--- Data & File Descriptions ---

A) Raw data files

Raw data includes the data measured by measurement devices/sensors from the field/lab (all csv in this project) in "Mahdavi & Siegel (2021) Indoor Air" (particle size distribution (PSD) phase). Due to data privacy, only a representative sample of the raw data (meaning one from each type) is shared (in the "Raw Data Files" folder).

See "Table 1.pdf" at the end of this document, "Data Pipeline and Code Configuration.pdf", and "list_of_files.xlsx" files to understand more about raw data.

B) Code files

Codes include all ".py" or ".ipynb" that read raw data or previously processed data to generate processed data, figures, calculations, or statistical analyses. There are four different code types used in this project:

- **Processing codes:** which process from raw data (stored in the "Processing Codes" folder).
- **Calculation codes:** which present numeric or statistical results from the processed data (stored in the main "Code" folder).
- **Visualization codes:** which aim to plot figures. This may come with additional processing of data to make the dataframe compatible for plotting (stored in the "Plotting Codes" folder).

See "Data Pipeline and Code Configuration.pdf", and "list_of_files.xlsx" files to see the list of code blocks. (One generic code titled "notion_correction" is not shown in "Data Pipeline and Code Configuration.pdf" but it is called by many code blocks).

C) Processed data files

Processed data includes the data generated after processing in data pipelines in "Mahdavi & Siegel (2021) IA". See "Data Pipeline and Code Configuration.pdf", and "list_of_files.xlsx" files to check the processed data lists and blocks. "Data Pipeline & Code Configurations.jpg" illustrates the entire data pipeline processing that generates processed data using codes from the raw data or previously processed data. Table 2 (at the end of this document) has a list of all these files.

A separate code block ("df_summary.ipynb") also presents a summary of some processed files. For data privacy purposes, not all the processed dataframes have been presented. Processed files are not shared as spreadsheet files either.

D) Plots

Plots include all the figures presented in "Mahdavi & Siegel (2021) IA" (PSD Phase) from the processed data. See "Data Pipeline and Code Configuration.pdf", and "list_of_files.xlsx" for more information. The plots aren't presented in separate files (e.g., jpg) but are illustrated in the same code file generating them (in Jupyter). See "Plotting Codes" for more information.

E) Other files

Any other file (mostly guidelines or descriptions) not classified above. A full list is available in "list_of_files.xlsx" (sheet "other_summary_files") provided in the repository.

Table 1 – Raw Data Summary (Mahdavi & Siegel (2021) IA – PSD Phase)

Item #	Raw Data Files & Name Conventions	Raw Data File Description	Variables/Columns	Variables/Columns Explanation
1	mastersizer_ida.csv mastersizer_bht.csv	Particle size distribution (PSD) of samples from institute for dust analyses (IDA) and Blue Heaven Technologies (BHT).	Record Number	A number that tracks the measurements of LDPS (#)
			Size bins (× 92)	Bin sizes varying from 0.1 to 3500 (total of 92 bins) (μm)
			Dx(10), Dx(50), Dx(90)	10^{th}, 50^{th}, and 90^{th} percentile sizes of the samples PSDs (μm)
			Sample Name	Name of the sample at the measurement time given by the user
			Measurement Date Time	Date and time of measurement
			File Name	Name of the file (including all measurement observations) given by the user
	{snn}.csv (× 4)	The four samples from filters going through 6 extraction cycles.	SOP File Name	Name of the Standard Operating Procedure (SOP) defined for the LDPS measurement
			Laser Obscuration	The obscuration of the sensor laser during measurement (%)
			Particle Refractive Index	Light Index of Refraction (IR) during the measurement (-)
	HUD_SD_{snn}.csv (× 3)	PSD for HUD project (Central Texas) samples for settled dust.	Particle Absorption Index	Light Index of Absorption (IR) during the measurement (-)
			Particle Density	Density of particle assumed in SOP (inputted not measured) (g/cm^3)
	{sin}_{xxx}_FD_VA.csv (× 6)	PSD for HUD project (Central Texas) samples for HVAC filter dust.	Ultrasonication Duration (SOP)	The duration of ultrasonication to the sample prior to testing (to avoid particle agglomeration during PSD measurement) (s)
			Ultrasonication Mode	Mode of ultrasonication
			Original Record Number	The original number tracking LDPS measurement prior to PSD recalculation for changing input parameters (#)
	low_high_obs_comp.csv	Low vs. high obscuration PSDs of test dust.	File Path	The path the mastersizer measurement records are saved in
			Specific Surface Area	Specific surface area of the particles (μm^2)

Table 2 – Processed Data Summary (Mahdavi & Siegel (2021) IA – PSD Phase)

Item #	Processed File Name	Processed File Description	Variables/Columns	Variables/Columns Explanation
1	natl_psd_master.xlsx hud_psd_master.xlsx	PSD of all1649 and HUD (Central Texas) samples	Size	Bin size of PSD (µm)
			Volume PSD properties of all 1649 samples (count, min, median, mean, max, and stdev)	Min, max, median, mean, and count of five runs of LDPS volume/count PSD measurements per bin (%, #)
2	natl_d_master.xlsx hud_d_master.xlsx	d-values of all 16949 and HUD (Central Texas) samples. d refers to the size under which a specific cumulative percentage of the distribution lies.	Sample Name	Name of the sample: 1649_{snn}_{X}_{yymmdd}_{in}_{in} {snn}: Filter Serial Number, {X}: Dust or Sieve Fractions, {yymmdd}, date where filters were extracted, {in}: initials of extractors.
			site	The site where the filter was deployed to collect dust: 1-20
			round	The round where filter was deployed (proxy for season): 1-4
			ft	Filtre types - 1: MERV8, 2: MERV8E, 3: MERC11, 4: MERV14
			Dx (0), Dx (10), Dx (25), Dx (50), Dx (75), Dx (90), Dx (100)	0, 10th, 25th, 50th, 75th, 90th, and 100th percentiles of the distribution samples
			stat	Statistical parameter of measurements: Median, geometric mean, geometric std, min, max, or count
			Ext_loc	IDA, BHT
3	natl_d_summary.xlsx	d-values of all 1649 samples (only after-sieve and only median values).	SN	Serial Number
			site	The site where the filter was deployed to collect dust: 1-20
			round	The round where filter was deployed (proxy for season): 1-4
			ft	Filtre types - 1: MERV8, 2: MERV8E, 3: MERC11, 4: MERV14
			Dx (0), Dx (10), Dx (25), Dx (50), Dx (75), Dx (90), Dx (100)	0, 10th, 25th, 50th, 75th, 90th, and 100th percentiles of the distribution samples
			No. Peaks	Number of peaks in the PSD of the sample
4	no_peaks.xlsx	Number of peaks in each PSD sample (1649 only)	SN	Serial Number
			No. Peaks	Number of peaks associated with SN of the sample
5	peak_locator.xlsx	Peak size of each PSD sample (1649 only)	SN	Serial number of the sample
			Fr	Dust fraction type (D: After-sieve, S: Sieve)
			Peak Size	The size at which the PSD peak takes place
14	laser_obs_master.xlsx	Obscuration level of each dust sample (1649 only)	SN	Serial number of the sample
			Fr	Dust fraction type (D: After-sieve, S: Sieve)
			Laser Obscuration	The obscuration measured during the laser diffraction measurement
15	low_high_obs_comp.xlsx	PSD of test dust samples with two different obscuration levels.	Size	Bin size of PSD (µm)
			Volume PSD properties of low obscuration and high obscuration samples	Min, max, median, mean, and count of five runs of LDPS volume/count PSD measurements per bin (%, #)