DATA MANAGEMENT AND SHARING PLAN
ELEMENT 1: DATA TYPE.

A. Type and amount of scientific data expected to be generated in the project: This project will generate data of multiple types and formats during the project. **X-ray crystallography**. ~50-60 crystal structures with associated raw experimental diffraction images (.img files). One complete dataset for a single structure may be ~30-50 GB, but may be compressed with standard lossless methods to facilitate storage and transmission. Processed data (text) files derived from the image files are on the order of 50-100 MB per dataset. The final refined coordinate mmcif file may be ~1-5 MB, the structure factor cif file can be ~5-50 MB, map coefficients (mtz format) are ~2-8 MB, and wwpdb structure validation reports (pdf and xml) will be ~3-5 MB. **Biophysical assays**. Three methods will be used: DSF using a Nanotemper Prometheus instrument, MST using a Nanotemper Monolith instrument, and ITC using a MicroCal ITC200 and an Affinity-ITC Auto. Raw data files are ~1-10 MB for each analysis. Each instrument uses it's own proprietary formats but the data will be exported to excel spreadsheets and CSV text files to facilitate sharing. We anticipate >150 compounds will be evaluated for binding to 2-3 different proteins with DSF, with smaller numbers of follow up experiments being done with MST and ITC. **Physiochemical profiling**. Compound solubility, pKa, and logD are measured. For compounds having appropriate spectral properties, a BioTek microplate reader will be used, otherwise mass spectrometry will be used as a readout. The instruments use their own proprietary file formats (Gen5 .xpt, Xcalibur RAW) but raw data will be exported to common formats (excel, CSV text, pdf, png) to facilitate sharing. Data for these studies is anticipated to be ~2-5 MB per compound and ~300-350 compounds will be evaluated.

ADME, PK, and MTD studies. Data from compound MDCK permeabilty, microsomal stability, in vivo pharmacokinetics and maximum tolerated dose studies will come from commercially available services (Charles River Laboratories). Data will be provided in pdf and excel formats, which will be converted to CSV text to facilitate sharing. We anticipate cell based evaluation of ~50-75 compounds (~1-5 MB of associated data per compound), and in vivo evaluation of ~1-2 compounds (~100-200 MB of associated data).

Biochemical assays. This will involve kinetic FLINT dose-response assays using a BioTek Synergy Neo2S plate reader (~1-3 MB data per assay). The proprietary xpt format will be exported to excel and CSV text for ease of sharing. Additional assays will use ^{32}P labeled phosphoprotein substrates and scintillation counting with a Beckman Coulter LS6500 (data will be in plain text format ~50-100 KB per assay). We anticipate ~1500-2000 FLINT assays and ~500-800 phosphoprotein assays. **Cell based assays.** This will involve: growth curves derived from cell counting (excel and CSV text), fluorescence microscopy (Nikon ND2, TIFF formats), SDS-PAGE gels (digitized in TIFF format), western blots (TIFF), and cell viability assays (BioTek xpt , excel, and CSV formats). We anticipate ~30-60 GB of data from these studies. **Proteomics.** LC-MS/MS/MS data will be collected from TMT-labeled phosphopeptides from inhibitor treated or vehicle treated HEK293 cells (WT + vehicle, WT + inhibitor, PPP5C -/- + vehicle, PPP5C -/- + inhibitor; 4 biological replicates each) with an Orbitrap Lumos instrument (XCalibur .RAW format). RAW files will be processed with Comet (.pep.xml, .pin.xml, .txt formats). We anticipate ~30-60 GB of data from ~1-2 inhibitors. **Chemistry/Synthesis**. ~300-350 compounds will be synthesized, purified and characterized using different analytical techniques. Major analytical techniques include Nuclear Magnetic Resonance (NMR) Spectroscopy, Mass Spectral analysis and HPLC. Proton NMR and Carbon 13 NMR chemical shift data for all compounds including intermediates and final compounds will be collected to validate compound identity and structure. Data associated with the proposed chemical syntheses in this project (including detailed experimental procedures and exported pdf files of the spectra) may require 5-6 GB of storage space. NMR raw data (proprietary Bruker FID format) will be converted to JCAMP-DX (non-proprietary ASCII) format to facilitate sharing and reuse.

B. Scientific data that will be preserved and shared, and the rationale for doing so: The data described above should allow researchers to reproduce our publications and will allow them to collect additional data in a similar way to extend our results. All data produced by key and/or technical personnel in the course of the project, or received by key personnel from commercial vendors without restrictions on sharing/reuse, will be preserved and shared with the following exceptions: laboratory notebooks, records of routine maintenance tasks, and data associated with troubleshooting or optimization of protocols.

C. Metadata, other relevant data, and associated documentation: In general, to facilitate the interpretation and reuse of the data, a README file and data dictionary will be generated and deposited into a repository along with all shared datasets. The README file will include method descriptions, instrument settings, RRIDs of resources such as antibodies, model organisms, cell lines, plasmids, and other tools (e.g., software, databases, services), and Protocol DOIs issued from protocols.io. A data dictionary will define and describe all variables in the dataset. The above will suffice for generalist repositories and for supplemental material deposited with publishers. However, much of the data generated by the project will be deposited in internet accessible specialist

repositories. We will ensure that all metadata required by these repositories is preserved and will be shared upon deposition in the manner proscribed by the repository. **X-ray crystallography data:** protein data bank (metadata will follow standards of the PDB Exchange Dictionary (PDBx). **Raw diffraction images:** Integrated Resource for Reproducibility in Macromolecular Crystallography (metadata include data collection details: project title, experiment date, site/beamline, equipment, X-ray wavelength, detector distance, oscillation width, number of frames). **Structures and identities of synthesized compounds, data from physiochemical profiling, biophysical, biochemical, and cell based assays, ADME data, pharmacokinetics, and MTD data** will be deposited with PubChem and adhere to PubChem's standards and guidelines for metadata, which will include detailed experimental protocols and lists of specific reagents and instrumentation. **Proteomics data** will be deposited with ProteomeXchange and adhere to MIAPE standards.

ELEMENT 2: ***Related Tools, Software and/or Code:*** Software for instrument proprietary formats: Origin (OriginLab), Gen5 (BioTek), TopSpin (Bruker), XCalibur (Thermo), Nano (TA Instruments), Panta (Nanotemper), NIS (Nikon). Specialized free or open source tools: Comet (GitHub): proteomics, DIALS (GitHub): raw diffraction images, CCP4 suite (GitHub): structure factors, ChimeraX (UCSF): structure coordinate files. When sharing data, to the fullest extent possible, we will export from proprietary to open or common formats, such as excel, CSV text, TIFF, png, pdf, and plain text. The specialist repositories we will use for structures, proteomics, and compound and assay data, are publicly accessible via the internet and will be able to facilitate data access, downloading, and some basic data analyses via the user's web browser.

ELEMENT 3: ***Standards:*** In accordance with FAIR Principles for data, we will, to the fullest extent possible, use open file formats (e.g. TIFF, CSV, TXT, PDF, mmcif etc.) and persistent unique identifiers such as RRIDs for resources (e.g., plasmids, antibodies, cell lines) and DOIs for protocols using protocols.io. For proteomics datasets, we will follow the minimum information about a proteomics experiment (MIAPE) standard. For crystallographic data, we will adhere to standards of the PDB Exchange Dictionary (PDBx).

ELEMENT 4: ***Data Preservation, Access, and Associated Timelines.***

A. Repository where scientific data and metadata will be archived: Research data and associated metadata suitable for specialized repositories will be made publicly available indefinitely through the following repositories (see above for discussion of which type of data is deposited where): PubChem, Protein Data Bank, ProteomeXchange, and the Integrated Resource for Reproducibility in Macromolecular Crystallography.

Other research data (and associated metadata) will be made publicly available indefinitely through deposition in a generalist institutional repository (JagWorks) at the University of South Alabama. Additionally, data may be published in journal articles and associated supplemental materials, which will also be deposited with PubMed Central in accordance with the NIH public access policy.

B. How scientific data will be findable and identifiable: Data will be discoverable via unique persistent identifiers. Protein Data Bank will associate a PDB code and DOI to each crystal structure dataset. The Integrated Resource for Reproducibility in Macromolecular Crystallography will assign DOIs to image sets. ProteomeXchage will assign PXD accession numbers. PubChem will assign compound IDs and/or substance IDs to synthesized compounds and assay IDs to assay datasets. JagWorks assigns a unique, persistent url to each dataset. PubMed Central assigns PMCIDs and DOIs pointing to published articles and associated supplemental material. These identifiers will be included in relevant publications and presentations to facilitate discovery.

C. When and how long the scientific data will be made available: All scientific data generated from this project will be made available as soon as possible, and no later than the time of publication or the end of the funding period, whichever comes first. The duration of preservation and sharing of the data will be a minimum of 10 years after the funding period.

ELEMENT 5: ***Access, Distribution, or Reuse Considerations.***

A. Factors affecting subsequent access, distribution, or reuse of scientific data: There are no anticipated factors or limitations that will affect the access, distribution or reuse of the scientific data generated by the proposal. Data will be released under a Creative Commons Zero (CC0) waiver.

B. Whether access to scientific data will be controlled: Controlled access will not be used. The data that is shared will be shared by unrestricted download.

C. Protections for privacy, rights, and confidentiality of human research participants: N/A

ELEMENT 6: ***Oversight of Data Management and Sharing:*** The lead PI will ensure that all personnel are trained on the DMS Plan elements and will be responsible for day-to-day oversight of team data management activities and data sharing. Plan compliance will be verified quarterly (at a minimum) by the PI. Broader issues of DMS Plan compliance, oversight, reporting, and revision will be handled by the lead PI in consultation with other key personnel and University officials.