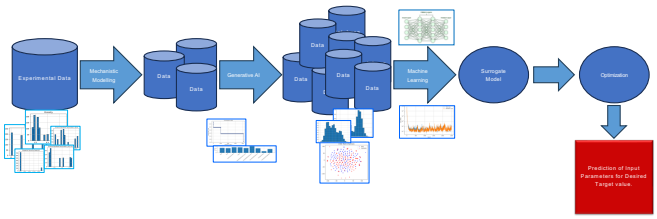
Model Driven Design of Formulated Products

Amir Arjmandi-Tash [1], Peyman Mostafaei [1], Daniel Markl[2], Rachel Smith [1]*
[1] The University of Sheffield, Department of Chemical and Biological Engineering, Sheffield, UK
[2] The University of Strathclyde, Strathclyde Institute of Pharmacy and Biomedical Sciences, Glasgow, UK

1. Introduction and Project Overview

Disintegration plays a crucial role in the performance of pharmaceutical, agricultural, and food products by breaking granules into smaller particles. This process increases the surface-to-volume ratio, facilitating the rapid release of active pharmaceutical ingredients (APIs) in the target environment.



The goal of this project is to develop models which link wet granulation process parameters with granule disintegration behaviour, enabling optimization for a desired performance.

2. Developed Workflow

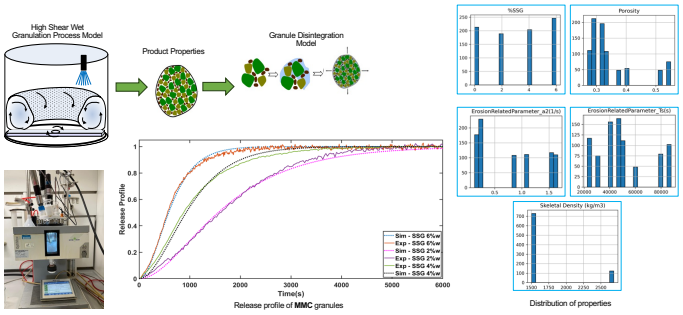
1. Experimental Data Collection: Gather real-world data pairs $D_{exp} = \{(x_i, y_i)\}_{i=1}^{n}$ to establish baseline observations.
2. Mechanistic Model Interpolation: Use a physics-based model to estimate outputs across the input space.
3. Data Expansion via Generative AI: Train a generator model to create new, realistic input samples from noise vectors.
4. Machine Learning Surrogate Model: Fit a predictive ML model using the expanded dataset to approximate disintegration behavior.
5. Inverse Design through Optimization: Search for optimal input $\hat{x}$ that leads to the target disintegration outcome.

Step	Process	Mathematical Representation
1	Experimental data	$\mathcal{D}exp = \{(x_i, y_i)\}_{i=1}^{n}$
2	Mechanistic Model interpolation	$\widehat{y_{mech}} = f_{mech}(x; \phi)$
3	Generative AI to expand data	$\hat{x} = G(z; \theta_G)$
4	ML model trained on expanded data	$\widehat{y_{ML}} = f_{ML}(x; \theta_{ML})$
5	Optimization using surrogate model	$\hat{x} = \arg\min_{\hat{x}} \mathcal{L}\left(f_{ML}(x), y_{target}\right)$

f_{mech}: Trained mechanistic model.
$\hat{y}_{mech}$: Disintegration values predicted via mechanistic model.
$\hat{x}$: New data generated via Generative AI model (G).
$\hat{y}_{ml}$: Disintegration values predicted via machine learning model.
f_{ml}: Trained machine learning model.
L: Loss function to be minimized during optimization.

3. Initial Database

Experimental data was obtained through multiple methods. Disintegration values were measured for granules based on microcrystalline cellulose (MCC) and formulated with sodium starch glycolate (SSG) as the disintegrant. In addition to disintegration, properties such as porosity, skeletal density, and erosion-related parameters were recorded. The granules were produced using a high-shear wet granulation process, resulting in a dataset for modeling disintegration behavior.

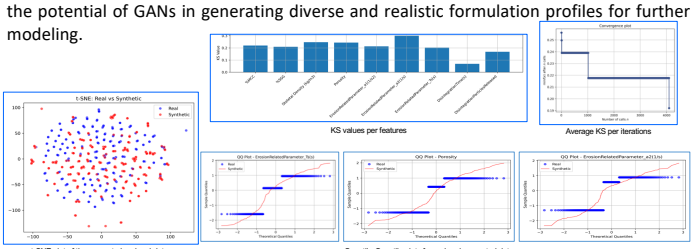


Release profile of **MMC** granules

Distribution of properties

4. Generative AI for Data Generation

A Generative Adversarial Network (GAN) was trained to generate synthetic formulation data for expanding the experimental dataset. GANs consist of two competing neural networks—a generator that creates fake data and a discriminator that tries to distinguish fake from real data. Through this adversarial training process, the generator learns to produce highly realistic data. In this study, new input combinations such as porosity and skeletal density were generated without needing to explicitly define them for each sample, addressing a key limitation of mechanistic models. The quality of the generated data was evaluated using the Kolmogorov–Smirnov (KS) test, which measures the maximum difference between the cumulative distribution functions (CDFs) of real and synthetic data. A KS value of 0.19205 was obtained, indicating a good alignment between the two distributions. Since values below 0.2 are considered valid, these results highlight the potential of GANs in generating diverse and realistic formulation profiles for further modeling.



t-SNE plot of the generated and real data: showcasing the high quality of the data generation process.

Quantile-Quantile plots for real and generated data

5. Forward ML Model Training

Several machine learning models, including Extra Trees and Gradient Boosting Regression, were trained to predict disintegration profiles. The Artificial Neural Network (ANN) performed best, with validation MSE between 10^{-4} and 10^{-5}. ANNs learn by adjusting weights through backpropagation to minimize prediction errors, making them well-suited for modeling complex relationships in formulation data.

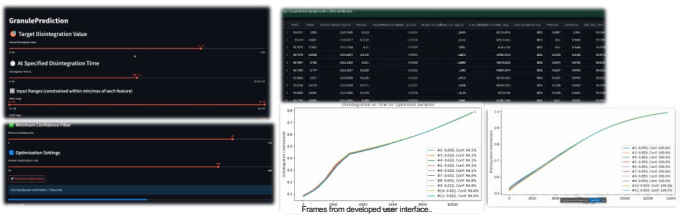
Train and validation loss for forward ANN model trained on generated data.

MSE values for trained models

Train and validation loss curves for optimized ANN model on initial database

6. Optimization and User-Interface

A user-friendly interface was built to implement Bayesian Optimization (BO) for inverse design. Given a user-defined disintegration value y^* at a target time t^*, BO uses the trained neural network model f_{ml} to predict outcomes and iteratively updates input features $\hat{x}$ by minimizing the loss $L(x)$, enabling efficient and automated formulation design.



Frames from developed user interface..

7. Conclusion

- A hybrid framework combining experimental data, generative AI, and machine learning was developed for inverse formulation design.
- GAN was used to generate realistic synthetic data (e.g. porosity, skeletal density), expanding the dataset beyond original measurements.
- A trained ANN achieved high accuracy in predicting disintegration profiles, with validation MSE between 10^{-4} and 10^{-5}
- Bayesian Optimization was integrated into a custom-built user interface to identify optimal input features that achieve user-defined disintegration targets.

8. Acknowledgment

We acknowledge excellent support for this work. Thanks to: Daniel Markl and group (University of Strathclyde), Poul Bach & Alexander Findeisen (Novyzymes), Christophe Grosjean & Michael Brozio (Syngenta), Joris Salari (Corbion), Sri Sharath Kulkarni & Daniel Sieber (DFE Pharma), Tristan Hessberger & Tobias Heß (Budenheim), Vanessa Havenith (SE Tylose), Bindhu Gururajan (Novartis), Brian Karim (Lincoln Electric), Omid Arjmandi Tash & Amir Esteghamatian (Pfizer), Maxime Touffet (Cargill), Nabhan Ahmed (Keurig Dr Pepper), Stefan Bellinghausen and the broader Siemens PSE team. Thank you to IFPRI & EPSRC for funding this research.