**Normalized Global / Phosphoproteomics Data for Clustering via PCA**

The datasets analyzed in this study included both **global proteomics** and **phosphoproteomics** measurements collected from two experimental cohorts of cNF (cutaneous neurofibroma) organoid samples. Each cohort comprised multiple tumor specimens derived from different patients.

For each dataset (global proteins and phosphosites), raw intensity values were preprocessed to ensure comparability across cohorts:

* **Missing values and zeros:** Zero measurements were replaced with missing values (NA) to prevent skew during normalization.
* **Feature filtering:** Proteins or phosphosites absent in more than 50% of samples were removed to reduce noise and improve robustness.
* **Normalization:** Remaining values were log2-transformed with a small offset (to prevent log(0)/undefined) and scaled using a modified z-score (median centering and median absolute deviation scaling). This approach places samples on a common scale while down-weighting outliers.
* **Cohort separation:** Each cohort was processed independently to account for technical differences, then merged into a single combined dataset.
* **Batch correction:** To address systematic differences between cohorts, ComBat (sva R package) was used, producing batch-adjusted abundance matrices for both global proteomics and phosphoproteomics.

The resulting normalized and corrected data provided a harmonized view of protein and phosphosite abundance across all patients and cohorts, enabling joint exploratory analysis such as PCA (Figures: phosphoCorrectedPCA.pdf and globalCorrectedPCA.pdf).

Figure: **phosphoCorrectedPCA.pdf**

This figure shows the results of a principal component analysis (PCA) performed on the phosphoproteomics data after normalization and batch correction using ComBat. Prior to correction, samples clustered primarily by cohort, reflecting a strong batch effect. After adjustment, the major sources of variation correspond more closely to biological factors, such as patient identity and tumor replicate. Each point represents an individual sample, colored by patient and shaped by tumor designation, enabling visualization of patient-specific clustering patterns. The correction substantially reduced separation by cohort, suggesting that technical batch effects were effectively mitigated and that the remaining signal more likely reflects underlying biological differences in phosphoproteome profiles.

Figure: **globalCorrectedPCA.pdf**

This figure presents a PCA of the global proteomics dataset following the same normalization and ComBat batch correction procedure. Similar to the phosphoproteomics analysis, initial clustering was dominated by cohort effects; however, correction aligned the data so that patient identity, rather than cohort, explained the primary axes of variation. Each sample is again colored by patient and shaped by tumor replicate, allowing assessment of within-patient reproducibility and between-patient divergence. The improved alignment indicates that batch effects were successfully minimized, and the corrected data provide a more reliable basis for downstream analyses of patient- and tumor-specific proteomic signatures.

**Biomarker Evaluation using Batch-Corrected Global / Phosphoproteomics Data**

For this biomarker evaluation analysis, we brought together four different data types collected on the same patient-derived samples:

1. **Drug sensitivity measurements** – single-dose viability values (and some full dose–response curves) across a large panel of compounds.
2. **RNA sequencing** – transcript abundance data (to be harmonized into long format in subsequent steps).
3. **Global proteomics** – normalized protein abundance values, batch-corrected across cohorts.
4. **Phosphoproteomics** – normalized and batch-corrected site-specific phosphorylation abundances.

To enable cross-modality analyses, all datasets were filtered to retain only those specimens present in both the drug response data and at least one molecular modality. This produced a shared set of samples suitable for correlation and biomarker discovery analyses.

* **Proteomics:** Both global and phosphoproteomic data were restructured into sample-by-feature matrices, where rows represent specimens and columns represent either proteins (global) or phosphosites (phospho). Abundance values correspond to batch-corrected, log-scaled measurements.
* **Drug sensitivity:** Viability data (expressed in relative units of surviving fraction at given doses) were reformatted into a specimen-by-drug matrix, with average values across replicates used where multiple measurements existed.
* **Filtering:** Only drugs with complete measurements across all available specimens were retained for certain analyses, yielding a consistent comparison set (“full drugs”).

Drug Count: 238

Drug list:

Abemaciclib, ABT-737, Adagrasib, Adavosertib, Afatinib, Alectinib, Alisertib, Alpelisib, Alvespimycin, Alvocidib, Anlotinib, Apitolisib, ARS-1620, Avapritinib, AVL-292, Avutometinib, Axitinib, Barasertib, Batoprotafib, Belinostat, Belzutifan, BI-3406, BI-847325, BI-D1870, Bimiralisib, Binimetinib, Birinapant, BLU9931, BMS-536924, Bosutinib, Brigatinib, Brivanib, Brukinsa, Brustaol, Cabozantinib, Calquence, Capivasertib, Capmatinib, CBL0137, CCT-018159, Cedazuridine, Ceralasertib, Cerdulatinib, Ceritinib, CFT8634, Cobimetinib, Copanlisib, CPI-613, Crenolanib, Crizotinib, Dabrafenib, Dacomitinib, Danusertib, Daporinad, Daprodustat, Dasatinib, Defactinib, Derazantinib, Digoxin, Dinaciclib, Dovitinib, Doxorubicin, Duvelisib, Eganelisib, Elimusertib, Enasidenib, Encorafenib, Enitociclib, Ensartinib, Entinostat, Entrectinib, Enzastaurin, Everolimus, Fadraciclib, Famotidine, Fedratinib, Fimepinostat, Fisogatinib, FRAX597, Futibatinib, Galunisertib, Ganetespib, Gefitinib, Geldanamycin, Gilteritinib, Glasdegib, GNE-617, GSK2256098, H 89 2HCl, Ibrutinib, Idasanutlin, IKK-16, Imatinib, INCB188053, INCB191856, Infigratinib, Ipatasertib, Ivosidenib, JNJ-7706621, Ketotifen, KO-947, KW-2478, Lapatinib, Larotrectinib, Lenvatinib, Lestaurtinib, Letrozole, Lifirafenib, Linsitinib, Lorlatinib, Losartan, Losmapimod, Lovastatin, Luminespib, LY231514, LY3023414, Manumycin A, Merestinib, Metformin, Methotrexate, Midostaurin, Mirdametinib, MK-2206, ML 210, MLN2480, Molibresib, Napabucasin, Navitoclax, Nedisertib, Nedometinib, Neratinib, Nilotinib, Nintedanib, Nirogacestat, NT219, NVP-BEZ235, Octreotide, Olaparib, Olutasidenib, Omipalisib, Onalespib, Onametostat, Onvansertib, Osimertinib, Pacritinib, Palbociclib, Pan-RAS-IN-1, Panobinostat, Pazopanib, PCNA-I1, Pelabresib, Pemigatinib, Pemrametostat, PF-562271, Pimozide, Pirtobrutinib, PLX-3397, Ponatinib, Pralsetinib, Prexasertib, PT2385, Quizartinib, Ranitidine, Rapamycin, Ravoxertinib, Regorafenib, Repotrectinib, Retaspimycin, Ribociclib, Ripretinib, RMC-6236, RO4929097, Rogaratinib, Romidepsin, Roxadustat, Rucaparib, Ruxolitinib, Sapanisertib, SAR405838, Seclidemstat, Selinexor, Selpercatinib, Selumetinib, Sertraline, Siremadlin, Sitravatinib, SNX-2112, SNX-5422, Sonidegib, Sorafenib, Sotatercept, Sotorasib, Subasumstat, Sulfasalazine, Sunitinib, Surufatinib, SW106065, Tadalafil, TAE226, TAK-243, Tanespimycin, Tazemetostat, TED-347, Tegavivint, Telaglenastat, Temozolomide, Temsirolimus, Tepotinib, THZ1, Ticlopidine, Tipifarnib, Tivantinib, Tivozanib, TK216, Tofacitinib, Tomivosertib, Tovorafenib, Tozasertib, Trametinib, Tretinoin, Trilaciclib, Triptolide, Ulixertinib, Umbralisib, UNC2025, Unesbulin, Vactosertib, Vandetanib, Vemurafenib, Venetoclax, Vismodegib, Vistusertib, Vociprotafib, Volasertib, Vorinostat, VS-6766, Y-27632, Zanzalintinib

**Exploratory Analysis of Drug Responses**

We performed initial exploration of the drug dataset to understand variability and efficacy across the compound panel.

**Correlation of molecular features with drug response**

* Spearman correlations were computed between every drug (response profile) and every molecular feature (protein or phosphosite).
* Significance values were estimated via permutation-based correlation tests, with multiple testing correction applied (FDR).
* Features were separated into **positive correlations** (higher abundance/phosphorylation associated with higher viability, i.e. resistance) and **negative correlations** (higher abundance/phosphorylation associated with lower viability, i.e. sensitivity).

**Figure: most\_efficacious.png**

This scatterplot highlights the subset of drugs that were most efficacious across the patient samples, defined as compounds with an average cell viability below 0.5 (i.e., less than 50% survival). Each point represents a drug, positioned by its mean viability (y-axis) and labeled along the x-axis. Point size reflects the variability in response across specimens, while point color indicates the number of samples tested. Several compounds demonstrate both strong overall activity and consistent performance across patients, suggesting broad-spectrum effectiveness.

**Figure: most\_variable.png**

This scatterplot highlights drugs with the highest variability in response across specimens, defined as those with a standard deviation greater than 0.15. Here, each point again represents a drug, with mean viability on the y-axis and drug identity on the x-axis. Point size encodes variability, and color denotes the number of samples measured. Unlike the most efficacious plot, these compounds are not necessarily the most potent but instead show strong heterogeneity between patients. Such drugs may provide the greatest opportunity for biomarker discovery, as differences in response are more likely to be explained by underlying molecular features.

**Figure: drug\_heatmap\_large.pdf**

This heatmap displays drug response values for the subset of drugs measured consistently across all specimens (“full drugs”). Rows correspond to patient samples and columns to drugs. Hierarchical clustering of both rows and columns highlights patterns of similarity, revealing groups of patients with shared sensitivity profiles as well as clusters of compounds with correlated activity. The visualization provides a global overview of drug response heterogeneity across the cohort, serving as a baseline reference for linking molecular features to therapeutic sensitivity.

**Figure: cor\_features\_by\_drug.pdf**

Bar charts summarizing, for each drug, the number of molecular features (proteins or phosphosites) that correlate with response at FDR < 0.25, separated by direction (features associated with **resistance** vs **sensitivity**) and faceted by data modality. These counts provide a quick sense of which drugs show the richest correlational signal.

**Functional enrichment of correlated features**

* Using the leapR package, correlated features were ranked and tested for enrichment against gene set collections..
* Enrichment was conducted in a **direction-aware** manner:
  + “Top” (positively correlated features) highlight pathways enriched among **resistance-associated features**.
  + “Bottom” (negatively correlated features, flipped in rank) highlight pathways enriched among **sensitivity-associated features**.
* Results were summarized at multiple levels:
  + **Pathways per drug:** Number of significant pathways identified for each drug, separated by resistant vs sensitive associations.
  + **Pathways across drugs:** Top 15 most recurrent pathways across the full drug panel, faceted by proteome vs phosphoproteome.
  + **Drug-specific profiles:** Bar charts of the top 15 pathways per drug and modality, annotated with significance stars (\* <0.05, \*\* <0.01, \*\*\* <0.001)

**Figure: pathways\_across\_drugs\_top15.pdf**

Pathway-level summary of enrichment results across the full drug panel. Bars show the top 15 pathways (per modality) most frequently enriched among significant feature–response correlations (FDR < 0.05), with direction indicating whether pathways are associated with relative **resistance** (features positively correlated with viability) or **sensitivity** (features negatively correlated with viability). This highlights recurrent biological programs linked to drug response.

**Figure: enriched\_pathways\_per\_drug\_wide.pdf**

Per-drug counts of enriched pathways (FDR < 0.05), separated by direction (resistant vs sensitive). Drugs with higher pathway counts show broader, more coherent biology in their correlational signatures and are promising targets for follow-up biomarker development.

**Generalized figures:**

Generalized pathway enrichment figures were generated for the **two most efficacious** drugs, the **two most variable drugs**, and **Onalespib.**

**Figure: pathways\_<Drug>\_global\_resistant\_top15.pdf**

Bar chart of the top 15 **protein** pathways whose higher abundance is **positively** correlated with viability for <Drug> (i.e., features associated with **relative resistance**). Bars are ranked by −log10(FDR), with stars marking significance thresholds. These pathways represent biological processes that tend to accompany diminished drug effect.

**Figure: pathways\_<Drug>\_global\_sensitive\_top15.pdf**

Bar chart of the top 15 **protein** pathways whose higher abundance is **negatively** correlated with viability for <Drug> (i.e., features associated with **relative sensitivity**). Ranking and significance annotation as above. These pathways capture biology that tends to track with improved drug response.

**Figure: pathways\_<Drug>\_phospho\_resistant\_top15.pdf**

Bar chart of the top 15 **phosphoproteomic** pathways whose phosphorylation levels are **positively** correlated with viability for <Drug> (**resistance**). Bars are ordered by −log10(FDR); stars indicate significance. These represent signaling programs that accompany reduced drug effect.

**Figure: pathways\_<Drug>\_phospho\_sensitive\_top15.pdf**

Bar chart of the top 15 **phosphoproteomic** pathways whose phosphorylation levels are **negatively** correlated with viability for <Drug> (**sensitivity**). Ordering and annotation as above. These reflect signaling changes linked to enhanced drug response.

**Significance in the `pathways` plots**. (False discovery rate is FDR)

**\*\*\*** FDR < 0.001

**\*\*** FDR < 0.01

**\*** FDR < 0.05

(no star) not significant at 0.05

**Note on Onalespib (and other uniformly potent drugs):**  
Onalespib was one of the most efficacious compounds in the panel, but because nearly all samples were highly sensitive, the variability across specimens was minimal. Correlation-based enrichment relies on inter-patient heterogeneity; without it, even strong biological effects cannot be linked robustly to specific pathways. This explains why Onalespib yields few significant pathways despite its overall potency.