



**University of
Nottingham**

UK | CHINA | MALAYSIA

A large, high-resolution image of the Earth as seen from space, showing the curvature of the planet and the blue oceans. The image is centered in the background of the slide.

RSE for Data-Driven AI Research

**Grazziela Figueredo
University of Nottingham**



**In times of Software Dev/Engineering living an
“existential crisis” ...**

**how have we advanced Research and Outputs
through the collective power of the University of
Nottingham RSE community**

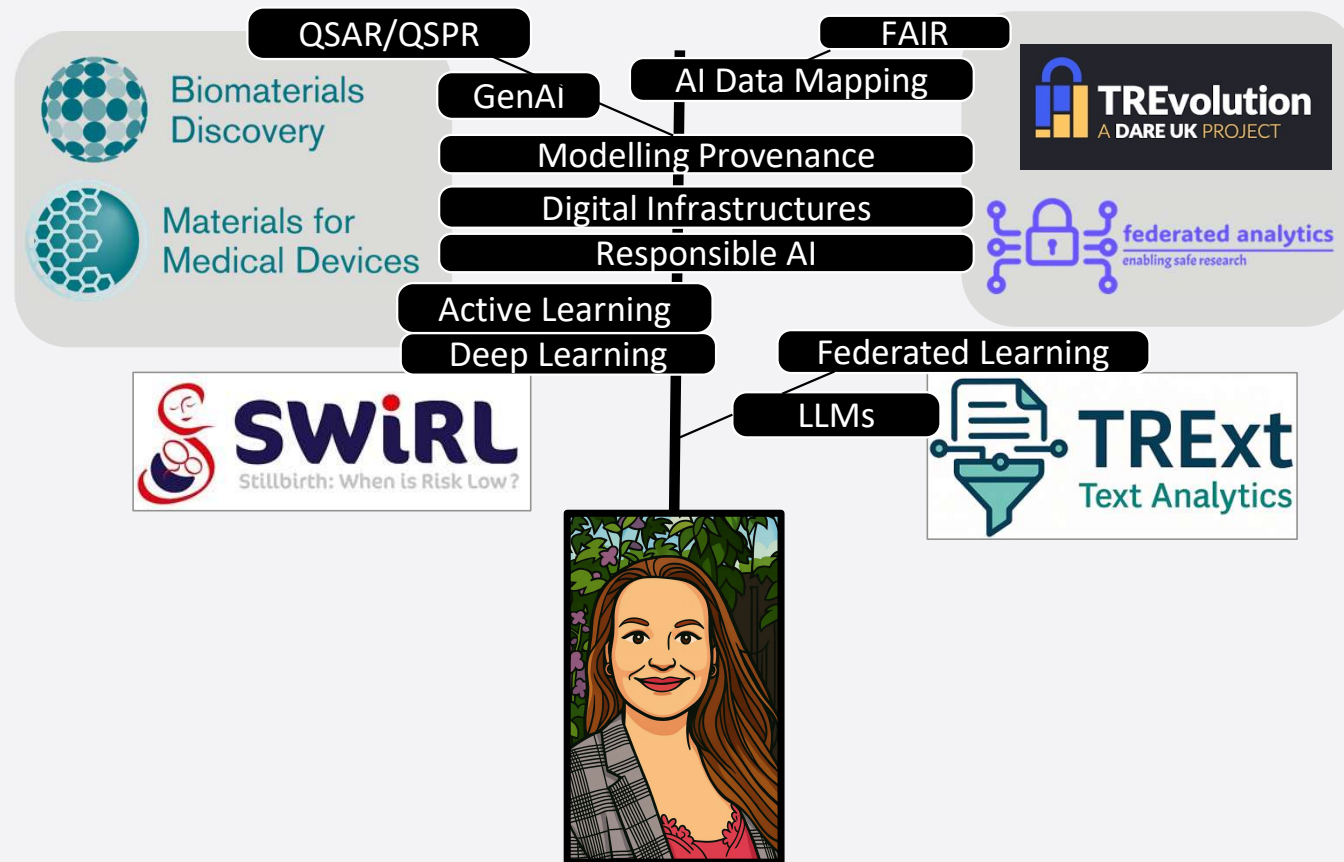


About Me

Grazziela Figueredo

Associate Professor in Health Data Science, The University of Nottingham

- Expertise
 - Data science and ML for medical and (bio)materials domains
 - Data science as a service:
 - Research into intelligent products
 - Knowledge transfer
- Worked at/with the DRS in Nottingham
- All projects benefitted from RSE input





Acknowledgements to the Team



Eduardo
Aguilar-
Bejarano



Rongjun
Dong



Jimiama M
Mase



James
Mitchel-White



Reza
Omidvar



Daniel Lea



Troy Kettle



Karthik
Sivakumar



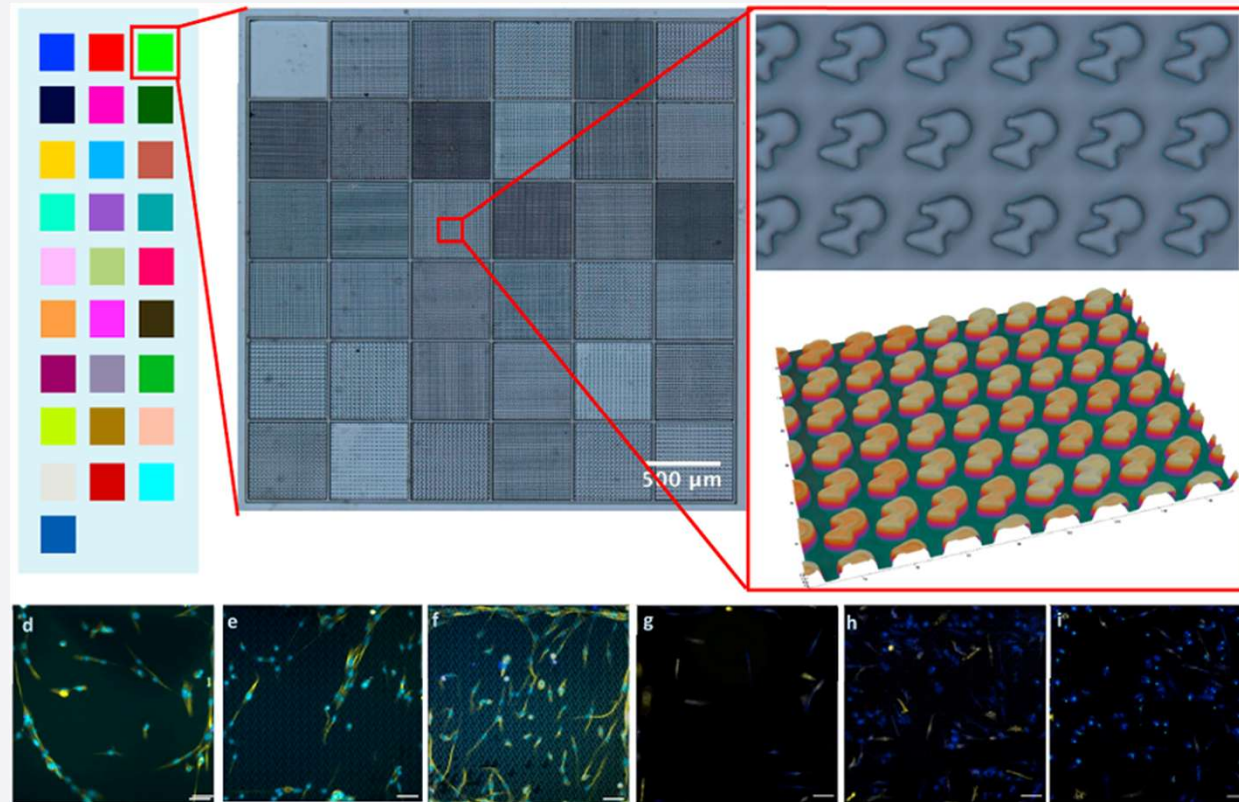
and thanks to the Digital Research Service at the
University of Nottingham



Biomaterials Discovery

High-throughput Screening

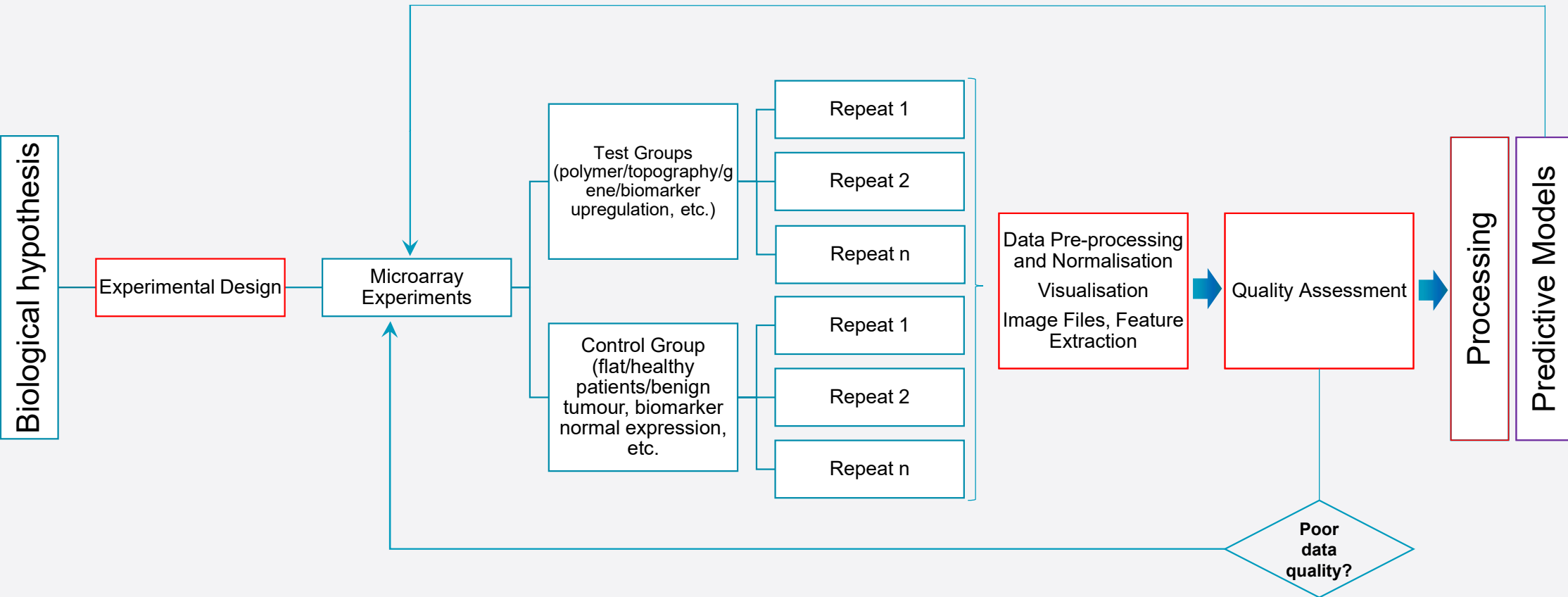
- Chemistries
- Topographies
- Various cellular responses
- Rich data



L. Burroughs et al. 2021. Discovery of synergistic material-topography combinations to achieve immunomodulatory osteoinductive biomaterials using a novel in vitro screening method: The ChemoTopoChip, *Biomaterials*, 271.



A Generic Analytics Pipeline





Overall Challenges in Data-Driven Discovery of Cell-Instructive Materials

Understanding interplay between chemistry, topography and biology

Understand synergy between chemistry and topographies

Mechanistic interpretation

Reliable FAIR (findable, accessible, interoperable, reusable) data

AI (the black box) as a bridge between materials and biology?

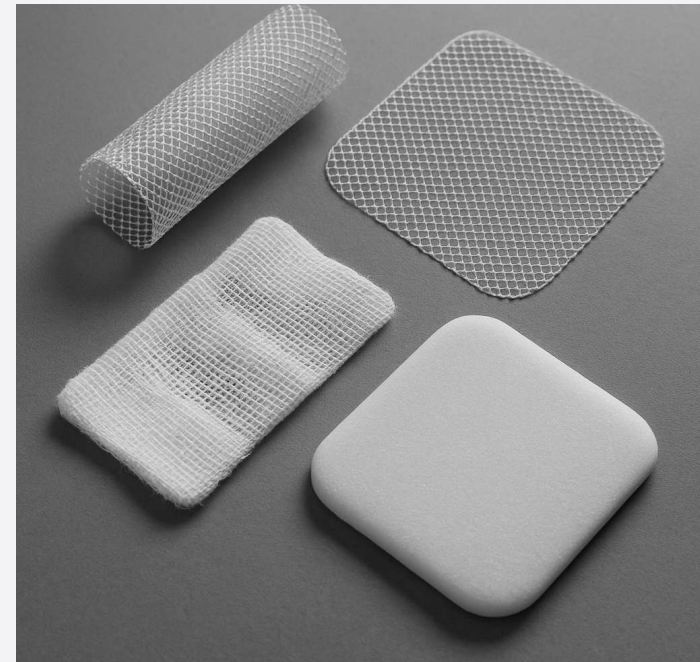
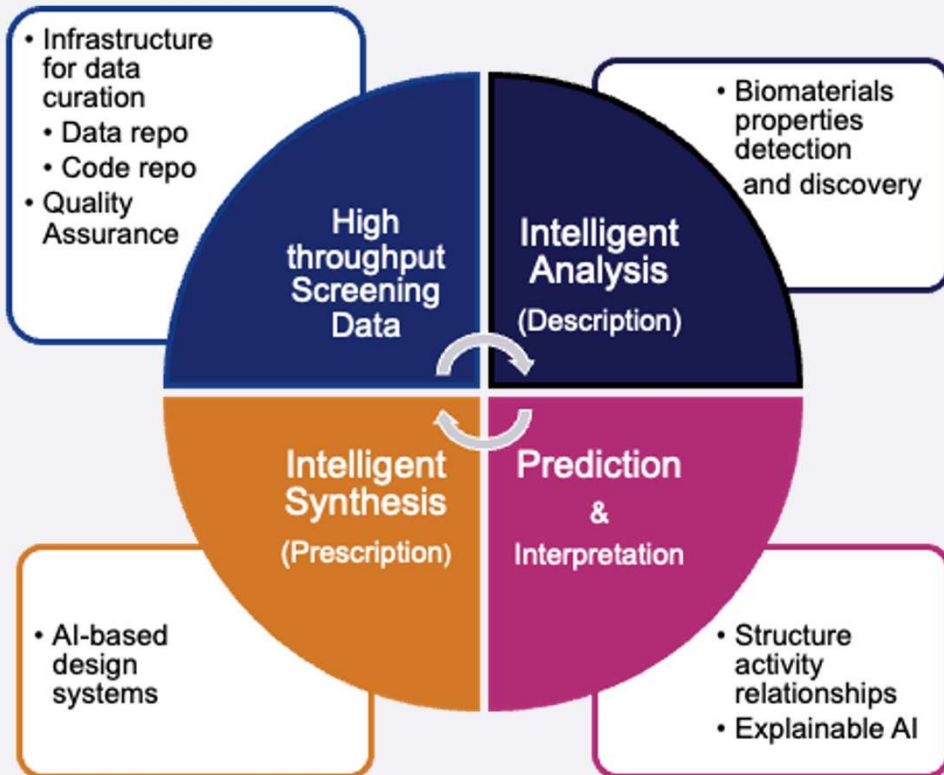
- AI Explainability
- AI Reliability
- FAIR models



Materials for Medical Devices



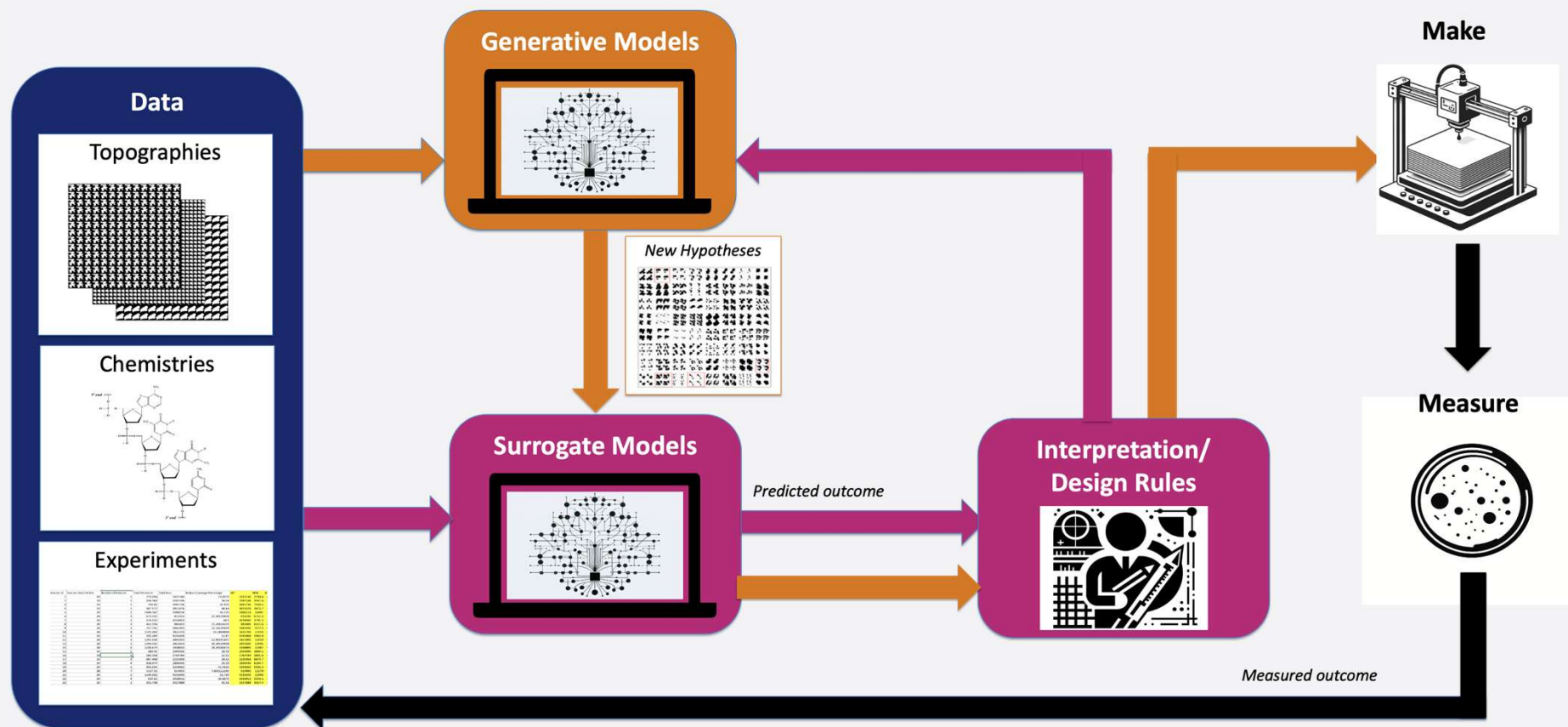
Digital infrastructure for intelligent analysis, interpretation and closed-loop synthesis of Biomaterials chemistries and topographies



Materials for Medical Devices



My Research within the Project Ecosystem





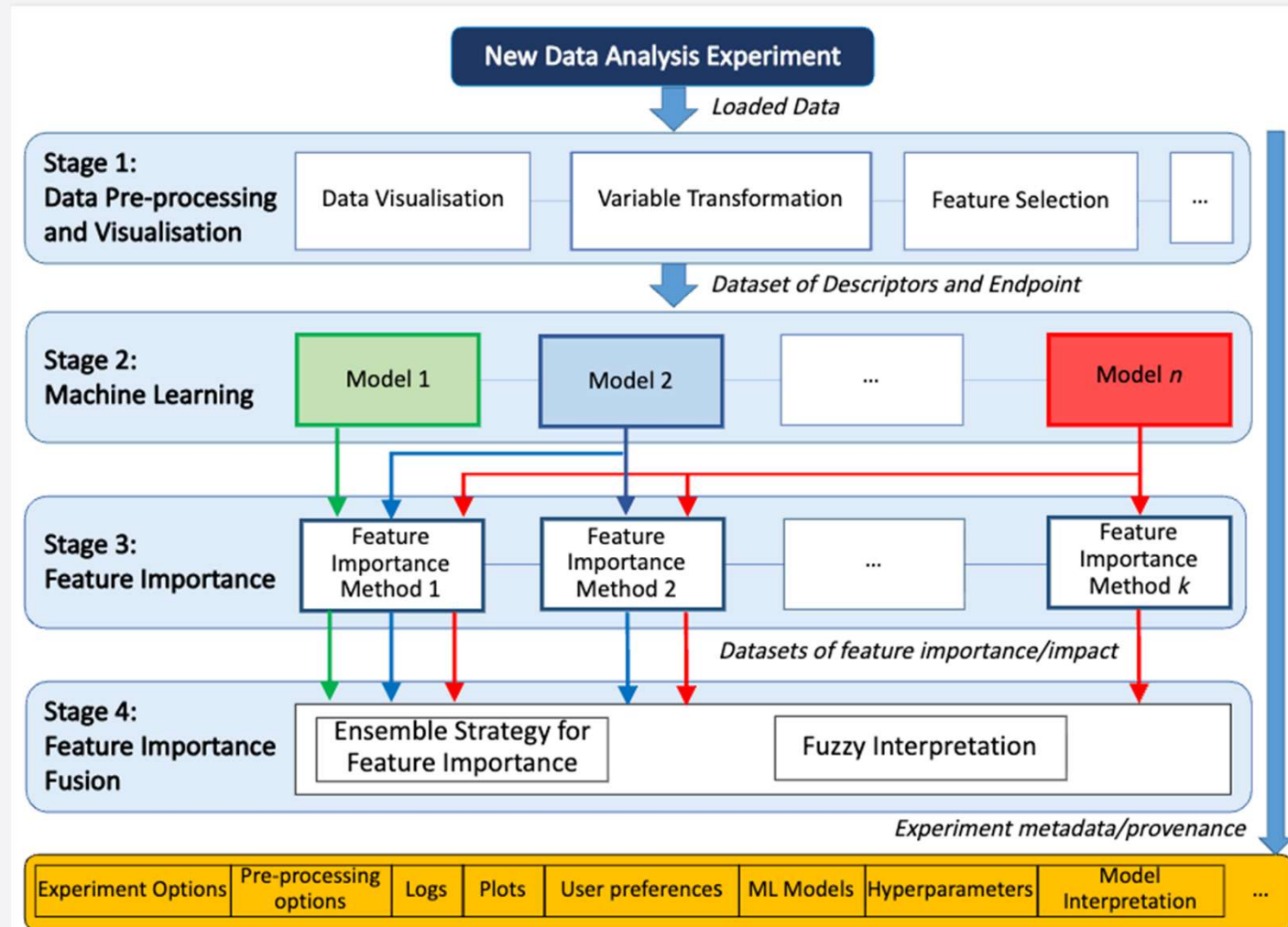
Tabular Data - Traditional Models



- Robust explainability
- Comprehensive design rules
- Analysis and modelling provenance

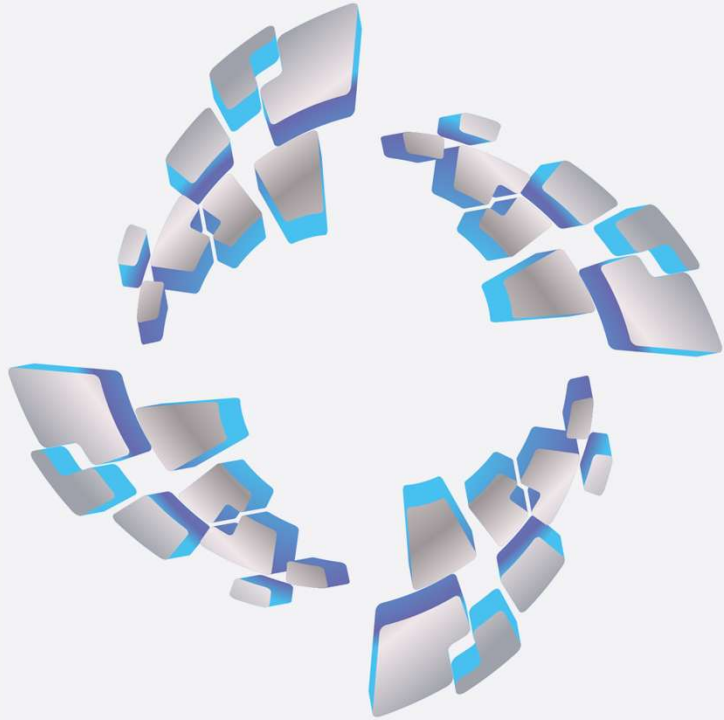
Rengasamy, D et al. Feature importance in machine learning models: A fuzzy information fusion approach, Neurocomputing, 2022, 511.

Rengasamy, D.; Rothwell, B.C.; Figueredo, G.P. Towards a More Reliable Interpretation of Machine Learning Outputs for Safety-Critical Systems Using Feature Importance Fusion. Appl. Sci. 2021, 11, 11854.





Tabular Data Solution - Helix



HELIX

README MIT license

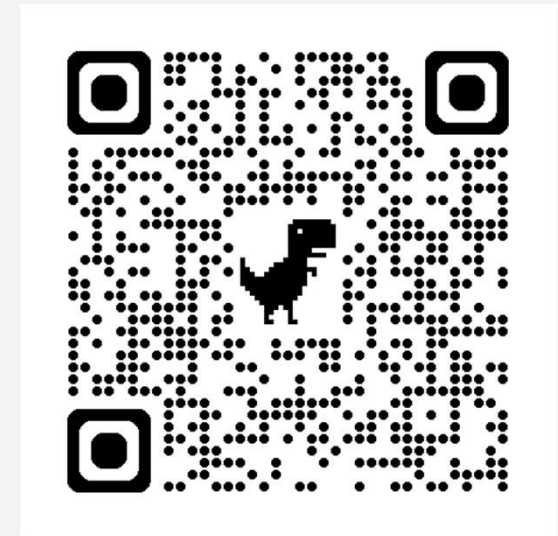
Helix: Python Toolkit for Machine Learning, Feature Importance, and Fuzzy Interpretation

LICENSE MIT PYTHON UV SCIKIT-LEARN MATPLOTLIB LINUX MAC OS

WINDOWS

ISSUES 7 OPEN BUILD DOCS PASSING PUBLISH DOCS PASSING CODE QUALITY PASSING

TESTS PASSING DOWNLOADS 25/MONTH



Aguilar-Bejarano, E., Lea, D., Sivakumar, K., Mase, J.M. et al. (2026) 'Helix 1.0: An open-source framework for reproducible and interpretable machine learning on tabular scientific data', Patterns. Available at: <https://doi.org/10.1016/j.patter.2026.101536>



Helix – Analysis and Modelling



Home

New Experiment

Data Preprocessing

Data Visualisation

Train Models

Feature Importance

Predict

View Experiments

New Experiment

Here you can create a new experiment. Once created, you will be able to select it on the Data Preprocessing, Data Visualisation and Train Models pages.

Create a new experiment

Give your experiment a name, upload your data, and click **Create**, located at the bottom of the page. If an experiment with the same name already exists, or you don't provide a file, you will not be able to create it.

Name of the experiment

/Users/pmzgf/HelixExperiments/ 

e.g. MyExperiment

Upload your data as a CSV or Excel (.xlsx) file.

Please note that the last column of the uploaded file should be the dependent variable.

Choose a file 



Drag and drop file here

Limit 200MB per file • CSV, XLSX

Browse files

Name of the dependent variable. **This will be used for the plots. As default, the name of the last column of the**



Helix Provenance



- Home
- New Experiment
- Data Preprocessing
- Data Visualisation
- Train Models
- Feature Importance
- Predict
- View Experiments**

2025_CarverAkmal_ABS_Lipids_Paper_SI

On this page, you can select one of your experiments to view.

Use the dropdown below to see the details of your experiment.

If you have not run any analyses yet, your experiment will be empty. Go to the sidebar on the **left** and select an analysis to run.

Select an experiment

2025_CarverAkmal_ABS_Lipids_Paper_SI
✕ ▾

- Test_Data2_Classification
- JosephManning_3Months_PostDischarge
- JosephManning_1Month_PostPCUDischarge
- JosephManning_6Months_PostDischarge
- Test_Data1_Classification
- Test_May_Data1_Regressionn
- RefactorEnsemble
- Test_May_Data1_Regression

random for...	Test	0.920 ± 0.0...	0.933 ± 0.0...	0.886 ± 0.1...	1.000 ± 0.0...	0.920 ± 0.0...
random for...	Train	1.000 ± 0.0...	1.000 ± 0.0...	1.000 ± 0.0...	1.000 ± 0.0...	1.000 ± 0.0...



Graph Representation of Polymers

Surrogate Models



Predicted outcome

Interpretation/
Design Rules

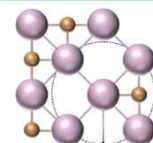


Polymer Modelling

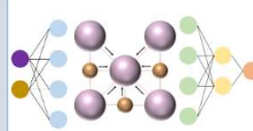
- Multiple representations
- Graph-based
- Transformer-based
- Etc
- Means for interpretation

Chemistry Modelling

Stage 1:
Data
Representation



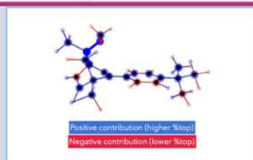
Stage 2: Graph
Neural Network
Design



Stage 3:
Prediction



Stage 4:
Interpretation



Chemical Fragments
Importance/Impact



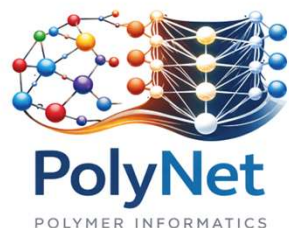
Graph Representation of Polymers - PolyNet



Eduardo Aguilar-Bejarano

PolyNet

python 3.9+ license MIT PyTorch 2.x PyTorch Geometric enabled RDKit supported Tests passing code style black
dependency management poetry status active development



PolyNet is a Python library for polymer property prediction using graph neural networks (GNNs) and traditional machine learning (TML). It provides a complete, configurable pipeline — from raw SMILES strings to trained models, evaluation metrics, result plots, and atom-level explainability — designed for polymer informatics research.

The library is split into two layers:

- **Core package (`polynet`)** — pipeline stages, model training/inference, featurisers, config schemas, and path/IO helpers. No Streamlit dependency; importable from any Python script or notebook.
- **Optional GUI (`polynet.app`)** — a Streamlit web application that wraps the same core stages. Install only when you want an interactive interface.





Unstructured Chemistry Data – PolyNet Data Representation

Welcome to PolyNet

- Create Experiment
- Representation
- Train Models
- Predict
- Explain Models
- Analyse Results

Create Experiment

In this section, you can create a new experiment by uploading a CSV file containing SMILES strings and the target variable you want to model. You can also select the columns that contain the SMILES strings and the target variable, as well as any additional features you want to include to your molecules.

Experiment name ⓘ

Please provide an experiment name.

Choose a CSV file ⓘ

Drag and drop file here
Limit 200MB per file • CSV Browse files

Please upload a CSV file to proceed.

Save Experiment



Unstructured Chemistry Data – PolyNet Model Interpretation

The screenshot shows a web browser window with the URL `localhost:8501/Explain_Models`. The browser tab is titled "Explain Models". On the left side, there is a navigation menu with the following items: "Welcome to PolyNet", "Create Experiment", "Representation", "Train Models", "Predict", "Explain Models" (which is highlighted), and "Analyse Results". The main content area is titled "Explain your models" and contains the following text:

In this section, you can explain the predictions of the GNN models. You can select the model you want to explain, the set of data you want to explain, and the specific datapoints you want to explain.

We offer different explanation levels, including explaining general trends in the model, explanation of the molecular embedding, and the explanation of specific datapoints.

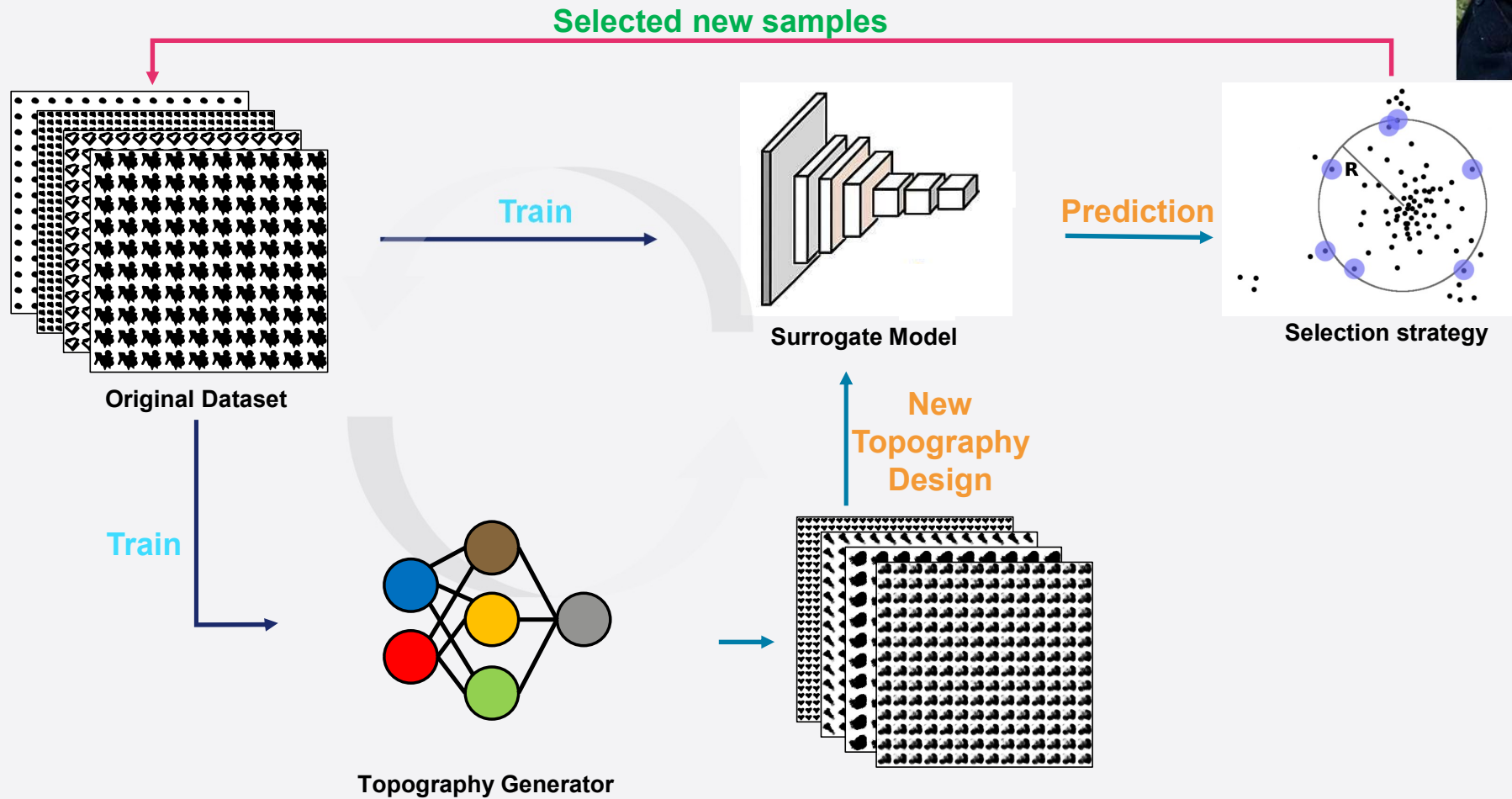
To explain the models or specific instances, you can select the explainability algorithm you want to use, the node features you want to explain, and the colors for the positive and negative explanations. The explanations will be displayed as plots.

For the molecular embedding, you can choose what method of dimensionality reduction to use to get a 2D projection from them. Further, you can select from different options to colour the projection plot, giving insights about how the model is organising the latent space.

Below the text, there are two dropdown menus. The first is labeled "Select an experiment" and has "demo" selected. The second is labeled "Model Results" and is currently empty.



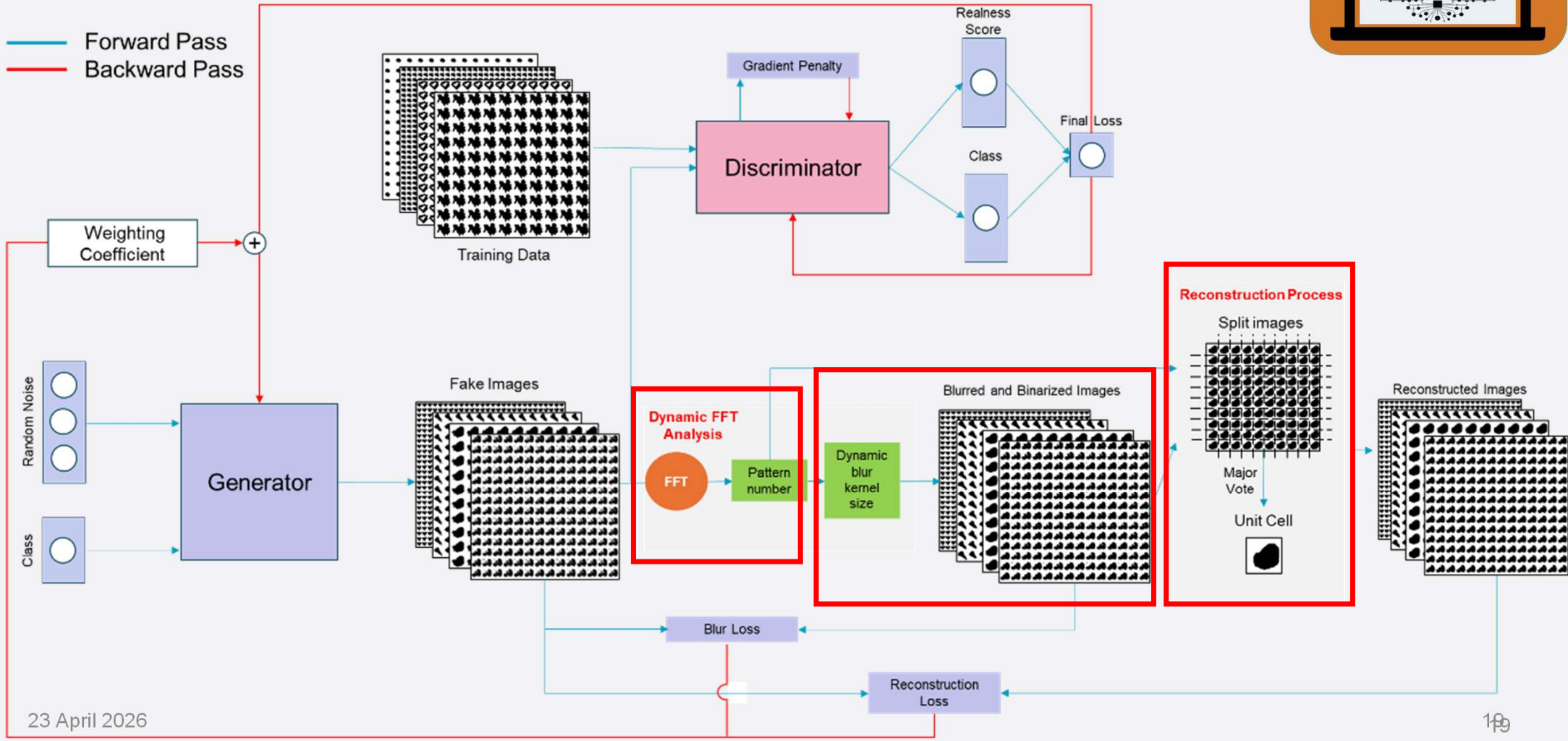
Proposed Solution – Design of Topographies





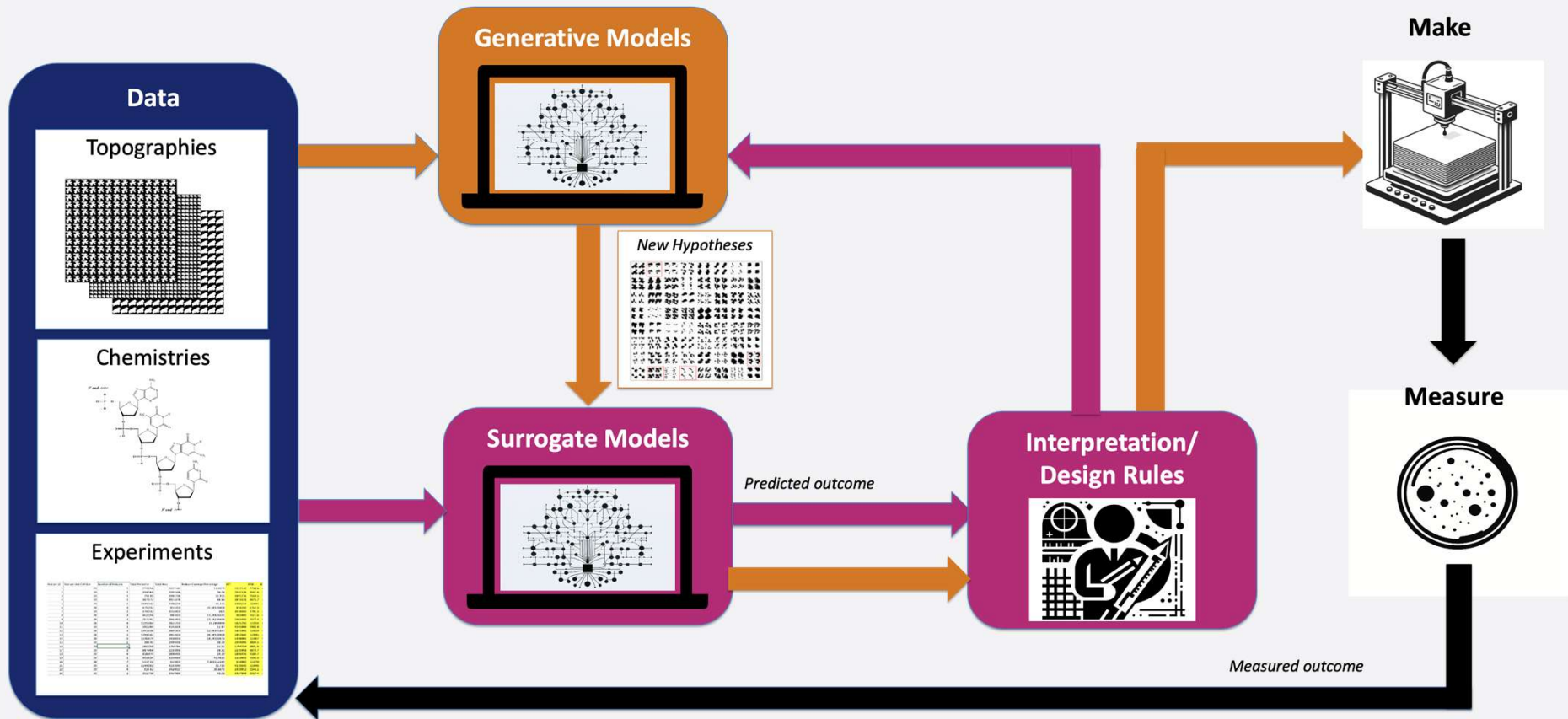
The Dynamic FFT Conditional Blurring Generative Adversarial Network (DF-ACBlurGAN)

Generative Models





Next steps: engineer the closed loop





Final Remarks

- Intelligent collaborative tools focused on
 - FAIR data and models
 - Explainable
 - Human-cooperation
 - Community driven (open source, with open backlog)
 - The value of RSE for FAIR close-loop research to accelerate discovery

Example of communities of practice in Nottingham fostering collaboration, peer-learning, software better fit for purpose, and sharing ideas and best practice



University of
Nottingham

UK | CHINA | MALAYSIA

Thank you