```
import pandas as pd
import glob


diann_report = pd.read_csv("../data/processed/run2/diannsummary/diann_report.tsv", sep="\t")
diann_report.head()
```

	File.Name	Run	Protein.Group	Protein.Ids	Protein.Names	Genes	PG.Quantity	PG.Normalis
0	250314_NS-10461_RJ_DIA_16_A.mzML	250314_NS-10461_RJ_DIA_16_A	P55011	P55011;G3XAL9	S12A2_HUMAN	SLC12A2	6719300.0	621055(
1	250314_NS-10461_RJ_DIA_16_B.mzML	250314_NS-10461_RJ_DIA_16_B	P55011	P55011;G3XAL9	S12A2_HUMAN	SLC12A2	7802160.0	601940(
2	250314_NS-10461_RJ_DIA_16_C.mzML	250314_NS-10461_RJ_DIA_16_C	P55011	P55011;G3XAL9	S12A2_HUMAN	SLC12A2	8079980.0	730301(
3	250314_NS-10461_RJ_DIA_24_A.mzML	250314_NS-10461_RJ_DIA_24_A	P55011	P55011;G3XAL9	S12A2_HUMAN	SLC12A2	9066360.0	766146(
4	250314_NS-10461_RJ_DIA_24_B.mzML	250314_NS-10461_RJ_DIA_24_B	P55011	P55011;G3XAL9	S12A2_HUMAN	SLC12A2	7172880.0	622965(

5 rows × 58 columns

Start by removing peptides with ambigous AAs

```
# Remove sequences with 'X' in them.
filtered = diann_report[~diann_report['Stripped.Sequence'].str.contains('X')]
filtered.to_csv("../data/processed/run2/diannsummary/diann_report_noX.tsv", sep="\t", index=False)
```

Rename precursor_mz to observed_mz

```
parquet_files = glob.glob("../data/processed/run2/mzmlstatistics/*.parquet")

for file in parquet_files:
    df = pd.read_parquet(file)
    if 'precursor_mz' in df.columns and 'observed_mz' not in df.columns:
        df = df.rename(columns={'precursor_mz': 'observed_mz'})
        df.to_parquet(file)
        print(f"Renamed 'precursor_mz' to 'observed_mz' in {file}")
    else:
        print(f"No change needed for {file}")
```

```
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_PIV_A_ms_info.pa
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_16_C_ms_info.par
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_IAV_B_ms_info.pa
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_16_A_ms_info.par
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_IAV_C_ms_info.pa
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_PIV_C_ms_info.pa
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_24_B_ms_info.par
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_24_A_ms_info.par
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_16_B_ms_info.par
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_24_C_ms_info.par
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_IAV_A_ms_info.pa
Renamed 'precursor_mz' to 'observed_mz' in ../data/processed/run2/mzmlstatistics/250314_NS-10461_RJ_DIA_PIV_B_ms_info.pa
```

Run in cmd line

quantmsioc convert-diann --report_path data/interim/diannsummary/diann_report_noX.tsv --qvalue_threshold 0.05 --mzml_info_folder data/interim/mzmlstatistics/ --sdrf_path data/interim/pipeline_info/sdrf.tsv --output_folder data/processed/ --output_prefix_file Becky

```
df = pd.read_parquet("../data/processed/run2/diann_convert-971b8c15-a02b-46ab-9012-5a779126e067.feature.parquet")
df.head()
```

	sequence	peptidoform	modifications	precursor_charge	posterior_error_probability	is_decoy	calc
0	HAVSEGTK	HAVSEGTK	None	2	0.258418	0	
1	HSHKPHPVETSQPSDK	HSHKPHPVETSQPSDK	None	4	0.762169	0	
2	RVEHHDHAVVSGR	RVEHHDHAVVSGR	None	4	0.000001	0	
3	KKPLGERPKDEDER	KKPLGERPKDEDER	None	4	0.003507	0	
4	HEEAPGHRPTTNPNASK	HEEAPGHRPTTNPNASK	None	4	0.000161	0	

5 rows × 29 columns

Explode the intensities column

```
# Explode the 'intensities' column so each dict becomes a separate row
df_exploded = df.explode('intensities').reset_index(drop=True)

# Expand the dictionaries in the 'intensities' column to new columns
intensity_df = pd.json_normalize(df_exploded['intensities'])

# Drop the old 'intensities' column and join the new columns
df_exploded = df_exploded.drop(columns=['intensities']).join(intensity_df)

df_exploded = df_exploded.rename(columns={'intensity': 'intensity', 'sample_accession': 'sample_accession'})

print(df_exploded[['sequence', 'sample_accession', 'intensity']].head())
```

```
             sequence              sample_accession     intensity
0           HAVSEGTK  250314_NS-10461_RJ_DIA_PIV_A  1.395730e+04
1   HSHKPHPVETSQPSDK  250314_NS-10461_RJ_DIA_PIV_A  2.703930e+04
2      RVEHHDHAVVSGR  250314_NS-10461_RJ_DIA_PIV_A  1.464260e+06
3     KKPLGERPKDEDER  250314_NS-10461_RJ_DIA_PIV_A  2.541140e+05
4  HEEAPGHRPTTNPNASK  250314_NS-10461_RJ_DIA_PIV_A  5.116560e+06
```

```
df_exploded.to_parquet("../data/processed/run2/feature_long.parquet")



# Load features and SDRF
features = pd.read_parquet("../data/processed/run2/feature_long.parquet")
sdrf = pd.read_csv("../data/processed/run2/pipeline_info/sdrf.tsv", sep='\t')

# from column 'assay name' remove the prefic 'Run ' and convert to integer
sdrf['assay name'] = sdrf['assay name'].str.replace('Run ', '')

# Rename SDRF columns to match ibaqpy expectations
sdrf_renamed = sdrf.rename(columns={
    'source name': 'sample_accession',
    'characteristics[biological replicate]': 'biological_replicate',
    'comment[fraction identifier]': 'fraction',
    'Factor Value[Treatment]': 'condition',
    'assay name': 'run'
})

# Select only the needed columns
cols_needed = ['sample_accession', 'condition', 'biological_replicate', 'run', 'fraction']
sdrf_for_merge = sdrf_renamed[cols_needed]

# Merge onto features
features_annotated = features.merge(sdrf_for_merge, on='sample_accession', how='left')

features_annotated.to_parquet("../data/processed/run2/feature_long_annotated.parquet")
```

```
df = pd.read_parquet("../data/processed/run2/feature_long_annotated.parquet")
df.head()
```

	sequence	peptidoform	modifications	precursor_charge	posterior_error_probability	is_decoy	calc
0	HAVSEGTK	HAVSEGTK	None	2	0.258418	0	
1	HSHKPHPVETSQPSDK	HSHKPHPVETSQPSDK	None	4	0.762169	0	
2	RVEHHDHAVVSGR	RVEHHDHAVVSGR	None	4	0.000001	0	
3	KKPLGERPKDEDER	KKPLGERPKDEDER	None	4	0.003507	0	
4	HEEAPGHRPTTNPNASK	HEEAPGHRPTTNPNASK	None	4	0.000161	0	

5 rows × 35 columns

Run features2peptides

ibaqpyc features2peptides -p data/processed/feature_long_annotated.parquet -s data/interim/pipeline_info/sdrf.tsv --remove_decoy_contaminants --remove_low_frequency_peptides -o data/processed/features2peptides_norm.csv --save_parquet

Run peptides2proteins

ibaqpyc peptides2protein -f data/external/Homo-sapiens-uniprot-reviewed-contaminants-decoy-202210.fasta -p data/processed/features2peptides_norm.csv -e Trypsin -n -t -r -i 2 -m human -c 200 -o data/processed/abs_quant.tsv --verbose

```
quant = pd.read_csv('../data/processed/run2/abs_quant.tsv', sep='\t')
quant.head()
```

	ProteinName	SampleID	Condition	NormIntensity	Ibaq	IbaqNorm	IbaqLog	IbaqPpb	MolecularWeight	
0	A0A024RBG1	250314_NS-10461_RJ_DIA_16_A	IAV	1.295169e+06	80948.068030	0.000040	5.603168	4010.222568	20421.303212	6
1	A0A096LP01	250314_NS-10461_RJ_DIA_16_A	IAV	3.347910e+06	185995.020129	0.000092	5.964463	9214.320308	10901.330247	3
2	A0AV96	250314_NS-10461_RJ_DIA_16_A	IAV	2.511659e+07	92002.155514	0.000046	5.658760	4557.849610	64058.234910	3
3	A0AVF1	250314_NS-10461_RJ_DIA_16_A	IAV	2.675556e+06	9290.125792	0.000005	4.662984	460.239176	64136.415847	4
4	A0AVI2	250314_NS-10461_RJ_DIA_16_A	IAV	8.774637e+05	4301.292643	0.000002	4.328561	213.088975	237782.462003	