Accelerare lo sviluppo di prodotti mediante apprendimento automatico su grafi da prove di formulazione

Manuel Dileo | Francesca Saraceni

L'Intelligenza Artificiale al servizio della ricerca chimica, 2 Ottobre 2025

intellico
From data to sustainable value

Connets LAB
Computer Science
UNIVERSITÀ DEGLI STUDI DI MILANO

FEDERCHIMICA
CONFINDUSTRIA

Work presented @ ECML PKDD 2024

We present a work developed within the framework of applied research activities carried out jointly by Intellico and the University of Milan (UniMi), and presented at the ECML PKDD 2024 conference.
The contribution focuses on the application of Machine Learning techniques to the formulation process, with the goal of supporting and enhancing industrial innovation

Connets LAB
Computer Science
UNIVERSITÀ DEGLI STUDI DI MILANO

+

intellico
From data to sustainable value

Research lab specializing in computer networks, data mining, and AI, advancing the analysis and modeling of complex systems (e.g., social, mobile, financial, and blockchain networks) with the goal of unveiling hidden patterns and driving innovation in data-driven intelligence.

AI-based solutions provider specialized in machine learning applied to the formulation process (cosmetics, food, and chemical industries), with the goal of reducing time-to-market

Business Needs Addressed

Clients now prefer a slightly different taste.
How can I change it quickly?

MARKET PRESSURE

Europe has blacklisted ingredient X. We have to re-formulate 50+ recipes.

REGULATORY PRESSURE

The price of Ingredient Y has gone through the roof. what do I substitute it with?

SUPPLY CHAIN PRESSURE

LENGTHY TIME TO MARKET

From 3 to 10 months

A WIDE RANGE OF POSSIBILITIES

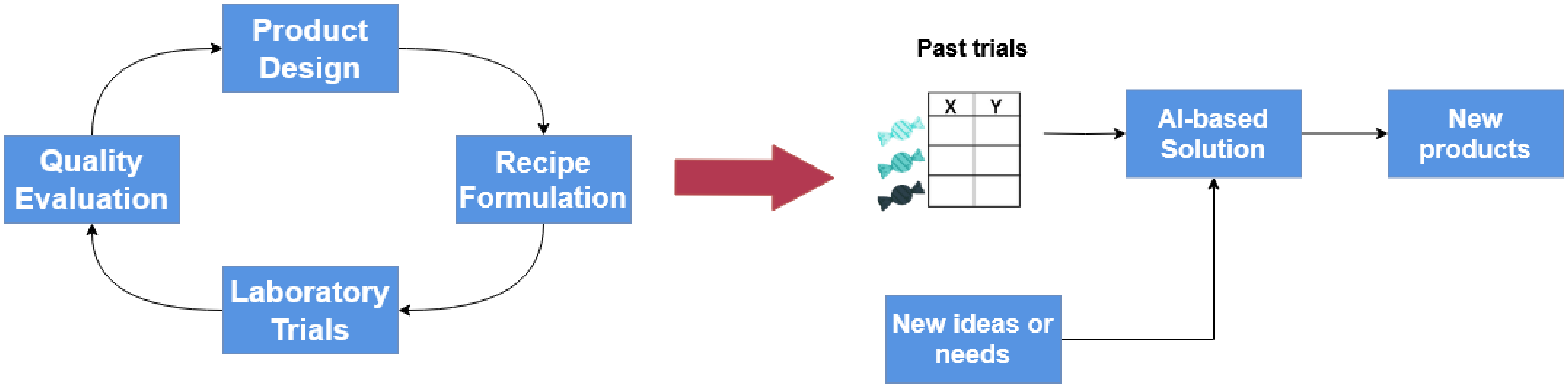
Today focus

- **Reduce Time-to-Market**
- Accelerate Material Discovery
- Optimize Manufacturing Processes
  – Real-time monitoring
  – Parameter optimization
- Streamline Regulatory Compliance

Product development

GOAL

TRADITIONAL APPROACH — **DATA-DRIVEN APPROACH**



Abandon an iteration-based process in favor of a data-driven approach

Product development

REACH THE GOAL

- Formulation trials modeled as tuples of product features *X* and target properties *y*
- Speed-up product development by predicting property values of new formulations: multiregression task.
- Allow transferable knowledge among between similar product lines.
- Provide explanations for identifying the most important features for reaching a desidered property.
- Visualize the results to identify new starting recipes and reduce the labor intensive experiments and waste of materials.

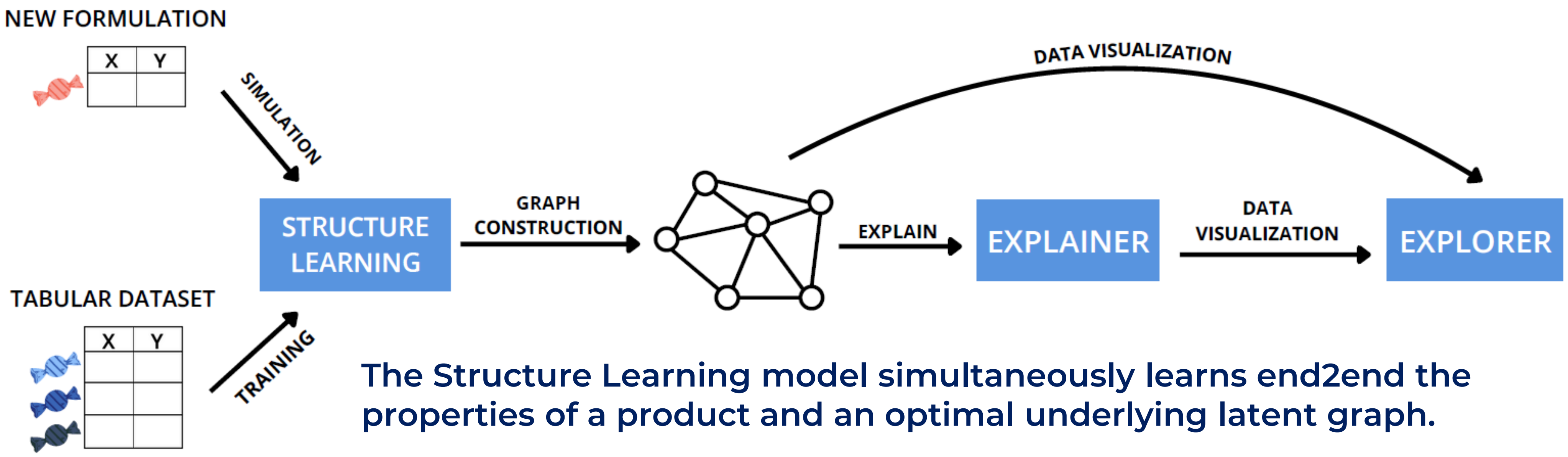
Food design

CASE-STUDY

- We collect two datasets in the context of product development for the food industry in collaboration with a multinational company
- Each formulation trial is characterized by the ingredients of the recipe and its physical properties (e.g. viscosity).
- The target properties are consumer-appealing characteristics such as malleability and toughness.

Our solution

METHODOLOGY

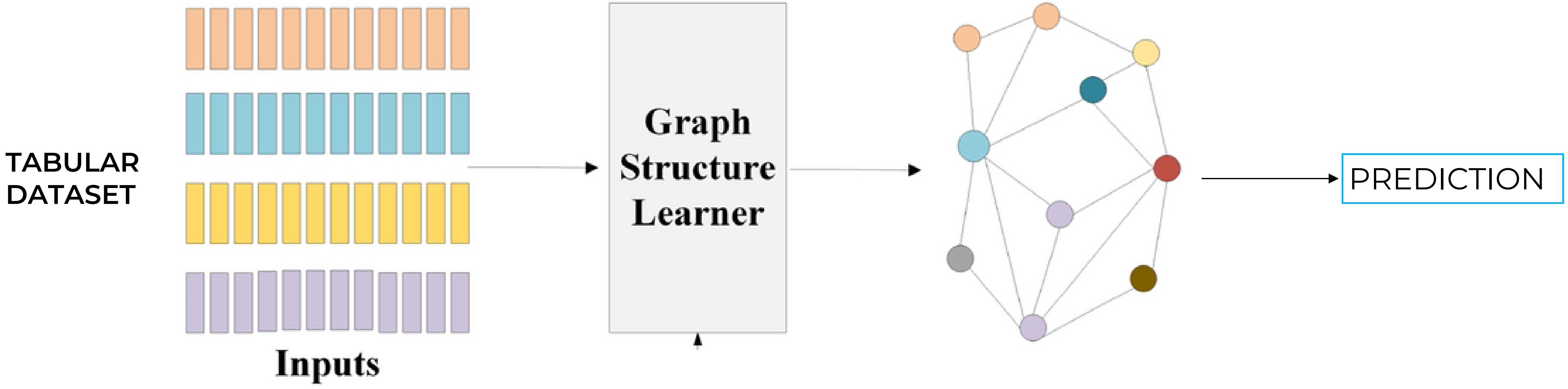


The Structure Learning model simultaneously learns end2end the properties of a product and an optimal underlying latent graph.

Explainer, Explorer and Simulation modules are offered through customized web applications

Structure learning

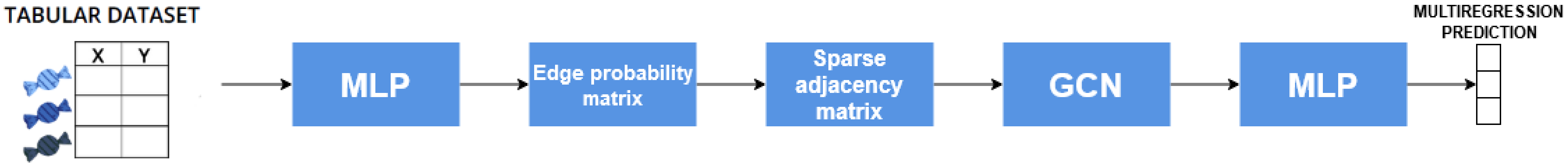
MAIN SCOPE



Structure Learning

METHODOLOGY

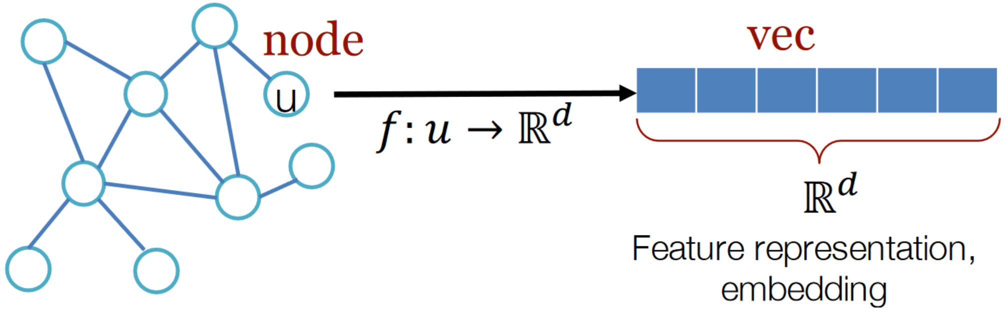
We adopt a graph machine learning approach to explore latent correlations among formulations and allow transferable knowledge between similar product lines.

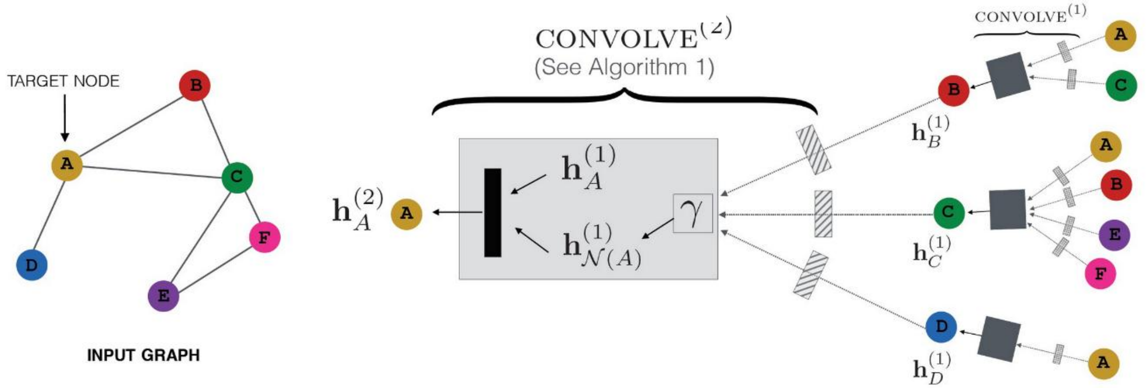


We leverage a differentiable graph module (DGM) to perform structure learning on formulation trials.

Graph Neural Networks (GNNs)

- Neural Networks that works naturally on graph-structured data.
- Automatic feature learning on graph with node attributes
- Node features combined with neighbour information





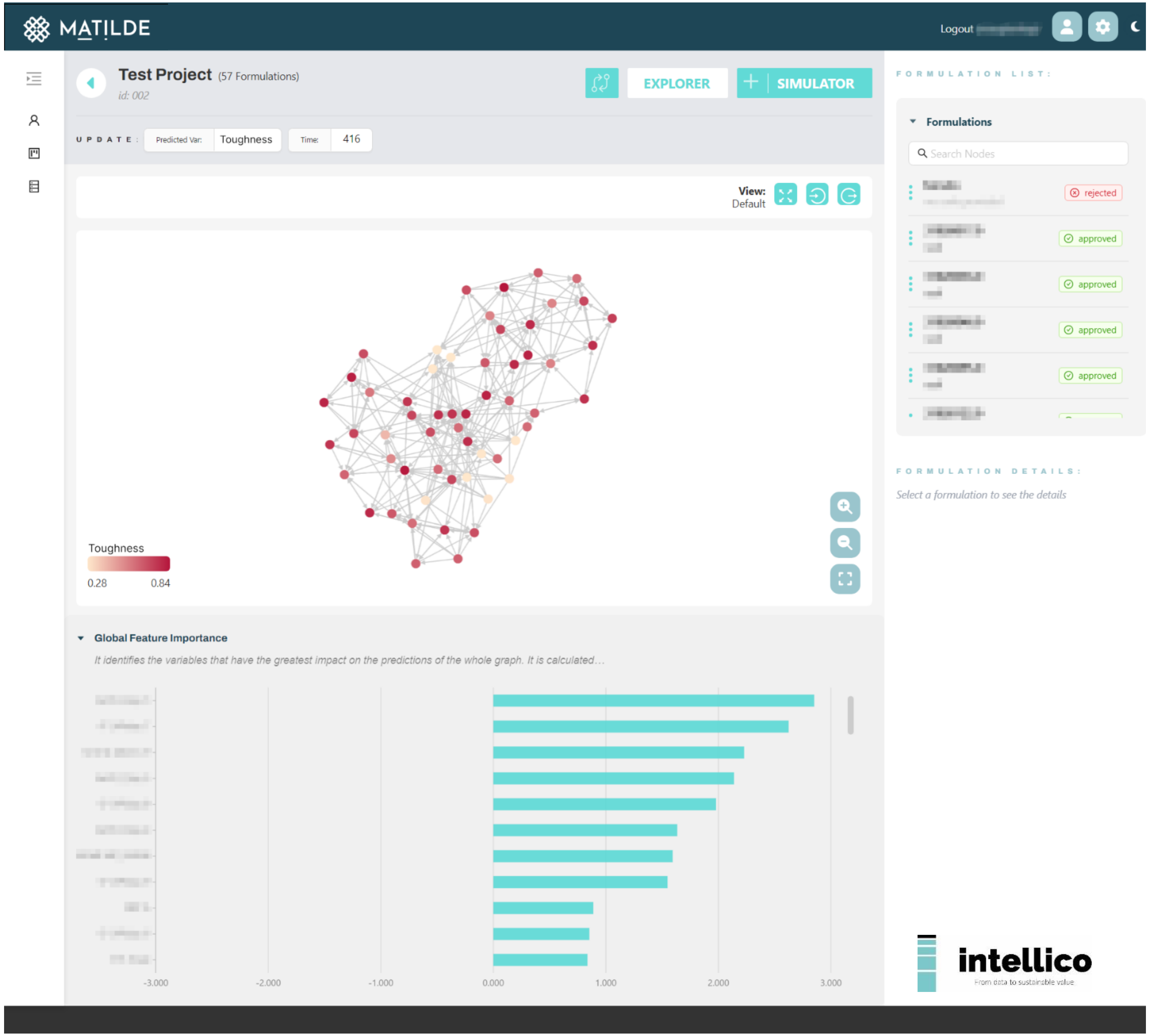
Representation Learning on Networks, snap.stanford.edu/proj/embeddings-www, WWW 2018

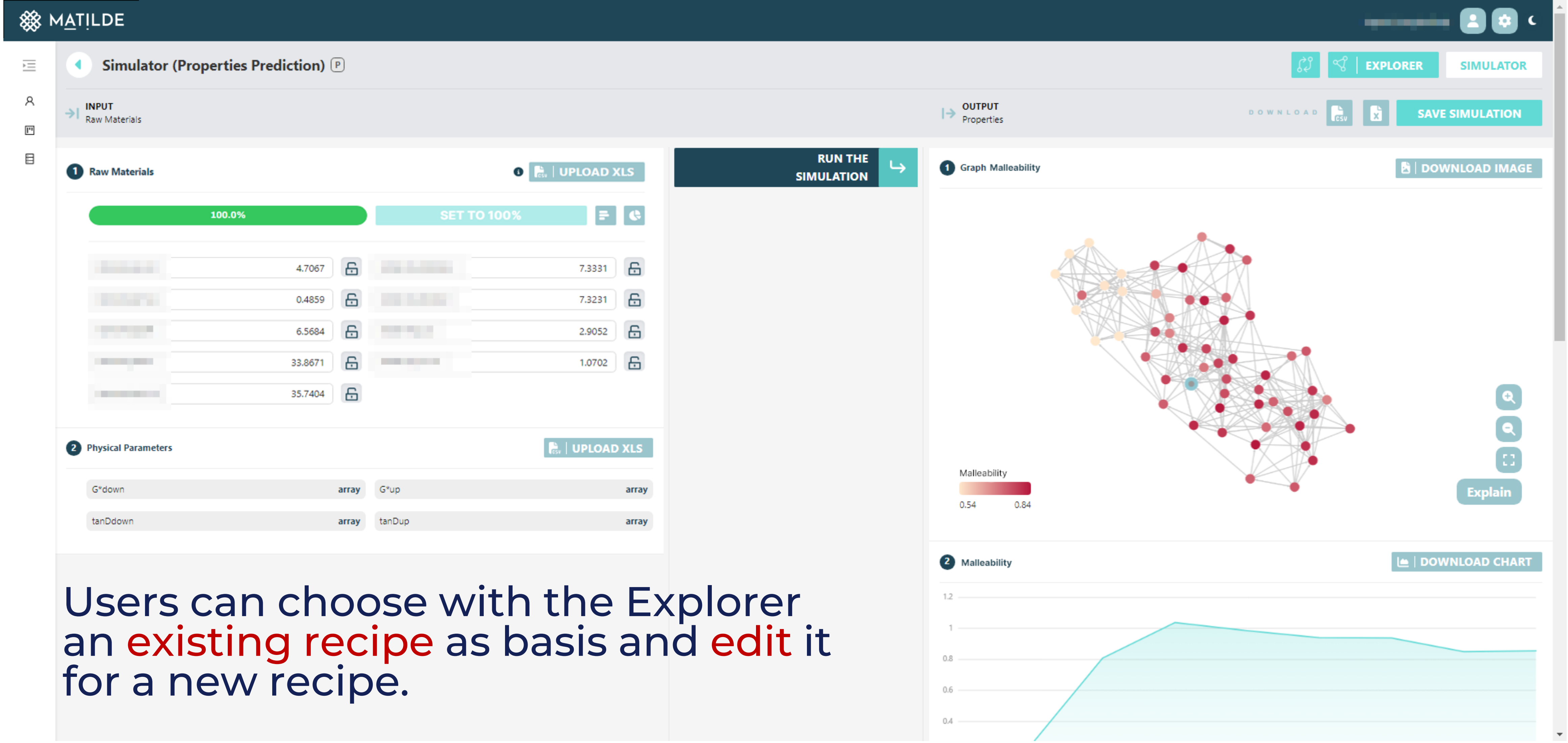
Explorer

Users can select a target node u to:

- See its features
- View its local explanations
- Obtain the subgraph of the top N nodes starting from u using D/BFS.

The app utilizes PageRank to prioritize the subgraph around the most important nodes first.





Users can choose with the Explorer an existing recipe as basis and edit it for a new recipe.

Run a simulation to obtain the predicted outcome and the relative explanation.

Save the simulation and update the model parameters.

intellico
From data to sustainable value

Baselines

EXPERIMENTS

We considered five different baselines:

- MLP: Process only tabular data
- GCN: use a fixed k-nn graph
- GAT: use a fixed k-nn graph
- Graph Transformer (GT): learn graph using attention over complete graph
- DGCNN: the graph is dynamically constructed using nearest neighbors in the feature space

Results

EXPERIMENTS

	Method	Malleability		Toughness	
		RMSE	MAE	RMSE	MAE
Standard approach	MLP	23.40 ± 18.30	21.60 ± 18.90	28.90 ± 16.20	23.80 ± 17.80
Fixed graph	GCN	67.60 ± 03.30	66.50 ± 03.30	70.70 ± 03.00	69.30 ± 03.20
	GAT	60.50 ± 09.20	59.10 ± 09.40	62.70 ± 09.40	60.80 ± 09.50
Structure learning	Graph Transformer	22.30 ± 07.30	20.10 ± 07.00	27.50 ± 03.40	24.80 ± 03.40
	DGCNN	30.60 ± 04.70	28.90 ± 04.90	32.60 ± 05.10	30.80 ± 05.40
	Our model	**14.22 ± 00.43**	**10.54 ± 00.46**	**12.48 ± 00.73**	**09.15 ± 00.29**

Models able to learn a latent graph structure perform better than models with a fixed *k*-NN graph topology

Ablation Study

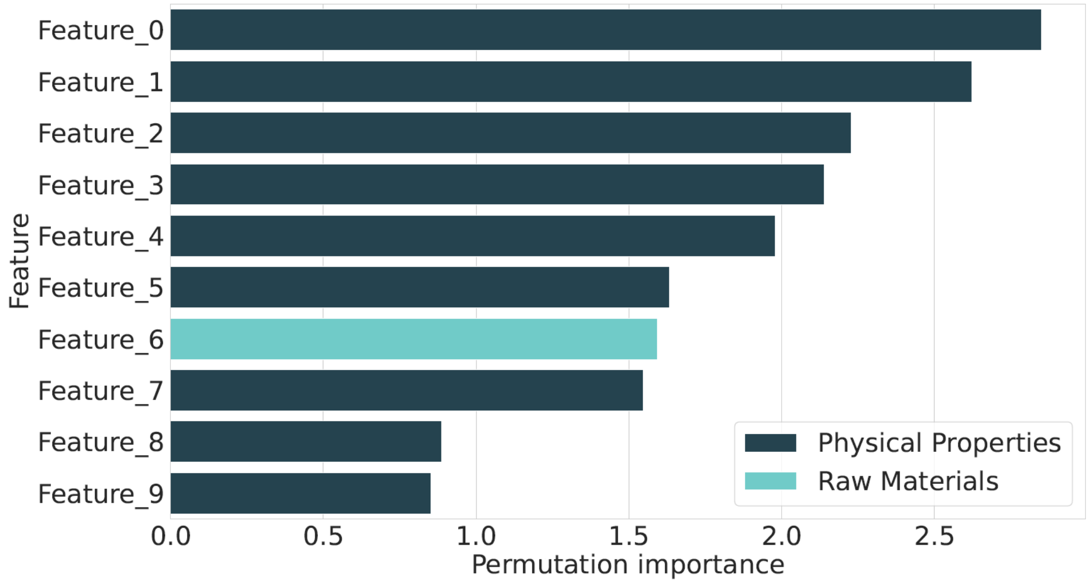
EXPERIMENTS

Model	RMSE
Our model	**12.48 ± 00.73**
Model w/o MLP	14.10 ± 03.05
Model w/o SL	28.07 ± 01.91
Model w/o GNN	28.90 ± 16.20

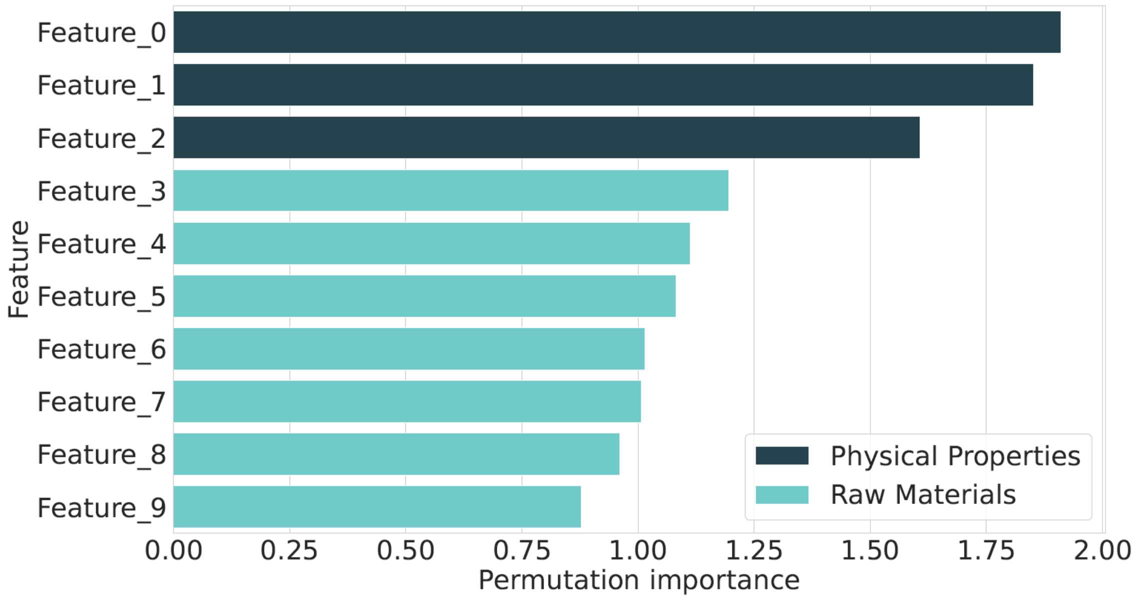
The Graph ML approach plays a crucial role in obtaining the best performance against the baselines

Global-level explanations

EXPERIMENTS



Malleability

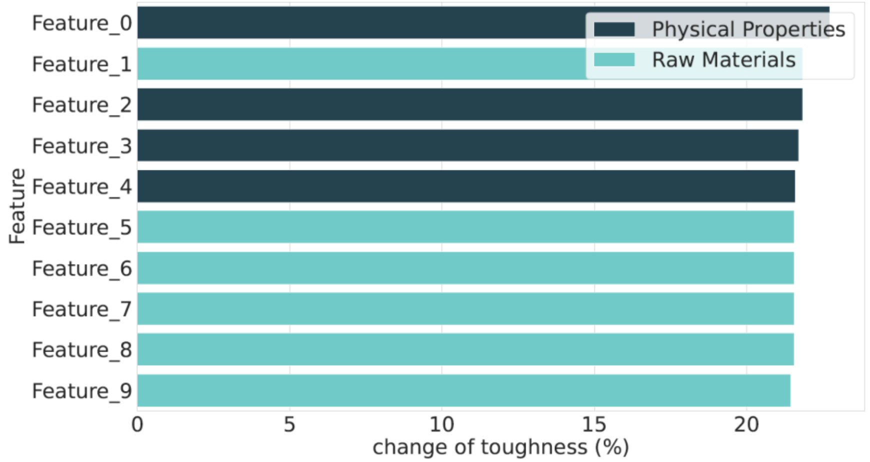


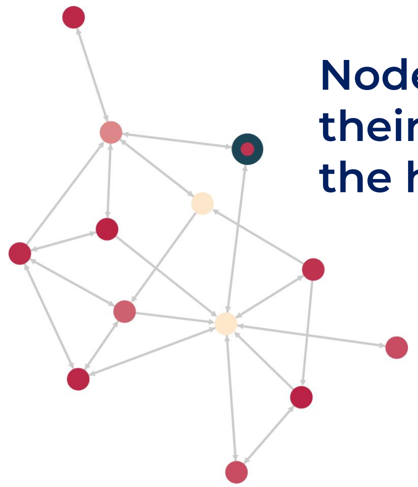
Toughness

Local-level explanations

EXPERIMENTS

Results allow the R&D department to identify new recipes and do the restock of raw materials.



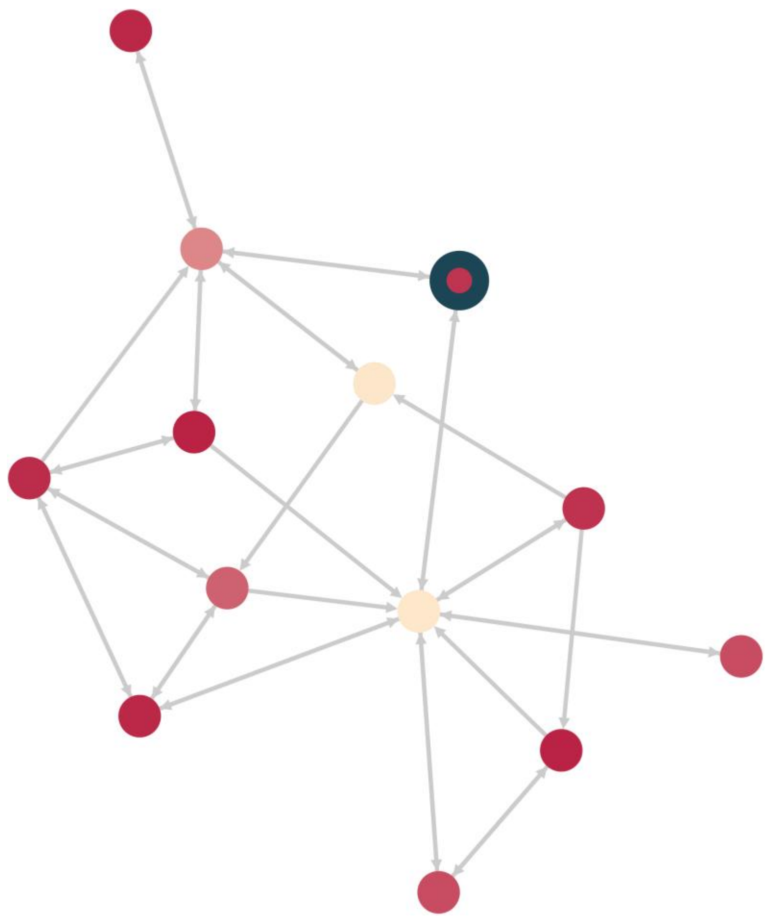


Nodes are colored based on their toughness: the darker, the higher.

The local-level explanations reduce up to 30% the waste of material of the company.

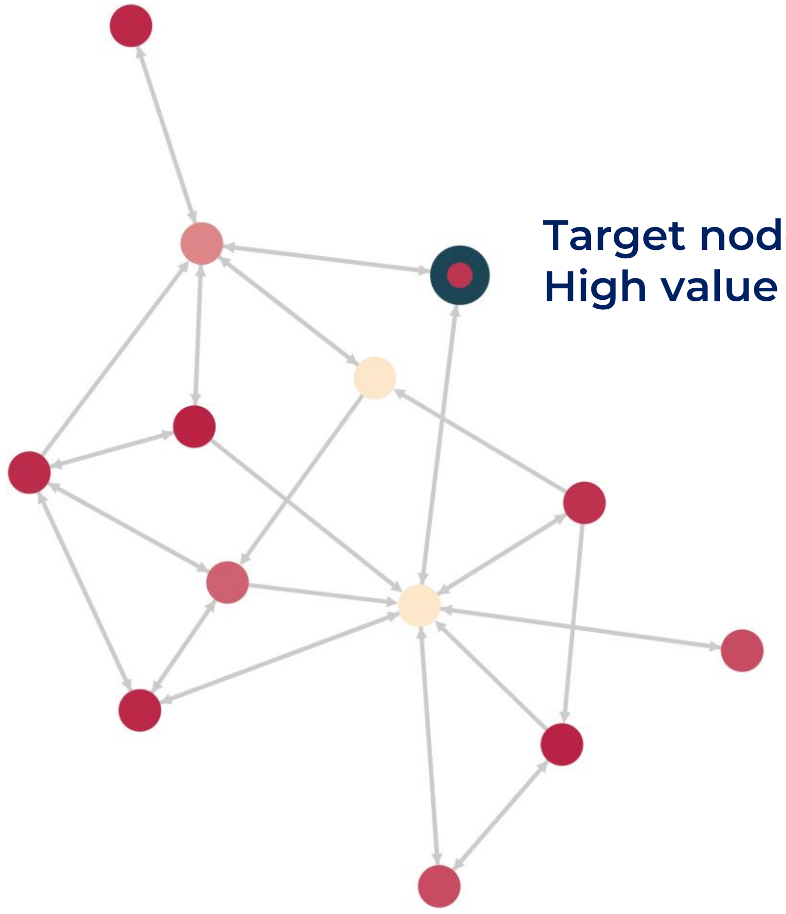
Local-level explanations

The subgraph gives an idea of the effect of recipe changes by exploring the relationships and comparing the formulations



Local-level explanations

The subgraph gives an idea of the effect of recipe changes by exploring the relationships and comparing the formulations

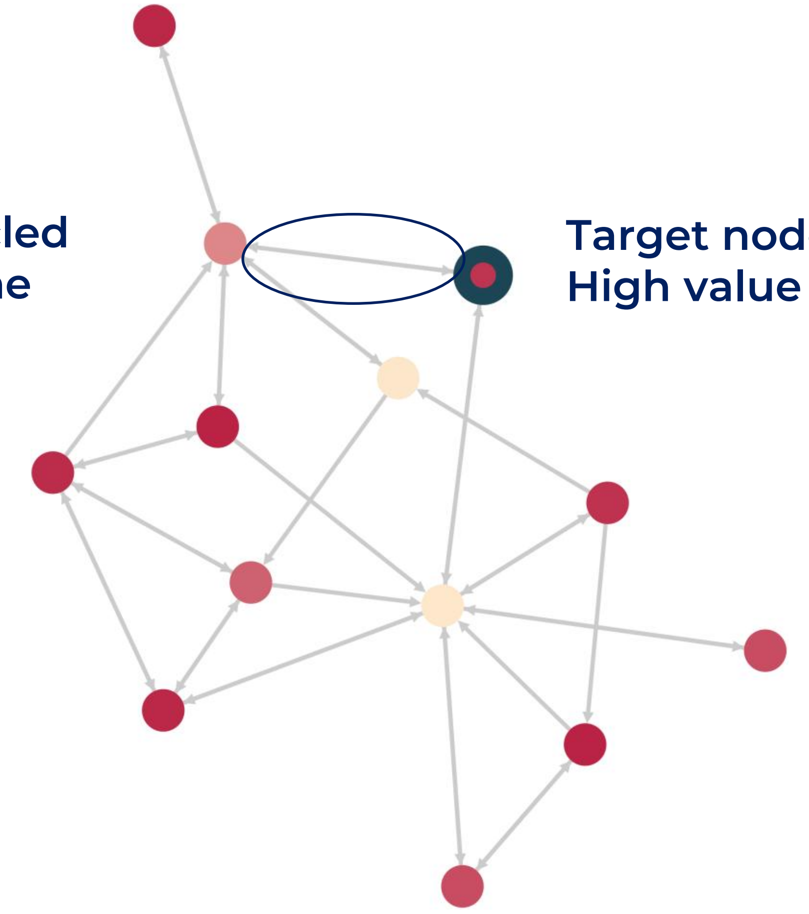


Target node
High value of toughness

Local-level explanations

The subgraph gives an idea of the effect of recipe changes by exploring the relationships and comparing the formulations

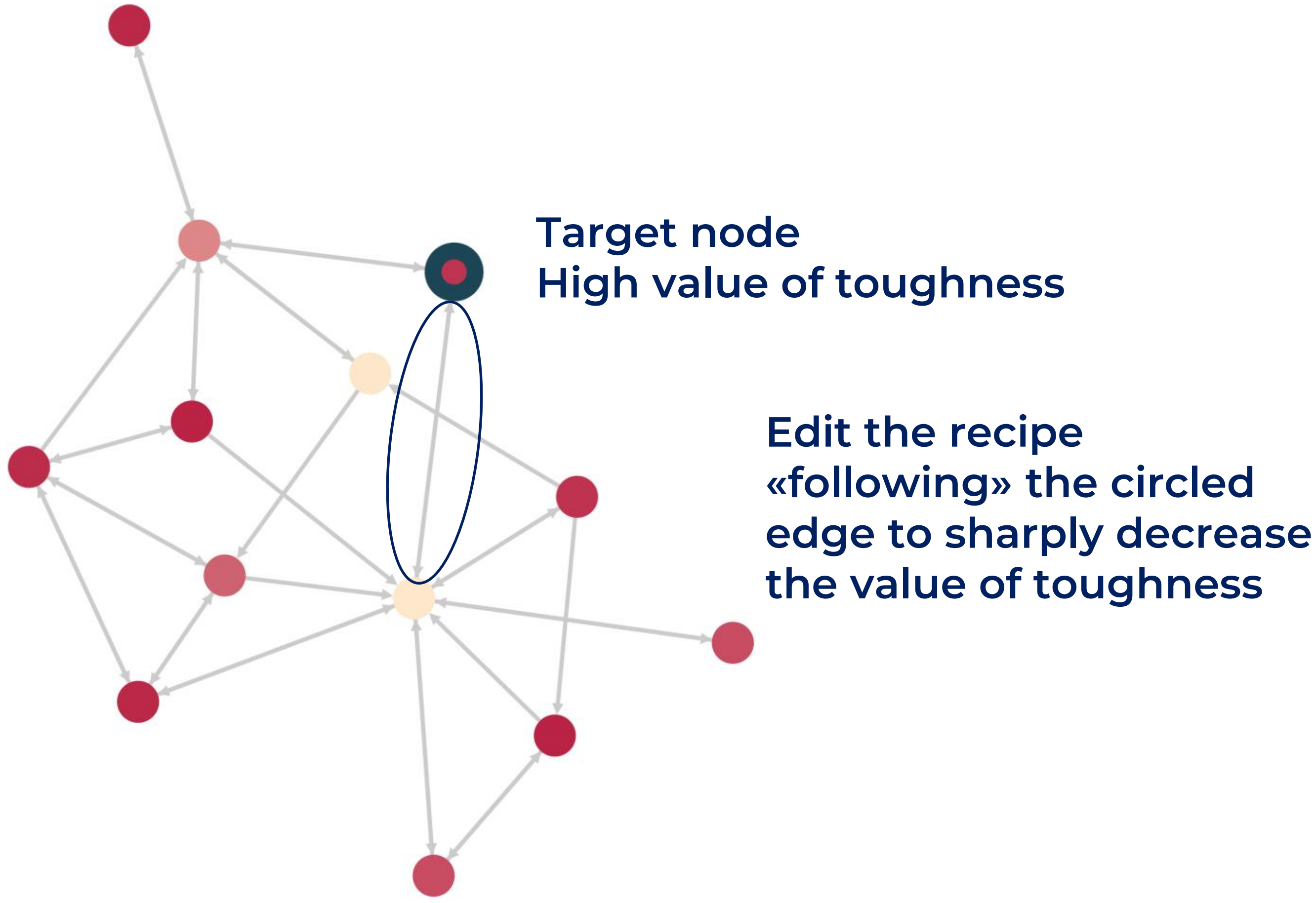
Edit the recipe «following» the circled edge to diminish the value of toughness



Target node
High value of toughness

Local-level explanations

The subgraph gives an idea of the effect of recipe changes by exploring the relationships and comparing the formulations



Conclusions

We propose a data-driven approach for fast product development

- A graph machine-learning model is trained to predict the desired properties of unseen formulations
- XAI techniques based on graphs, perturbation, and sensitivity analysis effectively support R&D in identifying new recipes for reaching a desired property.
- The solution is already deployed and offered through a web application

Thanks for your attention

Contacts
Francesca Saraceni - francesca.saraceni@intellico.ai
Manuel Dileo - manuel.dileo@unimi.it