

A curated dataset of microfluidic liposome formulations with cross-laboratory validation for machine-learning applications

This dataset documents how different formulation choices and microfluidic operating conditions influence the size and uniformity of liposomes produced using a controlled microfluidic system. The data were generated through a multi-step experimental workflow that included a screening phase, two response-surface optimization phases, and an independent cross-laboratory study. Together, these experiments provide an extensive view of how lipid composition and flow conditions shape the final properties of the liposomes.

Liposome formation is sensitive to both the ingredients used (such as lipid ratios) and the production settings (such as flow rates, flow-rate ratios, and buffer choice). Factors like mixing intensity, solvent-aqueous ratios, and chip geometry all influence how lipids self-assemble into vesicles. This dataset enables the systematic study of these relationships across a broad and diverse experimental space.

Introduction

The dataset package is composed of 5 subfolders:

- **code** contains Python code for preprocessing steps and examples of data exploration and usage, in particular
 - **logs**: directory containing logging information about the processing scripts in the current directory.
 - **data_gn_adder.py**: script to generate the dataset extension **data_extensions/formulations_extended_with_gn.csv**.
 - **data_smote_adder.py**: script to generate the dataset extension **data_extensions/formulations_extended_with_SMOTE.csv**.
 - **raw_data_checker.py**: script to validate and preprocess formulations in the **raw_data** directory.
 - **raw_data_slicer.py**: script to select the correct CHIP configuration from the files in the **raw_data** directory.
- **data** contains the preprocessed raw datasets, in particular
 - **formulations.csv** is the main cleaned dataset obtained by preprocessing and merging **initial_formulations_raw.xlsx** with **new_formulations_raw.xlsx**.
 - **wet_lab_validation.csv** contains independent wet-lab validation formulations obtained by preprocessing **wet_lab_validation_raw.xlsx**.
- **data_extensions** contains two artificially extended datasets obtained by adding Gaussian noise to **formulations.csv** (yielding **formulations_extended_with_gn.csv**) and by applying SMOTE interpolation (yielding **formulations_extended_with_SMOTE.csv**).
- **metadata** Contains documentation files such as
 - **features_bounds.json**: dictionary with the physical constraints of each feature.
 - **features_names_mappings.json**: dictionary containing standardized names for each feature and other data conventions.

- **features_descriptions.txt**: file with a brief description of each feature in the raw datasets.
- **raw_data** contains the raw spreadsheets compiled by the operators
 - **initial_formulations_raw.xlsx** was collected by one operator (operator A) using a specific instrument (equipment A) in a laboratory in Latina (Latium), Italy (laboratory A).
 - **new_formulations_raw.xlsx** was collected by operator A using an equivalent instrument (equipment B) in a laboratory in Rome, Italy (laboratory B).
 - **wet_lab_validation_raw.xlsx** was collected by another independent operator (operator B) using equipment A in laboratory A.

The structure of the dataset folder is shown below.

```
dataset
|
|_ code
|   |
|   |_ logs
|       |
|       |_ data_gn_adder.log
|       |
|       |_ data_smote_adder.log
|       |
|       |_ raw_data_checker.log
|       |
|       |_ raw_data_slicer.log
|
|   |
|   |_ data_gn_adder.py
|   |
|   |_ data_smote_adder.py
|   |
|   |_ raw_data_checker.py
|   |
|   |_ raw_data_slicer.py
|
|_ data
|   |
|   |_ formulations.csv
|   |
|   |_ wet_lab_validation.csv
|
|_ data_extensions
|   |
|   |_ formulations_extended_with_gn.csv
|   |
|   |_ formulations_extended_with_SMOTE.csv
|
|_ metadata
|   |
|   |_ features_bounds.json
|
```

```
|_ features_names_mappings.json
|
|_ features_descriptions.txt
|
|_ raw_data
|   |_ initial_formulations_raw.xlsx
|   |_ new_formulations_raw.xlsx
|   |_ wet_lab_validation_raw.xlsx
|
|_ README.md
```

Raw dataset description

Each row in the raw dataset represents one complete microfluidic production run. For every run, the dataset includes both the input settings controlled by the experimenter and the outputs measured afterwards.

These inputs and outputs vary across the following dimensions:

- **Formulation variables:** concentrations of four clinically relevant lipids that determine membrane composition and stability, in particular
 - **ESM:** Egg sphingomyelin concentration (mg/mL)
 - **HSPC:** Hydrogenated soy phosphatidylcholine concentration (mg/mL)
 - **CHOL:** Cholesterol concentration (mg/mL)
 - **PEG:** DSPE-PEG2000 concentration (mg/mL)
- **Process conditions** that affect mixing efficiency and lipid self-assembly inside the chip, in particular
 - **TFR:** the Total Flow Rate, indicating the speed at which the aqueous and organic streams were injected (mL/min).
 - **FRR:** the Flow Rate Ratio, showing the relative proportion of aqueous to organic phases (aqueous:organic).
 - **AQUEOUS** or **Aqueous medium:** the buffer environment used, encompassing two aqueous media commonly employed in liposome preparation (**MQ** and **PBS**).
- **Microfluidic hardware** including the variable
 - **CHIP:** either Droplet or Micromixer.

Note. Only the *Micromixer* produced reliable liposomes, so the cleaned dataset keeps only those entries.

- **Outputs** measured via Dynamic Light Scattering following the protocols in the associated article
 - **SIZE** or **Size:** the diameter of the liposomes (nm).
 - **PDI:** a measure between 0 and 1 of size uniformity (polydispersity index, where 0 means uneven size distribution and 1 means uniform size).
 - **OUTPUT** or **Formation:** a binary viability flag (removed in the cleaned dataset) indicating whether well-formed liposomes were obtained.

- **Auxiliary variables** including an alphanumeric **ID** for each formulation and a reference to the **Main Lipid** (either **ESM** or **HSPC**).

The raw dataset includes **three main experimental blocks**:

- **Seed dataset (called `initial_formulations_raw`, $n = 276$)** generated using a structured Design-of-Experiments approach, covering a wide range of formulation and flow conditions.
- **Cross-laboratory extension dataset (called `new_formulations_raw`, $n = 58$)** consisting of independently produced formulations in a second lab using the same equipment and protocols. This portion evaluates reproducibility, operator independence, and equipment consistency.
- **Independent wet-lab validation dataset (called `wet_lab_validation_raw`, $n = 12$)** including a small set of randomly chosen formulations recreated experimentally to test the predictive performance of machine-learning models trained on the main dataset.

Raw dataset preprocessing and cleaning

After applying the preprocessing pipeline — including checking physical limits, removing invalid chip conditions, correcting naming inconsistencies, and dropping redundant columns — the cleaned dataset, called **formulations**, contains **304 datapoints** corresponding to high-quality formulations obtained from the **combination of the preprocessed seed dataset and the preprocessed cross-laboratory extension dataset**. This dataset can be used for statistical or regression modelling, machine-learning tasks, analysis of formulation–performance trends, design-space exploration, and benchmarking inverse-design approaches.

The preprocessing workflow includes:

1. **Column standardization and fuzzy-matching correction**
 - 59 inconsistent labels automatically corrected
2. **Constraint checks** using the metadata files, for example verifying $PDI \in [0,1]$ and $SIZE \leq 10,000$
 - 5 outliers detected; corresponding entries removed
3. **Chip aggregation**
 - All Droplet-chip entries ($n = 25$) removed due to lack of liposome formation
4. **Dropping redundant columns**
 - **ID, ML, CHIP, OUTPUT**
5. **Optional augmentation**
 - SMOTE (25%)
 - Gaussian noise (25%)

All processing scripts are available in `code/preprocessing_scripts.py`.

Usage

Loading the dataset in Python

```
import pandas as pd

df = pd.read_csv("data/formulations.csv")
```

Exploring the dataset in Python

```
# Once the dataset has been loaded
# Inspect the first rows
df.head()
# Summary statistics
df.describe()
# Check distributions of key variables
df[['SIZE', 'PDI', 'TFR', 'FRR']].hist(figsize=(10, 6))
```

Example of a machine-learning task in Python

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor

# Once the dataset has been loaded
# Divide targets from features
X = df[['ESM', 'HSPC', 'CHOL', 'PEG', 'TFR', 'FRR']]
y = df['SIZE']
# Perform train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
# Try a simple regression model
model = RandomForestRegressor()
model.fit(X_train, y_train)
# Check the results
print("R squared:", model.score(X_test, y_test))
```