tanimoto distance	Figure2A_tanimoto.R	Tanimoto_figure2.csv
Figure 2c	Cell counts shown as heatmap form	figure2C_count_heatmap.R	raw_medians_fig2C.csv
Figure 2d	Cell counts shown as a scatterplot	figure2D_count_scatter.R	raw_medians_fig2C.csv
Figure 2e	Per well density curve of controls	Figure2E_dmsso_curves.R	cell_data_DMSO_figure2g.csv
Figure 2f	Cell cycle dose response plot of mitoxantrone treated cells	Figure2F_mitoxantrone.R	cell_data_mitoxantrone_figure2f.csv
Figure 2g	Density plots showing quartiles for a group of compounds	figure2G_quartiles.R	Data_fig2G.csv
Figure 3a	Heatmap of control well medians	figure3A.R	nucfeatures_medians_fig3A.csv
Figure 3b	Summary of two-way ANOVA test	figure3B_anova.R	anova_output_fig3b.csv
Figure 3c	Heatmaps showing positional adjustment	figure3C_positioneffect.R	Well_medians_fig3c.csv
Figure 3d	Cell cycle distributions	figure3D.R	cell_data_fig3D.csv
Figure 4a	Plots for replicate feature distributions, both treatment and control	figure4A.R	cell_data_figure4A_control.csv cell_data_figure4A_treatment1.csv cell_data_figure4A_treatment2.csv
Figure 4b	Plots of density curves and CDF curves for two data samples	figure4B.R	No data files required
Figure 4c	Plot for distribution of statistical scores	figure4C-D.R	control_data_fig4C.csv treatment_data_fig4C.csv
Figure 4d	Plot of sorted features	figure4C-D.R	treatment_replicate_test_allpanels_fig4D.csv control_replicate_test_allpanels_fig4D.csv
Figure 5a	Plot of replicate distributions R script Includes EMD calculation	figure5A.R	cell_data_figure5a_vincristine_control.csv
Figure 5b,c	Global control and individual treatment groups	figure5B-C.R	treated_cells_data_fig5B-C.csv control_cells_data_fig5B-C.csv
Figure 5d,e Supplementary Figure 4	Full feature EMD fingerprints of individual control samples.	figure5D-E_fullfeature.R radarchart2_new.R	raw_emd_profile174.csv
Figure 5f	Plot of Vincristine fingerprint	Figure5F_vincristine.R radarchart2_new.R	ScaledEMDprofile_69feats_785treatments.csv
Figure 6a	Heatmap of similarity among treatments	Figure6A_heatmap.R	fullfeature_EMDprofile_785treatments.csv
Figure 6b	Run the umap and plot phenotypic trajectories while projecting cell count onto UMAP.	line_segments.R Figure6B_umap.R	fullfeature_EMDprofile_785treatments.csv
Figure 7a	UMAP paths	line_segments.R figure7A_umap_paths.R	UMAP_dimensions_69feats_785treatments_new.csv
Figure 7b	Cell count dose response	figure7B.R	cell_counts.csv
Figure 7c	Cell cycle distributions	Figure7C_cellcycle_four.R	figure7c_treatments.csv figure7c_control.csv
Figure 7d	Radial plots (see Figure 5f)		

Supplementary Table 1: A summary of all R scripts and data files used for each figure in the manuscript. The files can be found on the GitHub within individual figure folders at <https://github.com/GunsalusPiano/EMD>.

a



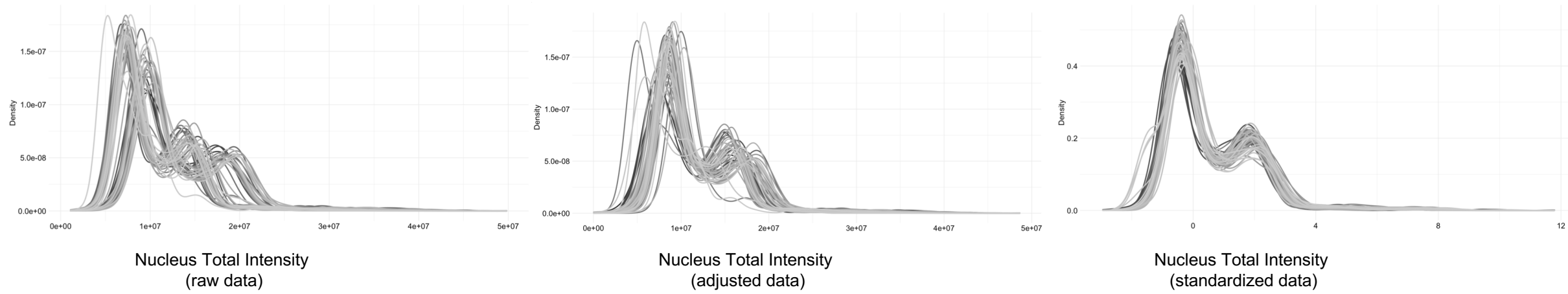
b



Supplementary Figure 1: Adjustment of positional effects and data standardization across different plates.

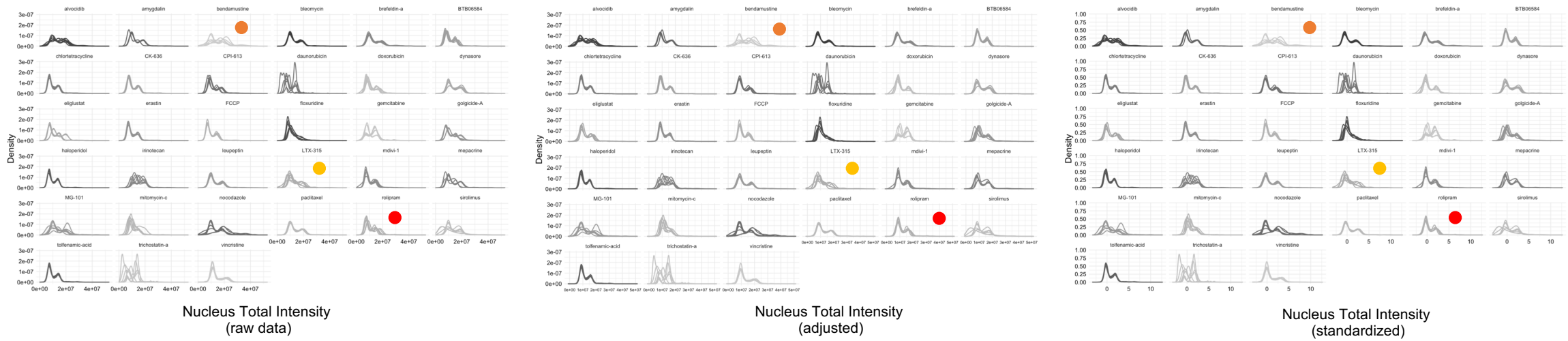
C

DMSO- control distributions

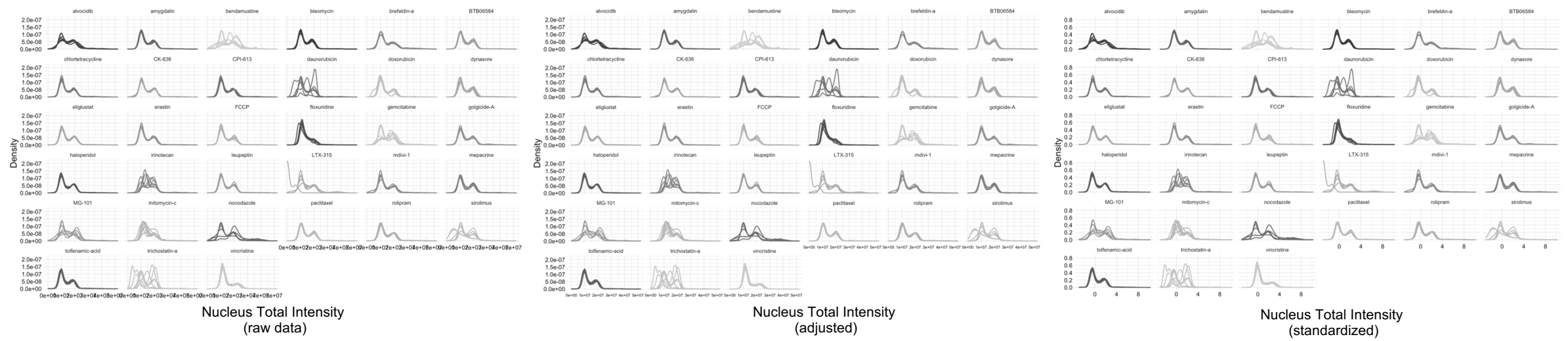


Supplementary Figure 1: Adjustment of positional effects and data standardization across different plates.

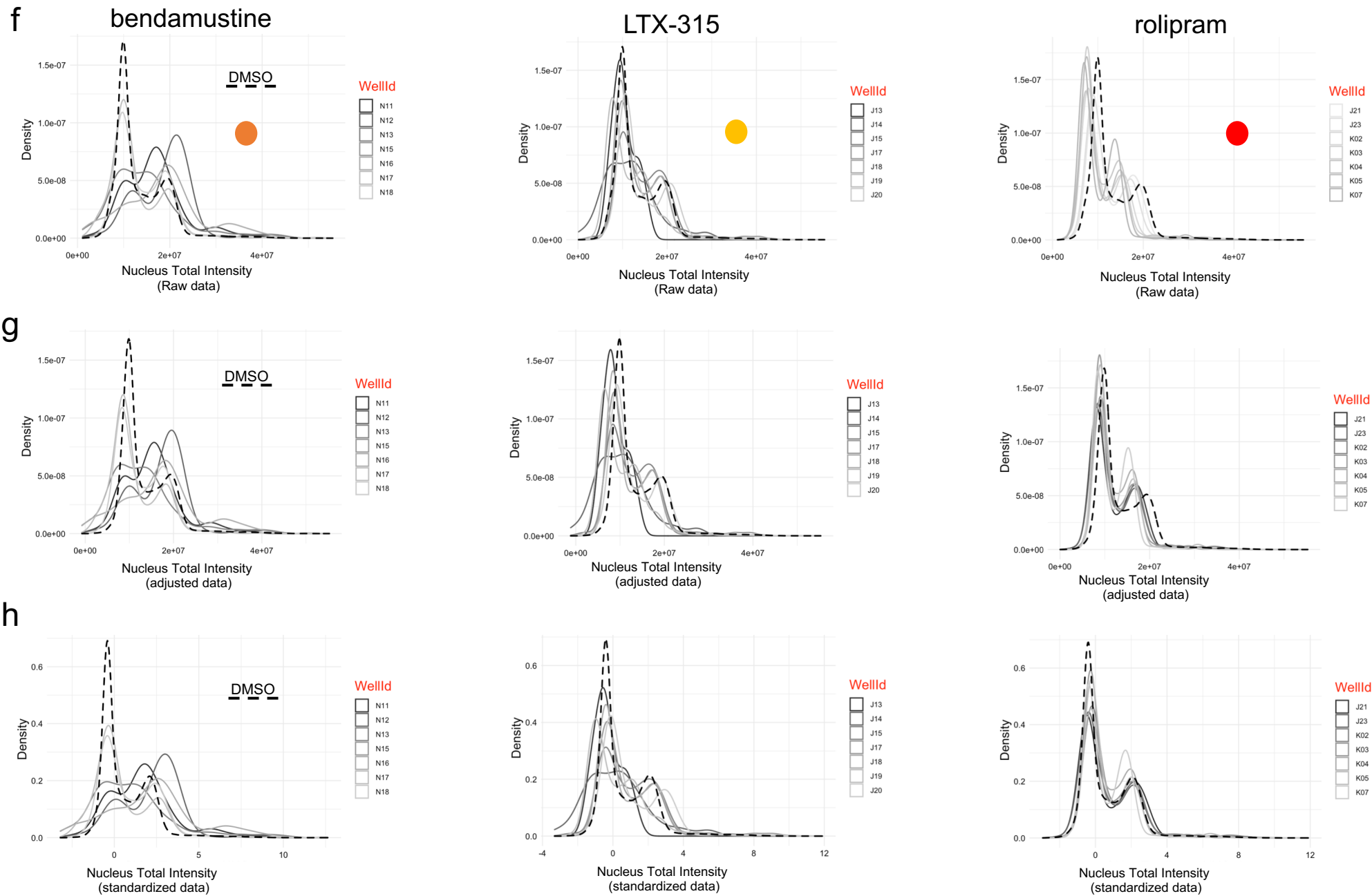
d Plate 1 replicate 1: Cell cycle feature under different chemical perturbations



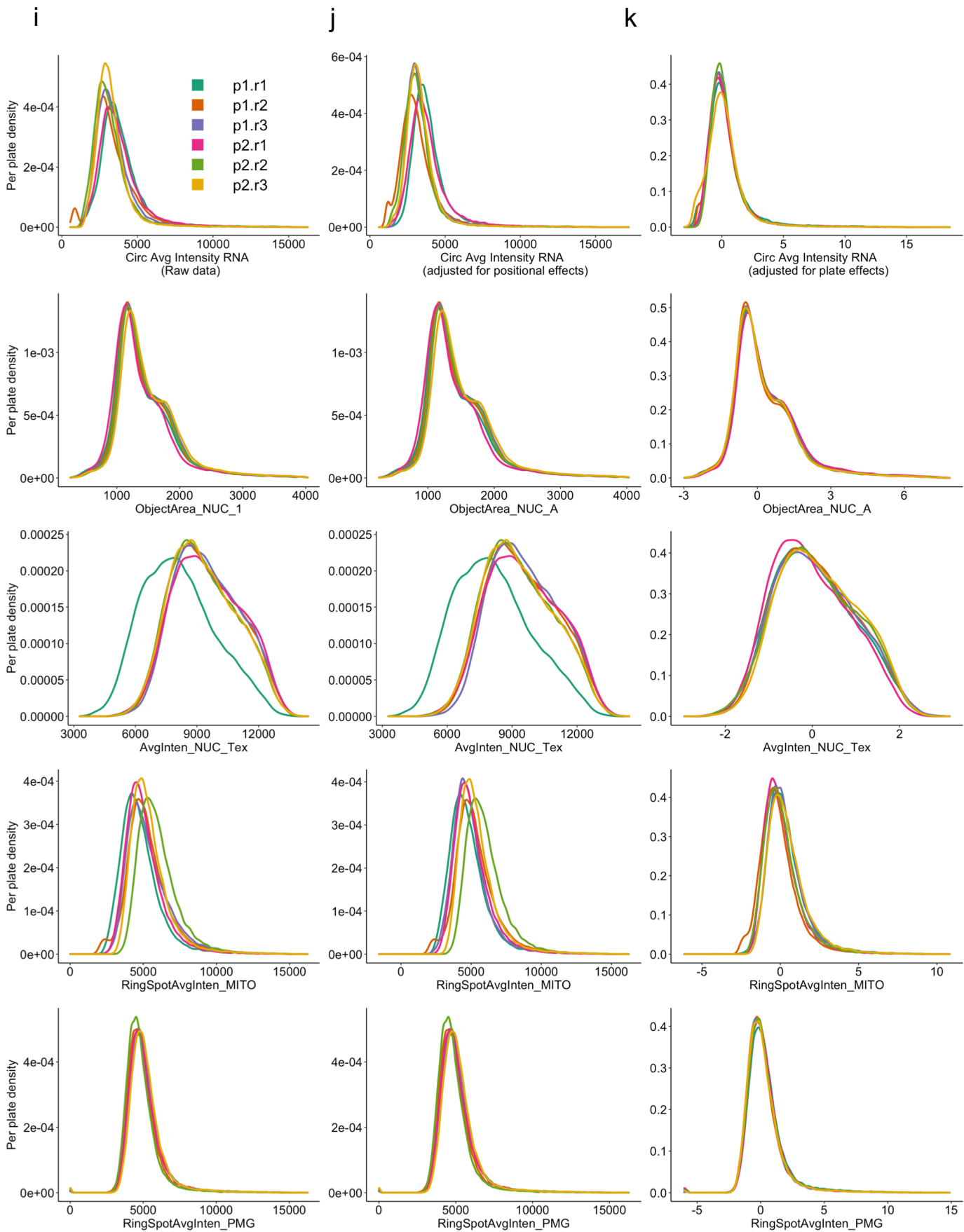
e Plate 1 replicate 3: Cell cycle feature under different chemical perturbations



Supplementary Figure 1: Adjustment of positional effects and data standardization across different plates.



Supplementary Figure 1: Adjustment of positional effects and data standardization across different plates.



Supplementary Figure 1: Adjustment of positional effects and data standardization across different plates.

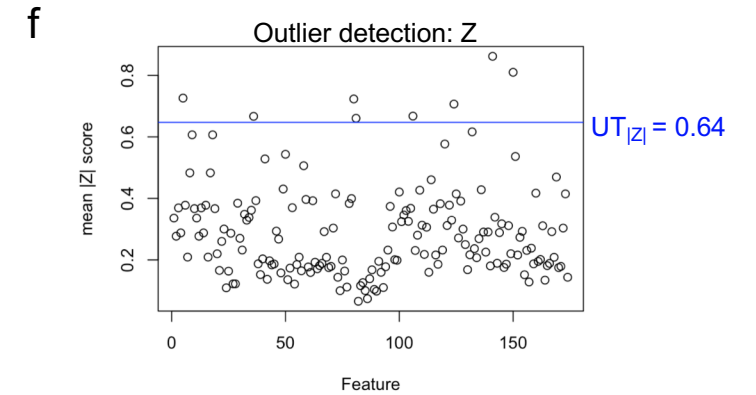
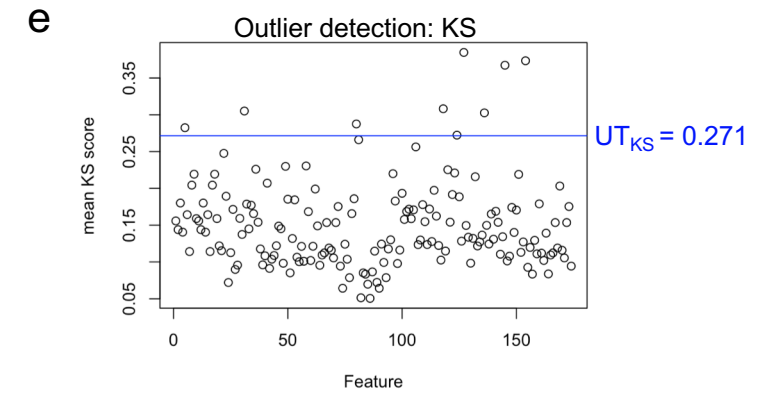
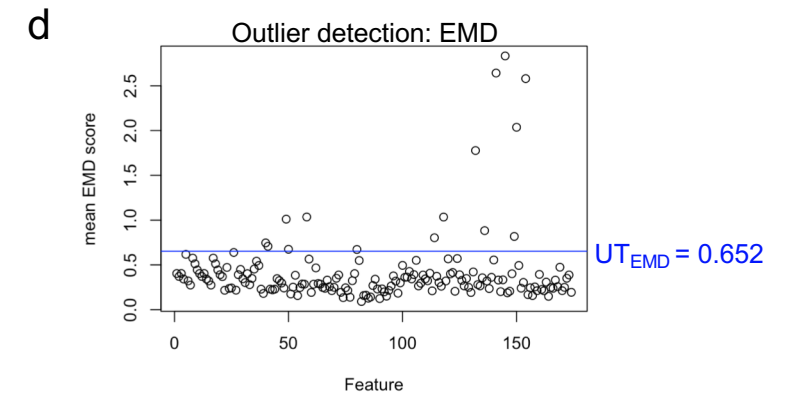
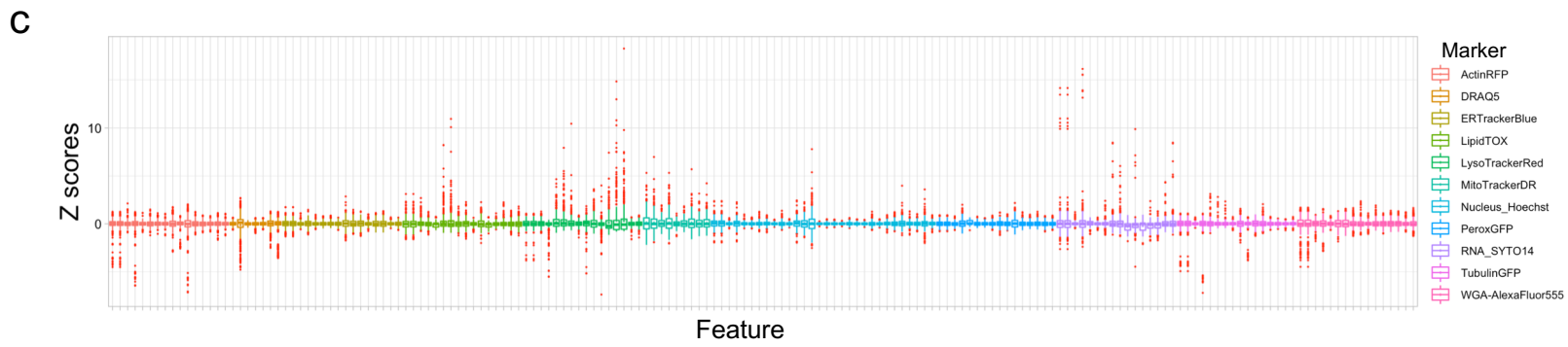
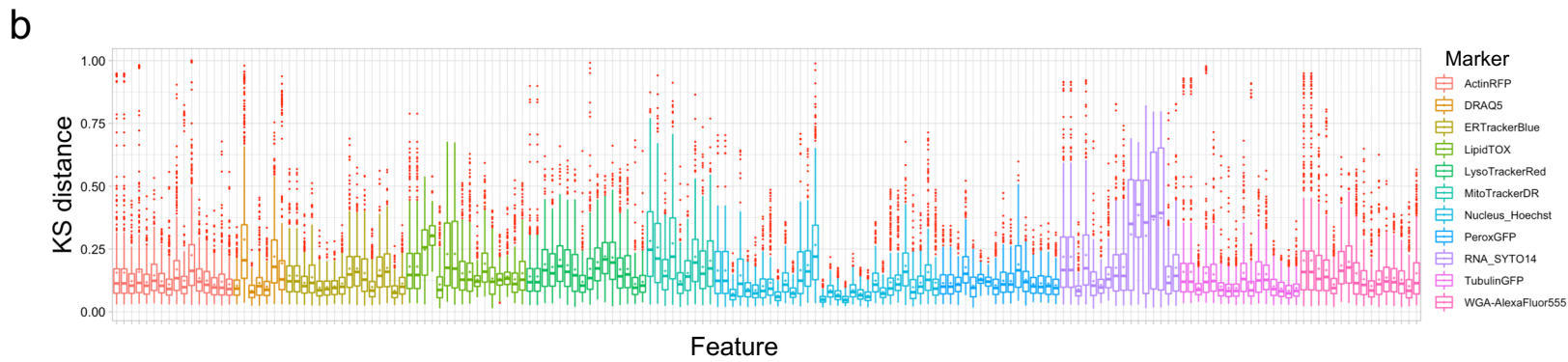
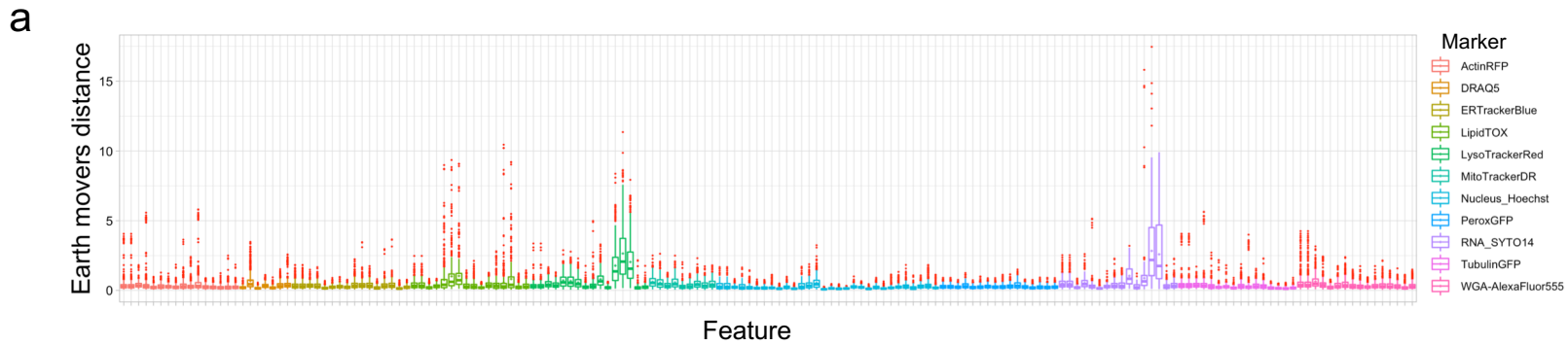
Supplementary Figure 1: Adjustment of positional effects and data standardization across different plates.

(a - b) Positional effect detection by two-way ANOVA is applied to control well medians, for all measured features and all plates. Plots show row (a) and column (b) dependencies for all control wells, ordered by decreasing significance (p-value increasing) for each assay panel. Features related to Intensity (gray) tend to be more sensitive to positional effects than non-Intensity (cyan) features related to area, texture, and shape.

(c) Per well distributions of raw, adjusted, and standardized DMSO-control cell populations from plate1 1 rep 1 (see Fig. 3d).

(d - e) Two replicate plates of cell cycle feature in response to 231 unique treatment conditions including 33 compounds at seven different concentrations. Three compounds bendamustine, LTX-315, and rolipram color labeled as orange (column 1), yellow (column 2) and red (column 3) demonstrate the data correction effects on cell feature distributions (see f - h).

(i - k) Correction for row and column positional effects and standardization of plate-to-plate variation for representative features (by row): RNA, Nucleus Area, Nucleus Texture, Mitochondria, Plasma membranes and Golgi (PMG).

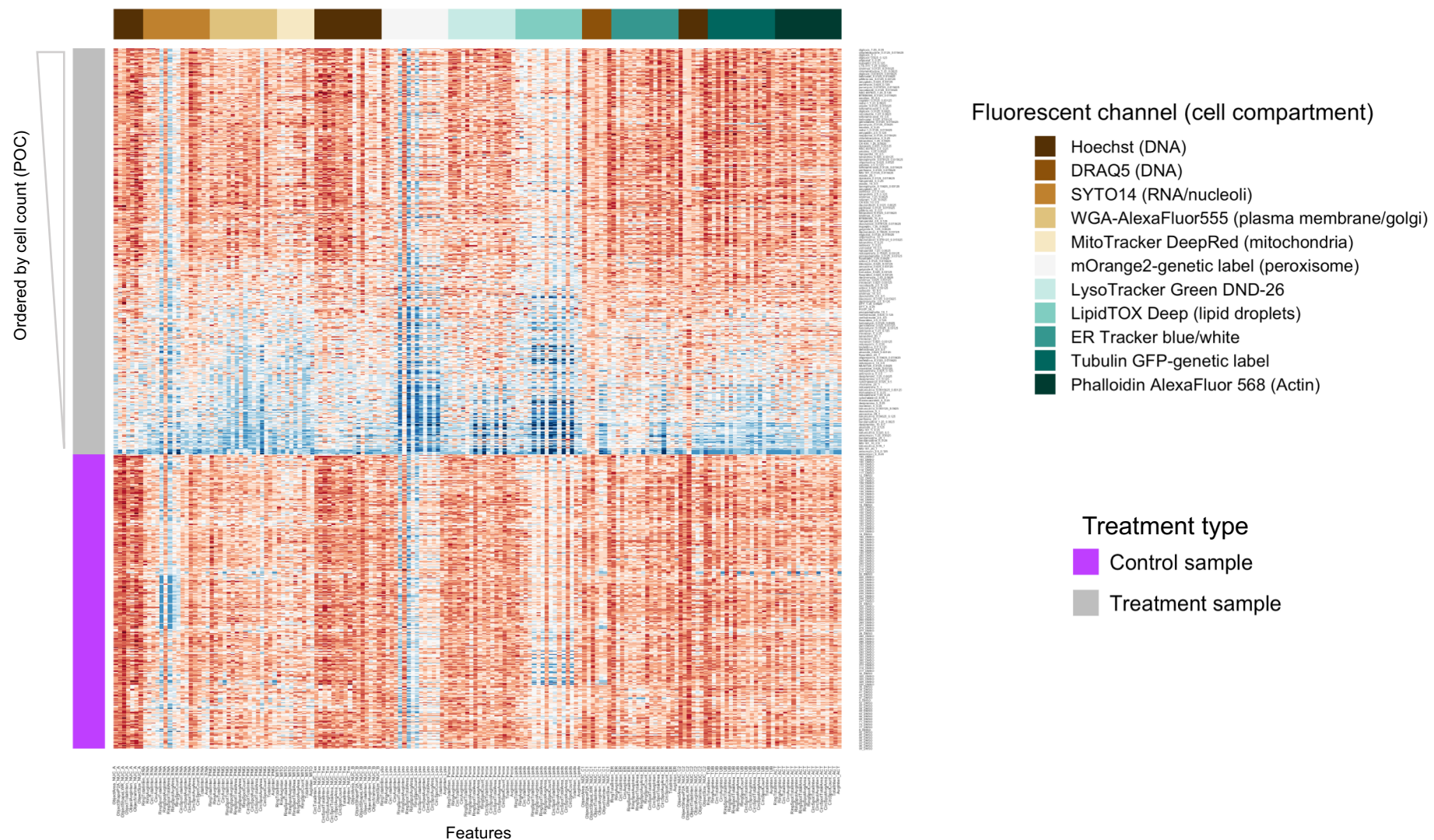
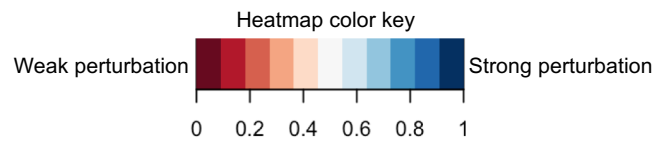


Supplementary Figure 2: Statistical distance measurement among replicates.

Supplementary Figure 2. Statistical distance measurement among replicates.

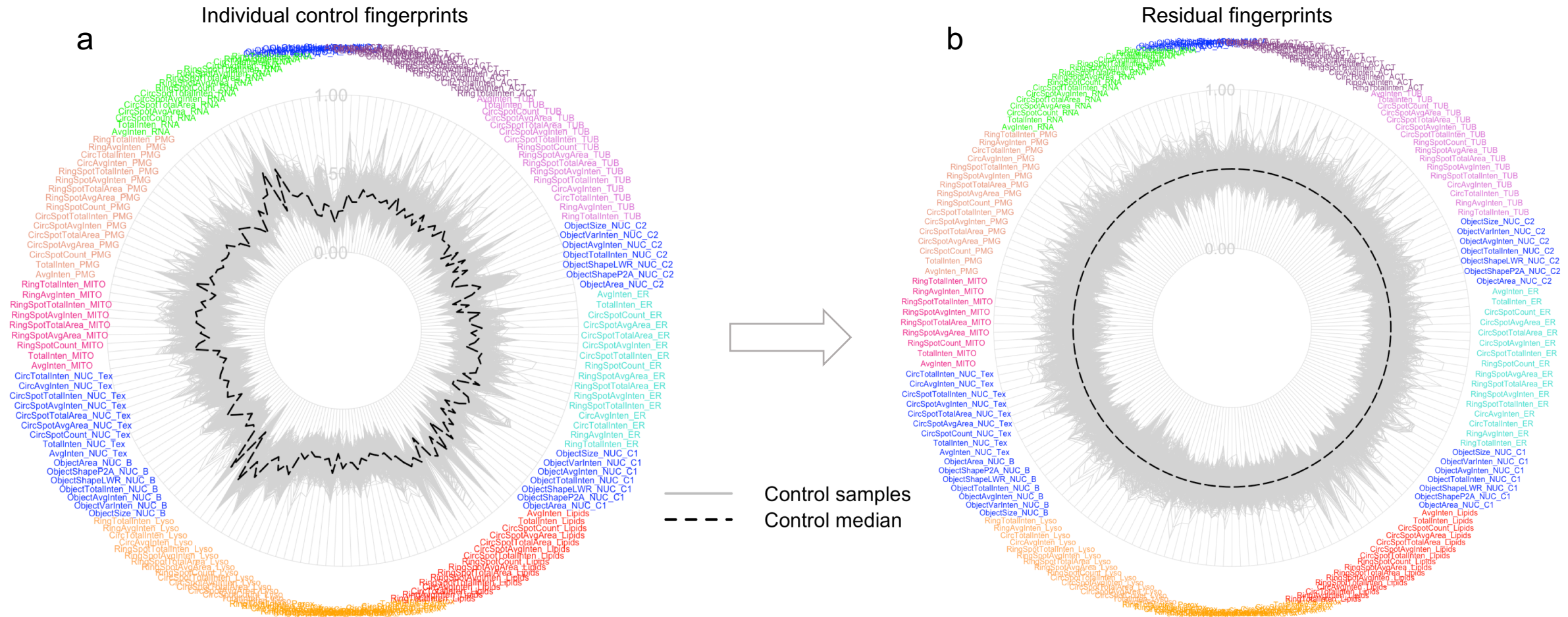
(a – c) Per-feature distributions for each of the three statistical metrics, color coded by cytological marker. (a) EMD scores are strictly positive and highly sensitive to feature distribution differences; (b) KS scores are bounded between 0 and 1; (c) robust Z-scores account for both positive and negative differences. EMD and robust Z-scores identify RNA (SYTO14), Lysosome (LysoTracker Green) channels (SYTO 14 and LysoTracker Green), and Lipid (LipidTOX) features with extreme outliers, whereas KS has difficulty discriminating noisy features.

(d – f) Outlier detection: Using per-feature averages and interquartile range (IQR) outlier detection (Upper threshold value (UT) = $1.5 \times \text{IQR} + \text{upper quartile}$, blue line) to identify features which fall too far from the expected range of values we find that EMD score (d) detects and separates outliers from the group more efficiently than KS (e) and robust Z-scores (f).



Supplementary Figure 3: Full feature cytological profile.

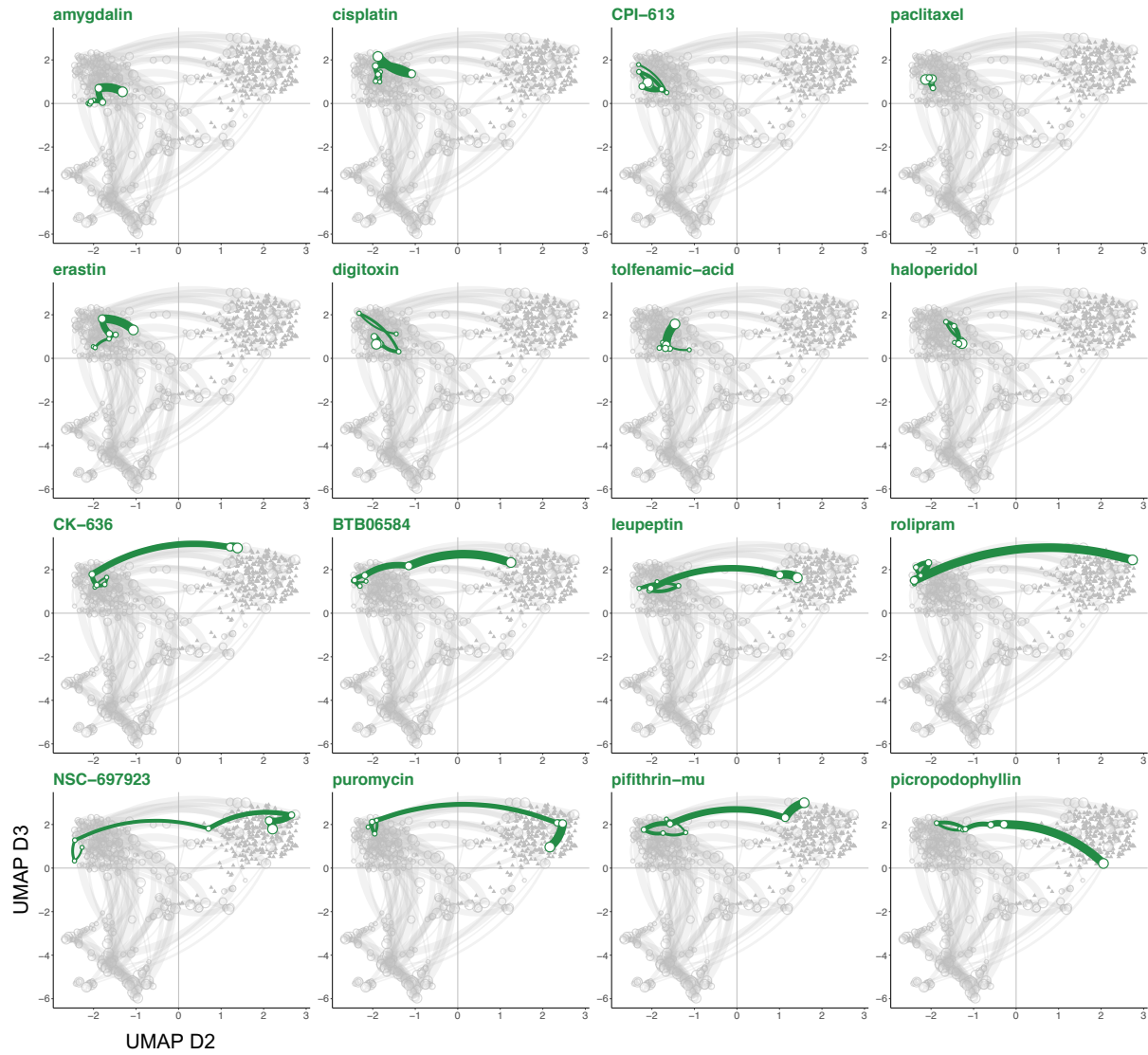
Heatmap summary of EMD profiles for all 455 treatments (grey rows); including 65 compounds at 7 concentrations each, and 330 DMSO-control samples (purple rows). The full profile is log transformed and features are min-max scaled to the range [0, 1]. Treatment profiles (grey rows) are further sorted by their cell count as percent of control (POC).



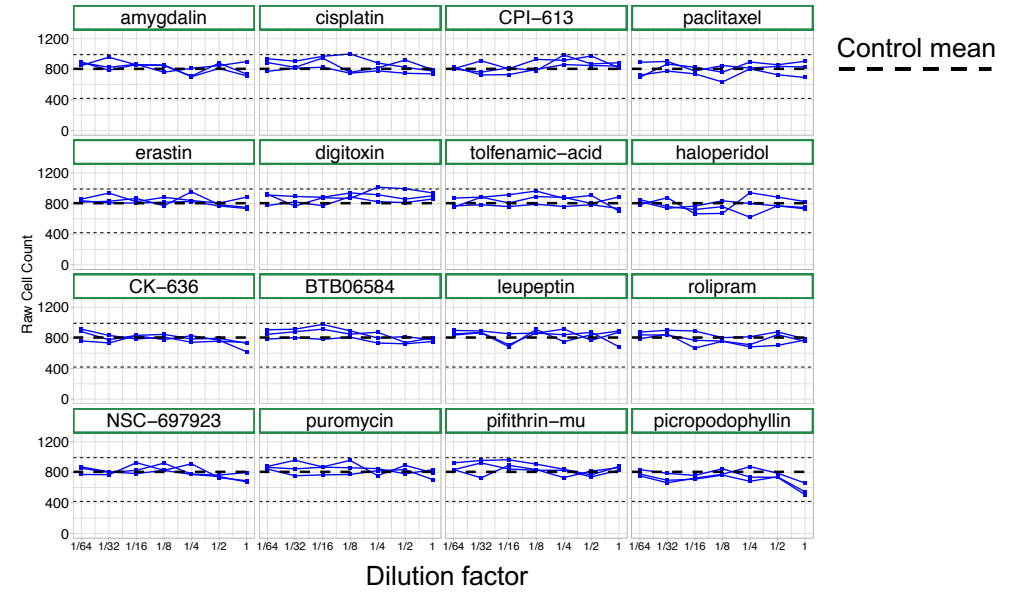
Supplementary Figure 4: Full feature EMD fingerprints of individual control samples.

(a) Radial plot of EMD scores spanning 174 measured features among individual controls (gray lines) and the median EMD score of all controls (black dashed line).
 (b) Radial plot of residual EMD scores of individual controls (gray lines) relative to the null control line (black dashed line). Residual score is defined as the difference between the score of the individual control and the median of all controls. Residual fingerprints naturally fluctuate around median zero (black dashed line), here the values have been offset by 0.5 to expand the plot for better visualization.

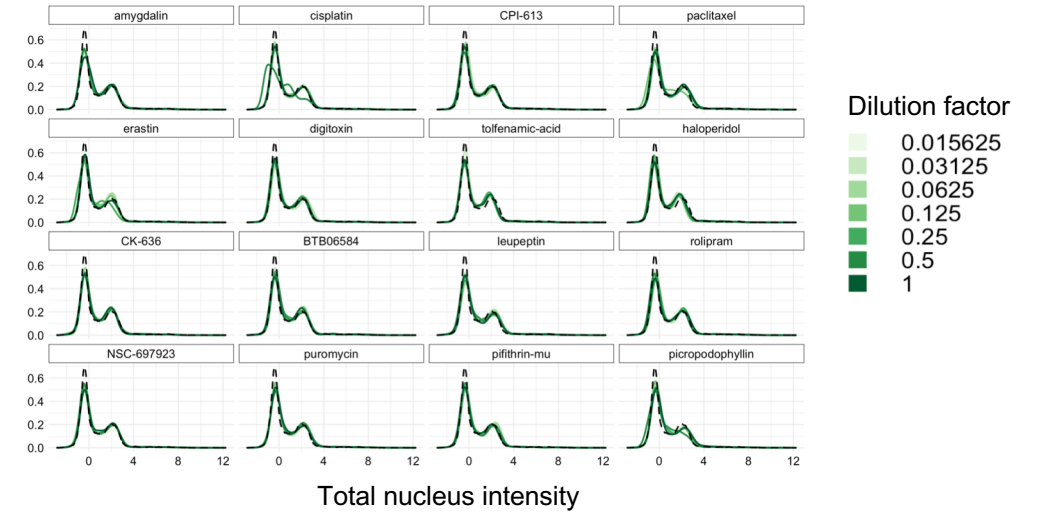
a Low stress compounds



b



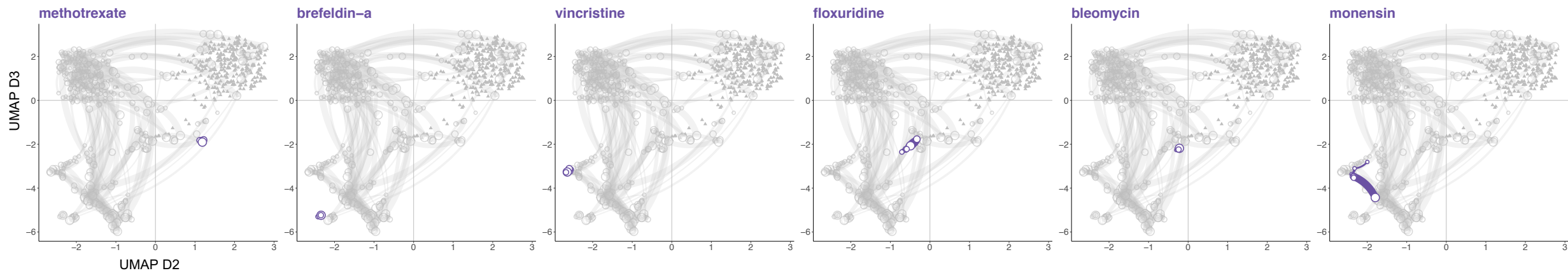
c



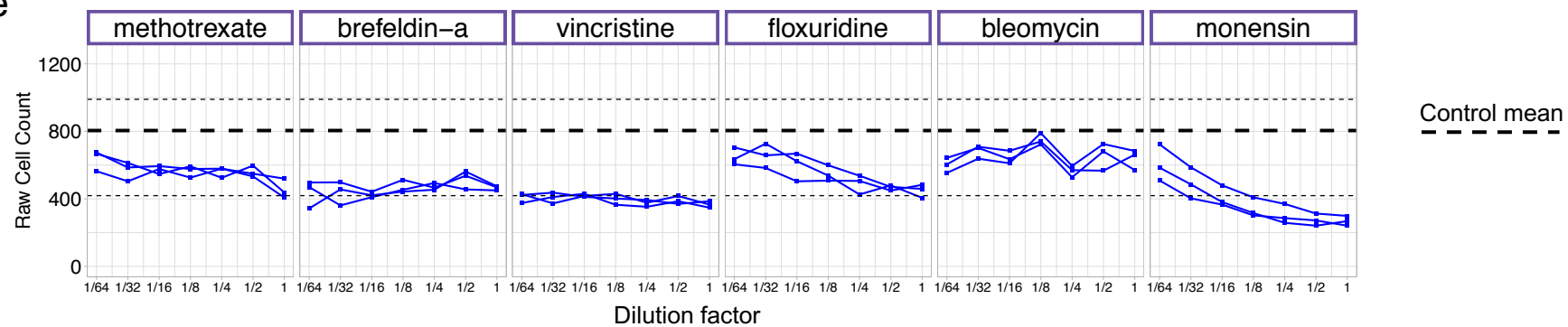
Supplementary Figure 5: Summary of phenotypes for 65 compounds.

d

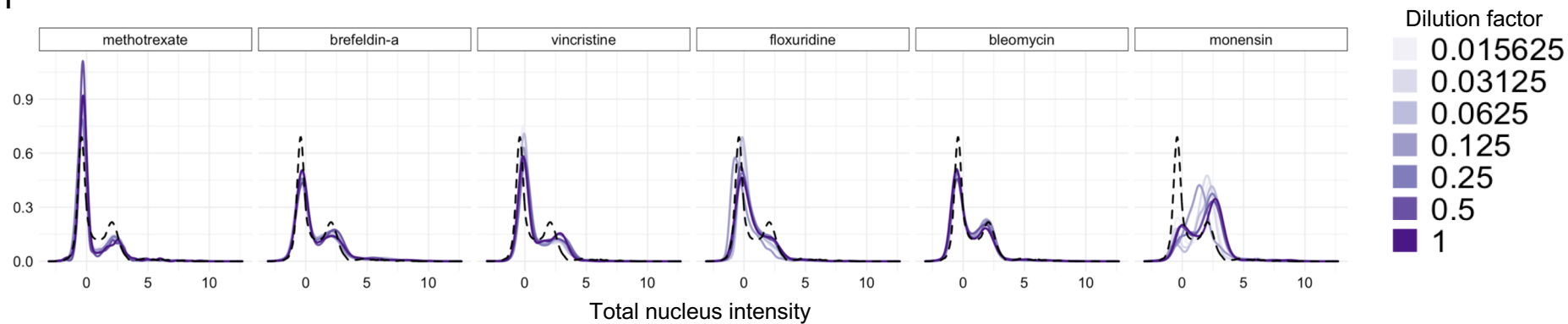
Active dose insensitive compounds



e

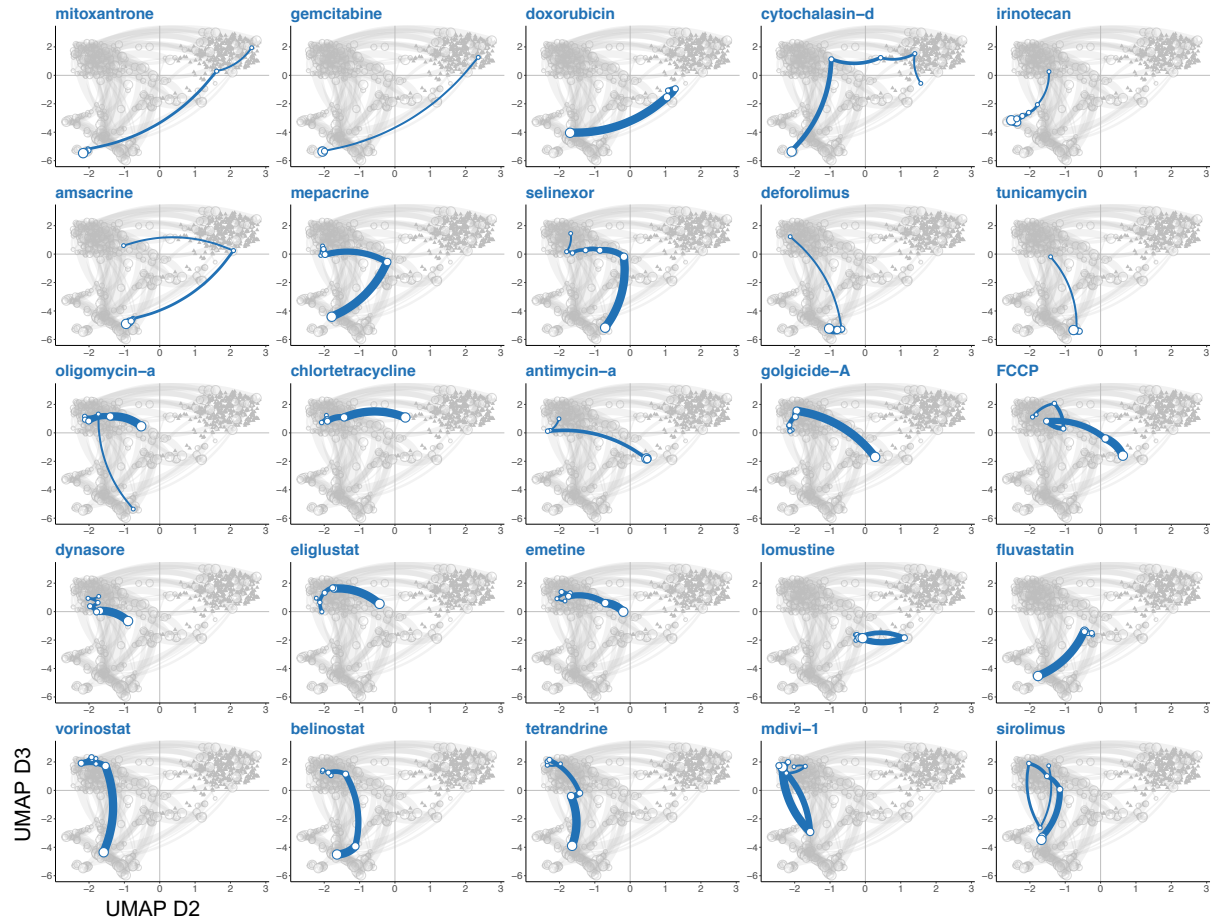


f

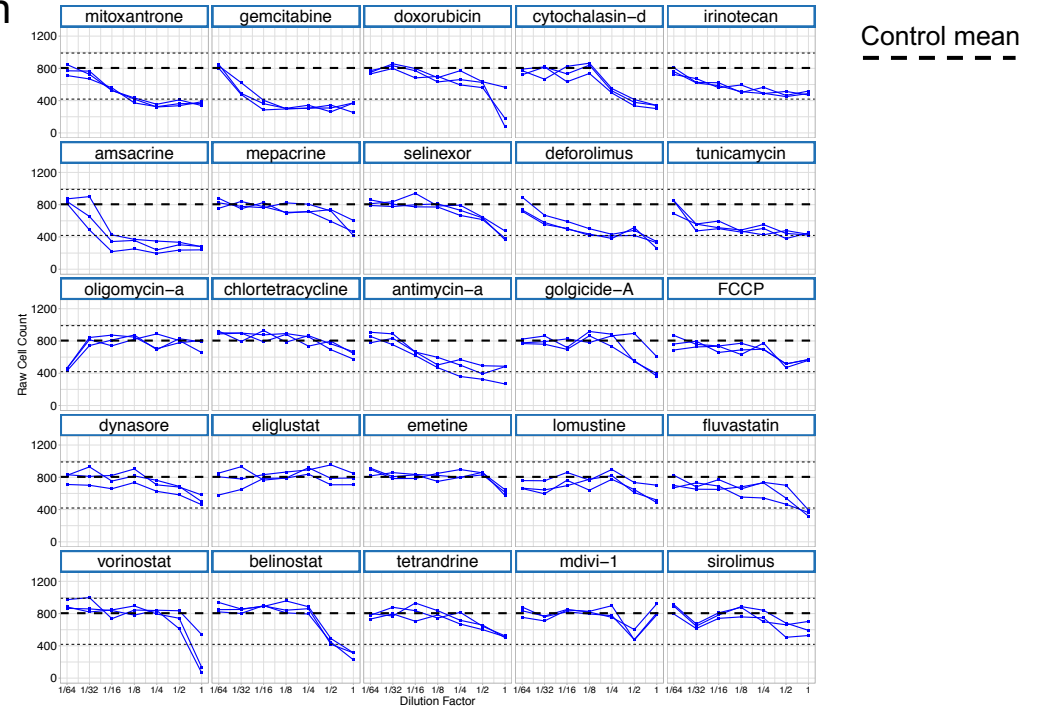


Supplementary Figure 5: Summary of phenotypes for 65 compounds.

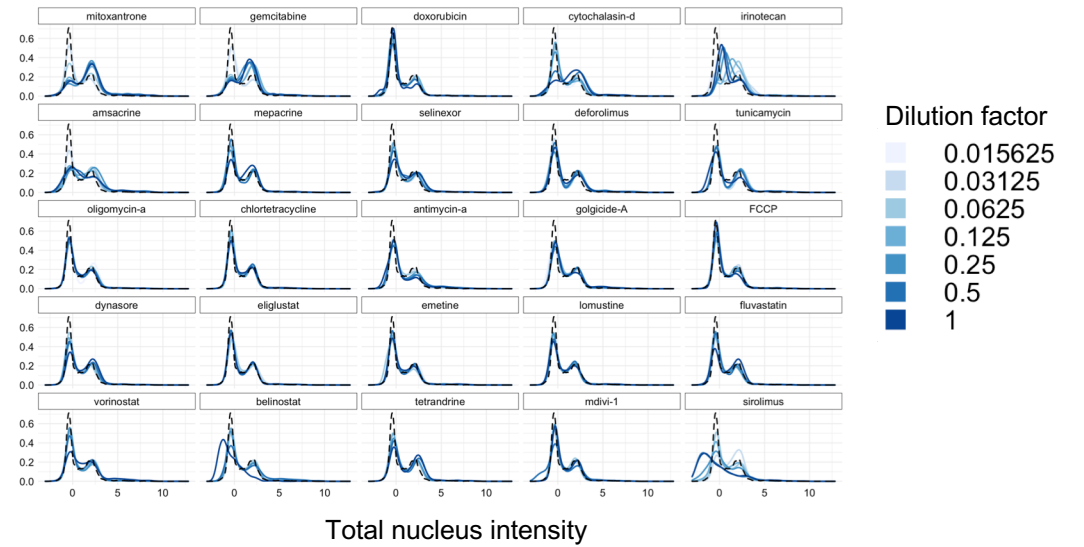
g Active dose-responsive compounds



h

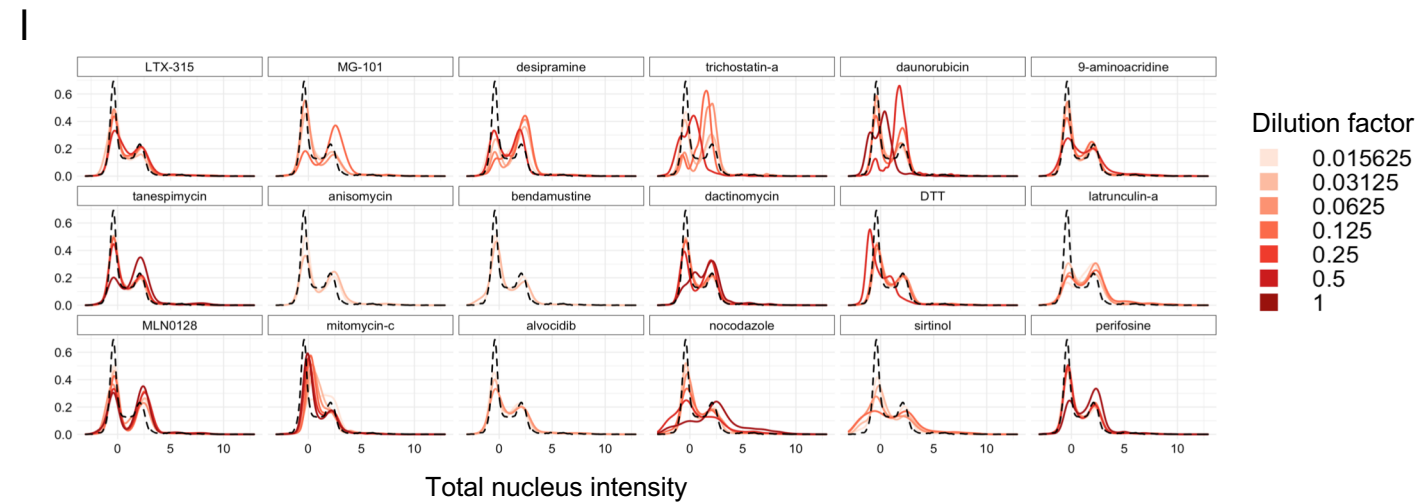
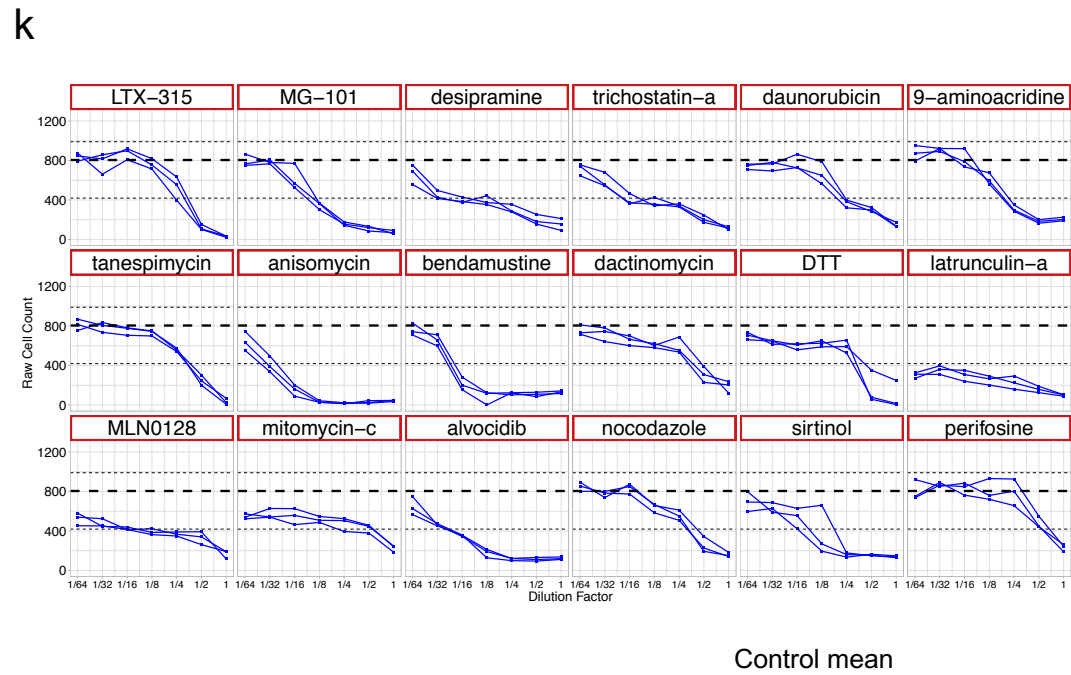
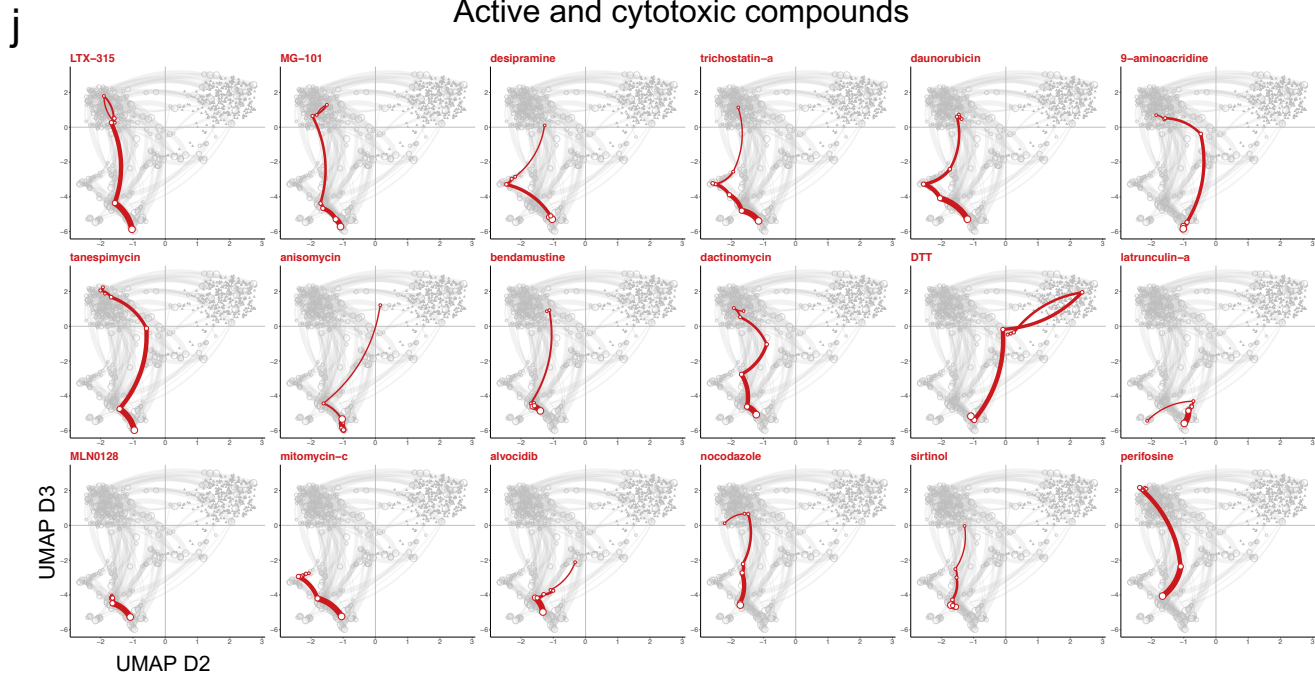


i



Supplementary Figure 5: Summary of phenotypes for 65 compounds.

Active and cytotoxic compounds



Supplementary Figure 5: Summary of phenotypes for 65 compounds.

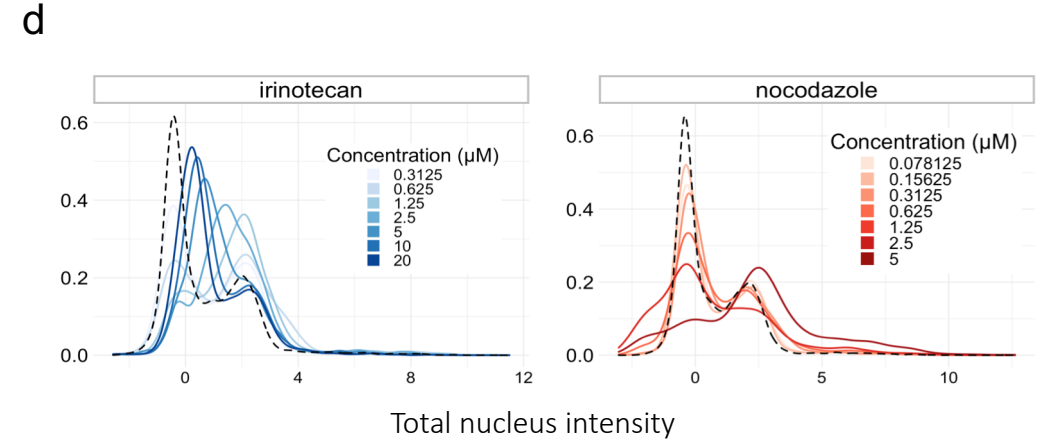
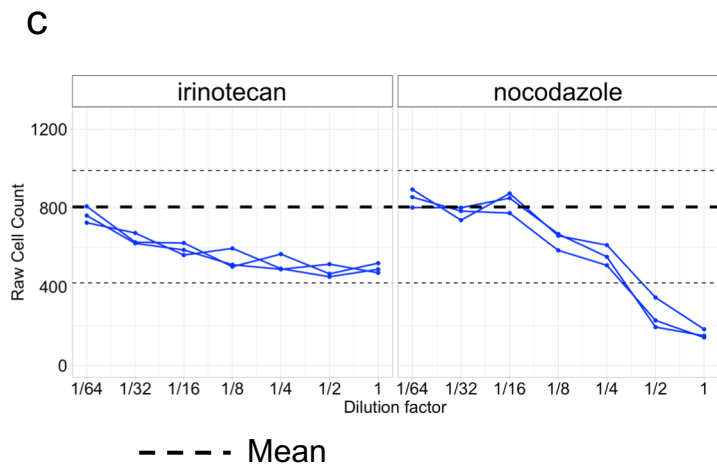
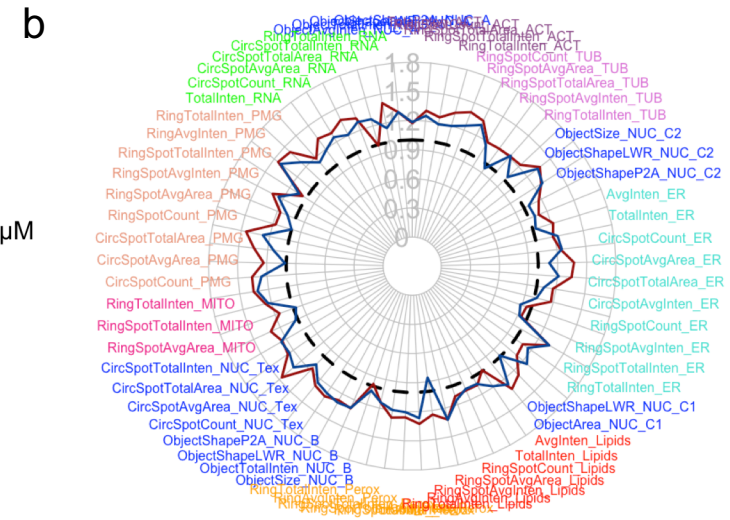
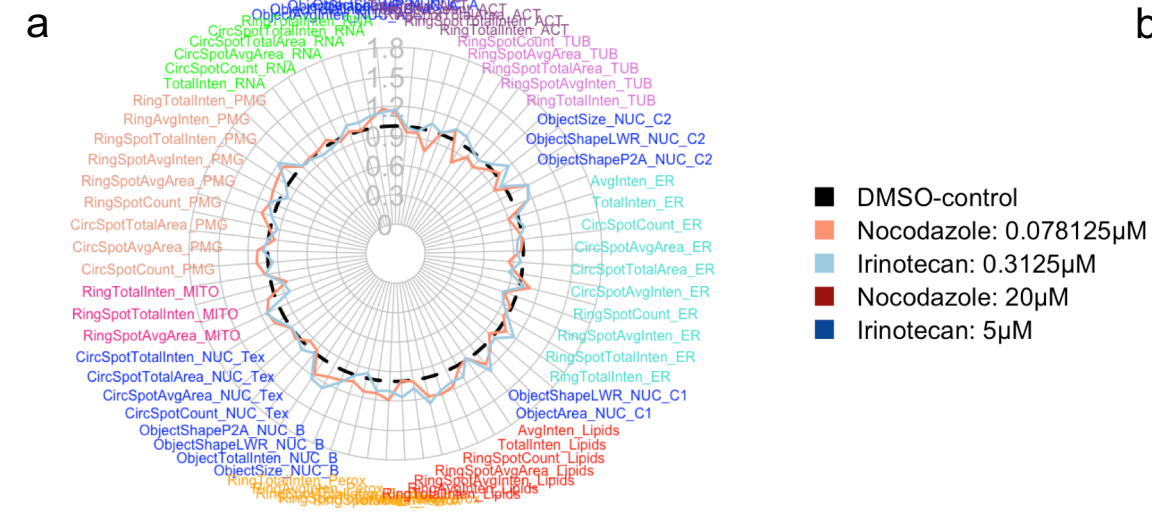
Supplementary Figure 5: Summary of phenotypes for 65 compounds.

(a – c) 16 compounds elicited little bioactivity in U2OS cells. **(a)** Phenotypic trajectories of the “low stress” group. Profiles for some compounds at higher concentrations are not well separated from controls in the first 3 UMAP dimensions. The 16 compounds showed little or no effect on cell counts **(b)** and cell cycle distributions **(c)** across concentration gradients. In **(b)**, each curve represents a replicate and three dashed lines represent max, min, and mean of control cell counts. In **(c)**, cell cycle is measured by total nuclear intensity; the global control is shown as a dashed line.

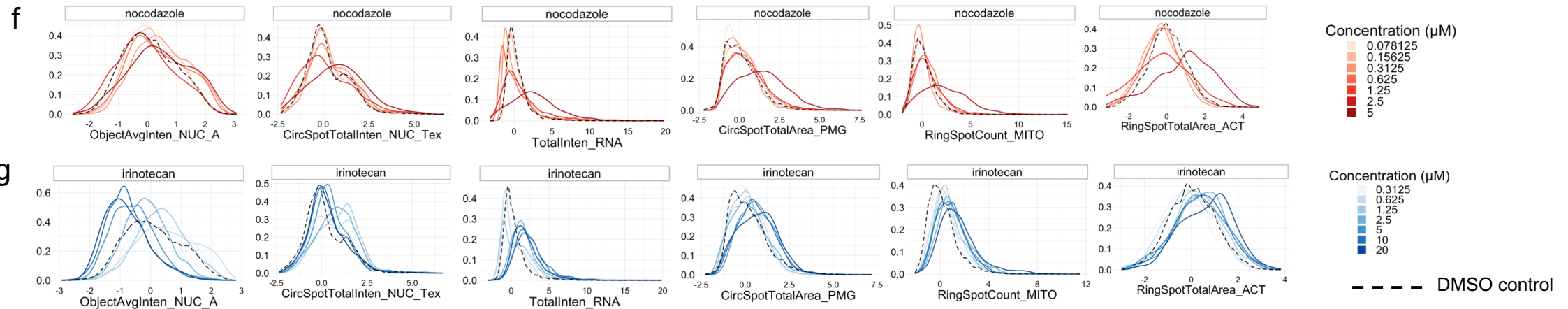
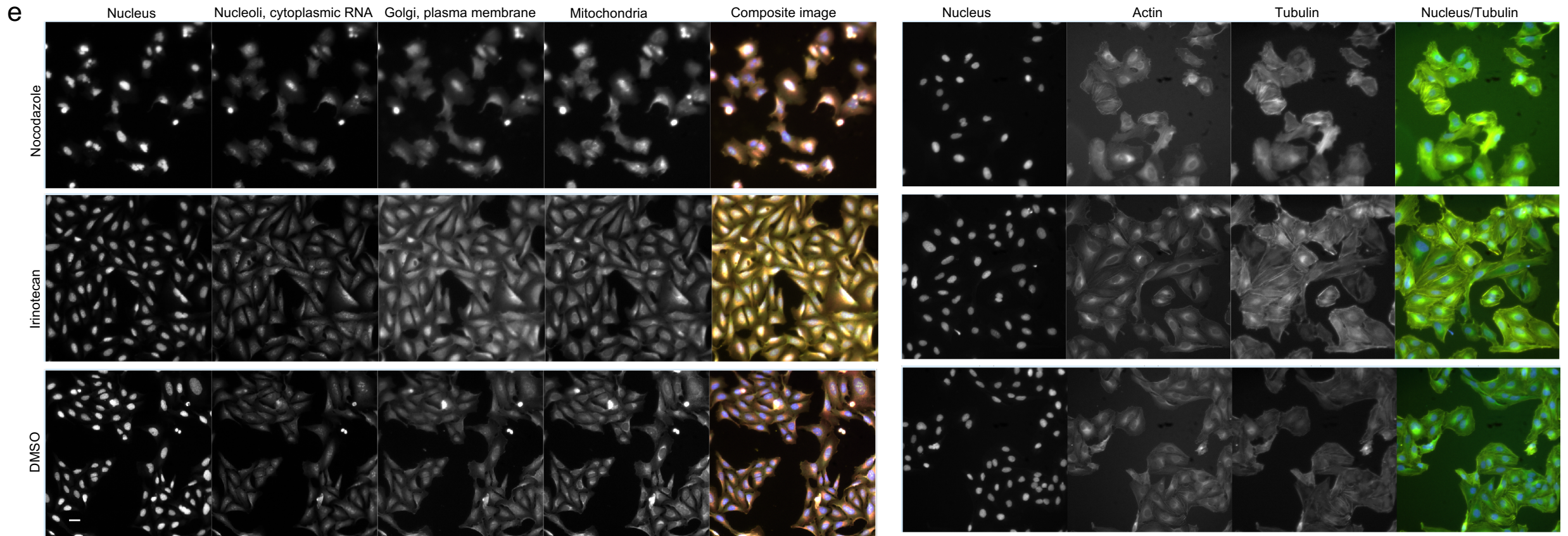
(d – f) 6 compounds with diverse MOA induce phenotypic activity largely independent of dosage/concentration. **(d)** Phenotypic trajectories of the “active (dose-insensitive)” group. These compounds show a distinctive phenotype from the control and low stress group, and the phenotypic response varies little with dosage. These compounds show mildly reduced cell counts in comparison to the control **(e)** and elicit differing effects on the cell cycle **(f)**. Visual cues for **(e, f)** are as in **(b, c)**.

(g – i) 25 compounds elicit dose-dependent phenotypes and are characterized as the “active (dose-responsive)” group. **(g)** Phenotypic trajectories of these compounds travel from the “low stress” region to other areas in the UMAP. Some, but not all, compounds show concentration-dependent effects on cell counts **(h)** and cell cycle responses **(i)**. Visual cues for **(h, i)** are as in **(b, c)**.

(j – l) 18 compounds show a dose-dependent phenotypic change and become toxic at high concentrations. **(j)** Phenotypic trajectories of the “active and cytotoxic” group. These compounds show strong decreases in cell counts with increasing concentration **(k)** and elicit diverse cell cycle responses **(l)**. Due to extreme cytotoxicity, cell cycle distributions are not available for some compounds at higher concentrations. Visual cues for **(k, l)** are as in **(b, c)**.



Supplementary Figure 6: Divergent phenotypes.

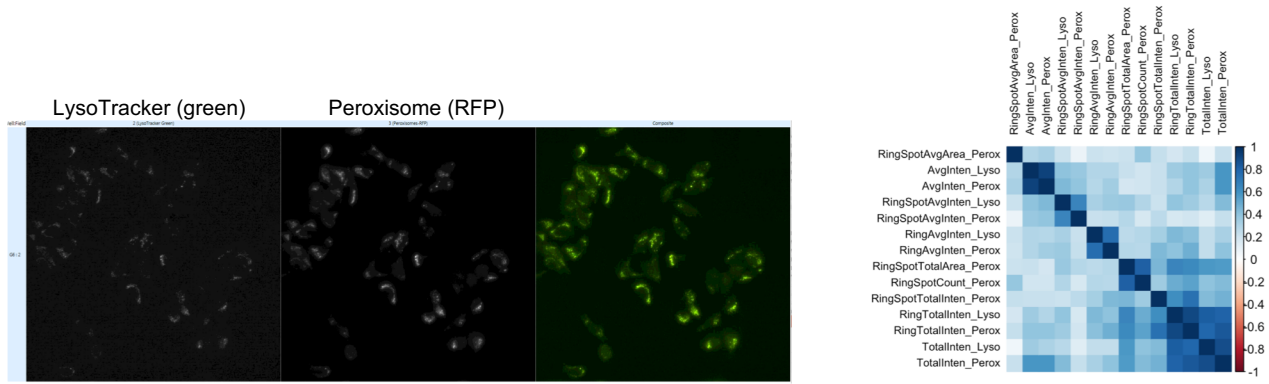


Supplementary Figure 6: Divergent phenotypes.

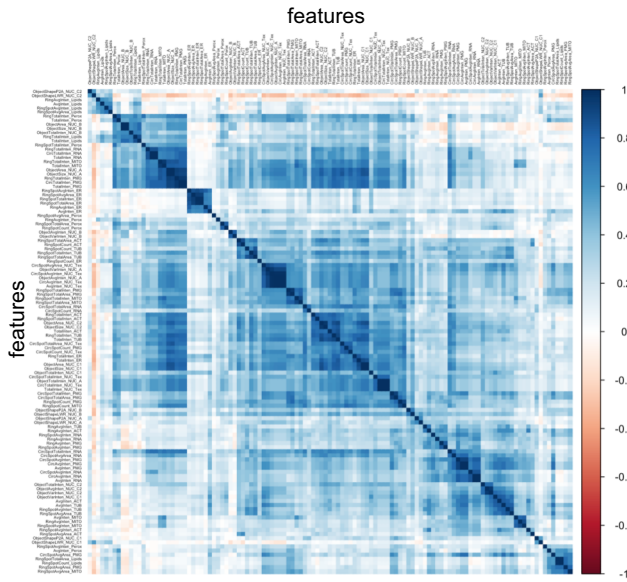
Supplementary Figure 6: Divergent phenotypes.

Phenotypic fingerprint of both Nocodazole and Irinotecan superimposed at their (a) lowest and (b) highest concentrations. (c) Cell count and (d) cell cycle dose response reveal increased levels of stress across the concentration gradient, with diverse cell cycle effects. Cell count is measured from three replicate plates, dashed lines represent the max, min and mean of the DMSO control. (e) Raw biological images of cells treated with 20uM of Nocodazole (top row), 5uM of Irinotecan (middle row), and DMSO (bottom row). Columns display a selection of channels imaged in the multi-panel assay. Scale bar: 20 μ m. (f) Cell feature distributions display diverse phenotypic responses to Nocodazole and (g) Irinotecan.

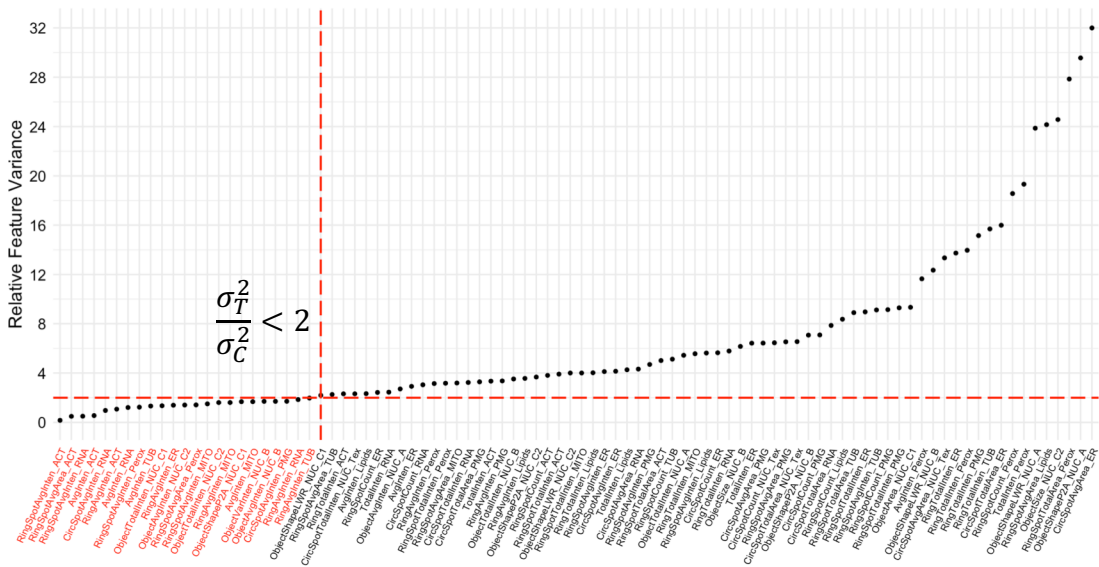
a



b



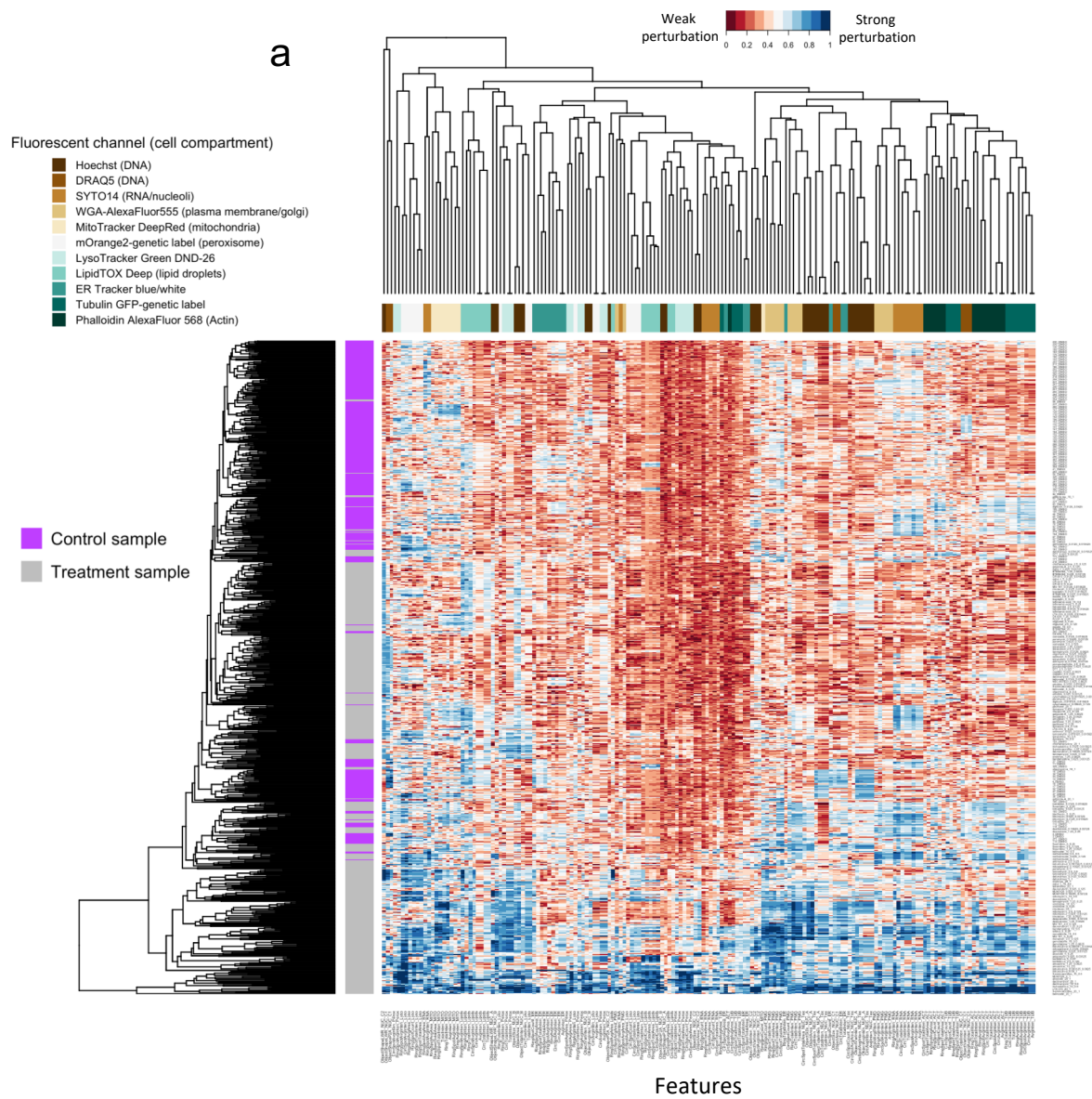
c



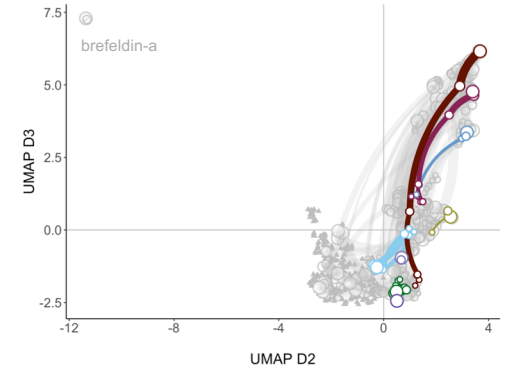
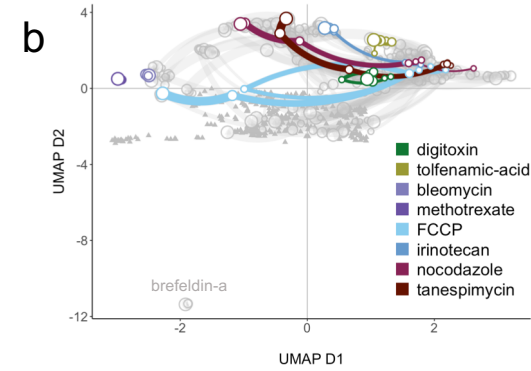
Supplementary Figure 7. Feature reduction.

Supplementary Figure 7. Feature reduction.

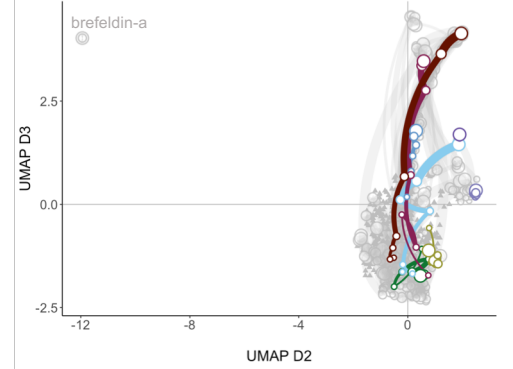
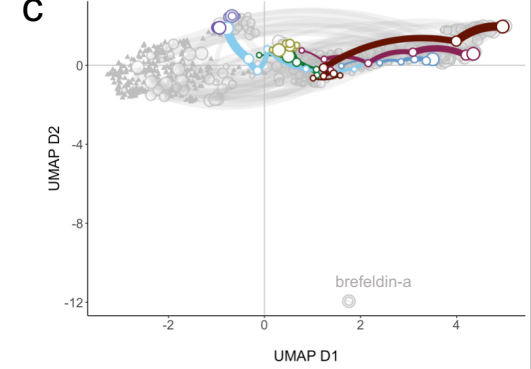
(a) The lysosome and peroxisome reporters are highly correlated due to overlap in their fluorescent emission spectra arising from the weak lysosomal staining by LysoTracker Green. The right-hand panel shows the linear correlation between the parameters of these two channels. (b) Linear correlation analysis of all remaining features. For any pair of highly correlated features (correlation coefficient > 0.9), the feature showing a larger mean correlation with all other features is removed. (c) Relative feature activity: per-feature variance is calculated for both control and treatment samples. Features whose variance among treatments is less than twice the variance of the control are deemed “inactive”.



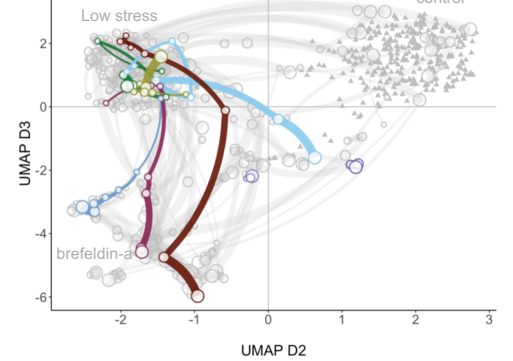
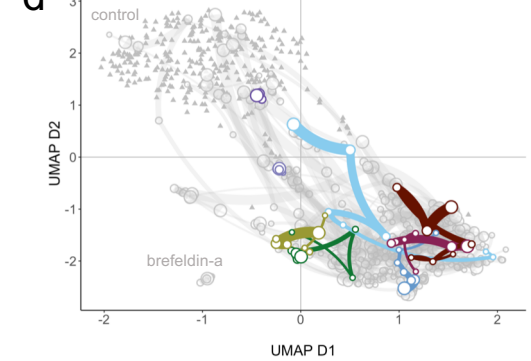
b UMAP: unprocessed data and full feature set



c UMAP: processed data and full feature set



d UMAP: processed data and optimally reduced feature set



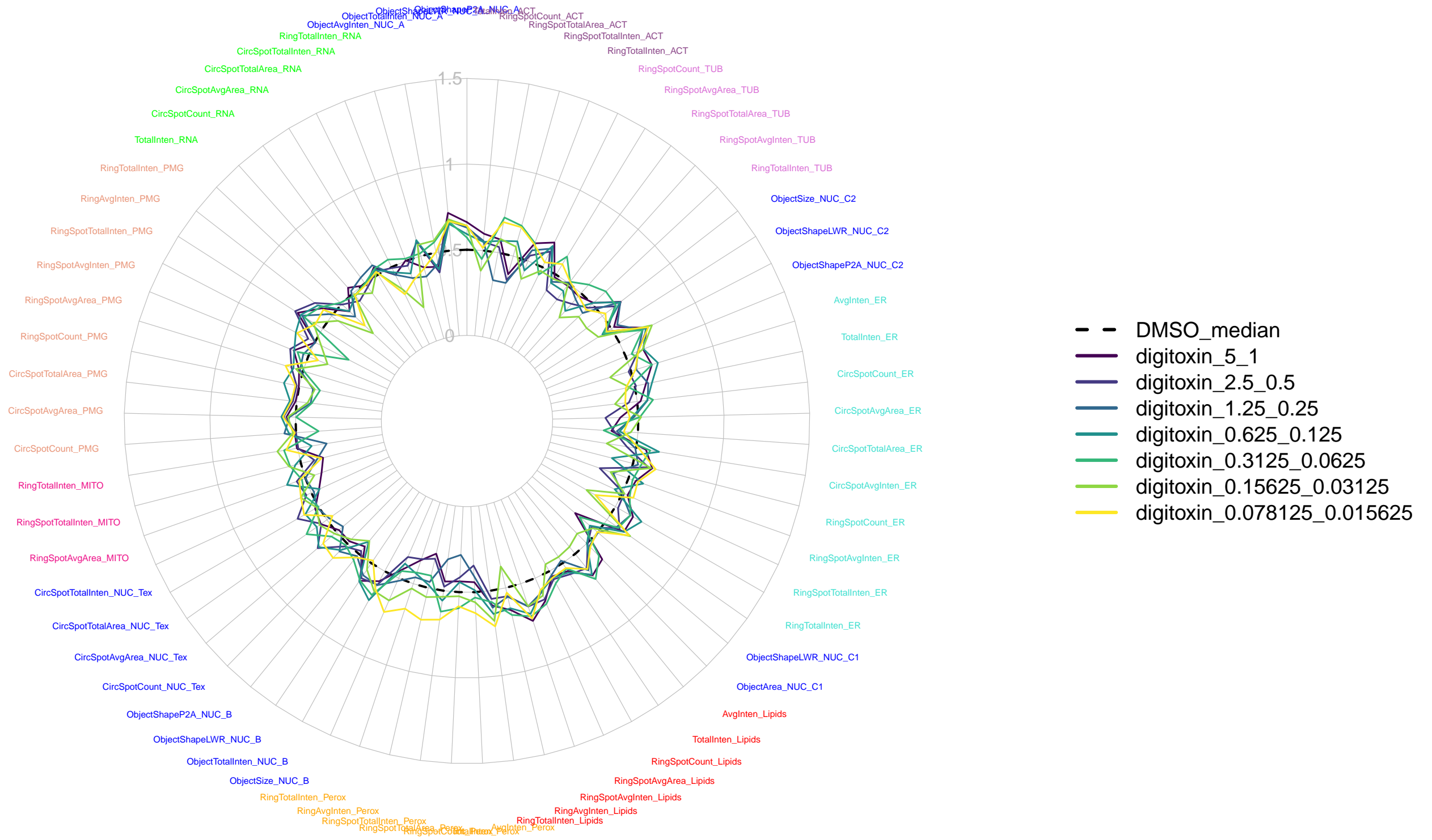
Supplementary Figure 8: Comparison of cytological profiles

(a - b) EMD profiles without adjusting for positional effects and plate to plate variation. (a) Heatmap visualization of similarity by hierarchical clustering shows several treatment groups (grey) clustered with the controls (purple). Dimension reduction by UMAP of (b) unprocessed EMD profiles and (c) processed full feature EMD profiles strongly separates the brefeldin-a cluster from other treatments. (c) UMAP of processed full feature profile also fails to separate the low stress cluster from the control when compared to the (d) UMAP of the processed and optimally reduced feature profile (described in Fig. 7a).

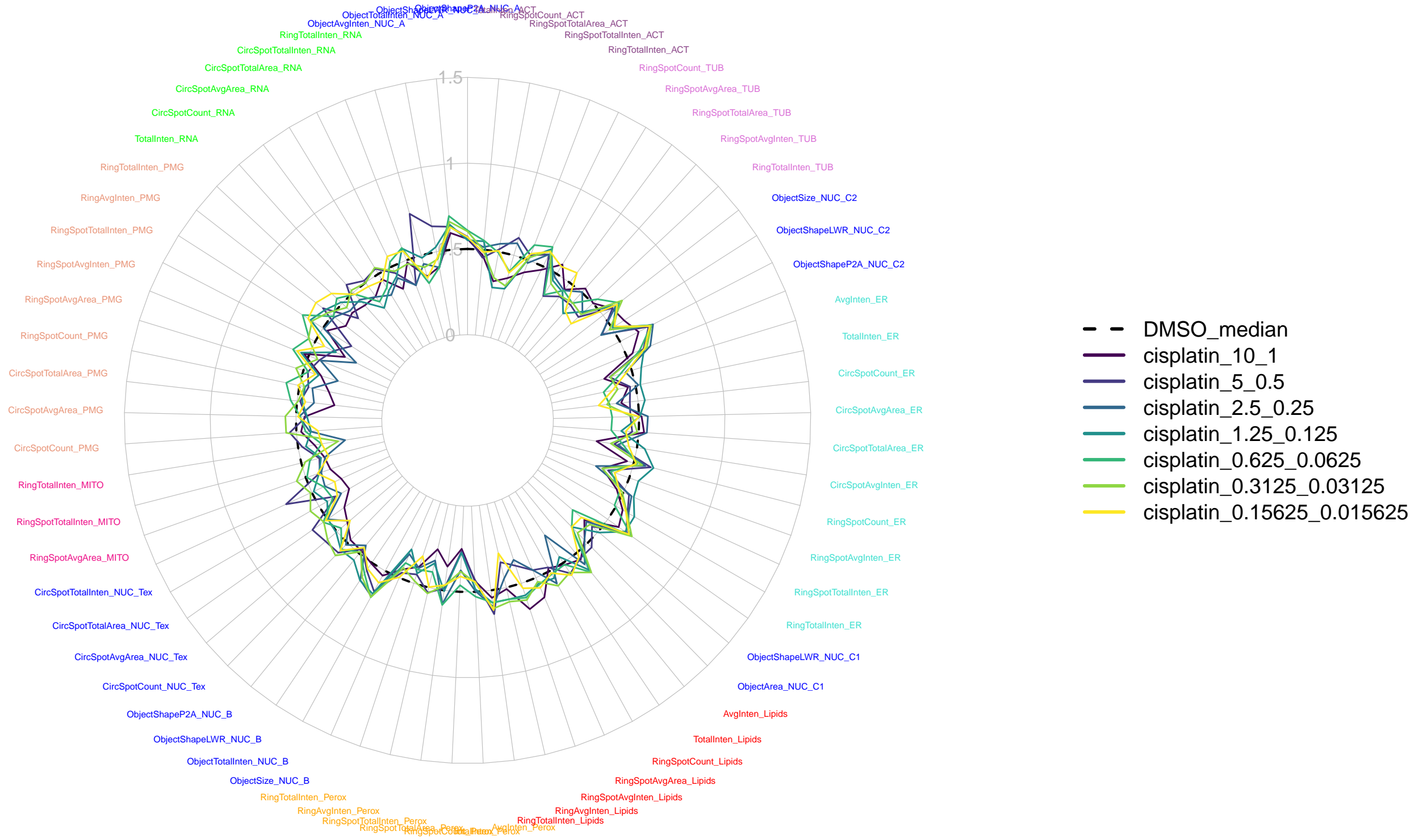
Supplementary Figure 9: Radial plot fingerprints of 65 compounds

Radial fingerprints of residual EMD scores for 65 compounds at multiple concentrations, relative to the control median (black dotted line). Labels for the reduced 69 feature set are color-coded by corresponding cellular marker. Per compound legends include compound name, concentration and dilution factor (*compound_concentration_dilution factor*). Lowest and highest concentrations are colored yellow and dark purple respectively.

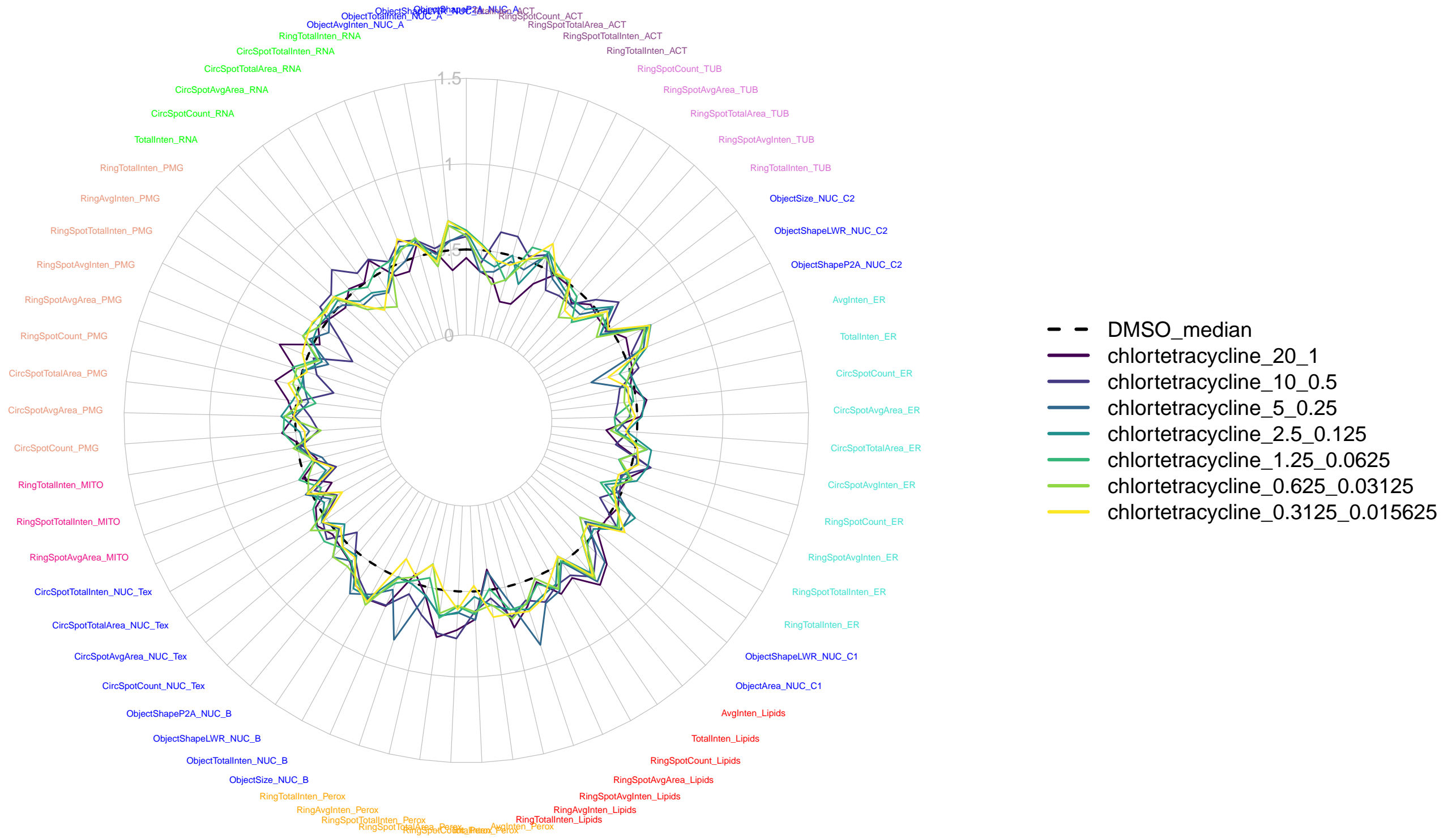
Compound: digitoxin



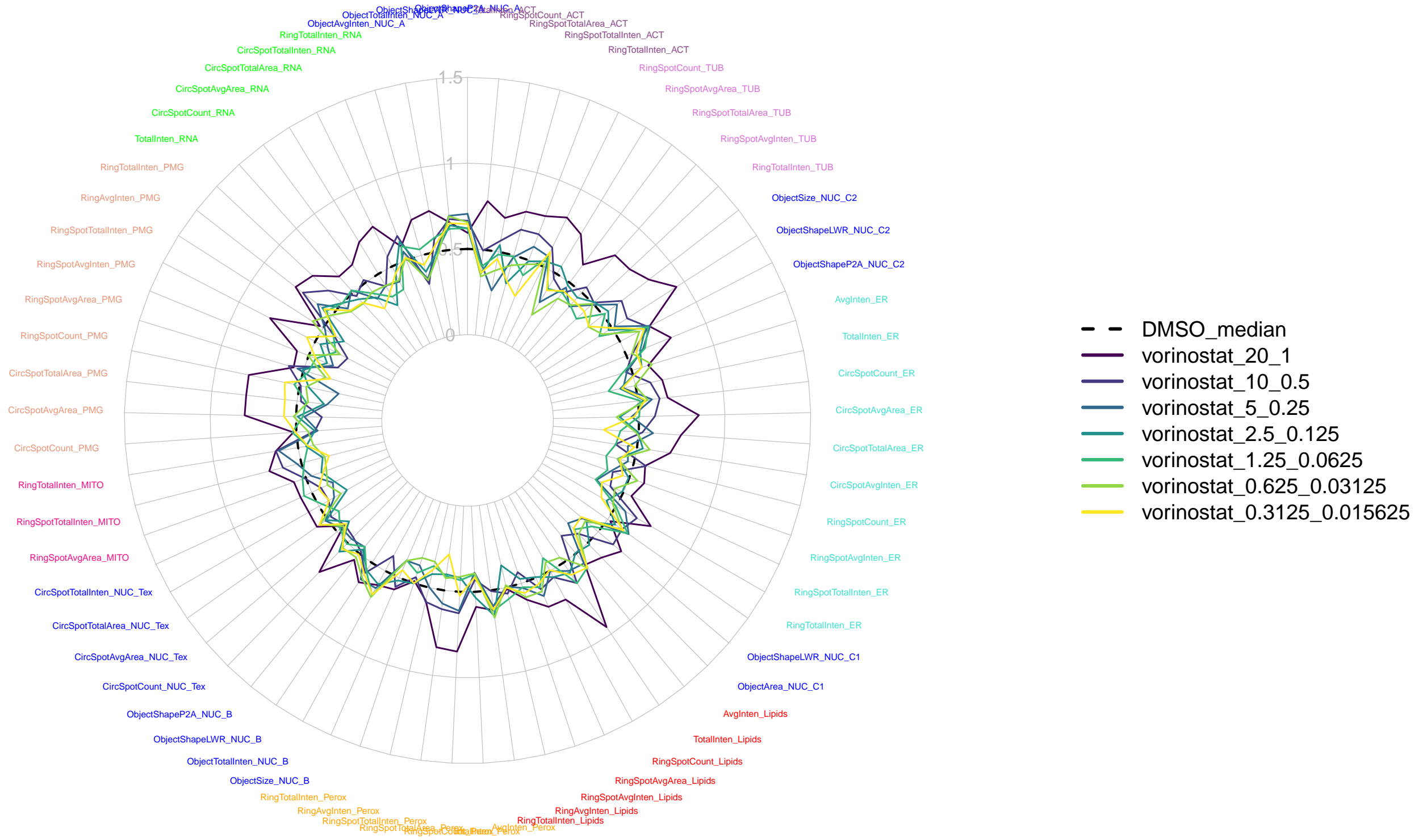
Compound: cisplatin



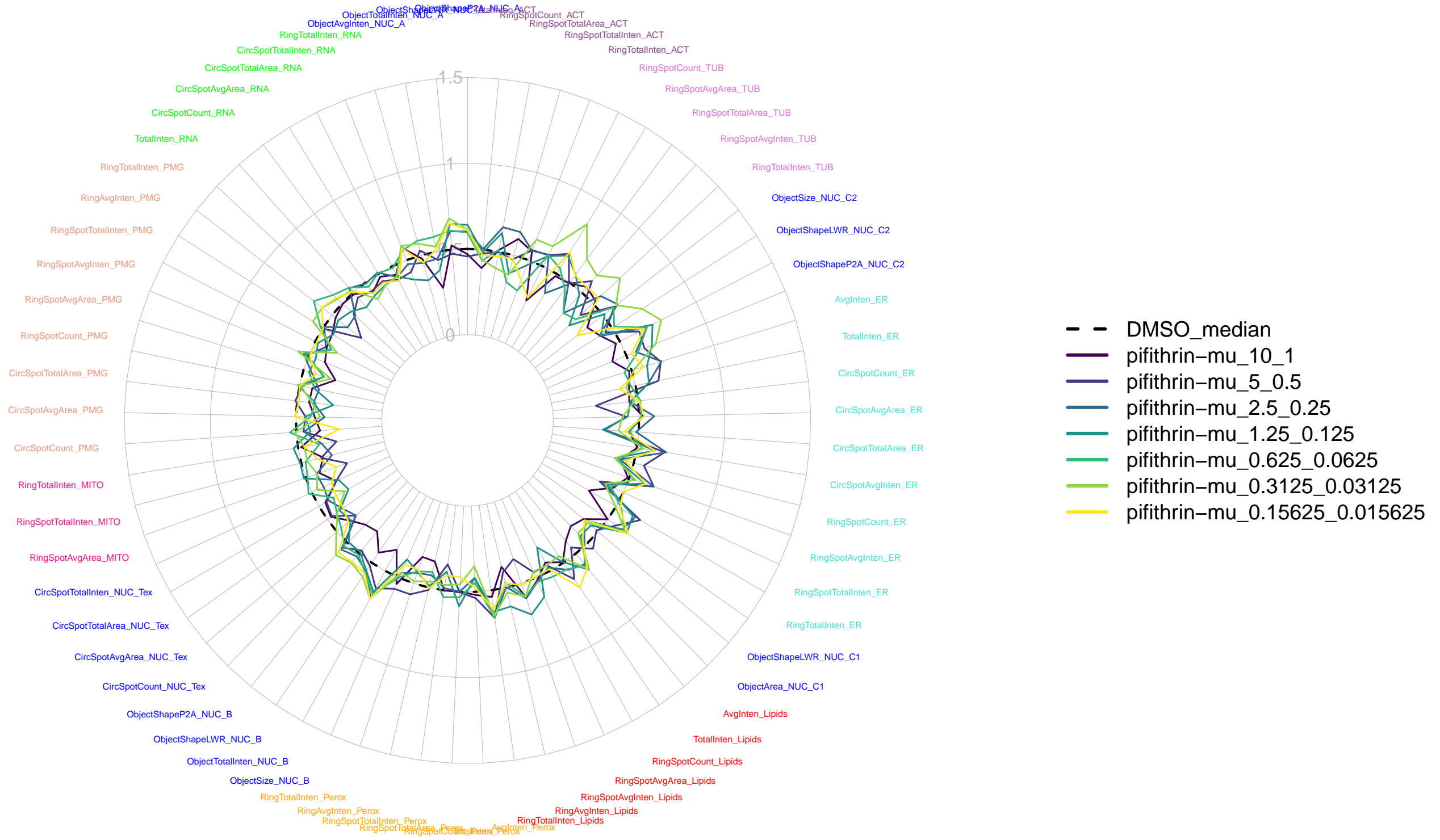
Compound: chlortetracycline



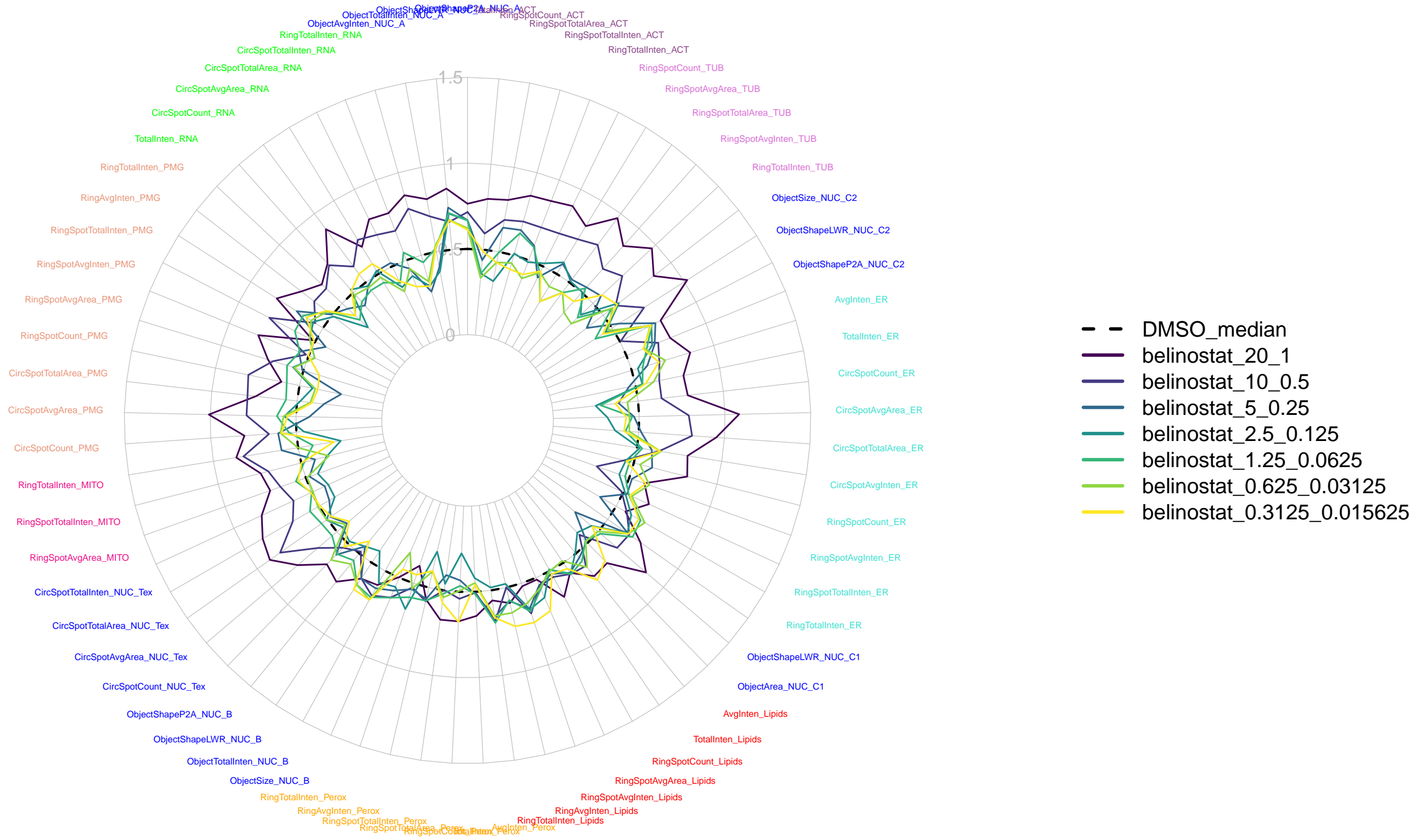
Compound: vorinostat



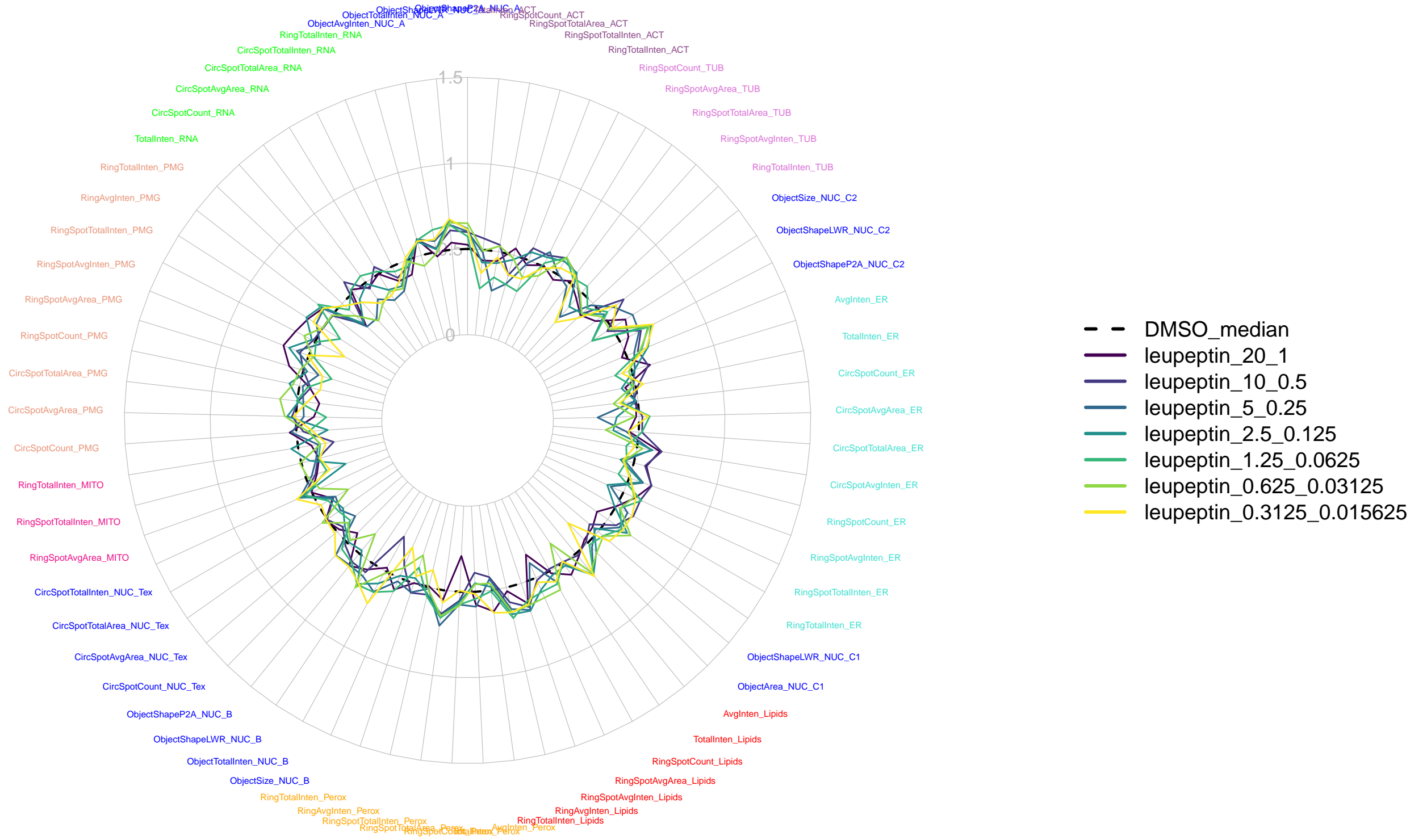
Compound: pifithrin-mu



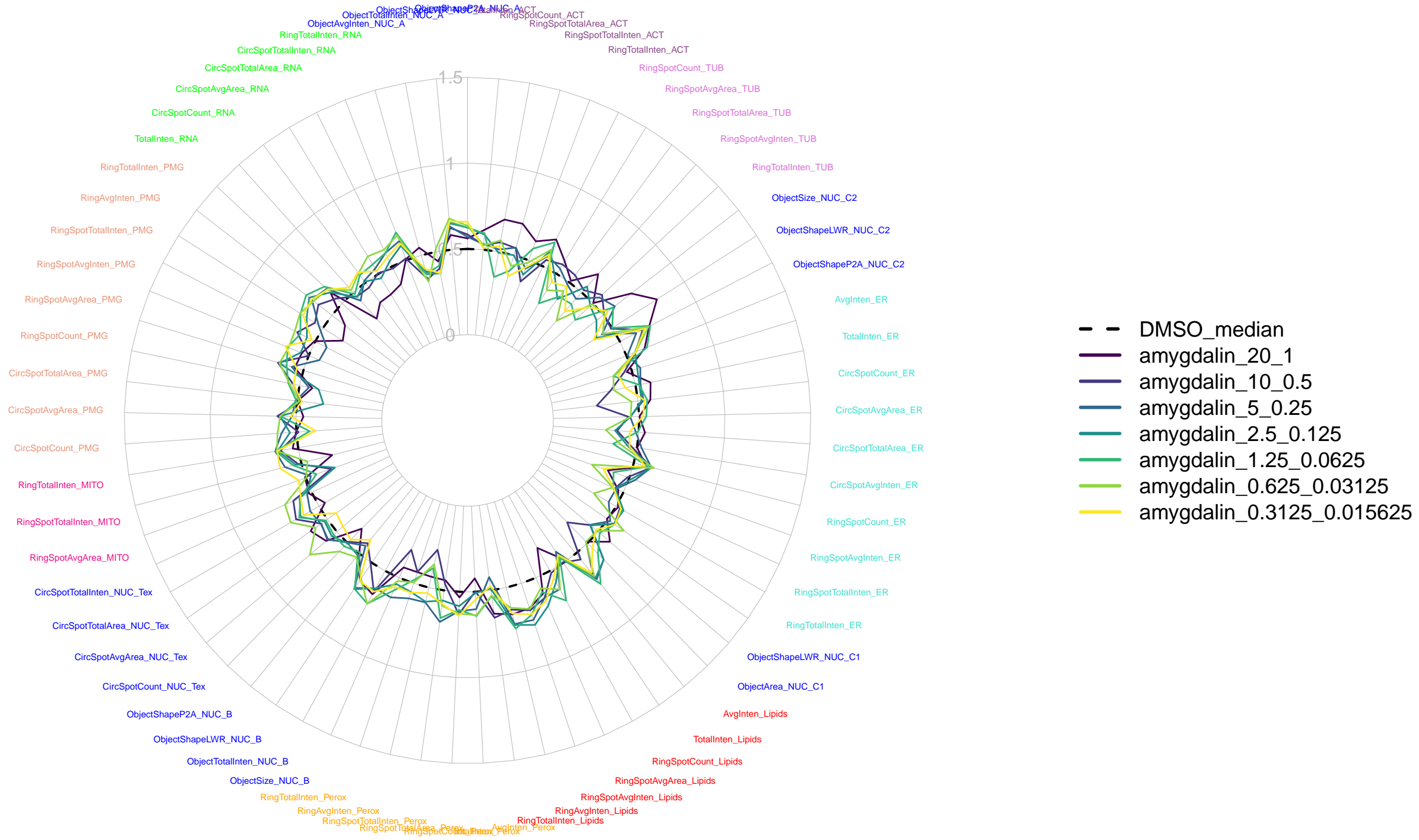
Compound: belinostat



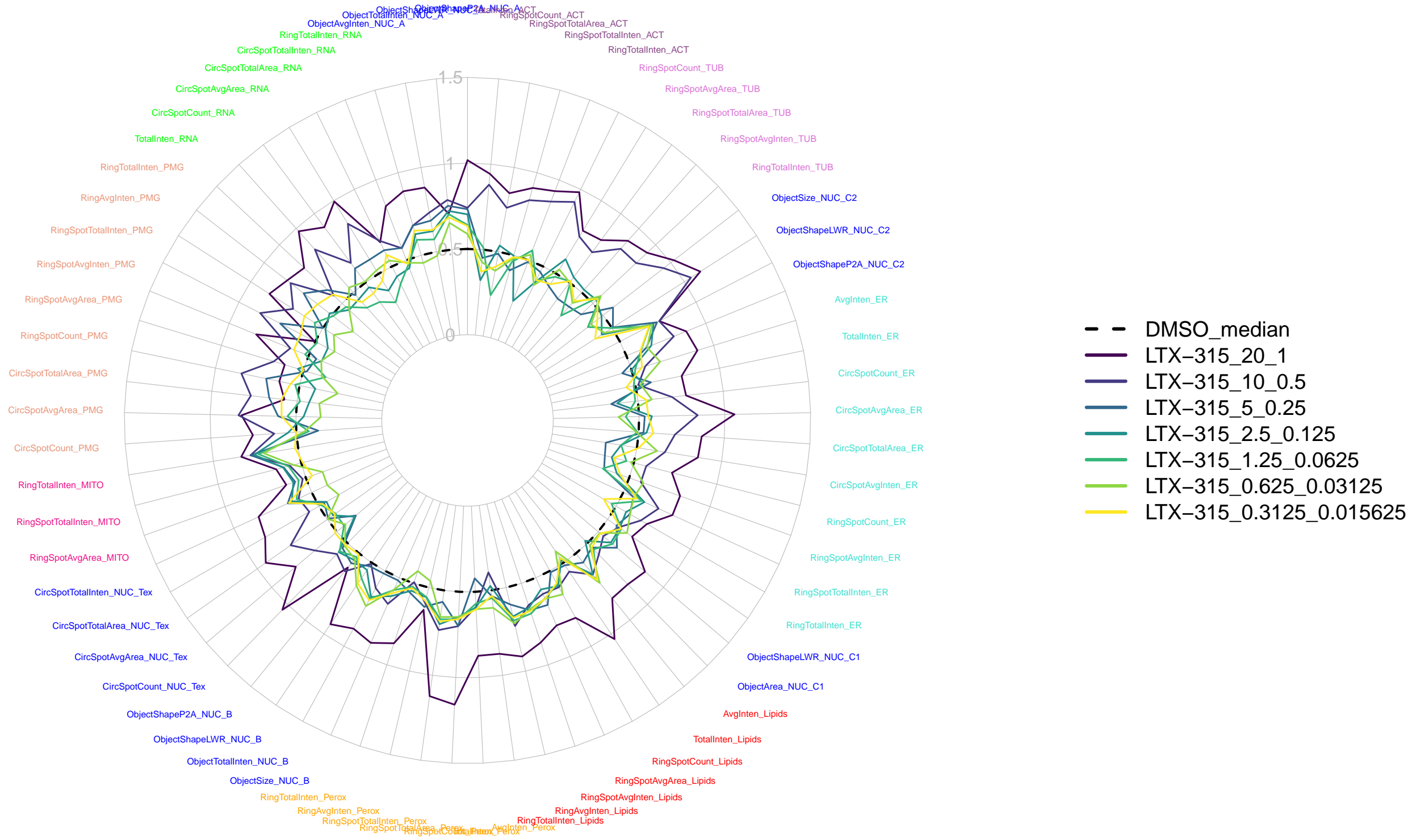
Compound: leupeptin



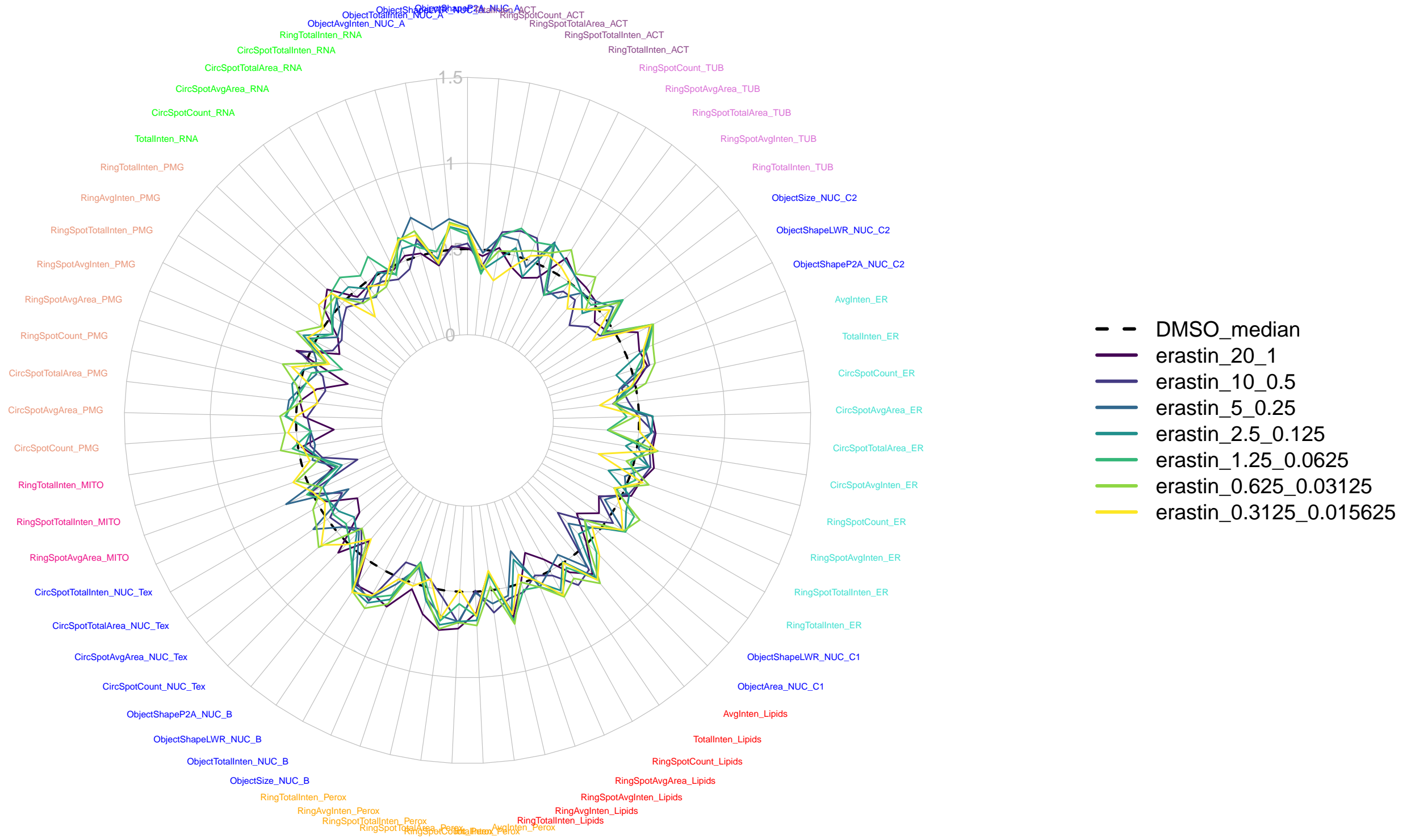
Compound: amygdalin



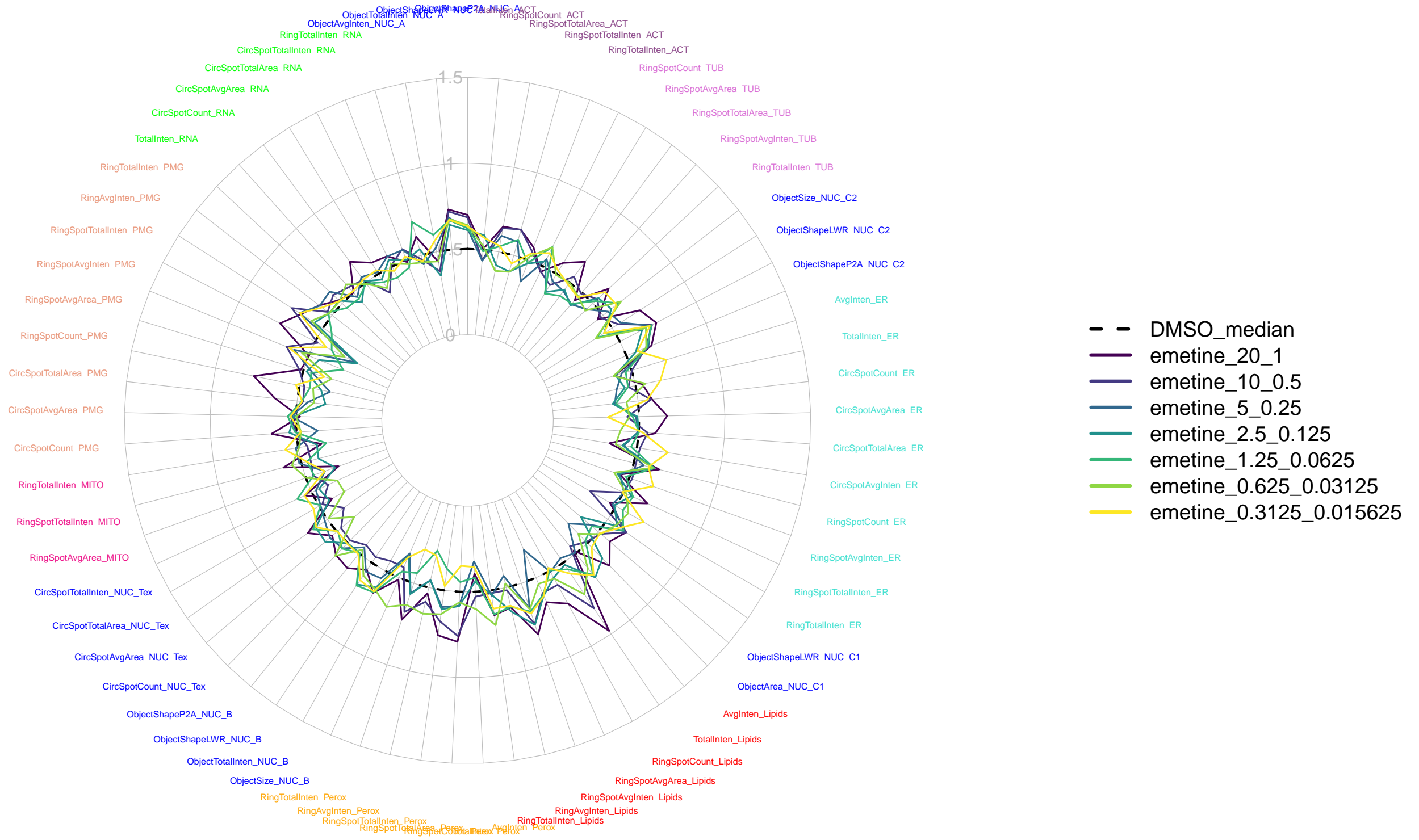
Compound: LTX-315



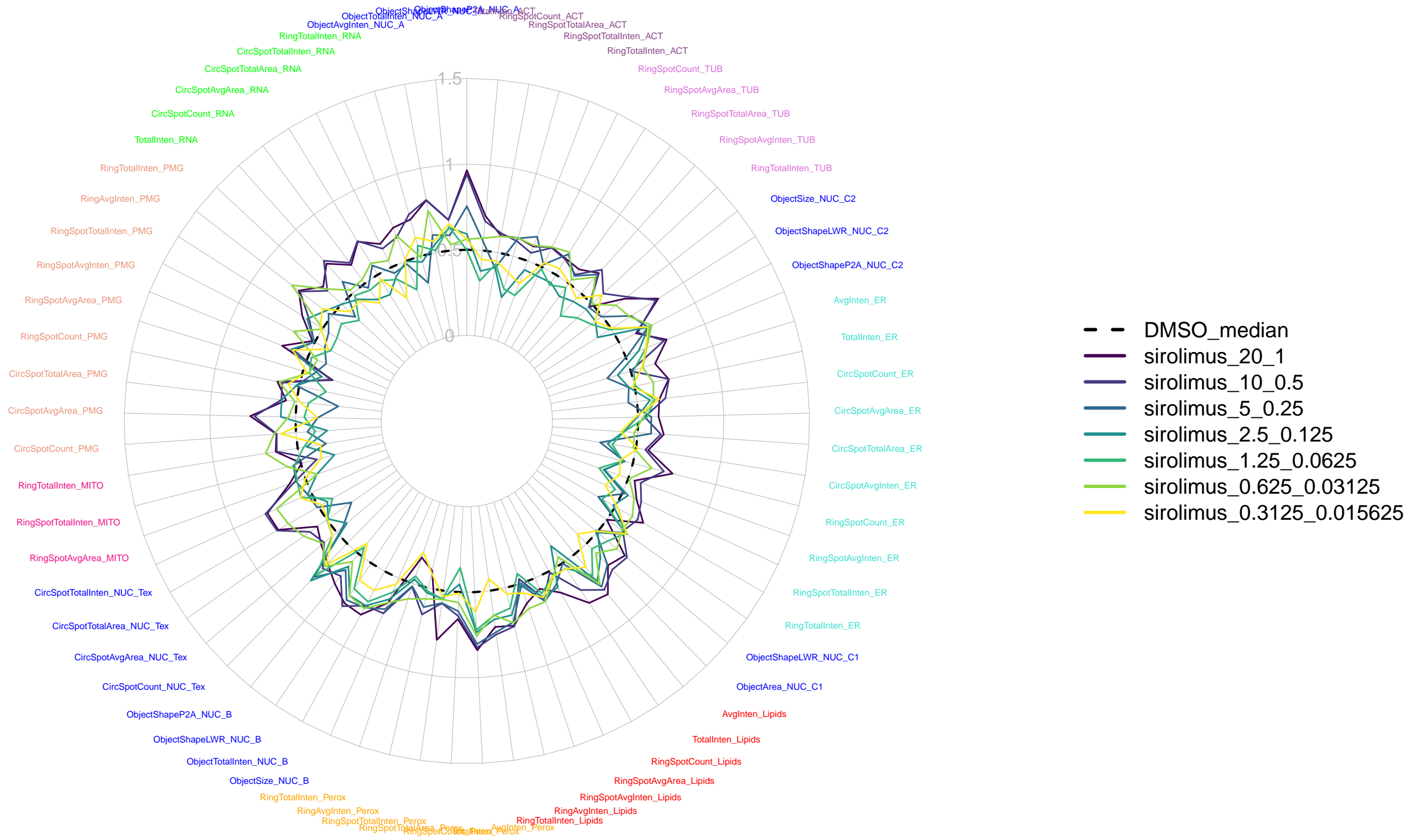
Compound: erastin



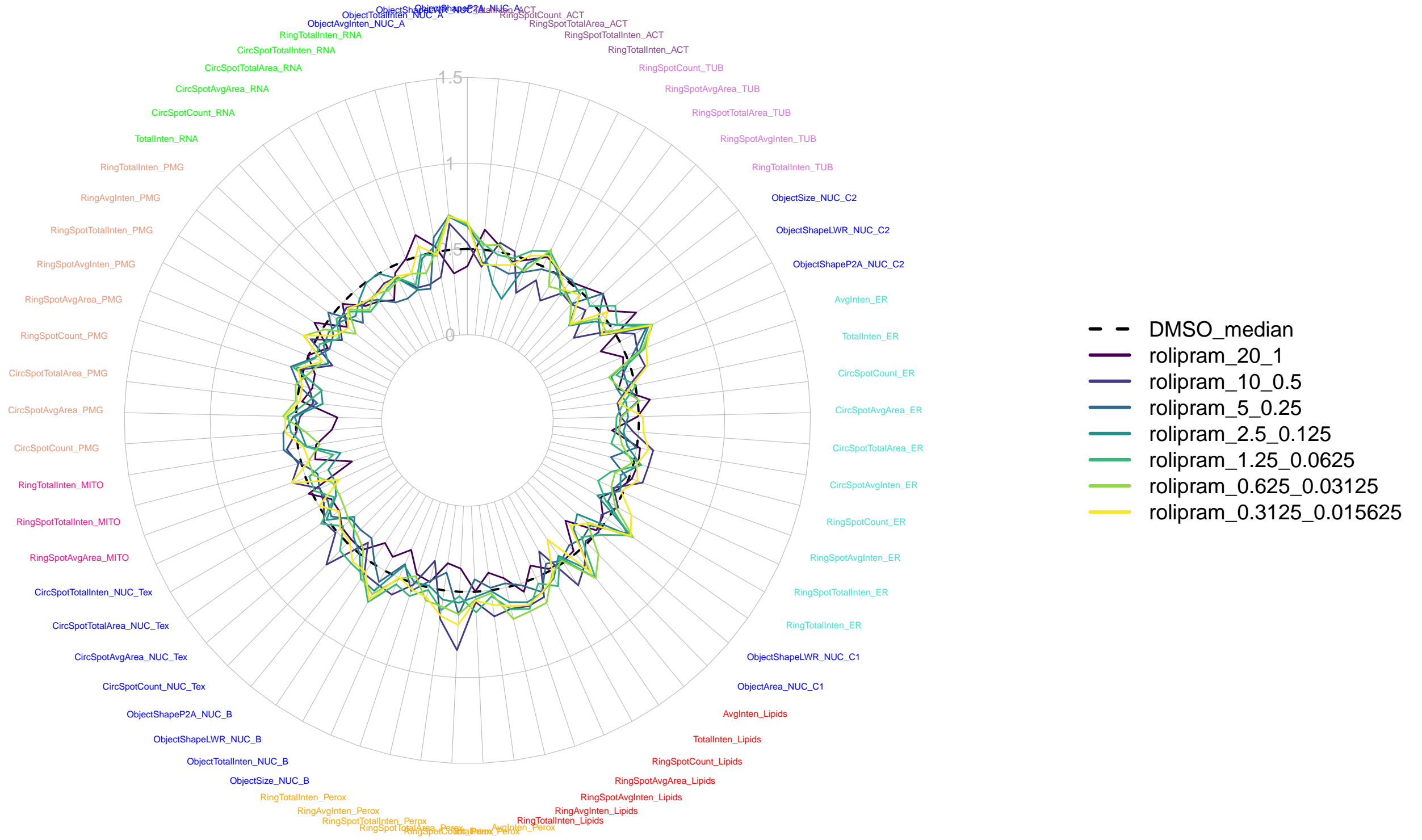
Compound: emetine



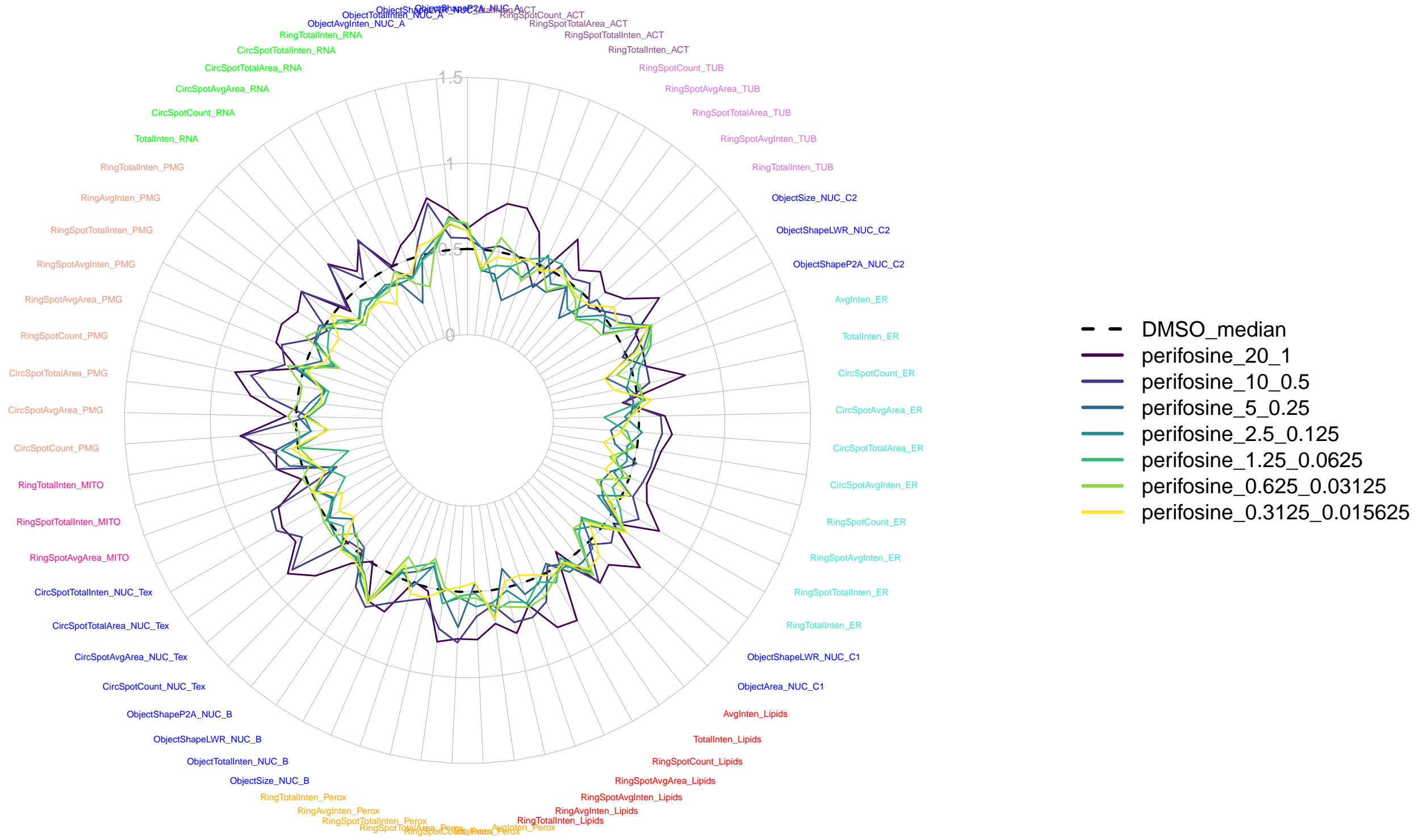
Compound: sirolimus



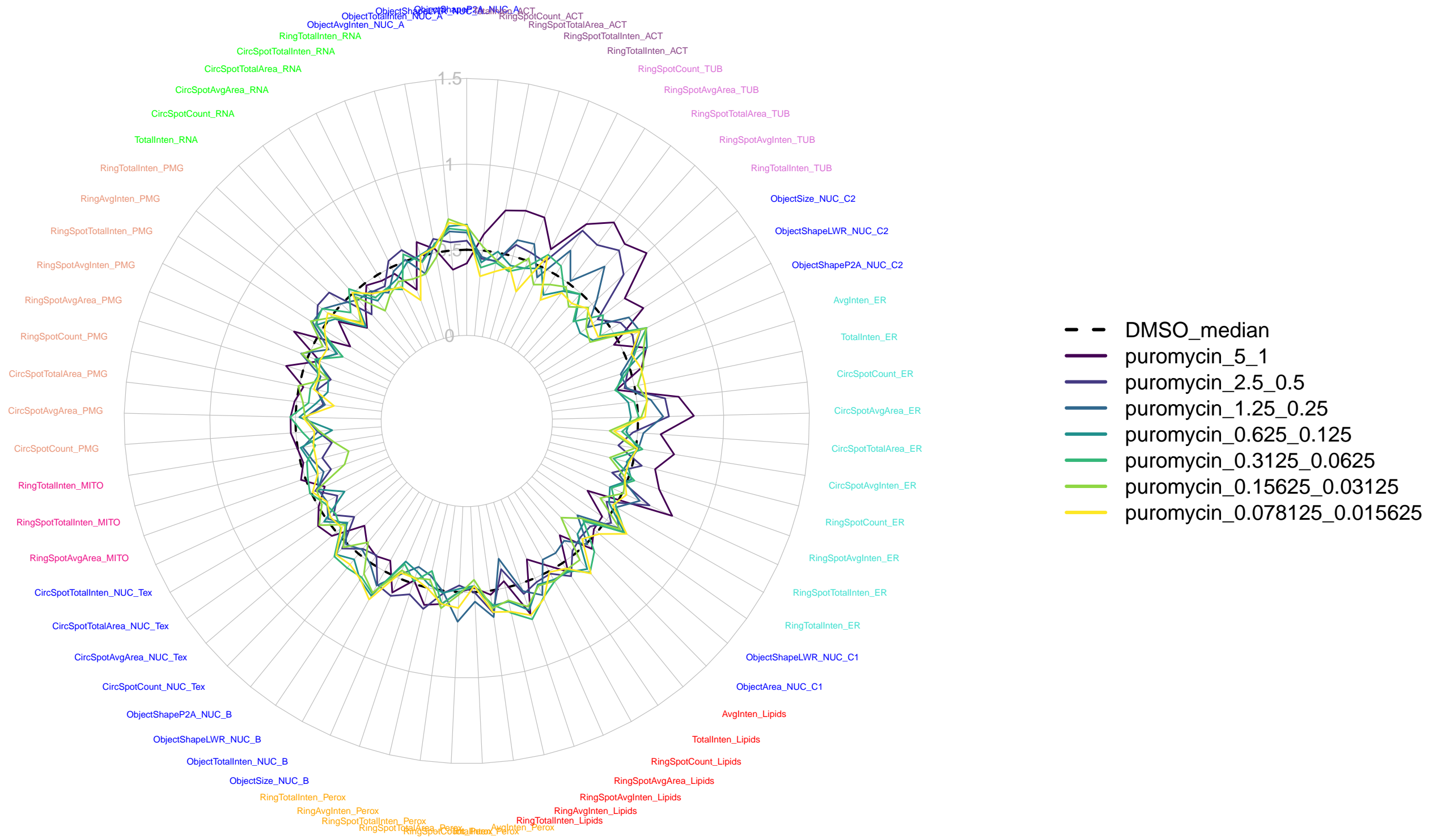
Compound: rolipram



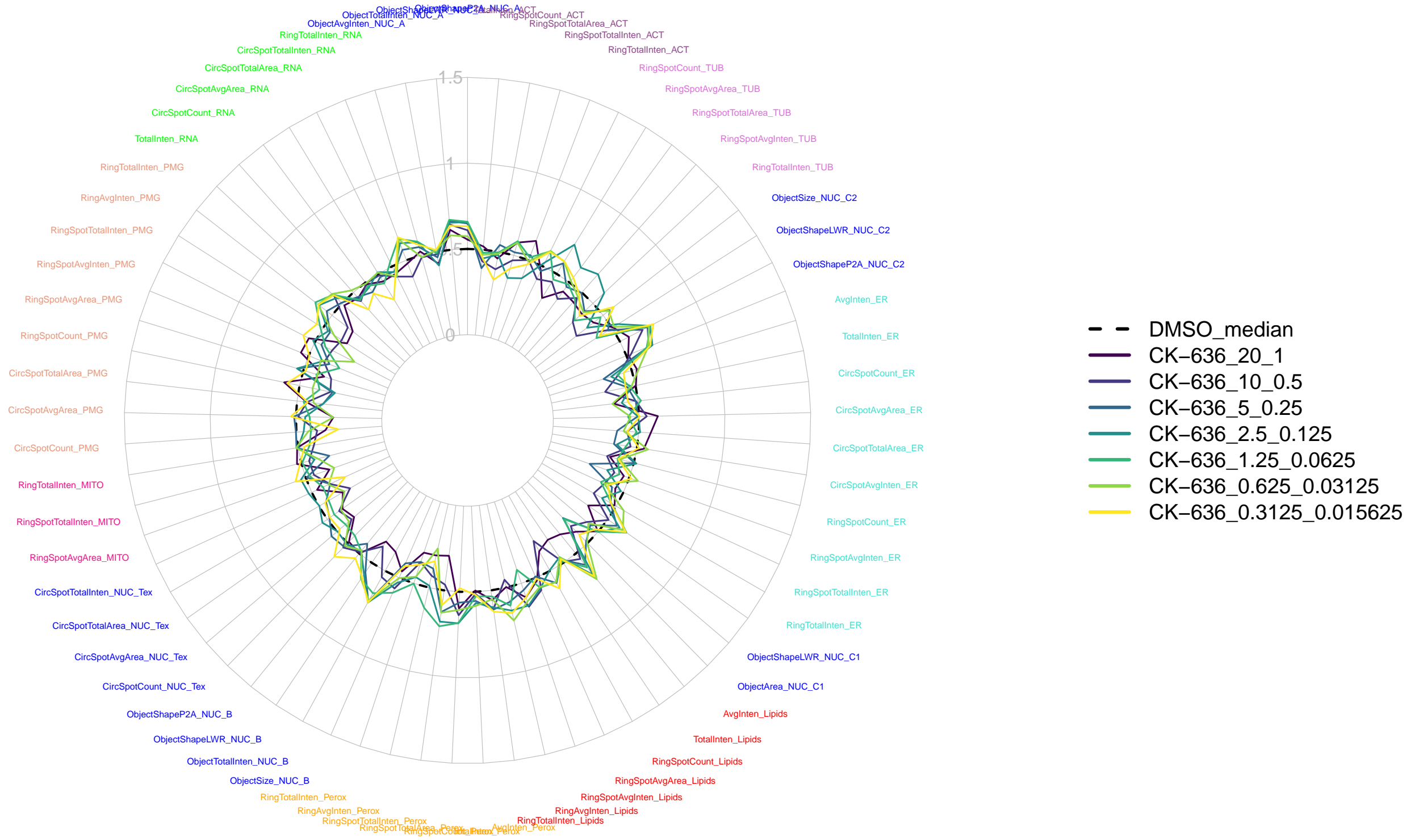
Compound: perifosine



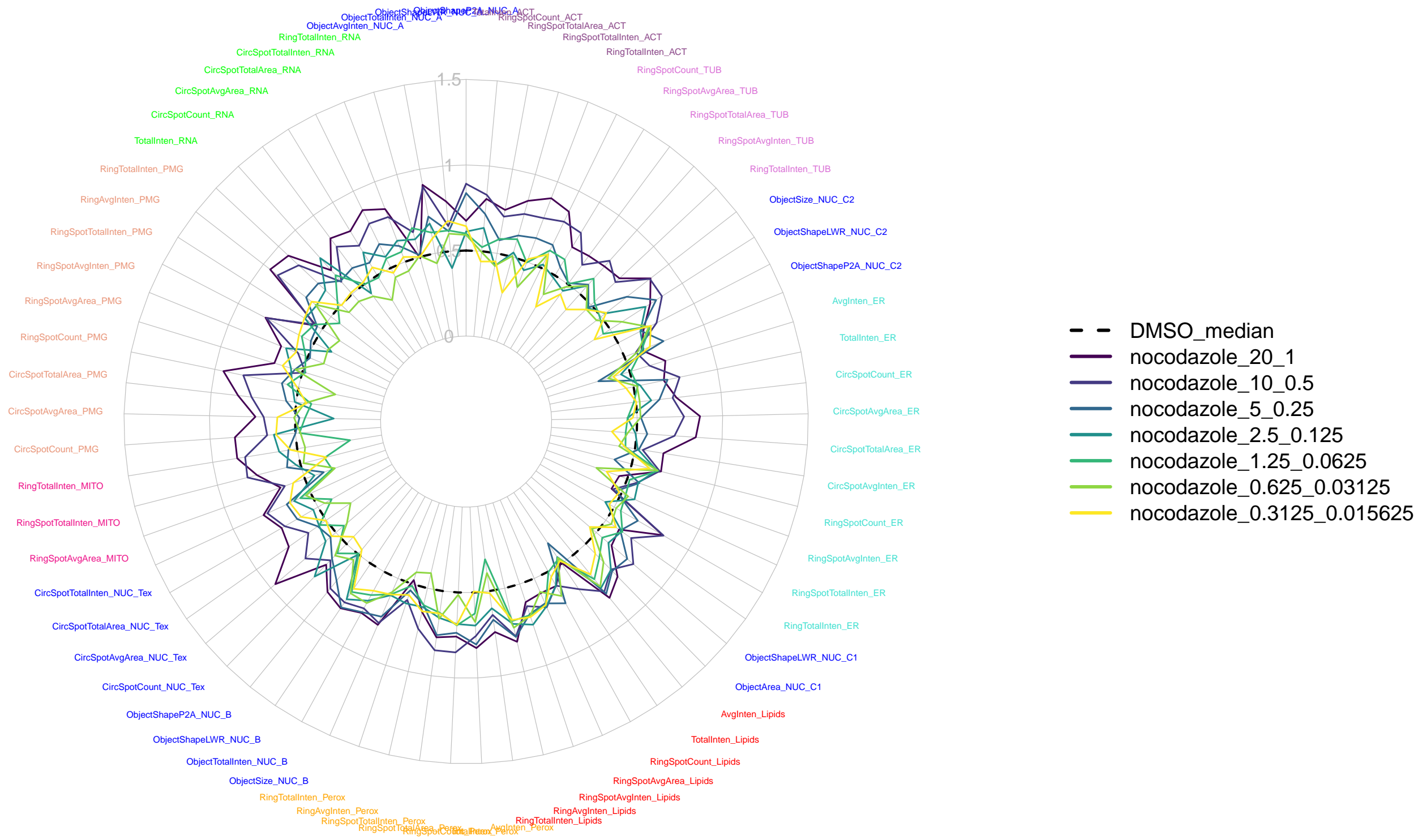
Compound: puromycin



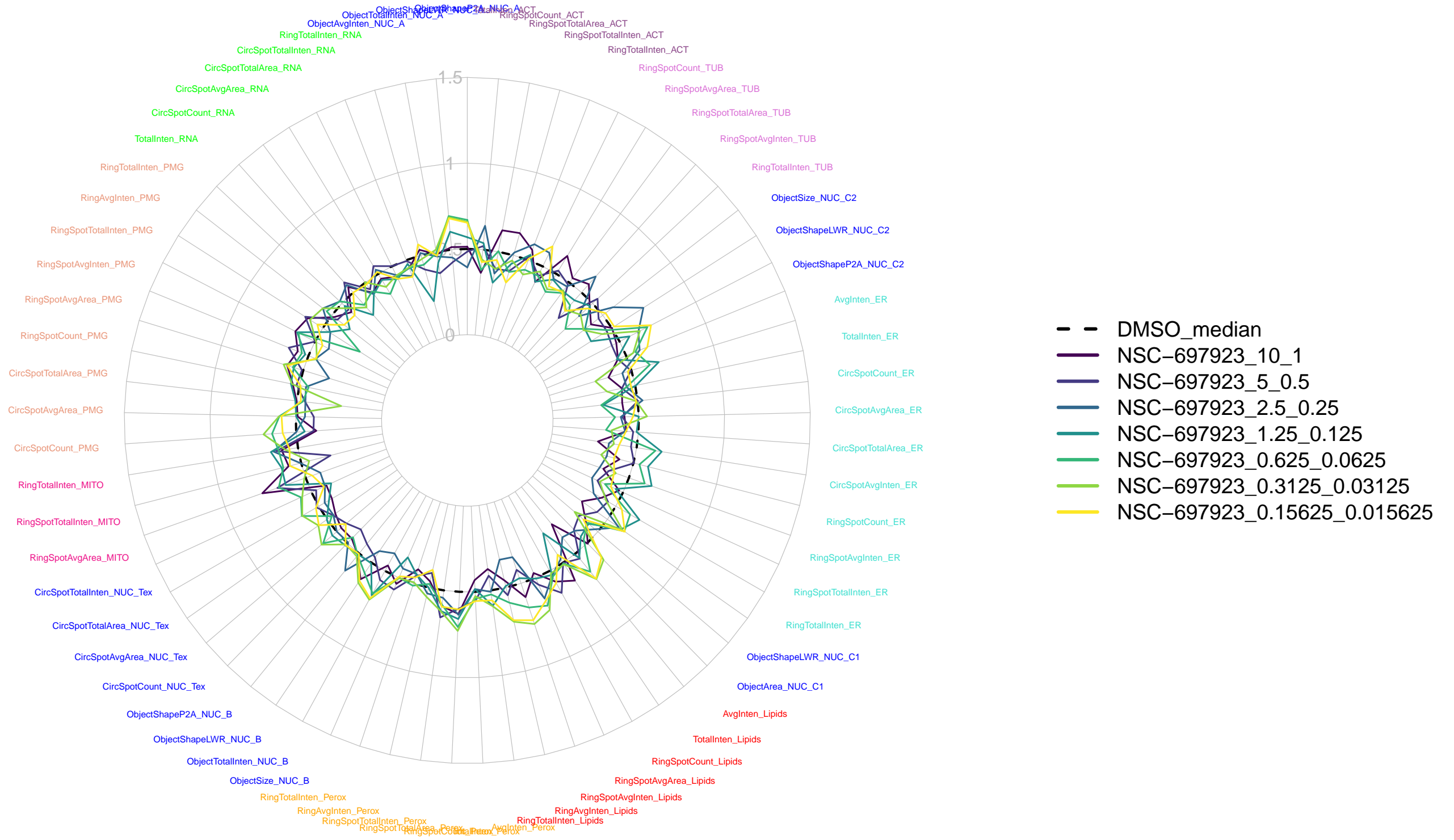
Compound: CK-636



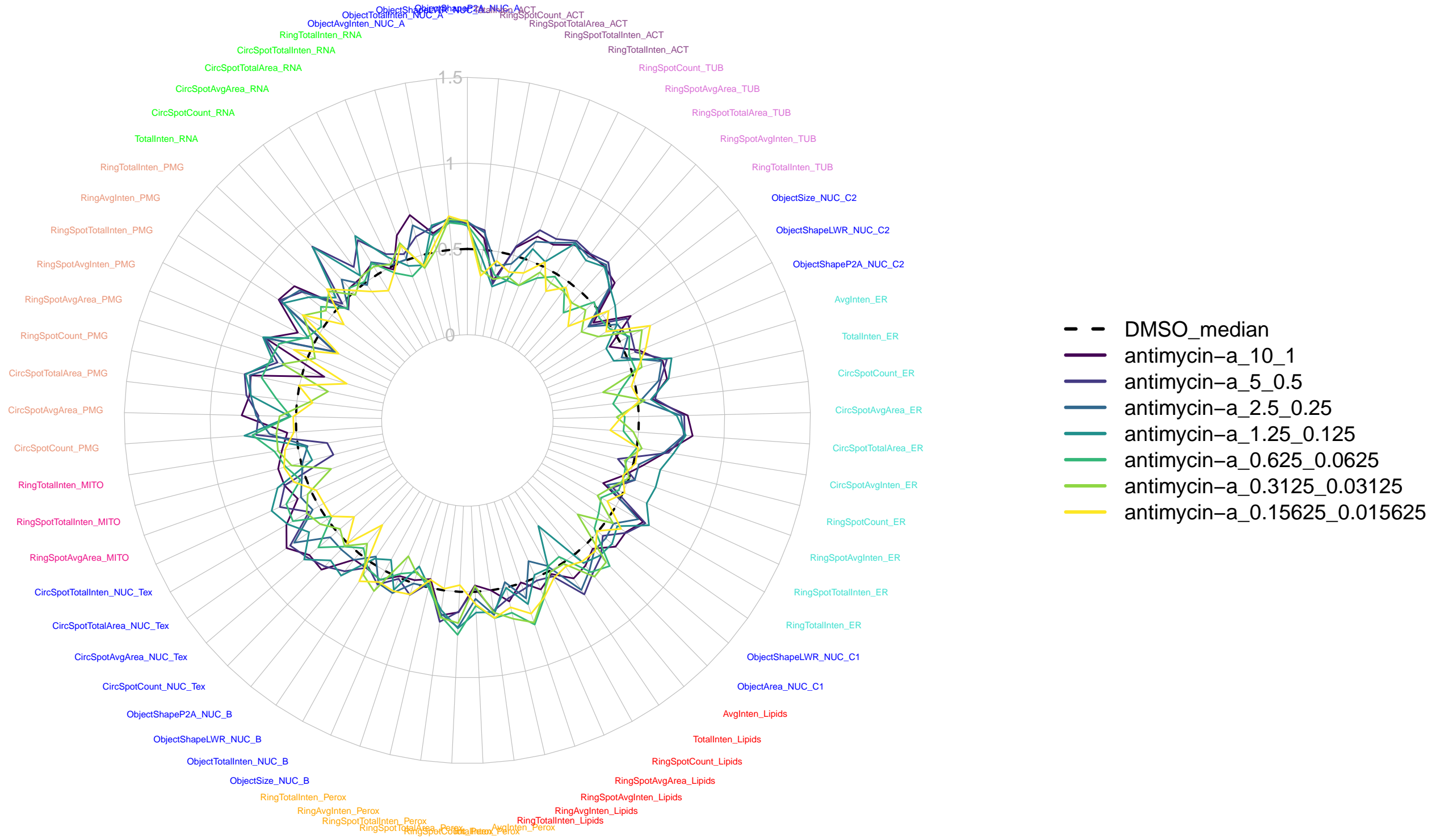
Compound: nocodazole



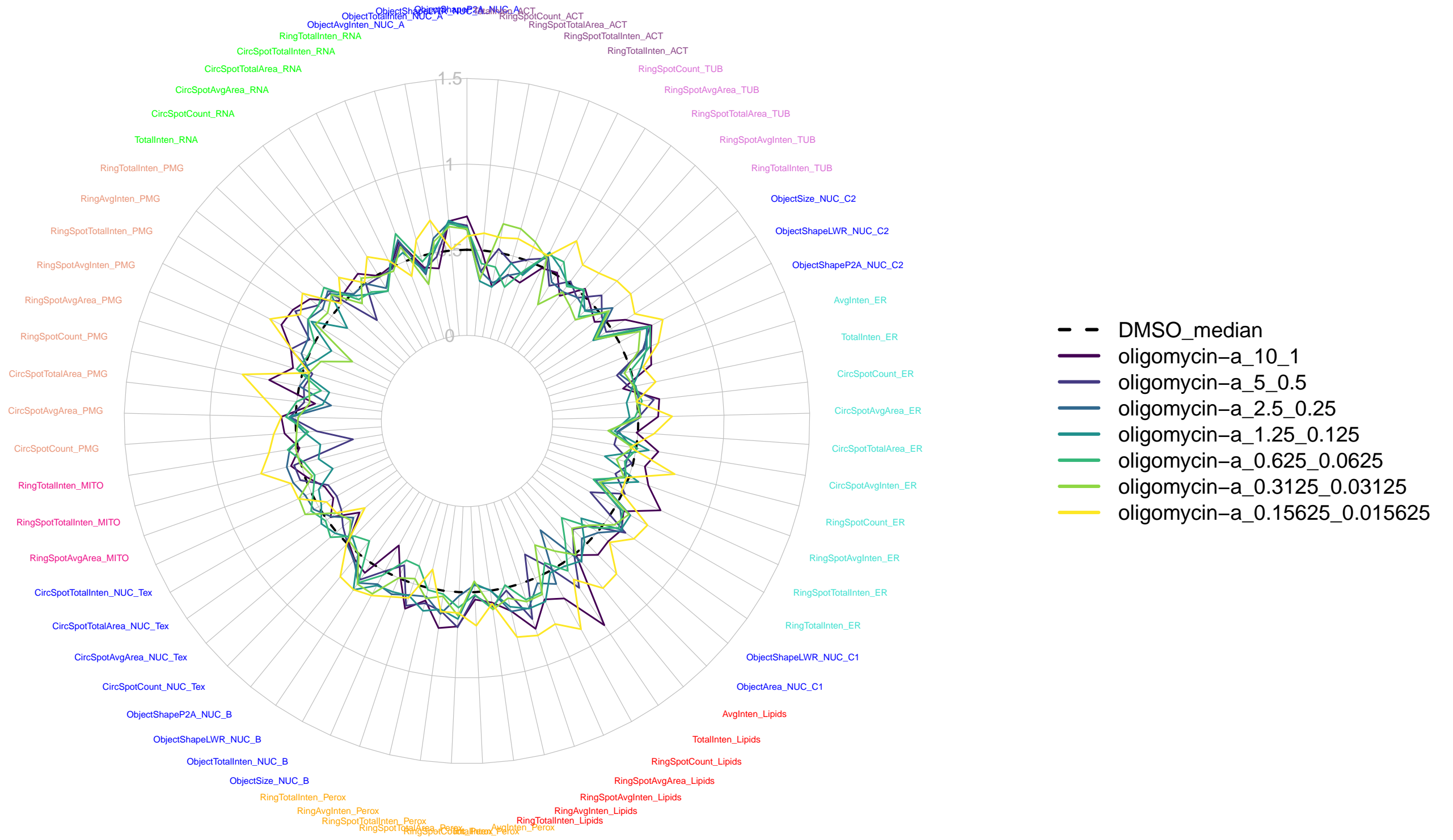
Compound: NSC-697923



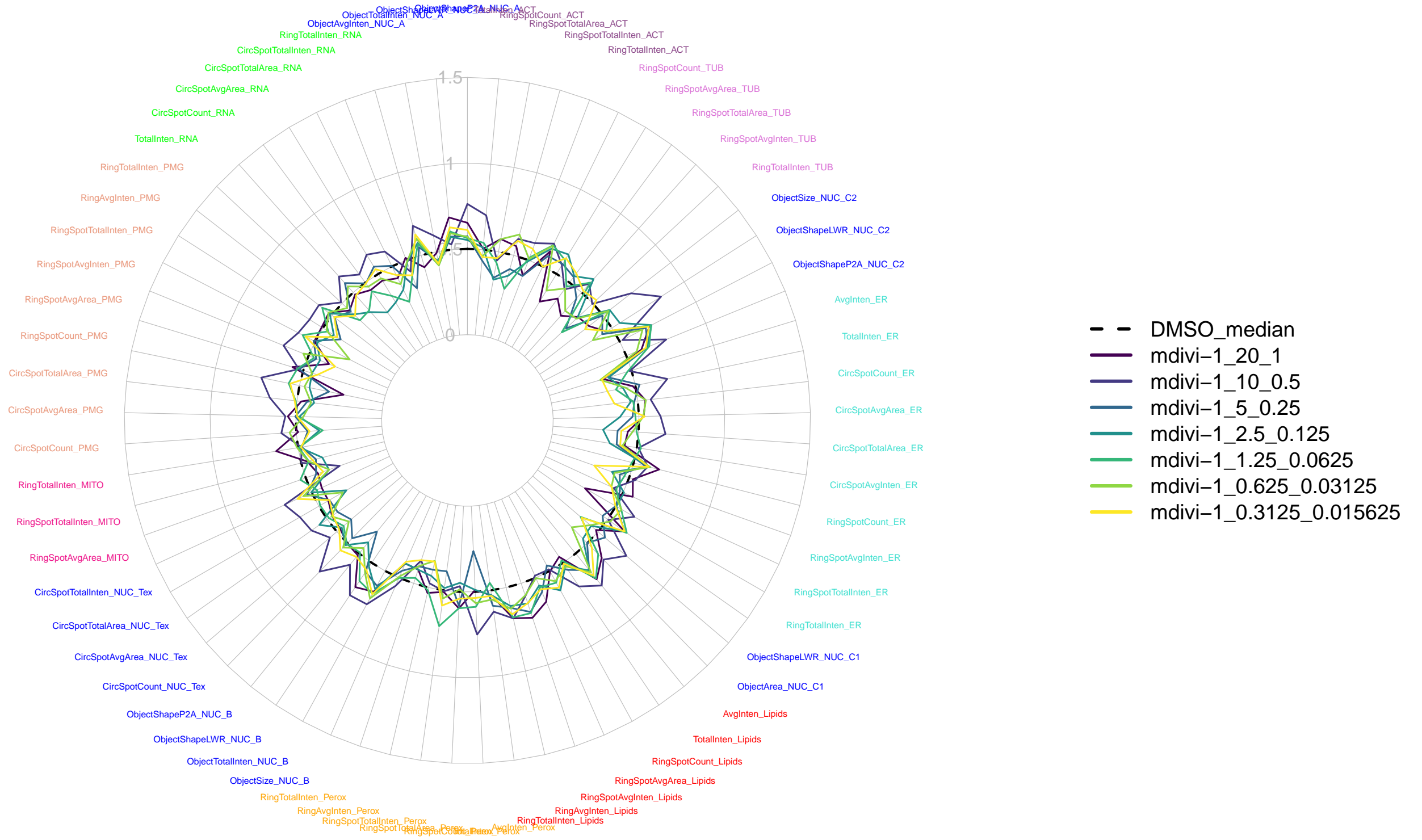
Compound: antimycin-a



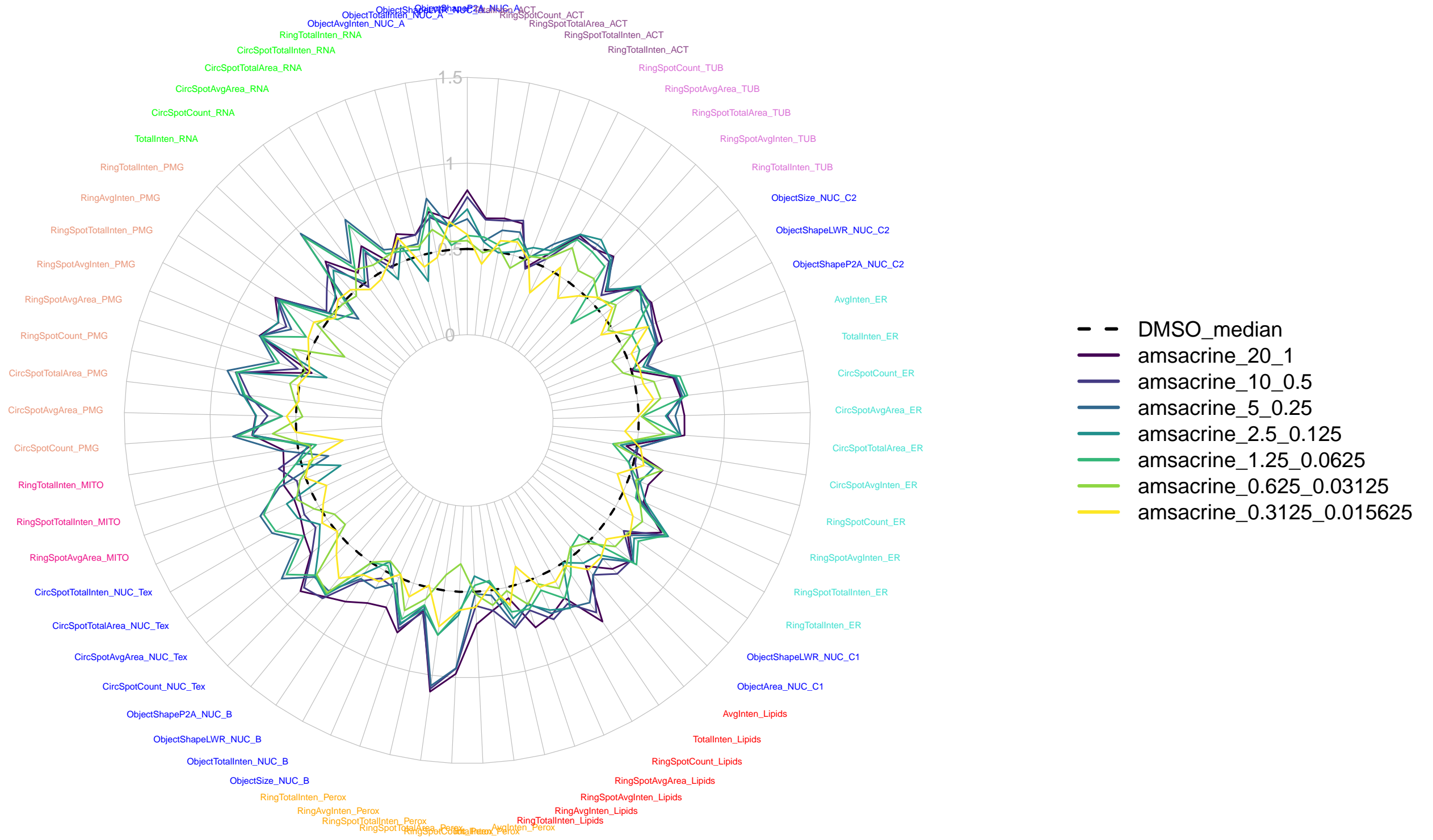
Compound: oligomycin-a



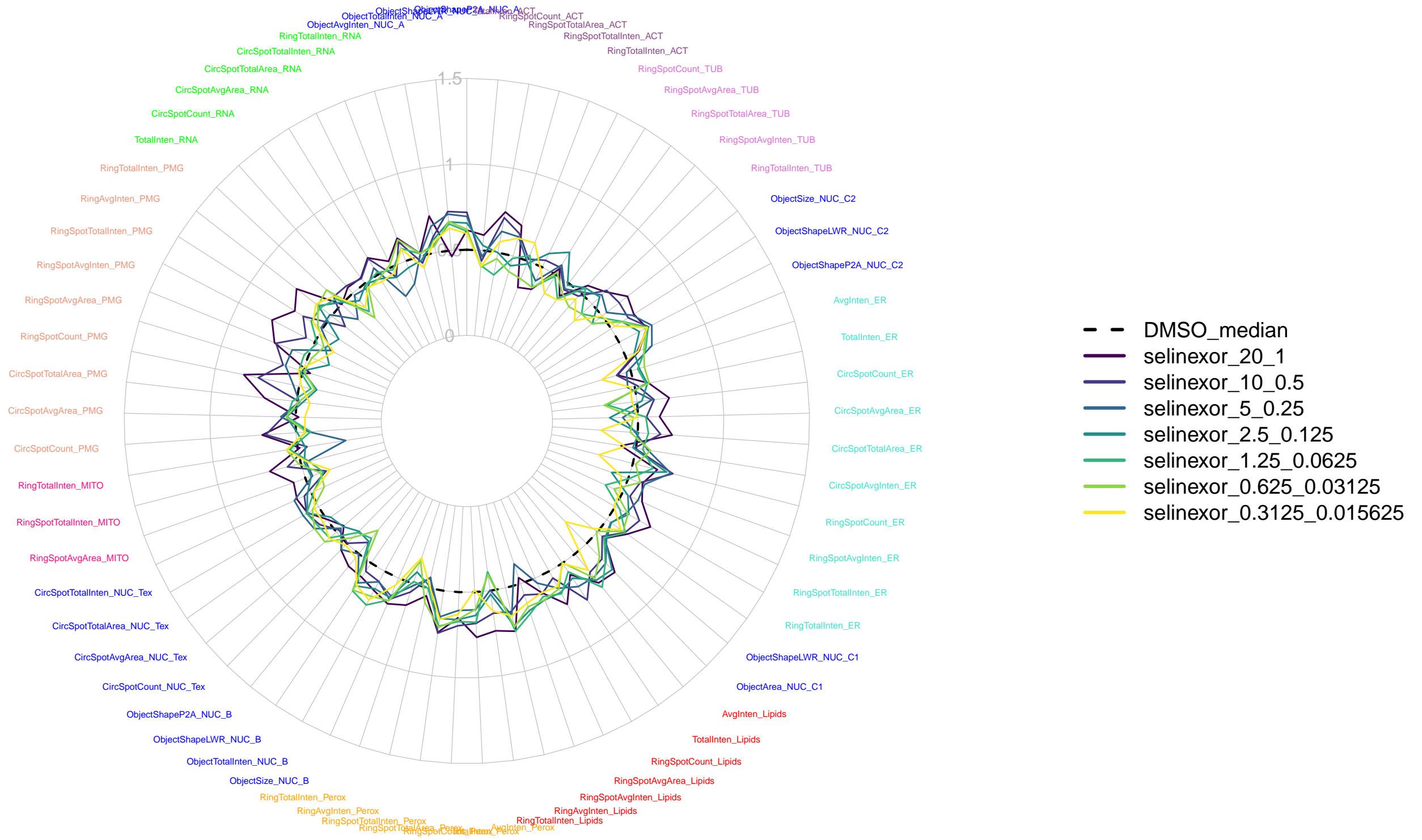
Compound: mdivi-1



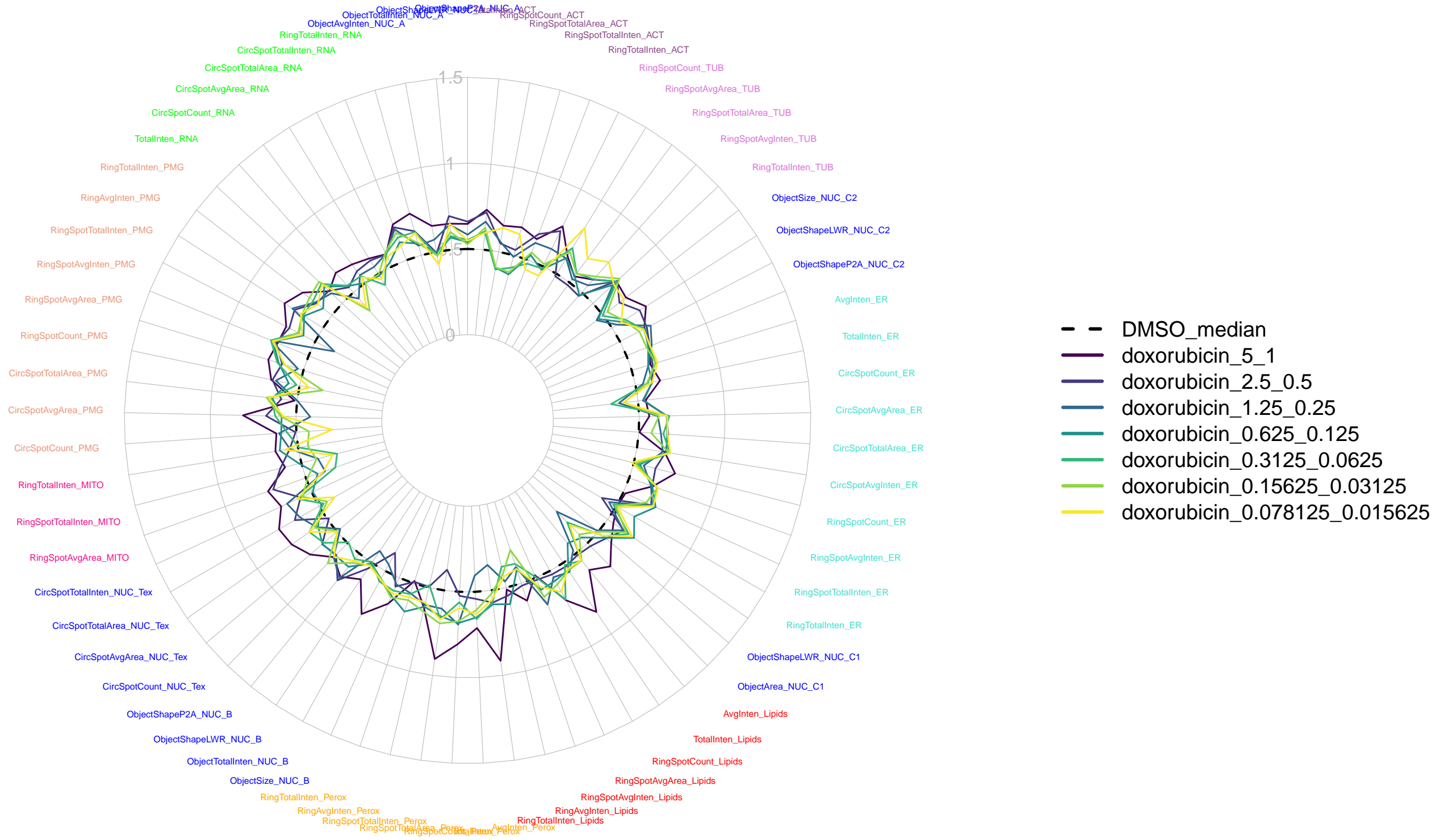
Compound: amsacrine



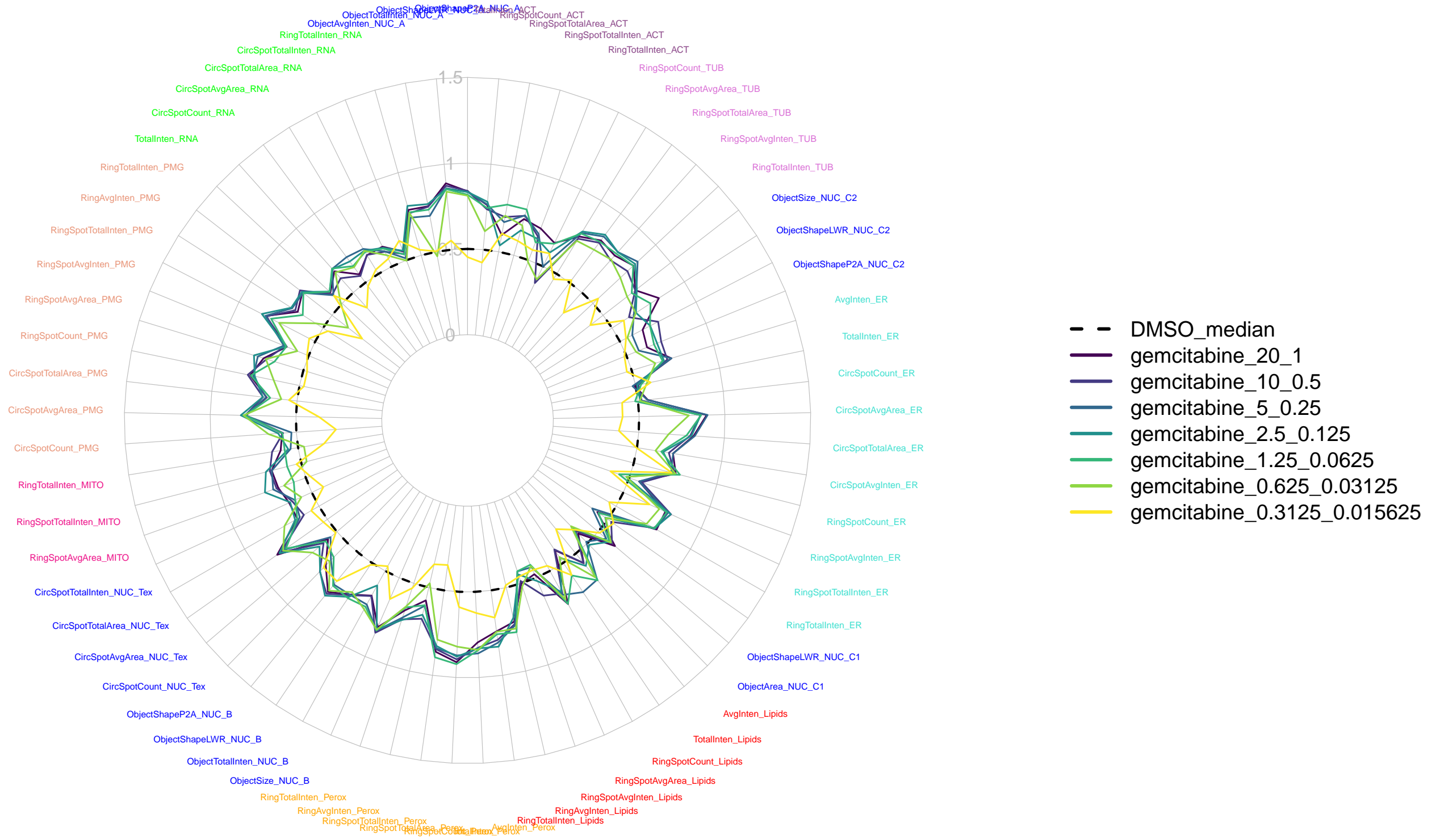
Compound: selinexor



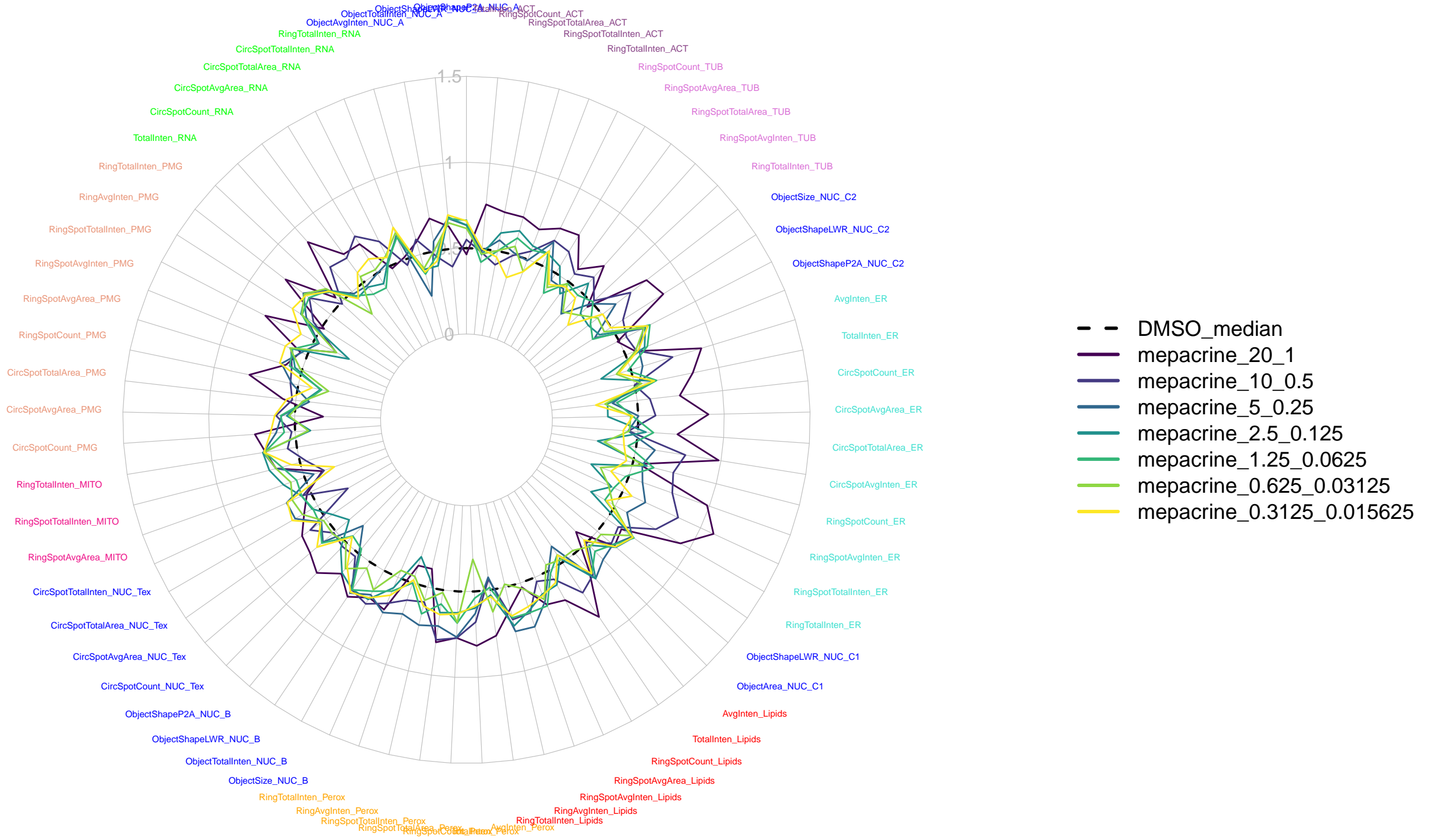
Compound: doxorubicin



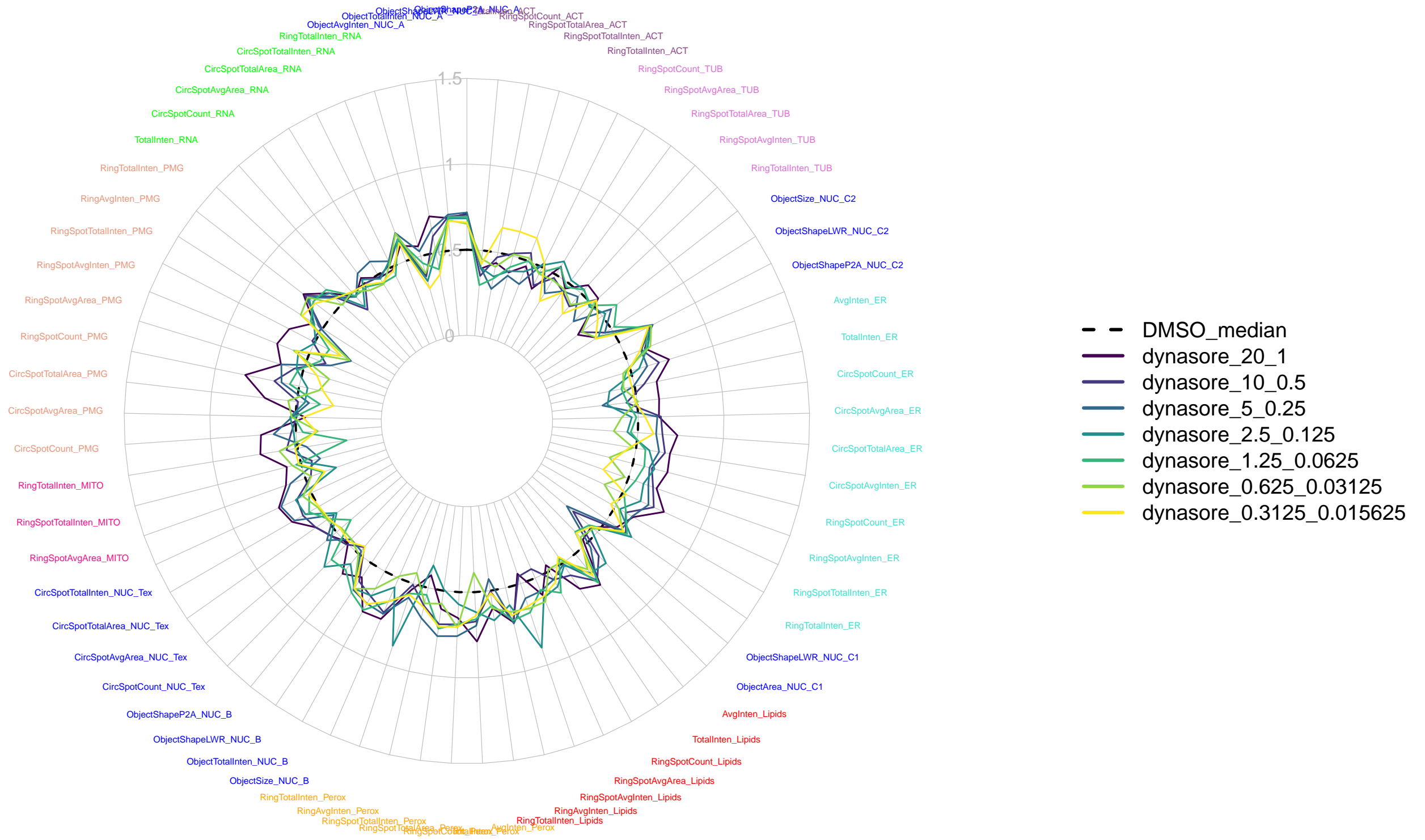
Compound: gemcitabine



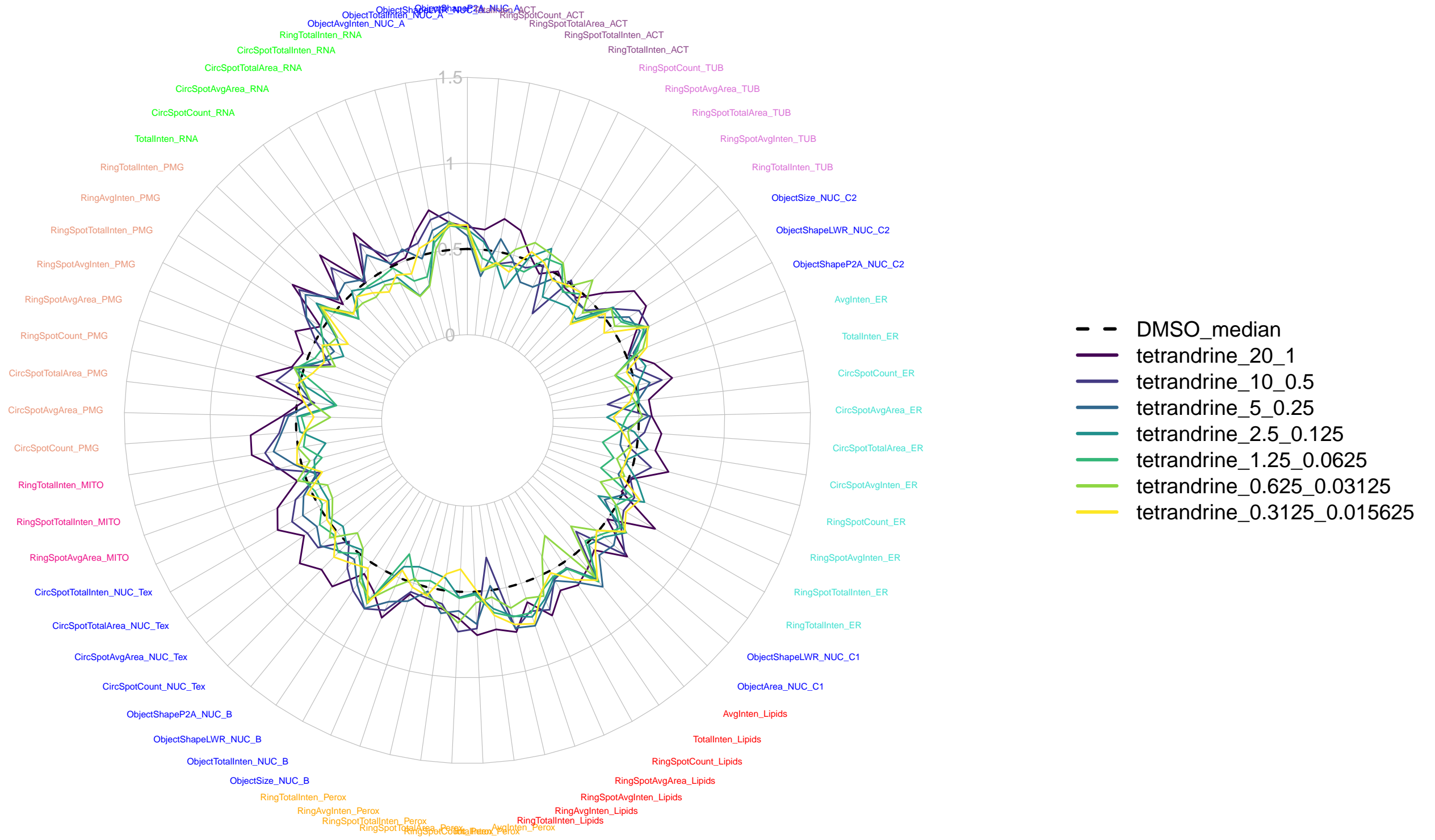
Compound: mepacrine



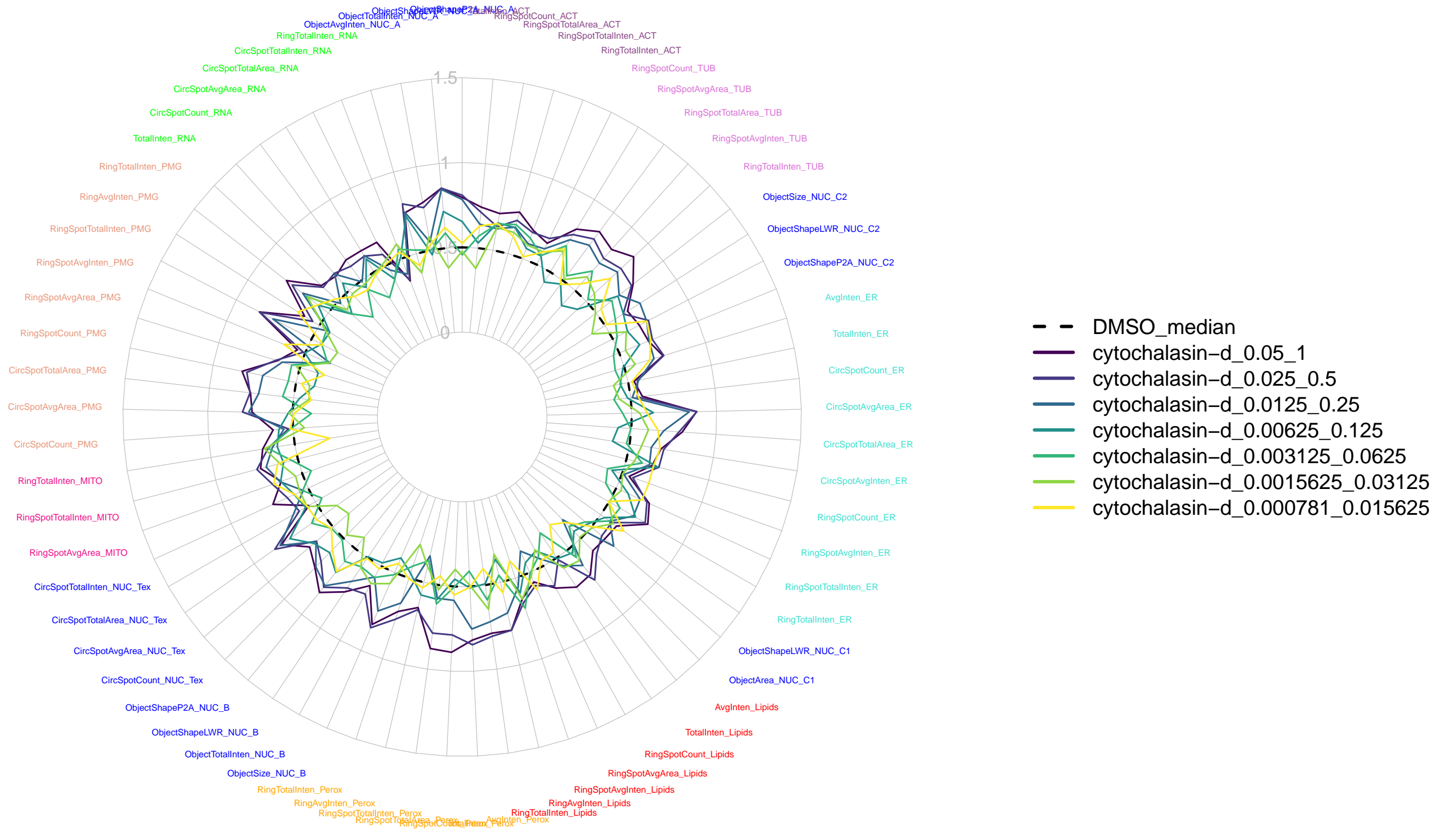
Compound: dynasore



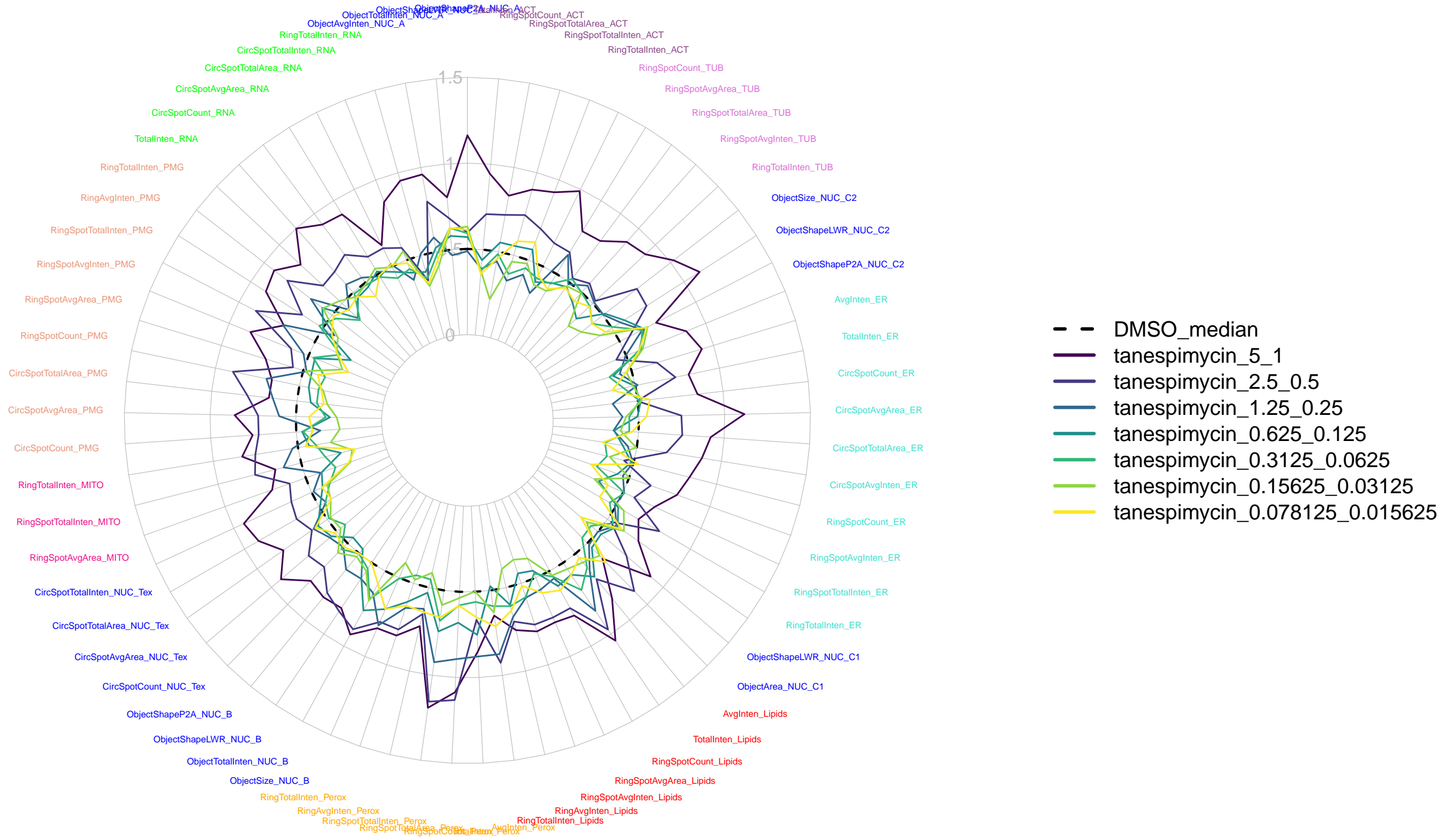
Compound: tetrandrine



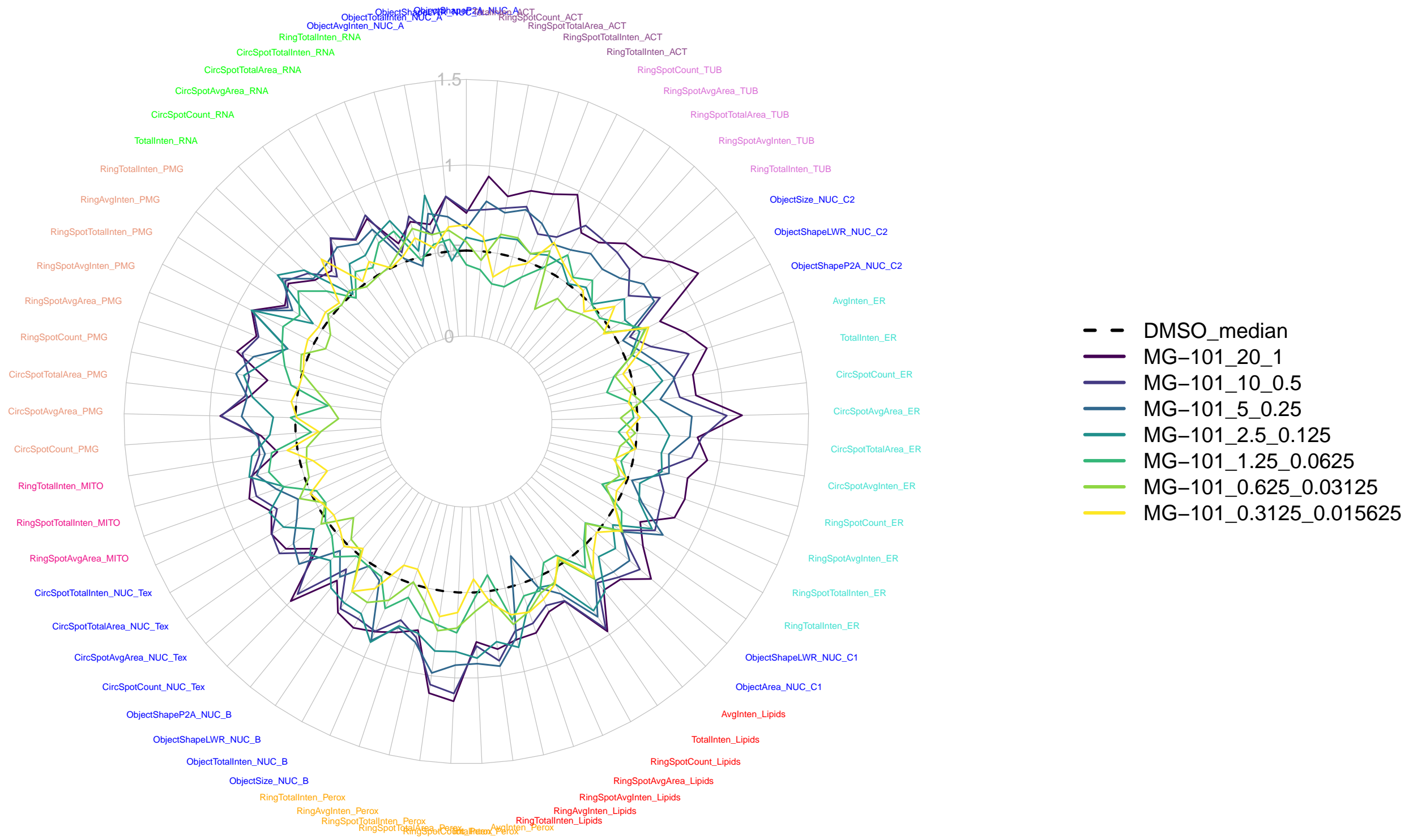
Compound: cytochalasin-d



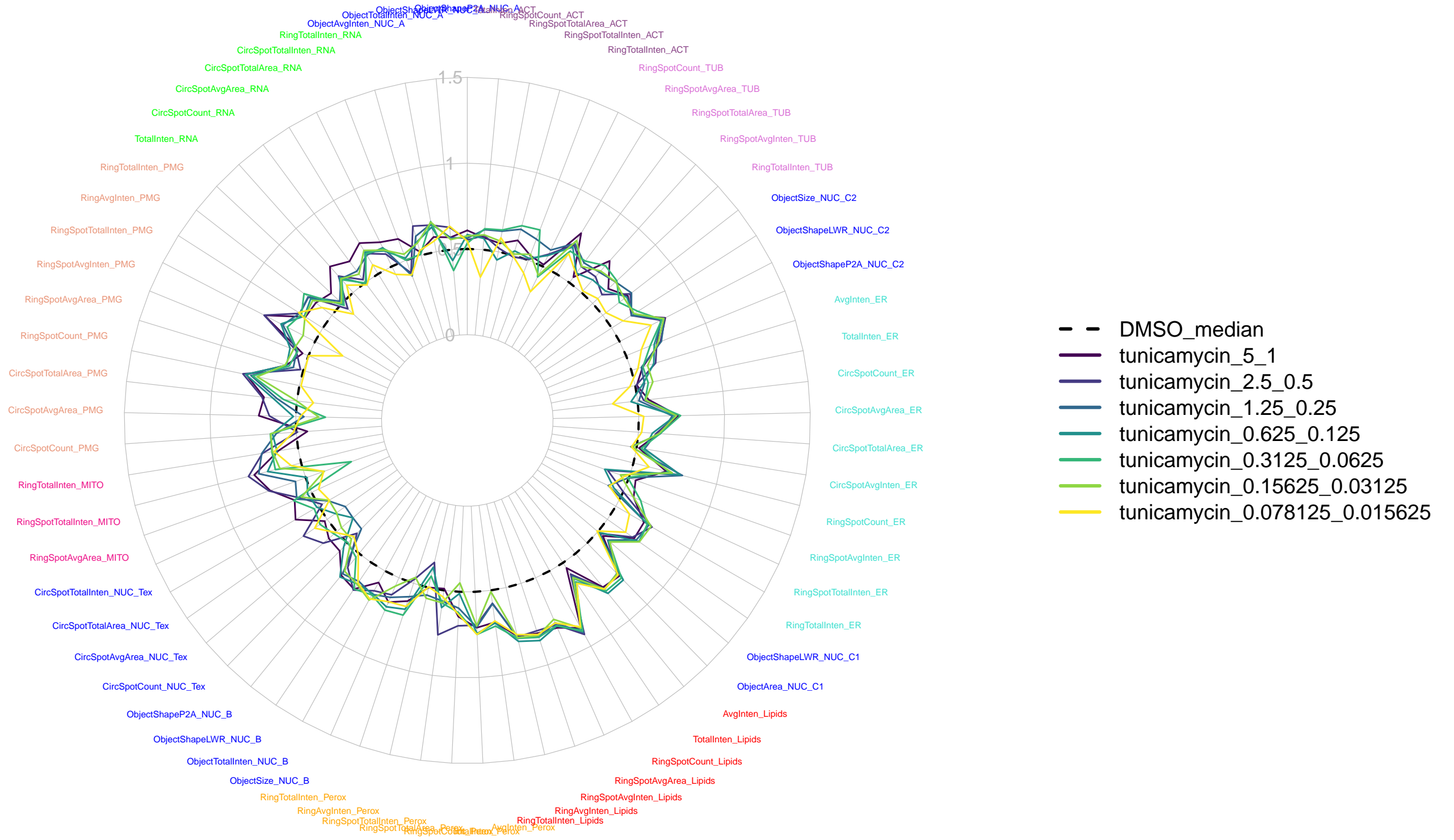
Compound: tanespimycin



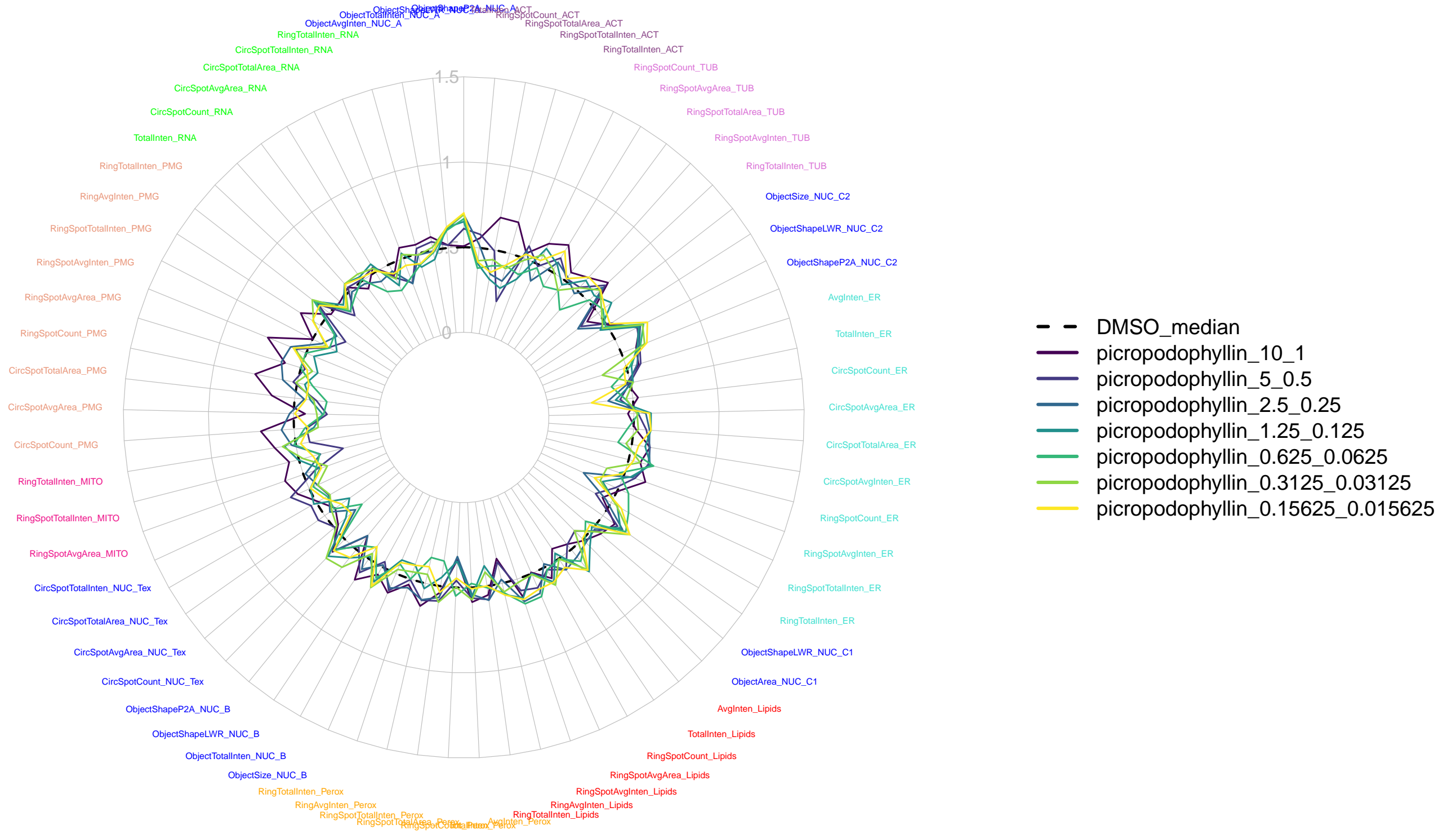
Compound: MG-101



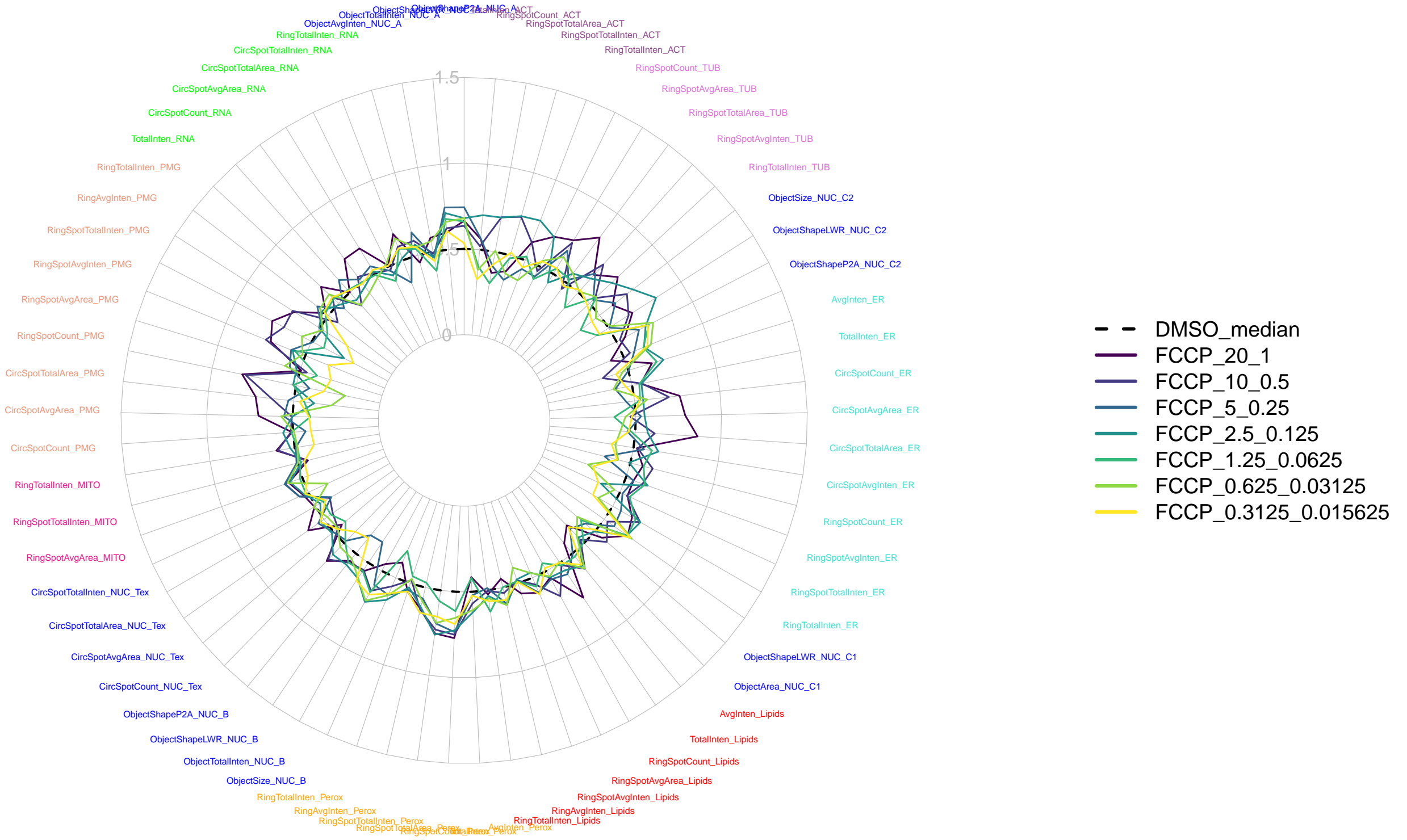
Compound: tunicamycin



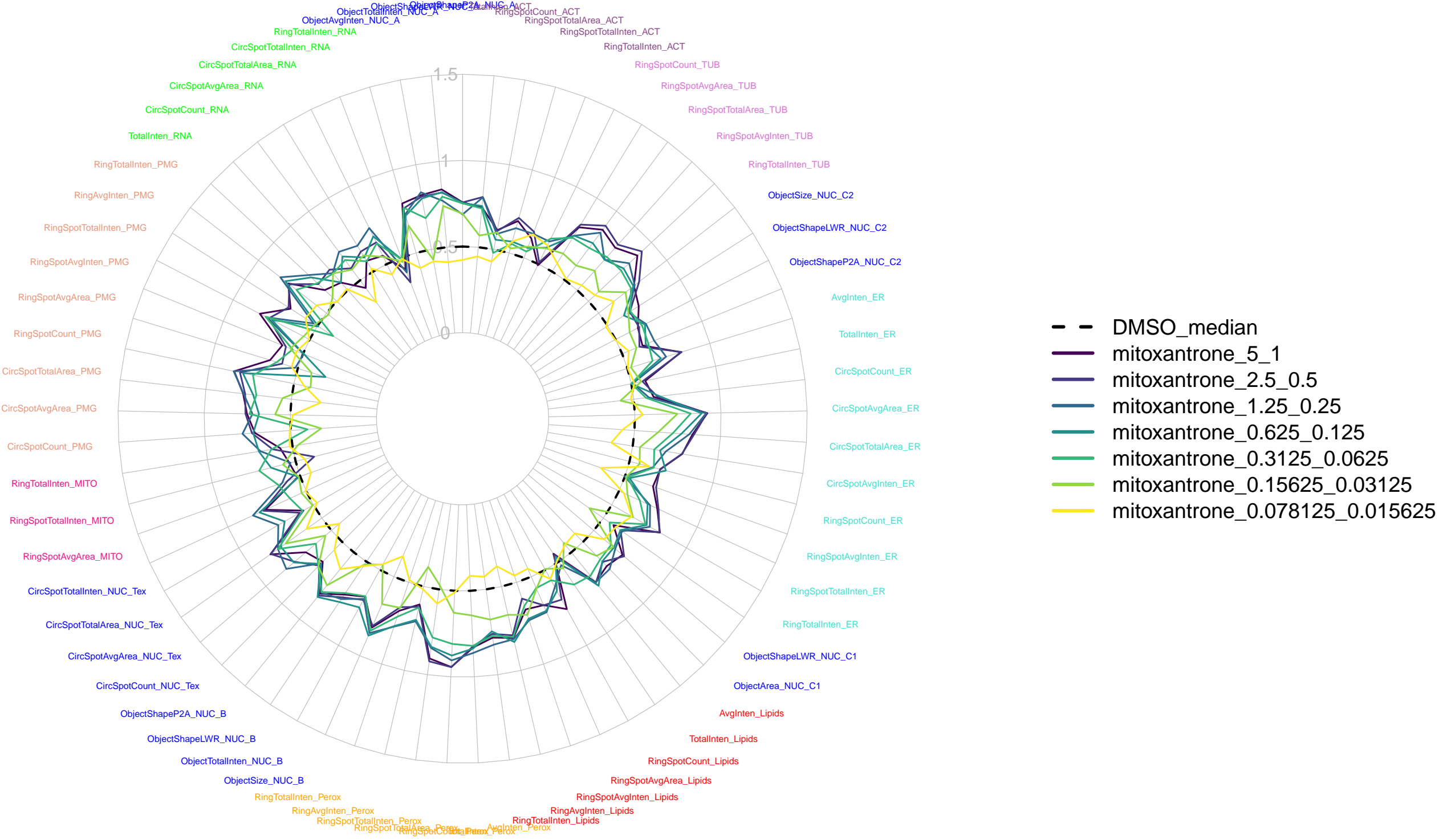
Compound: picropodophyllin



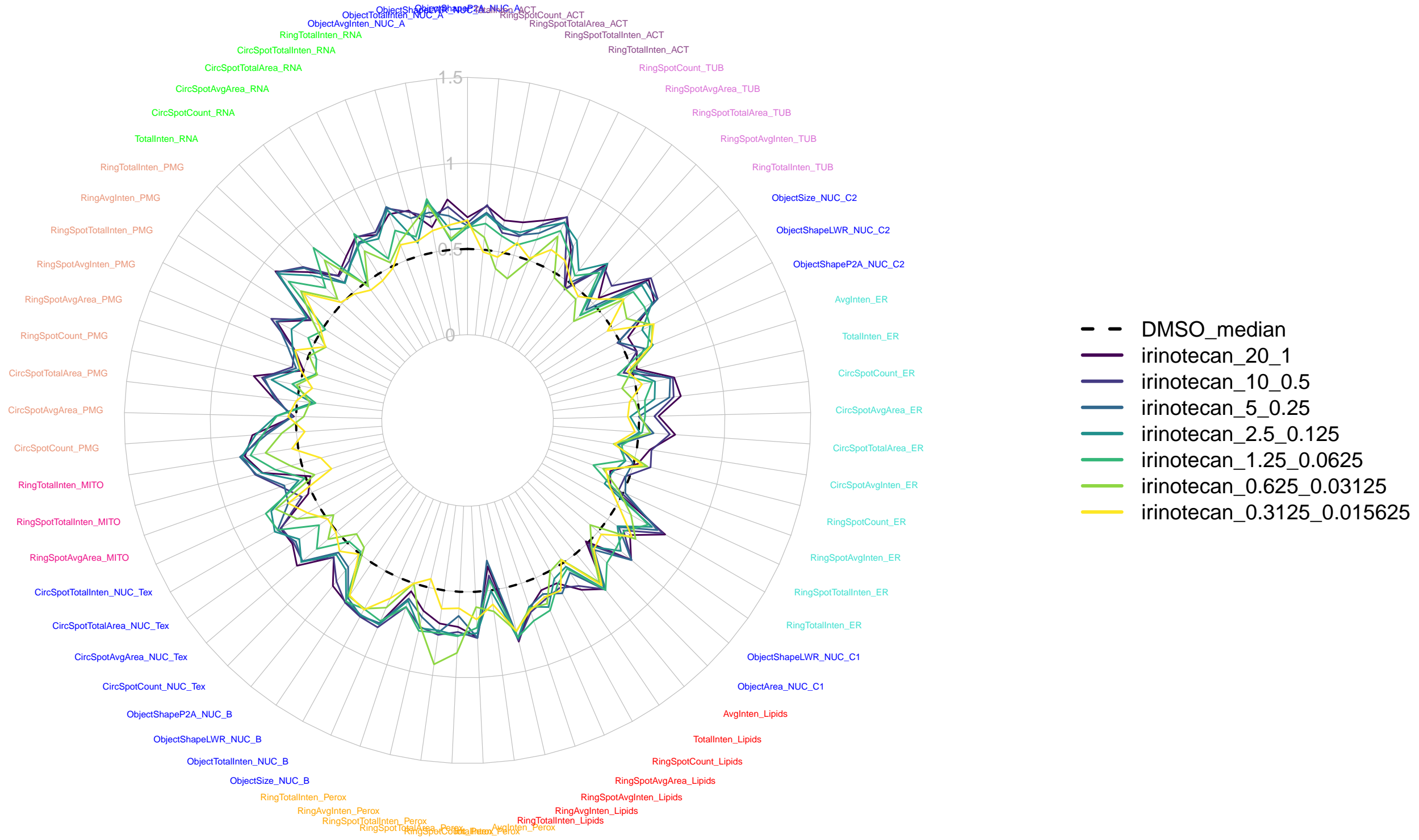
Compound: FCCP



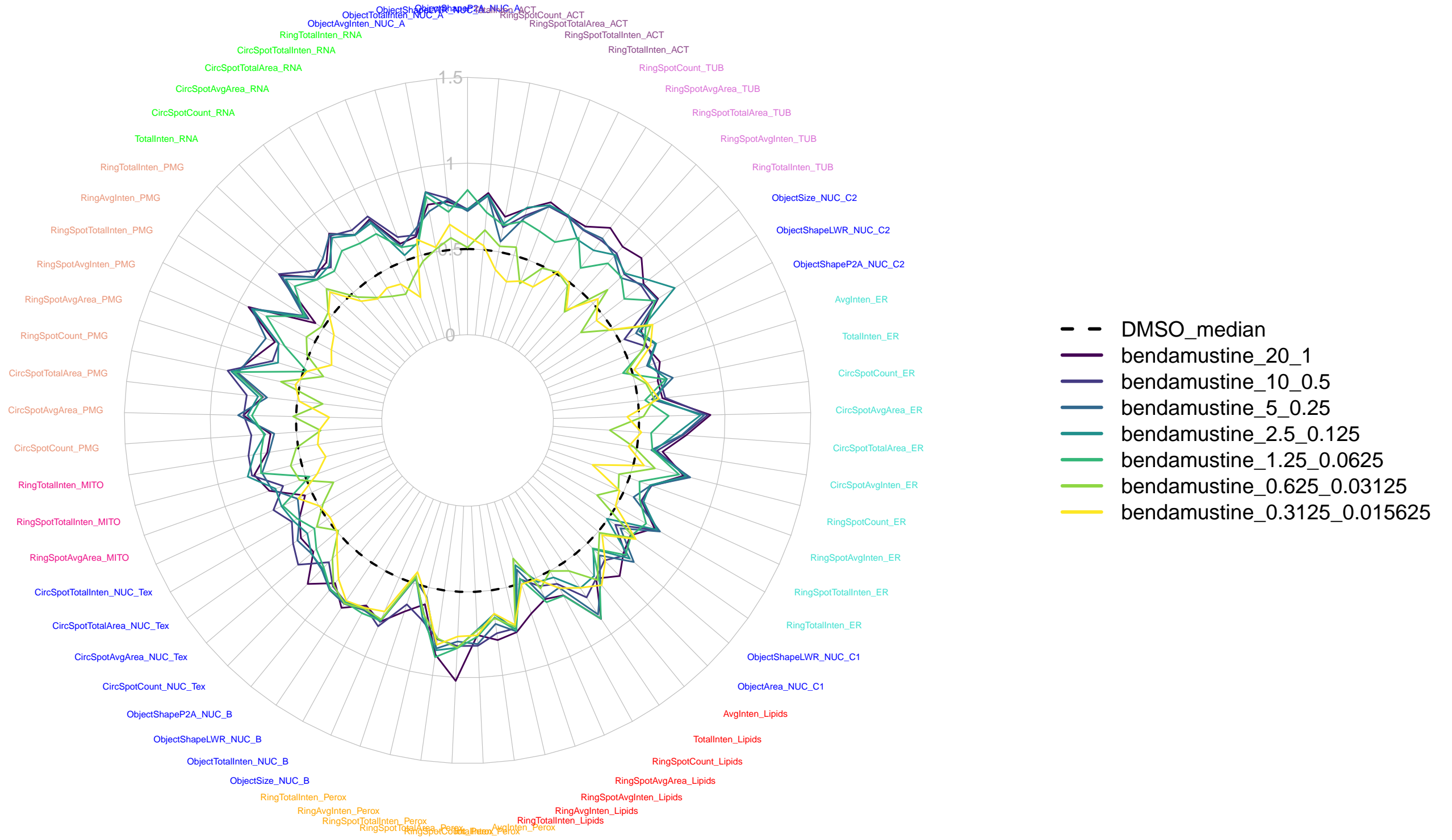
Compound: mitoxantrone



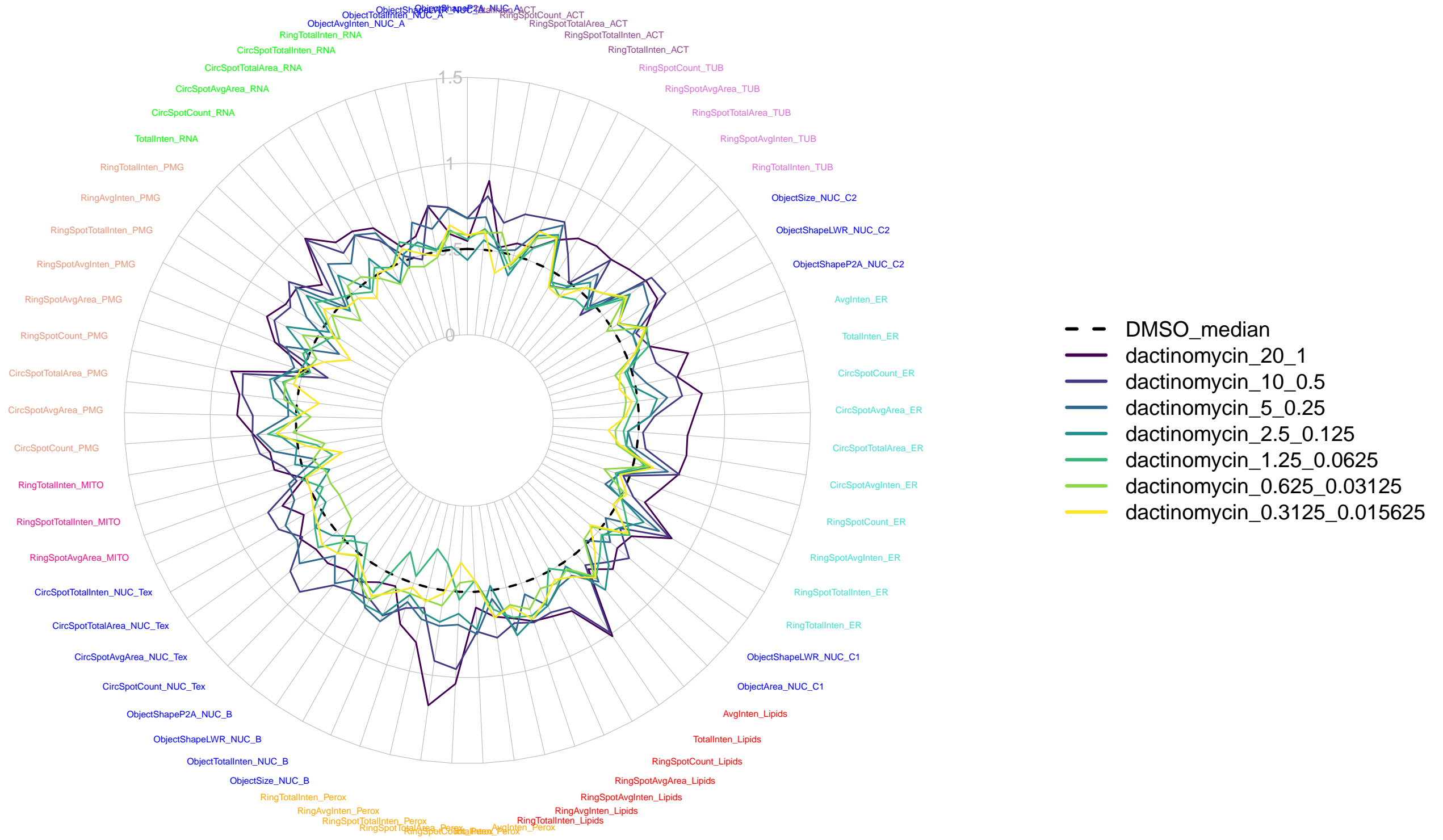
Compound: irinotecan



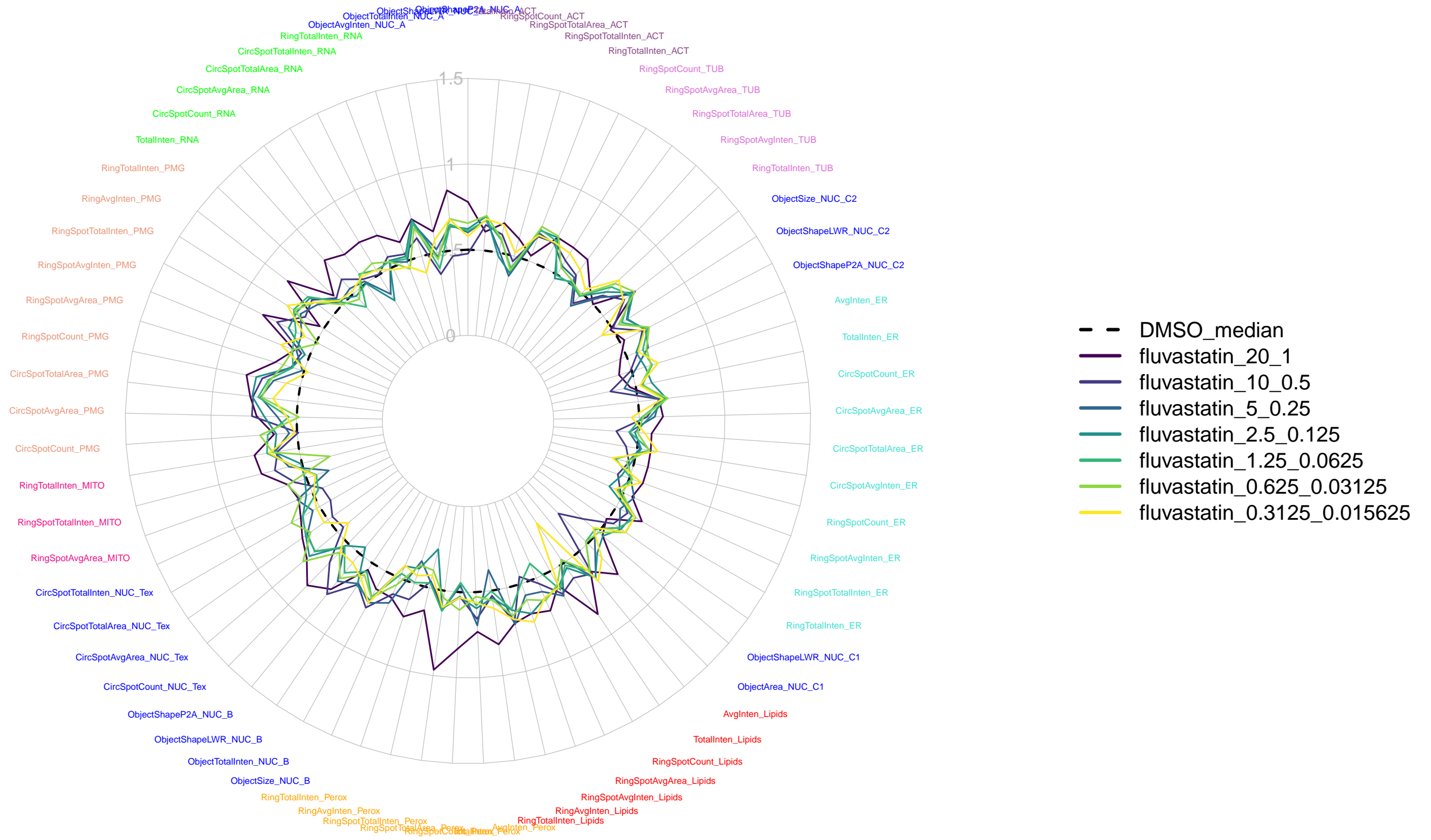
Compound: bendamustine



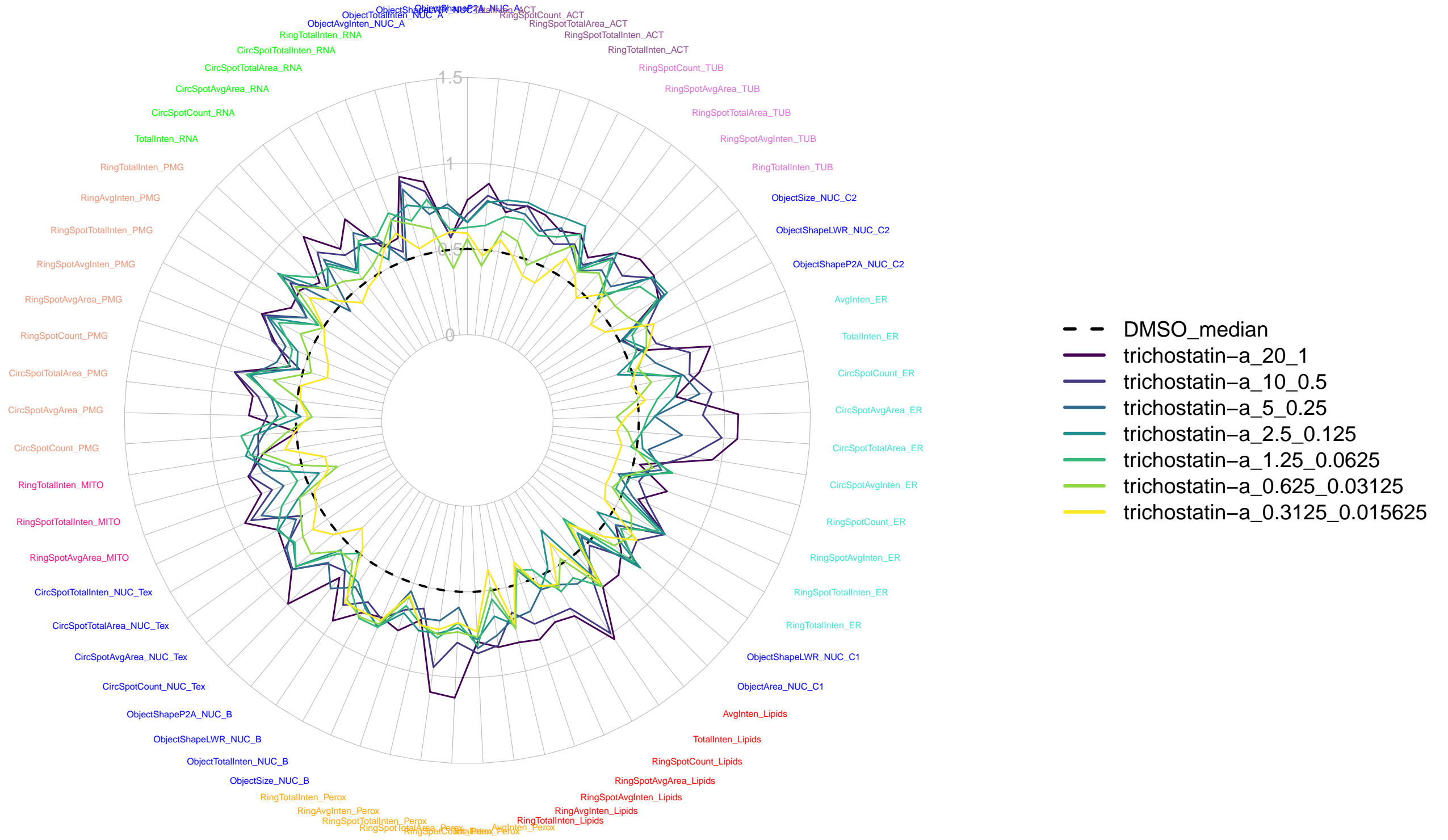
Compound: dactinomycin



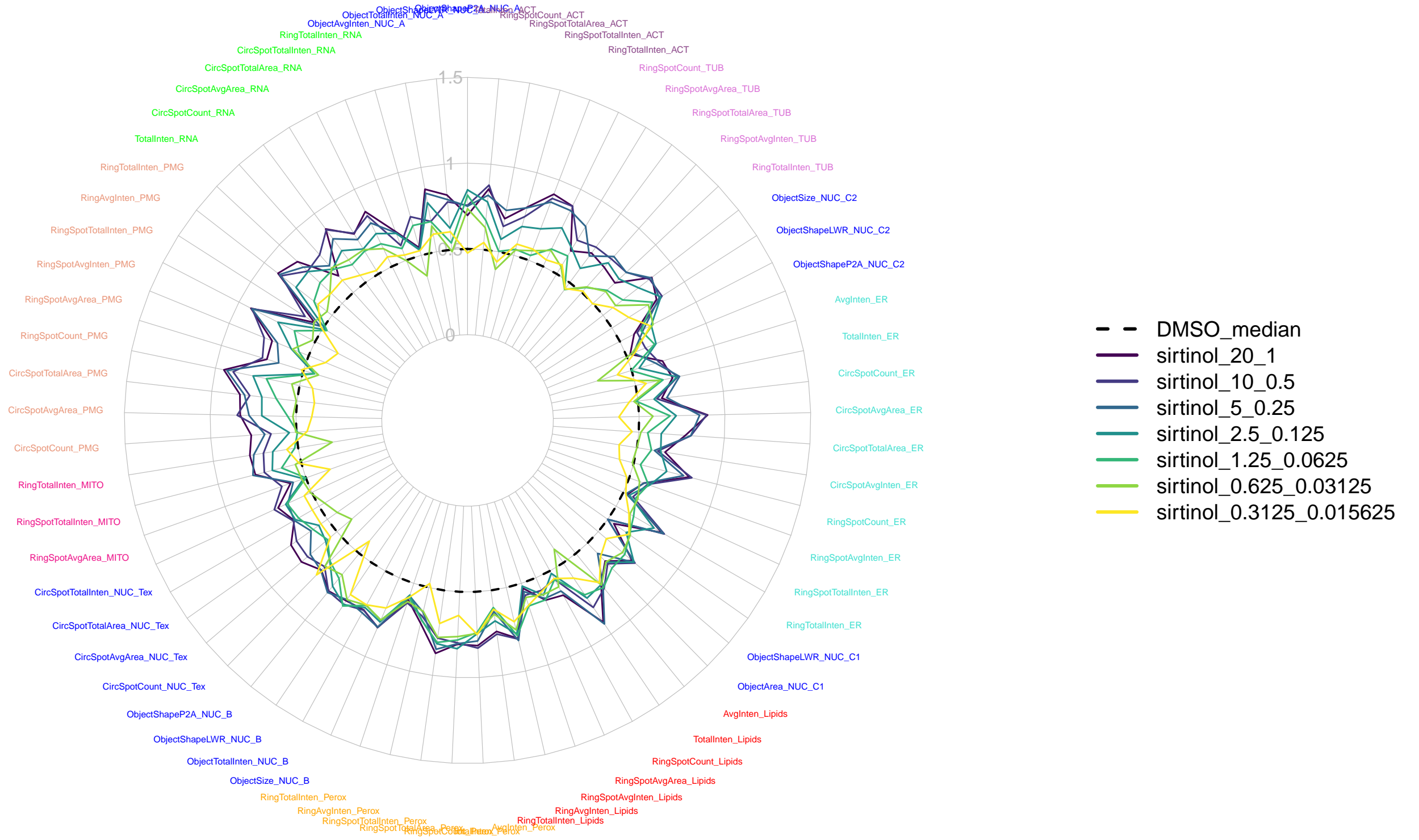
Compound: fluvastatin



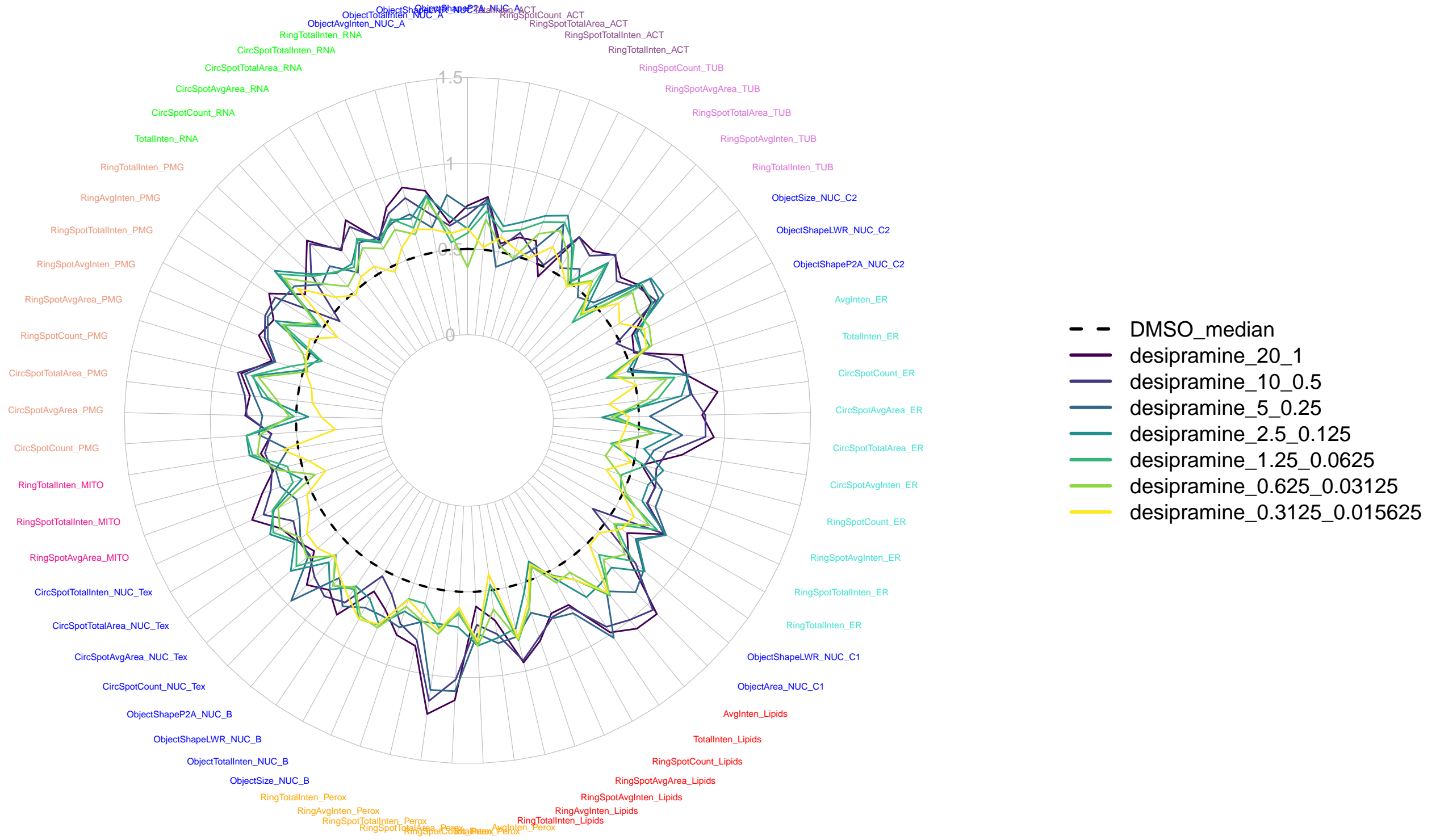
Compound: trichostatin-a



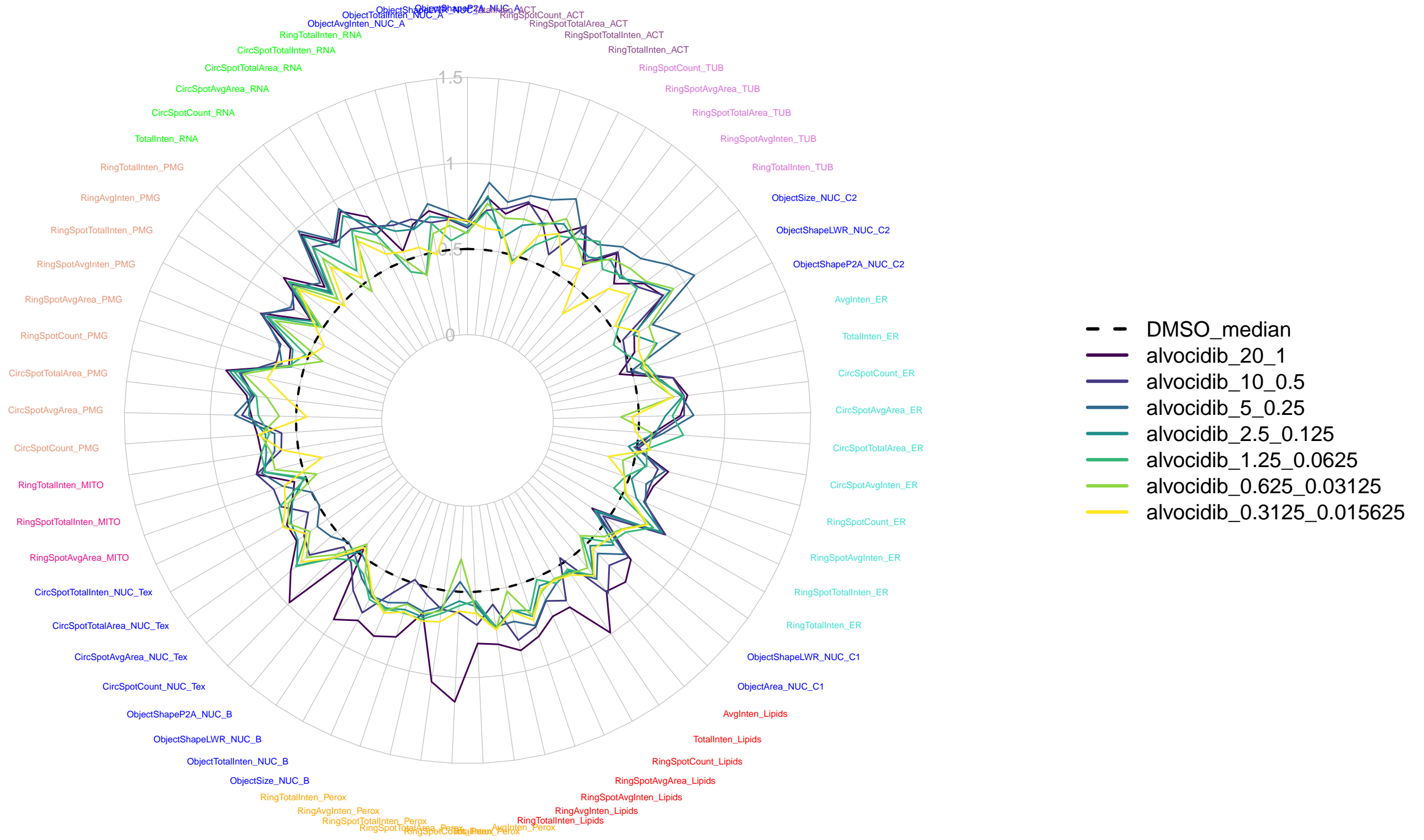
Compound: sirtinol



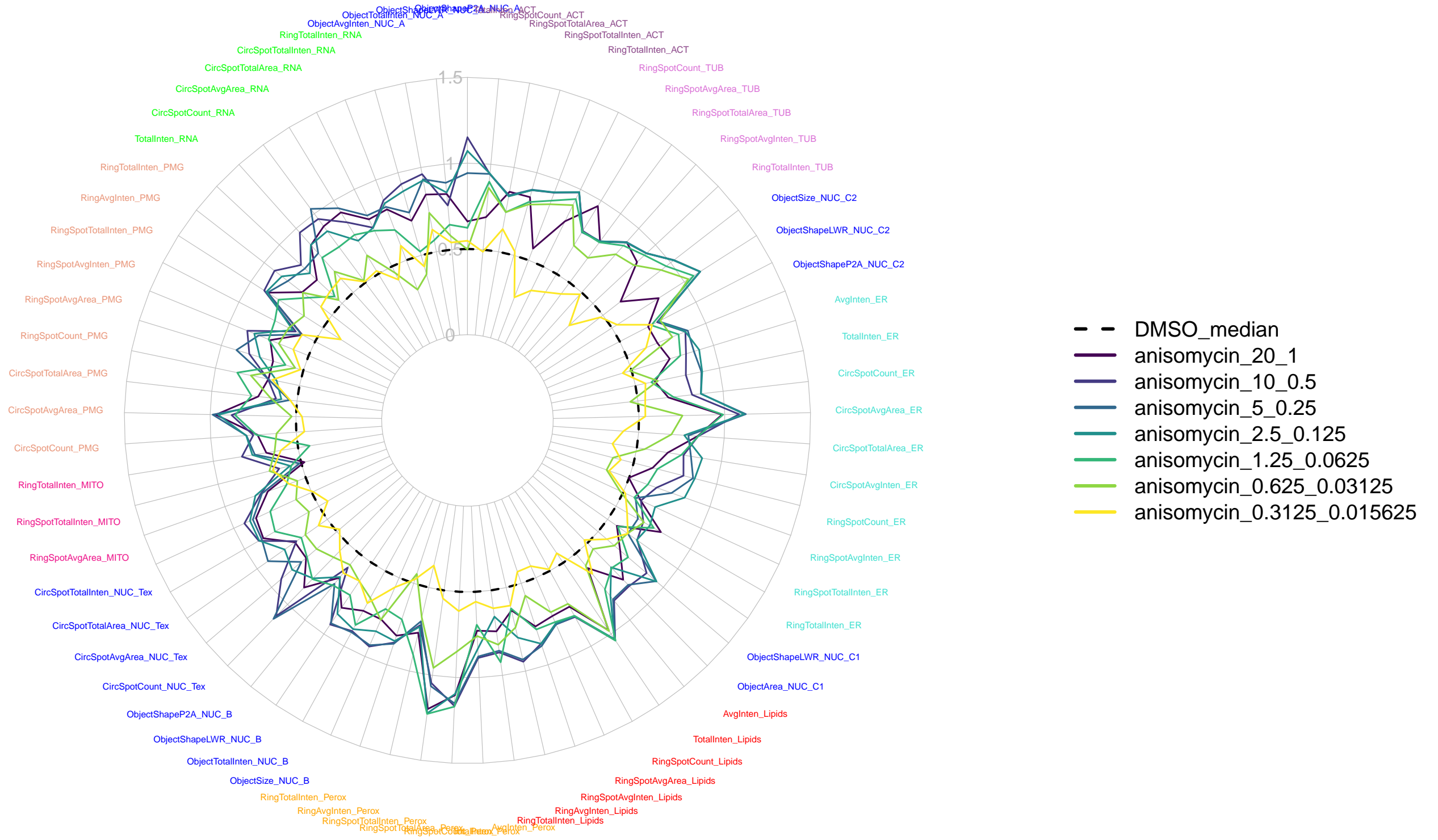
Compound: desipramine



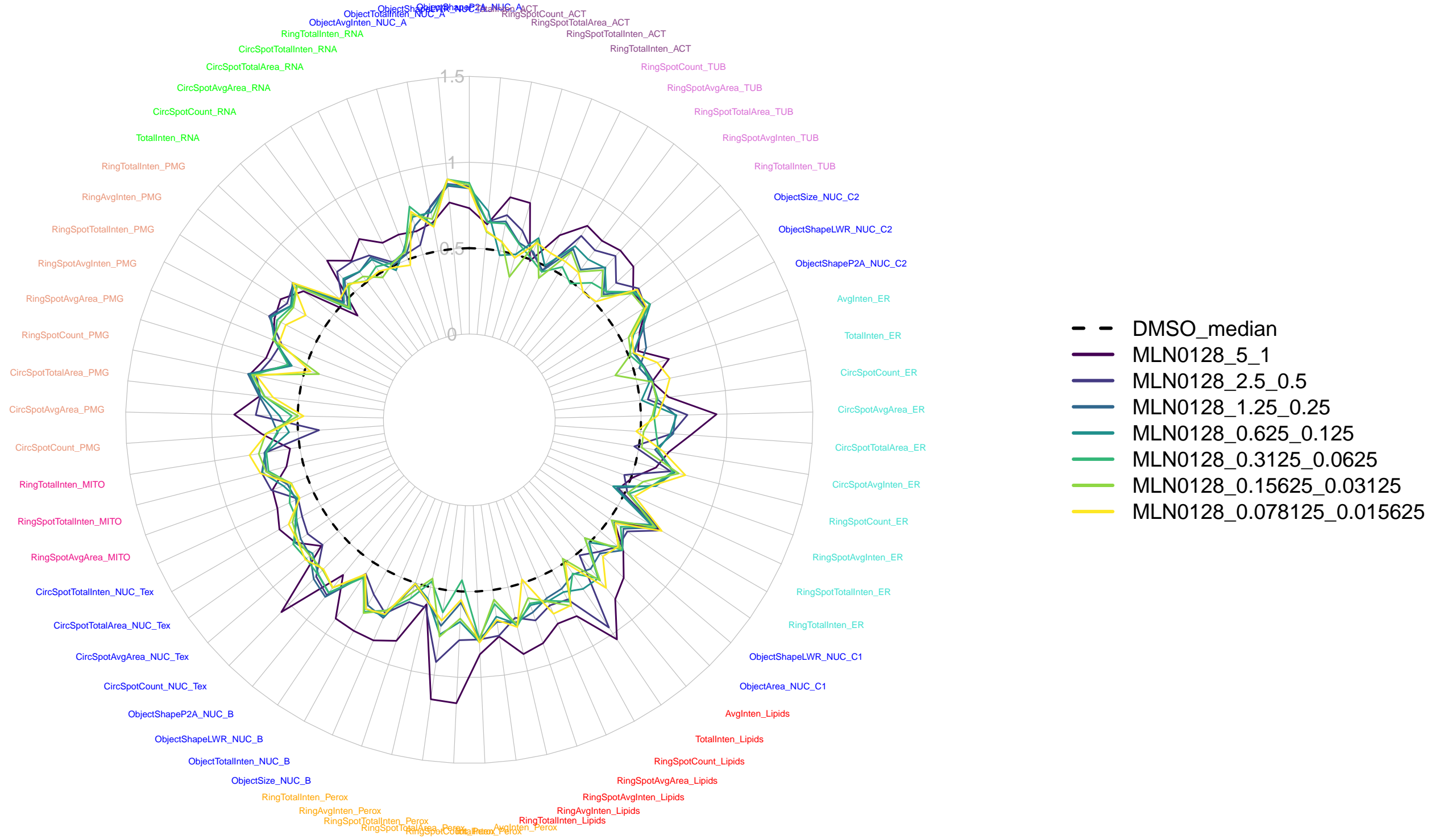
Compound: alvocidib



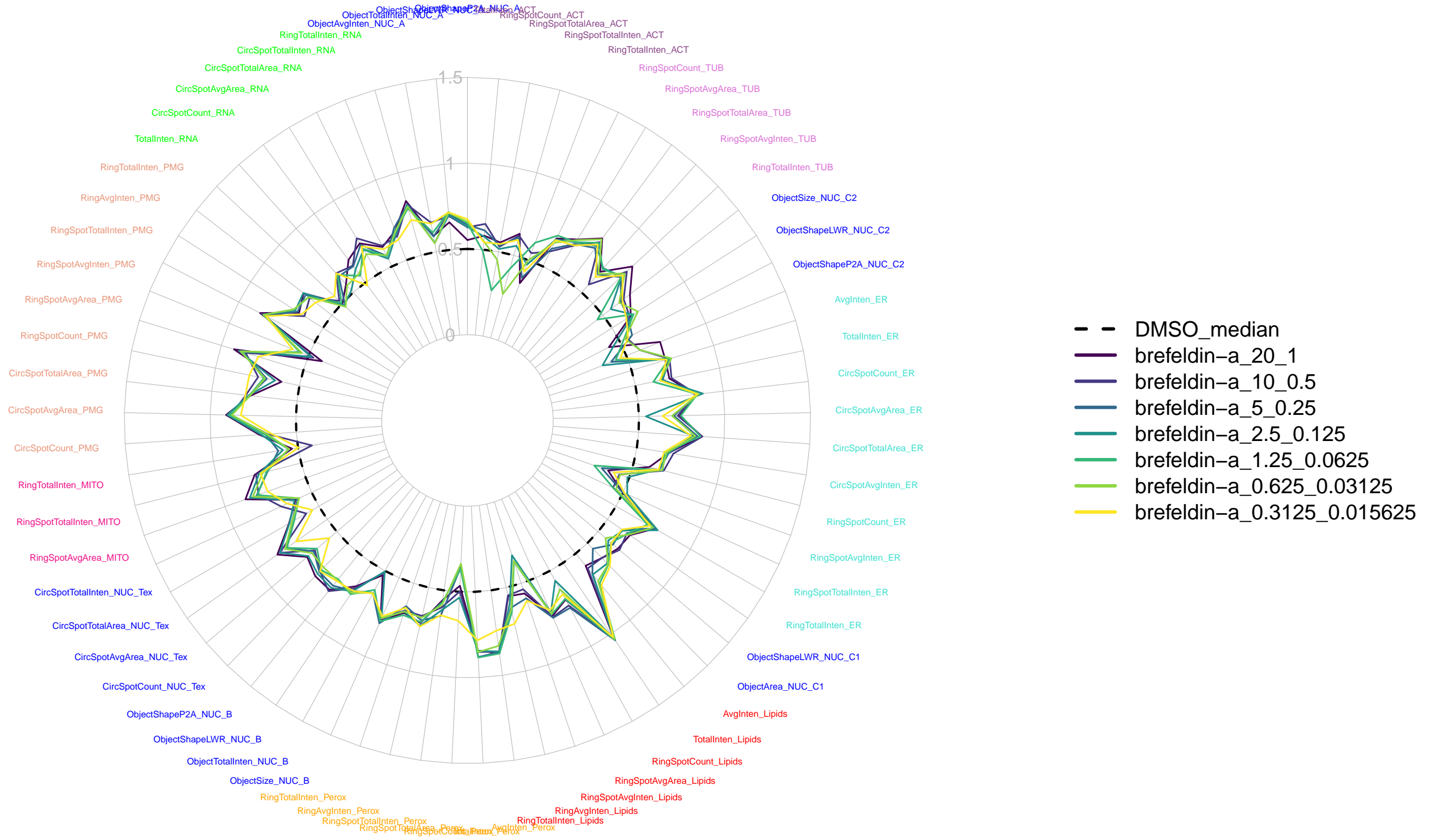
Compound: anisomycin



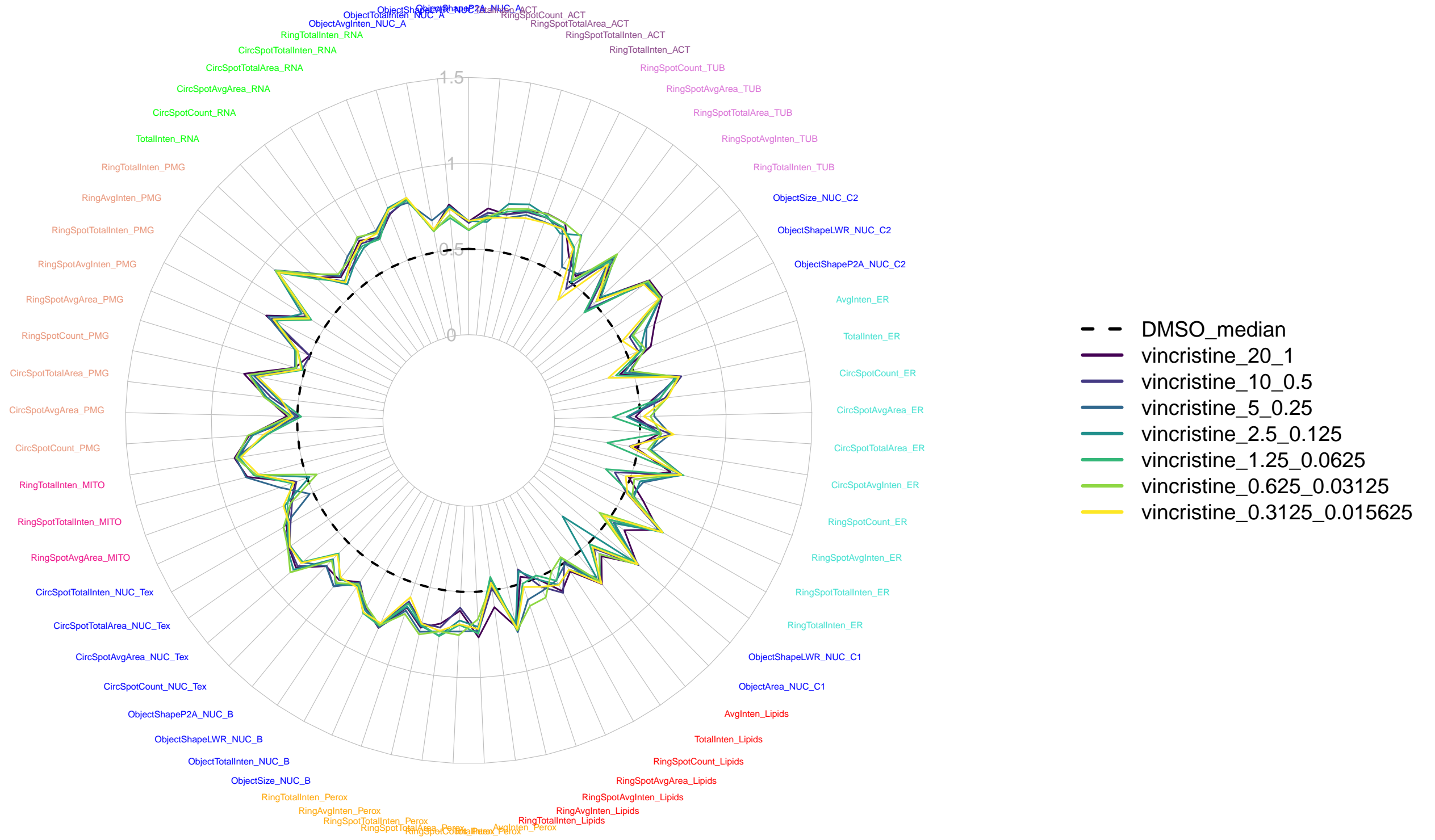
Compound: MLN0128



Compound: brefeldin-a



Compound: vincristine



Compound: latrunculin-a

