Université de Strasbourg

EDSC École Doctorale des Sciences Chimiques

Chémoinformatique STRASBOURG

Chimie de la MATIÈRE COMPLEXE UMR 7140 - CMC

# Combinatorial Library Neural Network (CoLiNN) for Combinatorial Library Visualization without Compound Enumeration

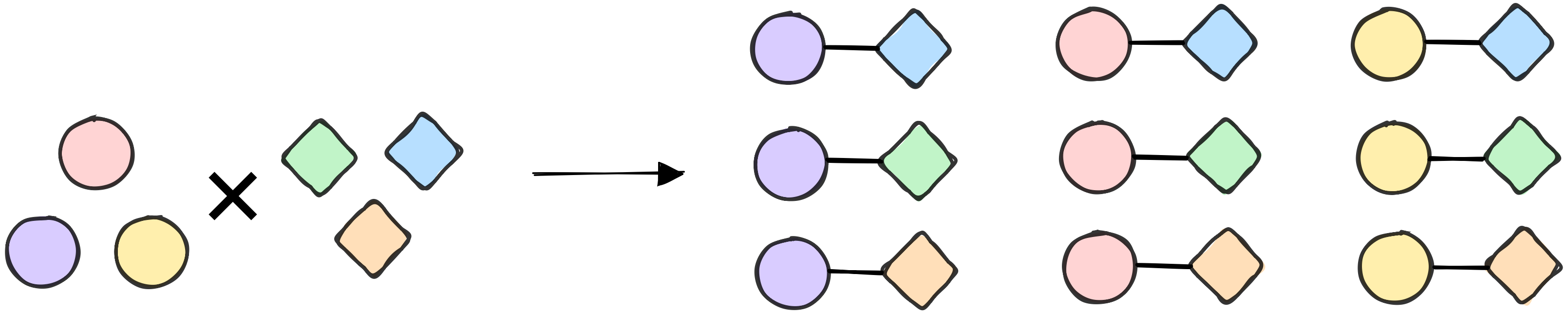
**Dr. Regina PIKALYOVA**

September 20th 2024

ICANN 2024, Lugano, Switzerland

# Combinatorial Chemistry

**Combinatorial chemistry** involves reaction of some or all combinations of diverse reagents according to a common synthetic scheme:



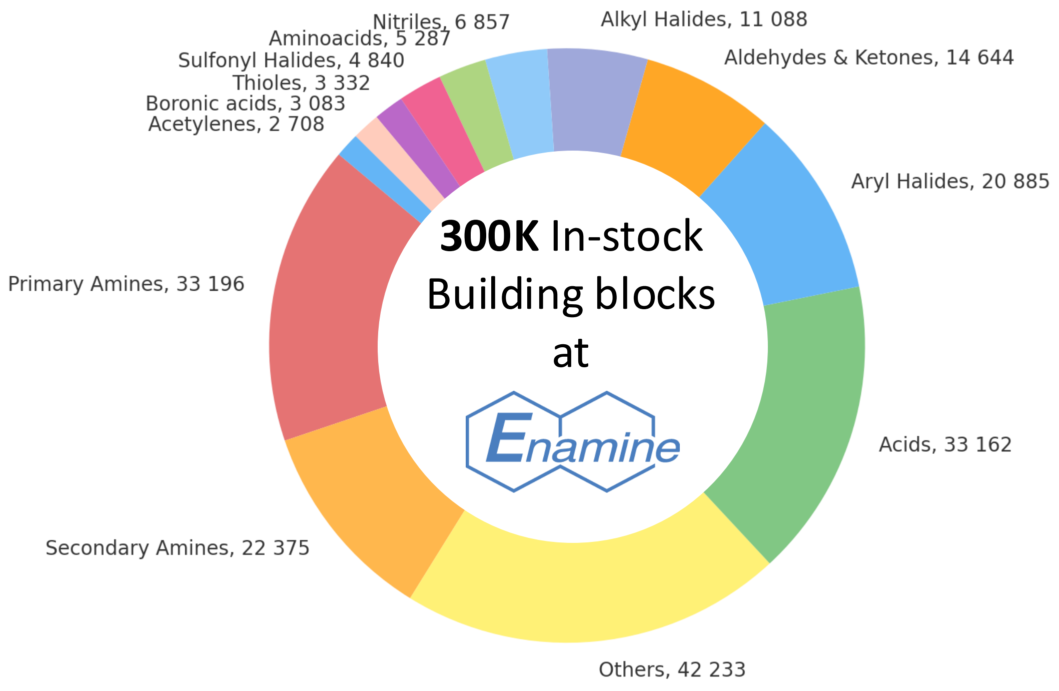
3 x 3 Building Blocks

9 Products

Goodnow et al. A handbook for DNA-encoded chemistry: theory and applications for exploring chemical space and drug discovery, John Wiley & Sons, 2014.

# Combinatorial explosion problem

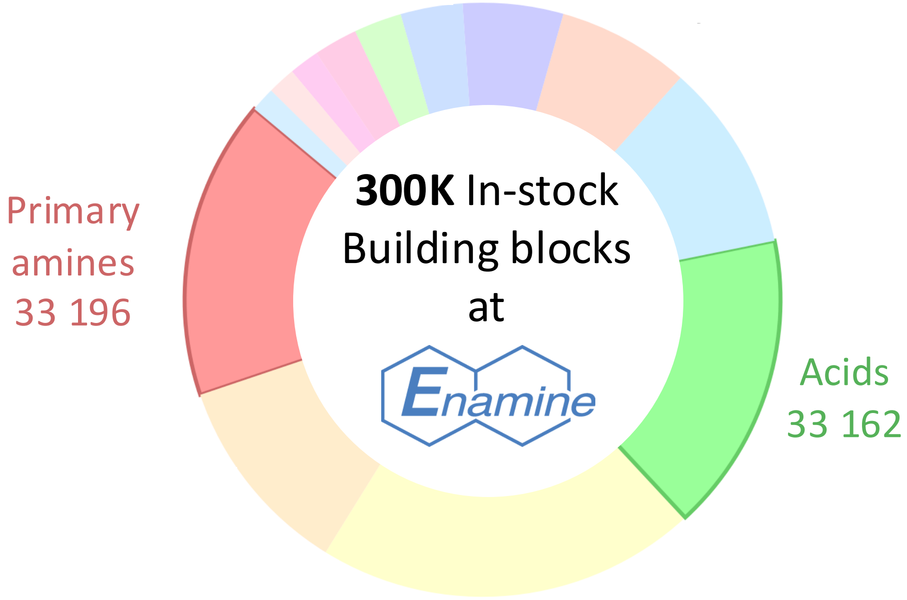
Abundance of commercial building blocks allows to create ultra-large combinatorial compound libraries



https://enamine.net/building-blocks/building-blocks-catalog, acessed 16.09.2024

# Combinatorial explosion problem

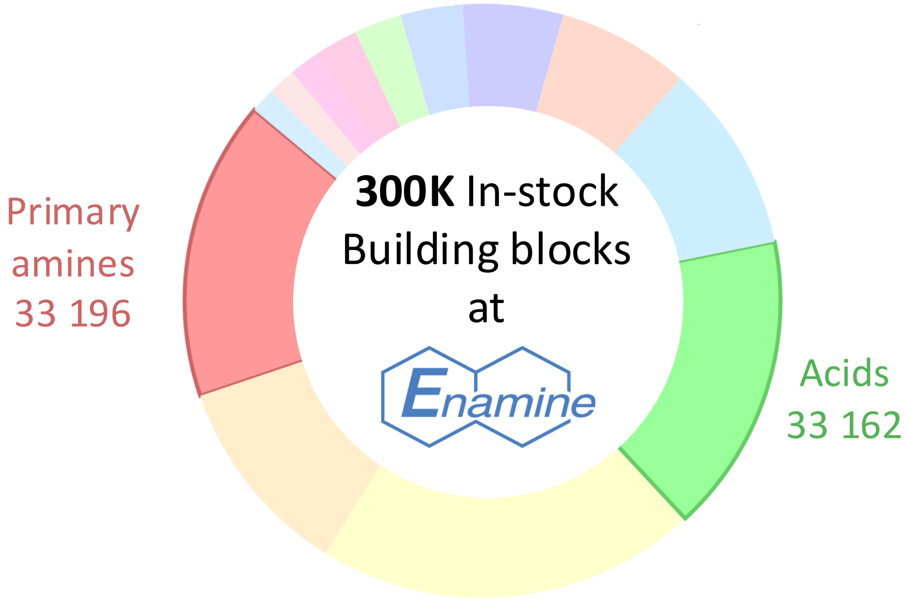
Abundance of commercial building blocks allows to create ultra-large combinatorial compound libraries

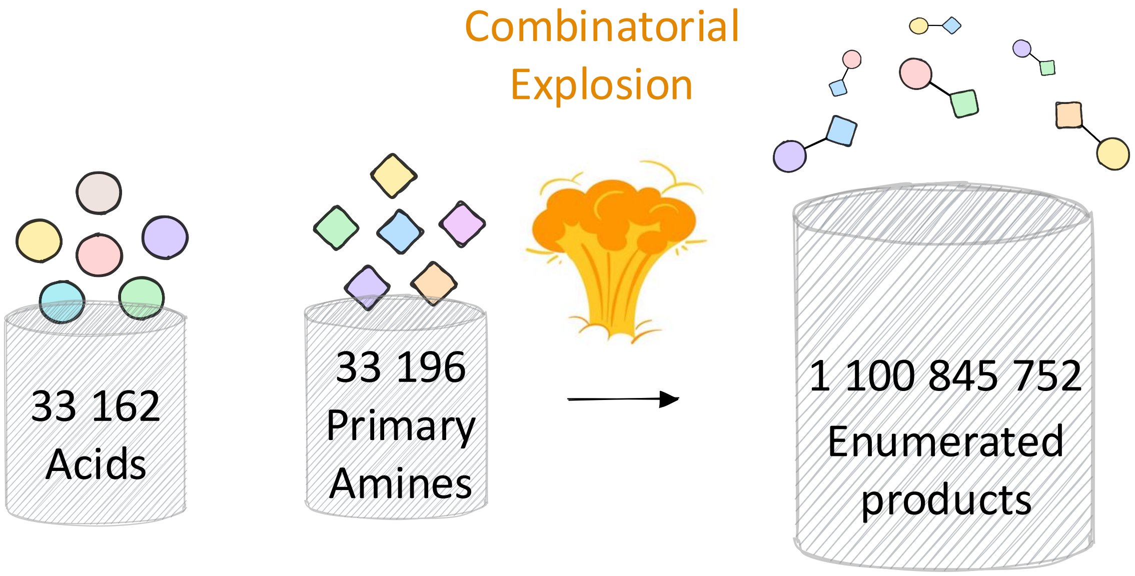


https://enamine.net/building-blocks/building-blocks-catalog, acessed 16.09.2024

# Combinatorial explosion problem

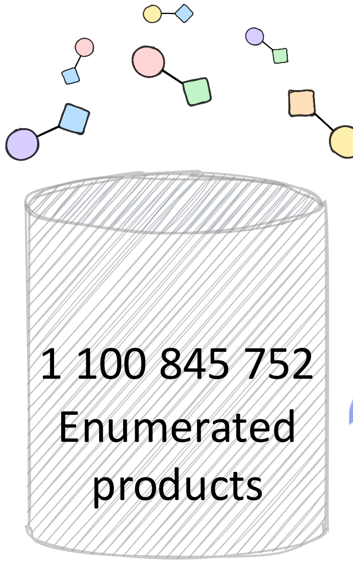
Abundance of commercial building blocks allows to create ultra-large combinatorial compound libraries





https://enamine.net/building-blocks/building-blocks-catalog, acessed 16.09.2024

# Potential of combinatorial libraries



1 100 845 752 Enumerated products

**Advantages:**

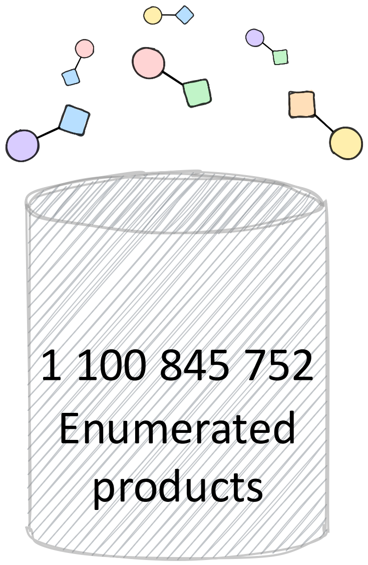
- ~ $10^{12}$ compounds can be synthesized
- Fast experiments
- DNA-encoding allows to screen all compounds at once

↓

Fast exploration of previously uncharted chemical space

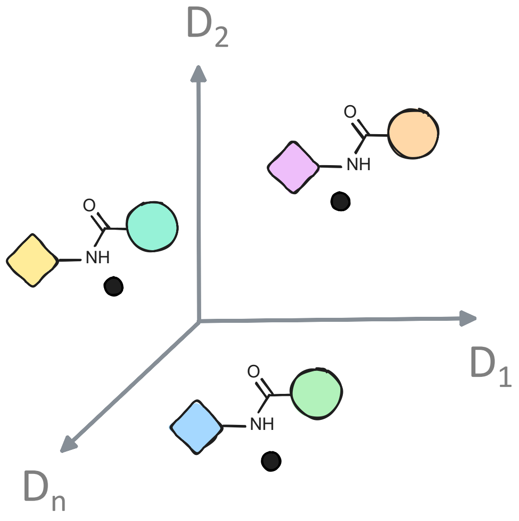
# Analysis of Combinatorial Libraries

Interpretable navigation of the chemical space of a combinatorial library using dimensionality reduction



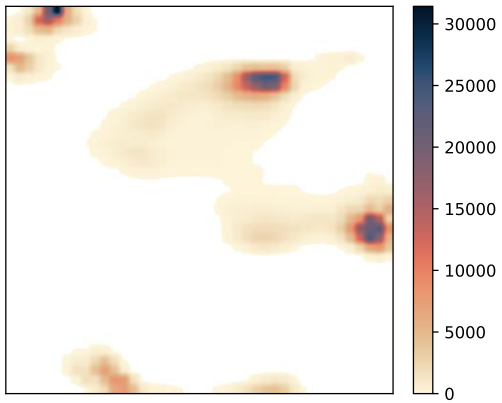
1 100 845 752 Enumerated products

Representation in the N-dimensional descriptor space

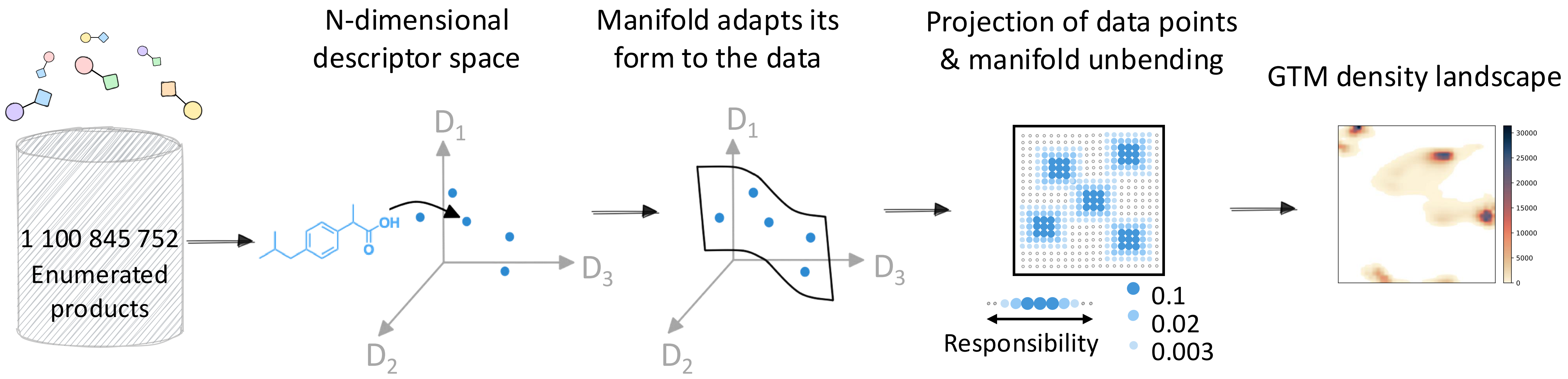


Dimensionality reduction

Chemical space map



# Generative Topographic Mapping



1 100 845 752 Enumerated products

N-dimensional descriptor space

Manifold adapts its form to the data

Projection of data points & manifold unbending

GTM density landscape

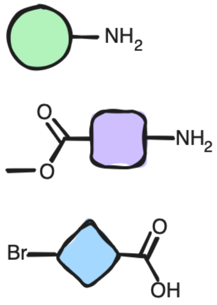
Responsibility

● 0.1
● 0.02
● 0.003

- ✓ Intuitive navigation
- ✓ Fast comparison to other libraries
- ✓ Big Data compatibility

# Workflow of combinatorial library analysis

Building blocks:



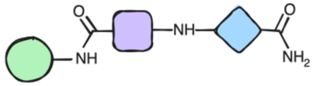
Reactions:

Rxn 11

Rxn 25

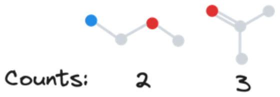
Rxn 18

1. Enumeration

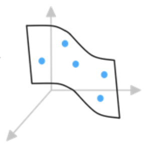


2. Standardization

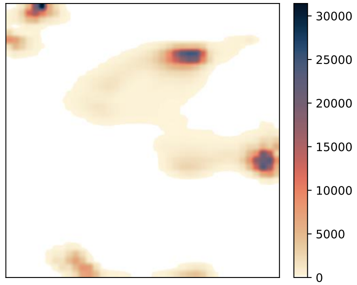
3. Descriptor calculation



4. Dimensionality reduction



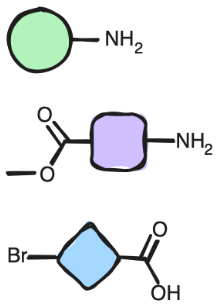
Chemical space map



# Workflow of combinatorial library analysis

Building blocks:

Reactions:

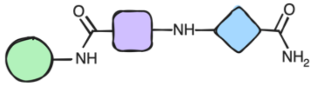


Rxn 11

Rxn 25

Rxn 18

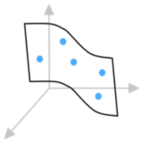
1. Enumeration



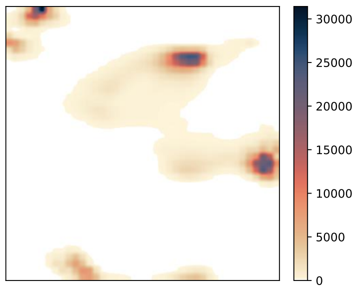
2. Standardization

3. Descriptor calculation

4. Dimensionality reduction

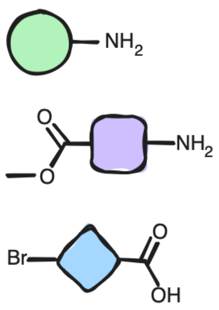


Chemical space map



# Workflow of combinatorial library analysis

Building blocks:

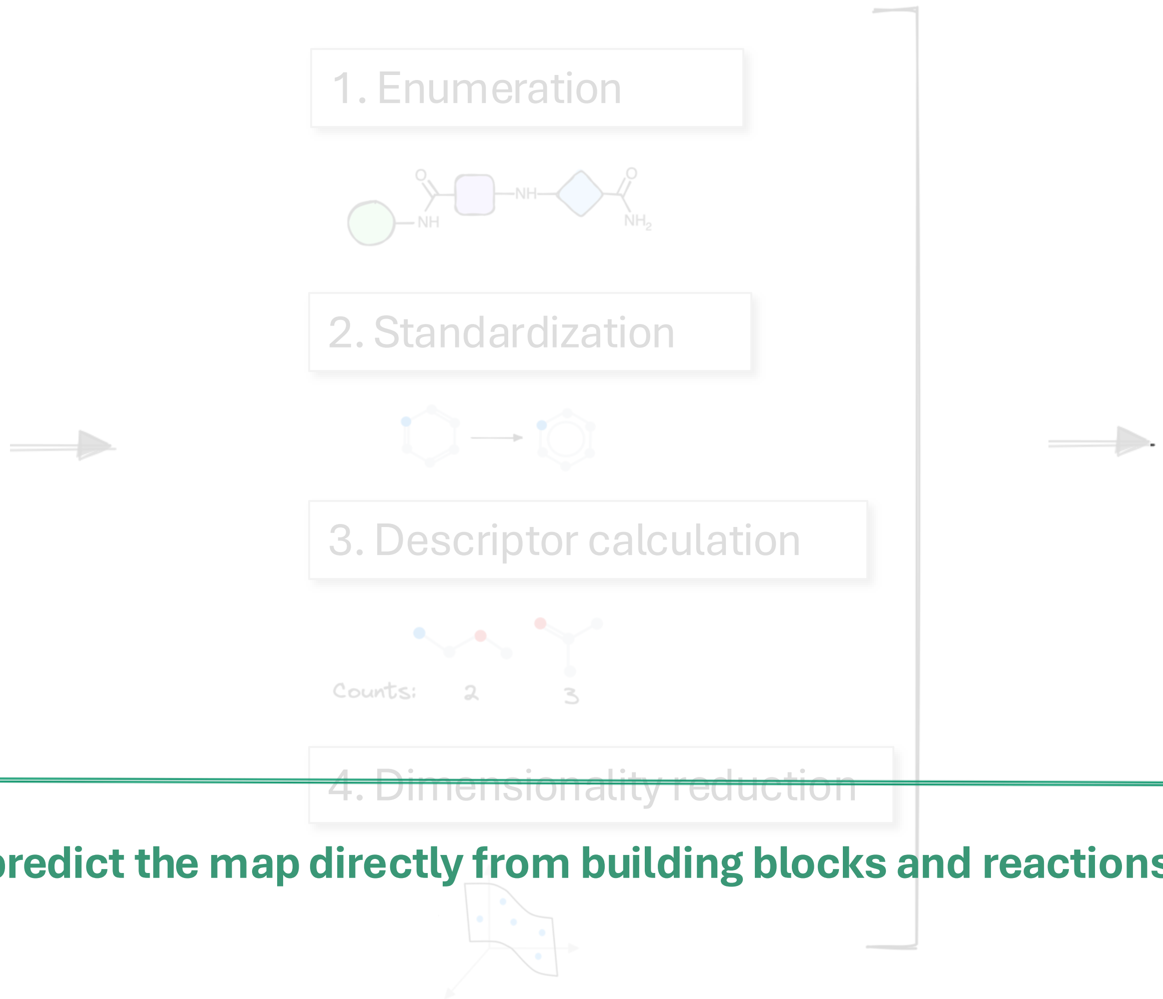


Reactions:

Rxn 11

Rxn 25

Rxn 18



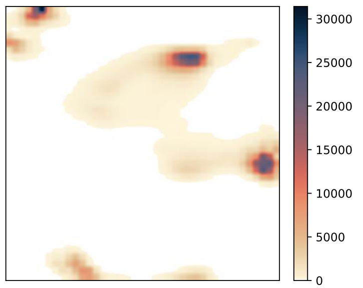
1. Enumeration

2. Standardization

3. Descriptor calculation

4. Dimensionality reduction

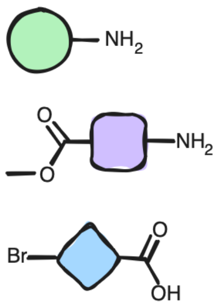
Chemical space map



How to predict the map directly from building blocks and reactions?

# CoLiNN – Neural Network for combinatorial library analysis without compound enumeration

Building blocks:



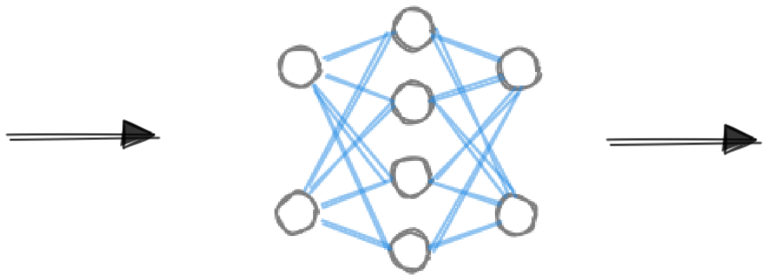
Reactions:

Rxn 11

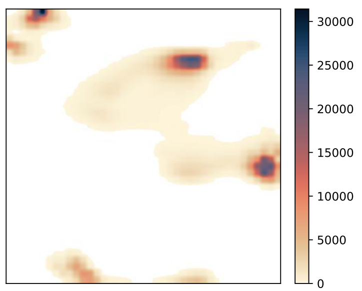
Rxn 25

Rxn 18

Combinatorial Library Neural Network (CoLiNN)



Generative Topographic Map of the combinatorial library



CoLiNN **skips the enumeration step**, making the process of combinatorial library visualization **faster and simpler**

# CoLiNN – Neural Network for combinatorial library analysis without compound enumeration

1. **Building Block Embedding Creation**

2. **Reaction Embedding Creation**

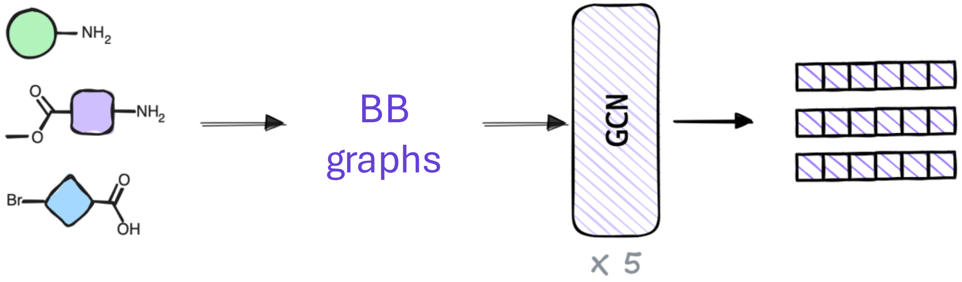
3. **Responsibility vector prediction**

# CoLiNN – Neural Network for combinatorial library analysis without compound enumeration

### 1. Building Block Embedding Creation

### 2. Reaction Embedding Creation

Building blocks:



BB graphs

GCN

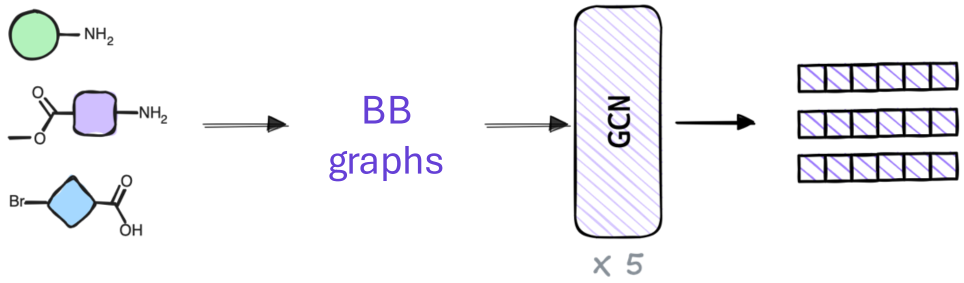
× 5

### 3. Responsibility vector prediction

# CoLiNN – Neural Network for combinatorial library analysis without compound enumeration

## 1. Building Block Embedding Creation

Building blocks:



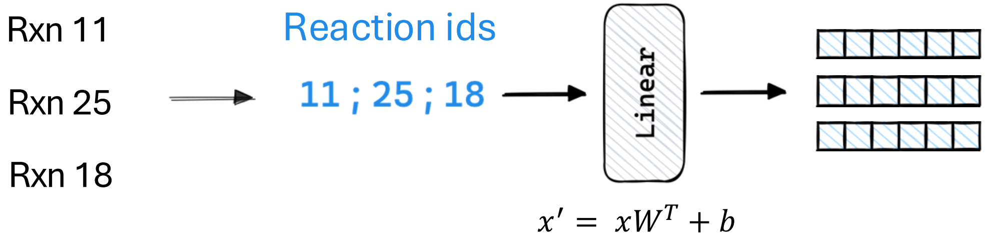
BB graphs

GCN

× 5

## 2. Reaction Embedding Creation

Reactions:



Rxn 11

Rxn 25

Rxn 18

Reaction ids

11 ; 25 ; 18

Linear

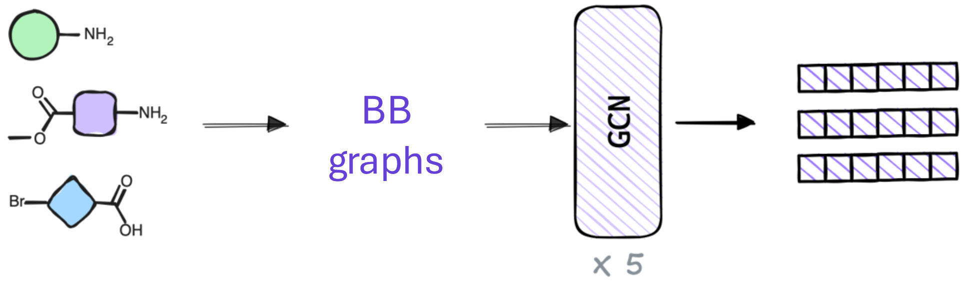
$$x' = xW^T + b$$

## 3. Responsibility vector prediction

# CoLiNN – Neural Network for combinatorial library analysis without compound enumeration

## 1. Building Block Embedding Creation

Building blocks:



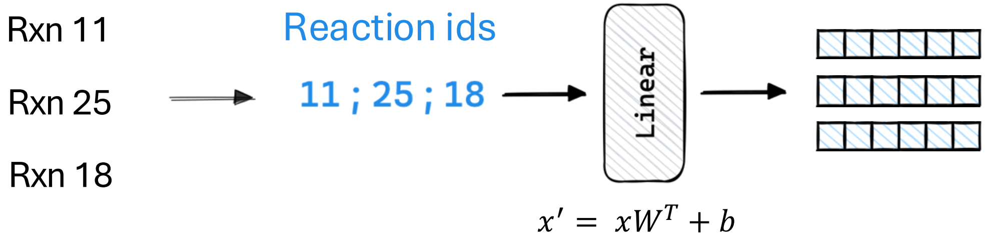
BB graphs

GCN

× 5

## 2. Reaction Embedding Creation

Reactions:



Rxn 11

Rxn 25

Rxn 18

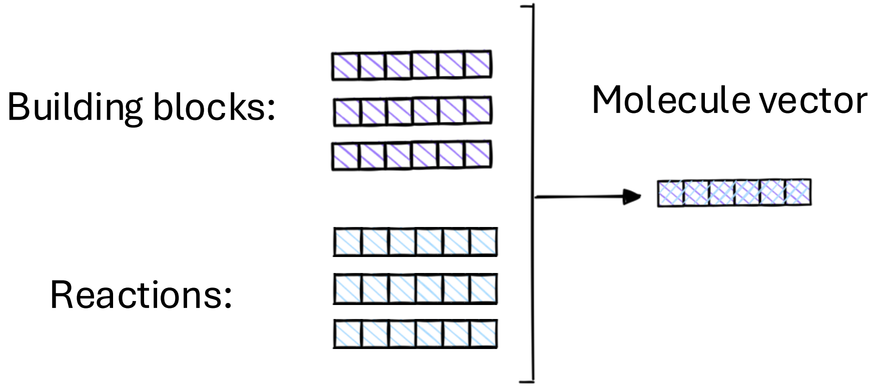
Reaction ids

11 ; 25 ; 18

Linear

$x' = xW^T + b$

## 3. Responsibility vector prediction



Building blocks:

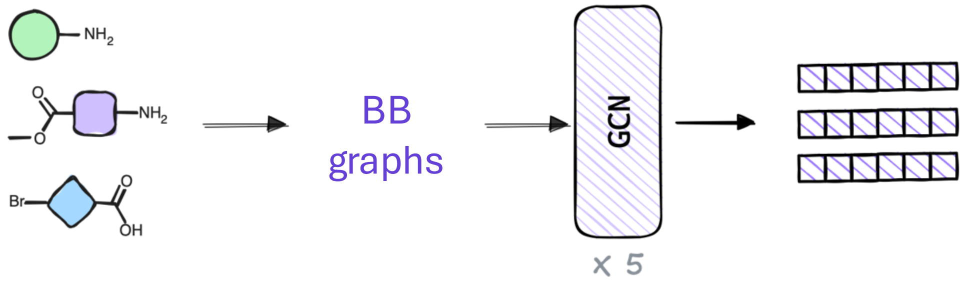
Reactions:

Molecule vector

# CoLiNN – Neural Network for combinatorial library analysis without compound enumeration

## 1. Building Block Embedding Creation

Building blocks:



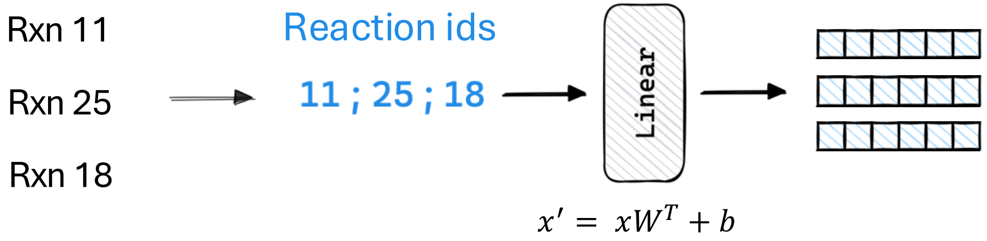
BB graphs

GCN

× 5

## 2. Reaction Embedding Creation

Reactions:



Rxn 11

Rxn 25

Rxn 18

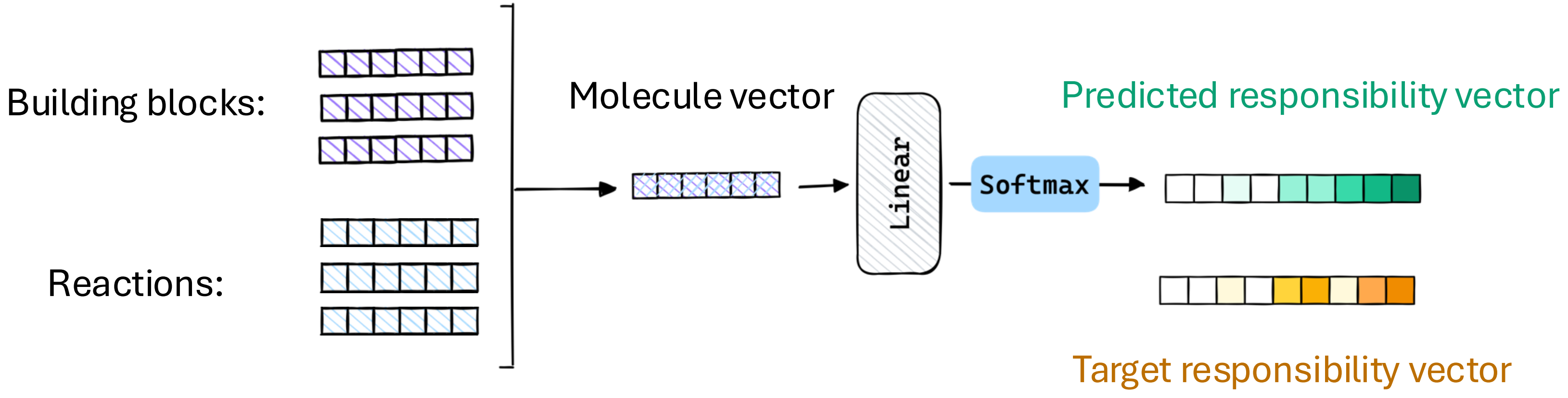
Reaction ids

11 ; 25 ; 18

Linear

$x' = xW^T + b$

## 3. Responsibility vector prediction



Building blocks:

Reactions:

Molecule vector

Linear

Softmax

Predicted responsibility vector

Target responsibility vector

**Loss:**

Kullback Leibler divergence = 0.7

0 – perfect match

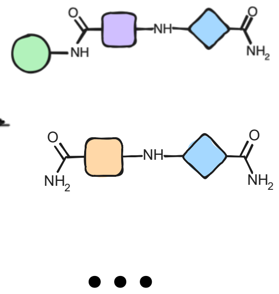
∞ - worst case scenario

# Training set for general-chemistry sensitive CoLiNN model

A general CoLiNN model was trained on 388 DELs based on diverse reaction schemes:

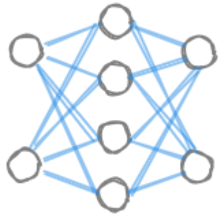
**Training set :** 388 DELs

**3M 880K compounds**



| Input to CoLiNN: | Target Resp. Vector: |
|---|---|
| Reaction ids, BB graphs | 🟦🟦🟦🟦🟦 |
| Reaction ids, BB graphs | 🟧🟧🟧🟧🟧 |
| • • • | • • • |

General CoLiNN



**Optimal training time & performance:**

Training time: 13 h

Validation KL div. loss: 0.79

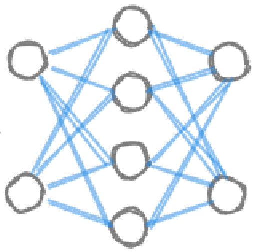
# CoLiNN – Gain in time

**Input:**

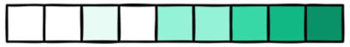
**Physical time needed for prediction:**

Reaction ids & BB graphs

**CoLiNN**



Predicted responsibility vector



0.055 ms/molecule/GPU

✓ **7000-fold acceleration compared to traditional workflow**

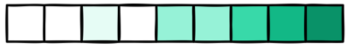
**VS**

**Traditional workflow:**

1. Enumeration
2. Standardization
3. Descriptor calculation
4. Projection on the map

Building Block SMILES

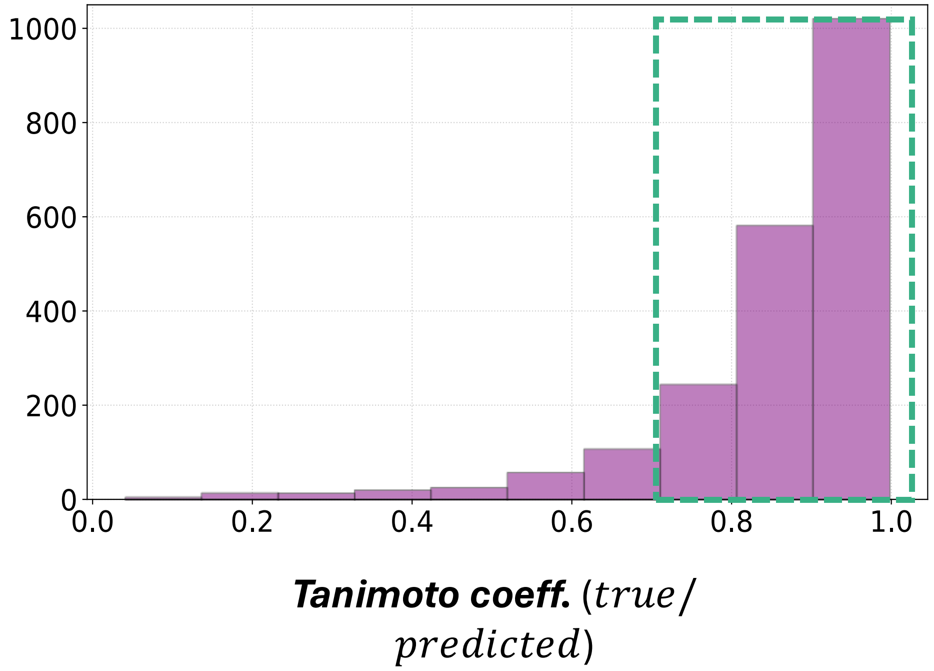
Responsibility vector



395 ms/molecule/CPU

# Predictions for test set DELs

Similarity between predicted and true maps of 2089 DELs from the external test set:

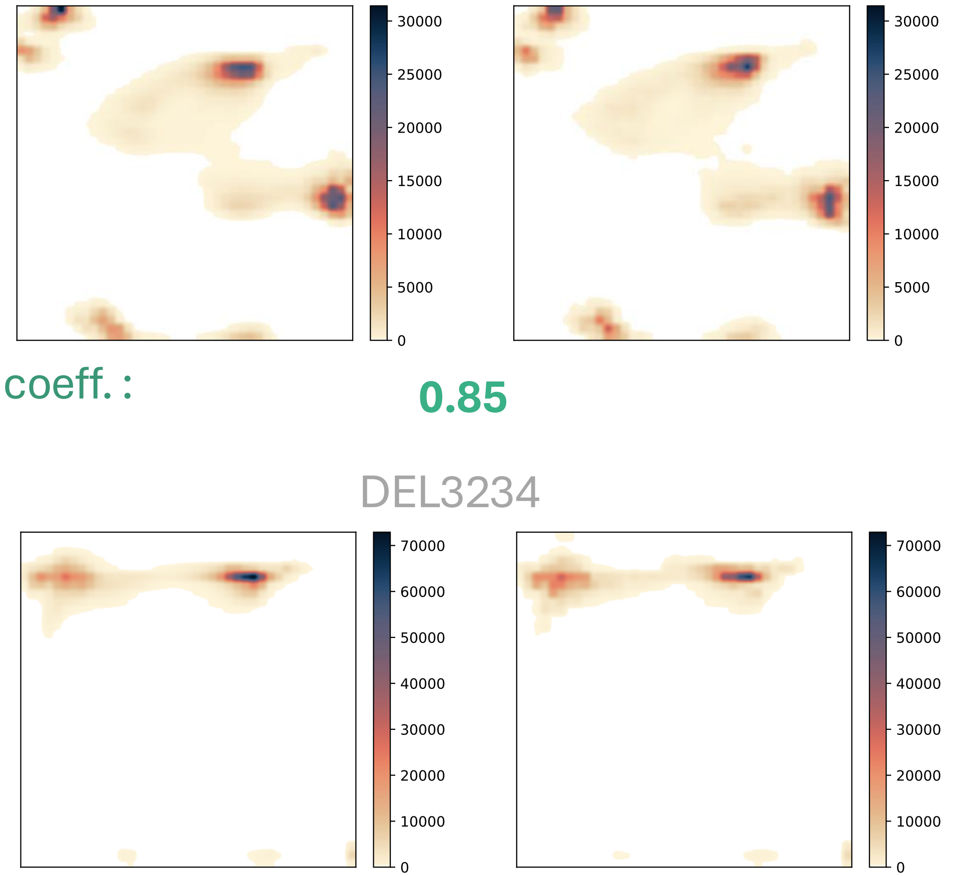


For the majority of test set DELs the predicted maps are nearly identical to the true ones

GTM

CoLiNN

DEL3827



Tanimoto coeff. :

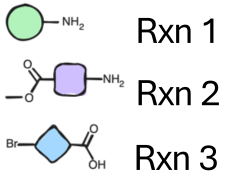
0.85

DEL3234

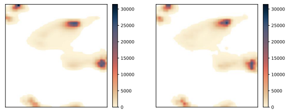
0.92

# Conclusions

1. CoLiNN predicts compound projections on the GTM using only their building blocks and reactions, **skipping the compound enumeration**



2. The **predicted maps** for external test set DELs **are very similar to the true GTM-derived ones**



3. CoLiNN achieves **7000-fold acceleration** compared to the enumeration-based workflow

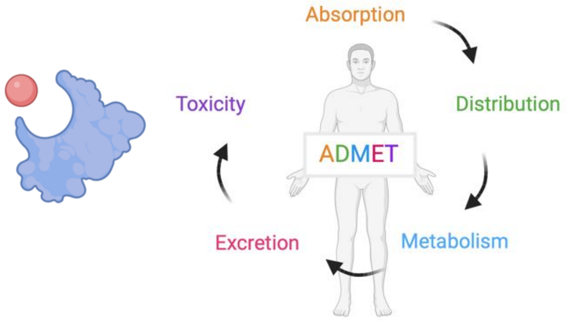
**7000-fold faster**

# Perspectives

- Different reaction representation

SMIRKS / Reaction SMARTS / CGR

- Prediction of molecular properties without structure enumeration



# Thank you for your attention!

Université de Strasbourg

EDSC École Doctorale des Sciences Chimiques

Chémoinformatique STRASBOURG

Chimie de la MATIÈRE COMPLEXE UMR 7140 - CMC

**Acknowledgements:**

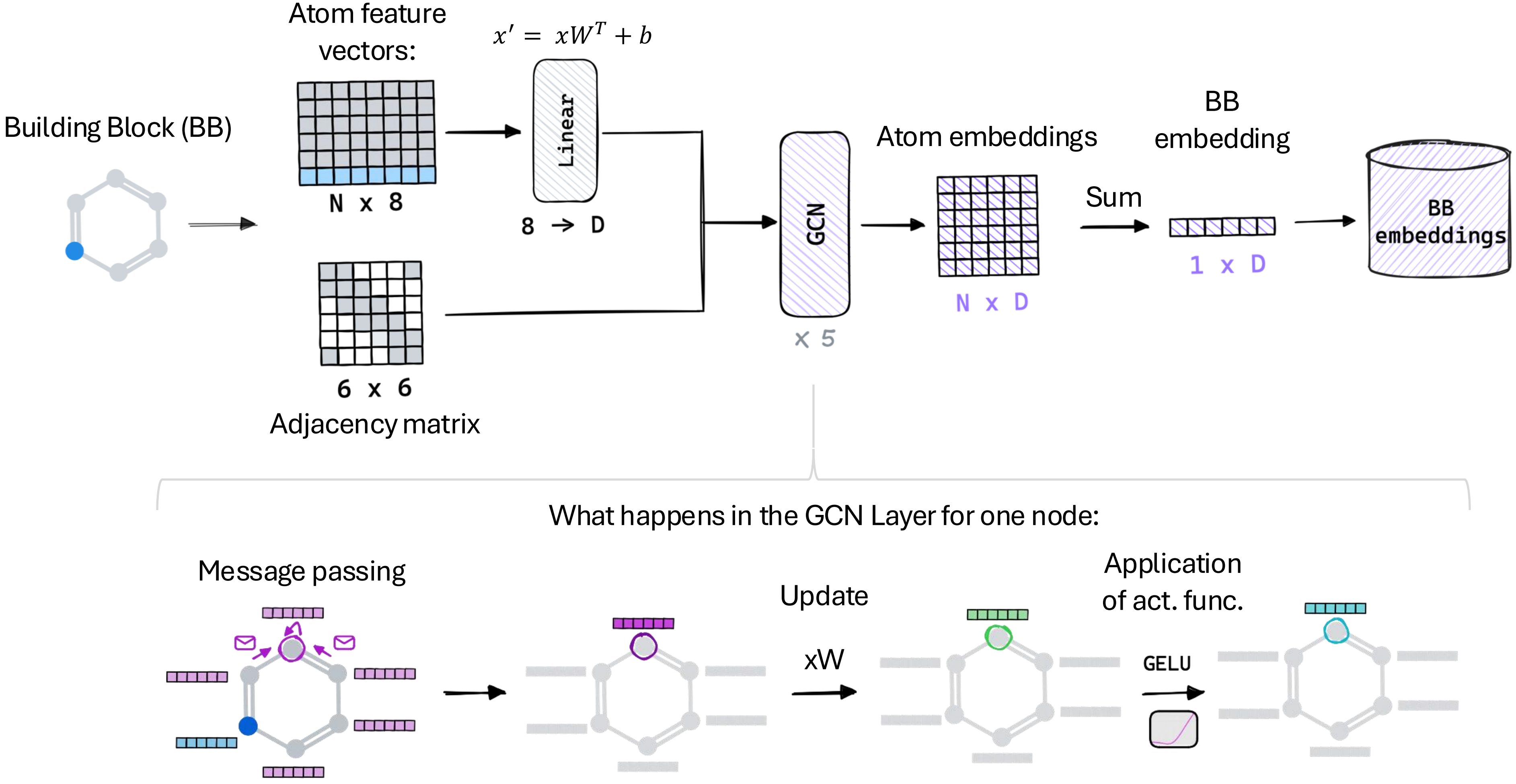
Dr. Tagir Akhmetshin Prof. Alexandre Varnek Dr. Dragos Horvath

# BB graphs' and embeddings' creation

During CoLiNN training we save BB graphs and later we save BB embeddings that will be further used for inference

N – Number of atoms (nodes)

D – Vector dimension



Building Block (BB)

Atom feature vectors:

$x' = xW^T + b$

Linear

N x 8

8 → D

6 x 6

Adjacency matrix

GCN

x 5

Atom embeddings

N x D

Sum

BB embedding

1 x D

BB embeddings

What happens in the GCN Layer for one node:

Message passing

Update

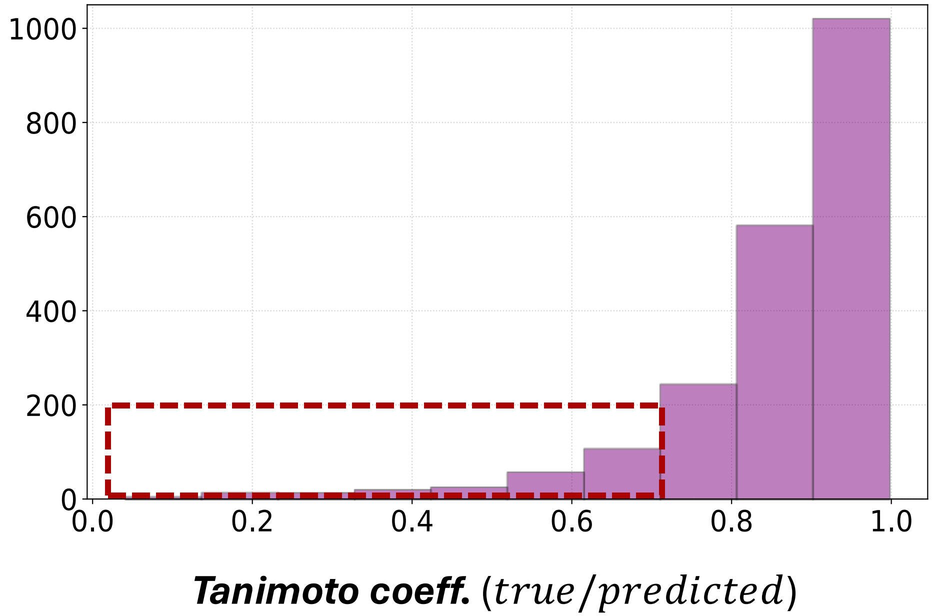
xW

Application of act. func.

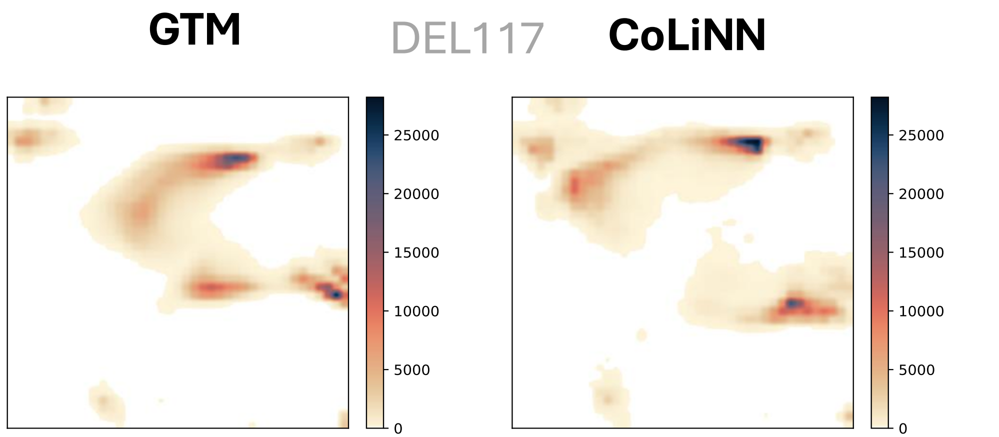
GELU

# Predictions for test set DELs

Similarity between predicted and true maps of 2089 DELs from the external test set:

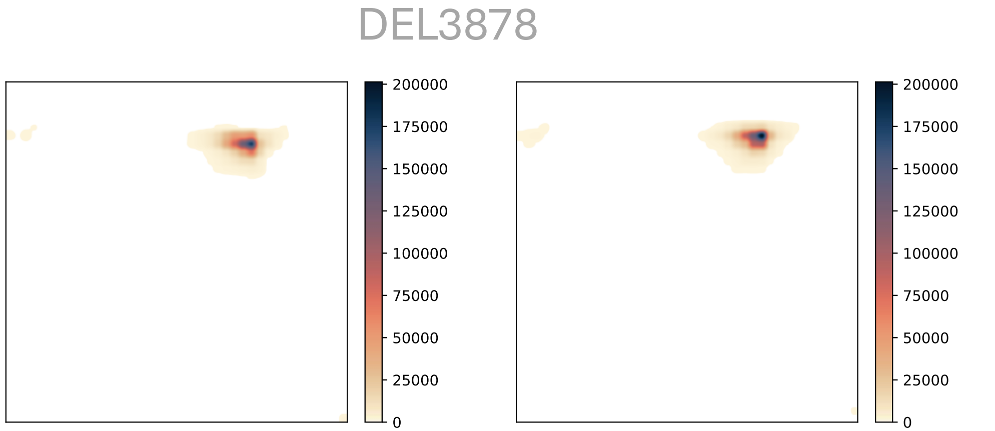


For some DELs predictions are of low quality



Tanimoto coeff. :

**0.07**



**0.56**

**Pikalyova R.**, T. Akhmetshin et al., et al. ChemRxiv (2024).

# GELU Activation function

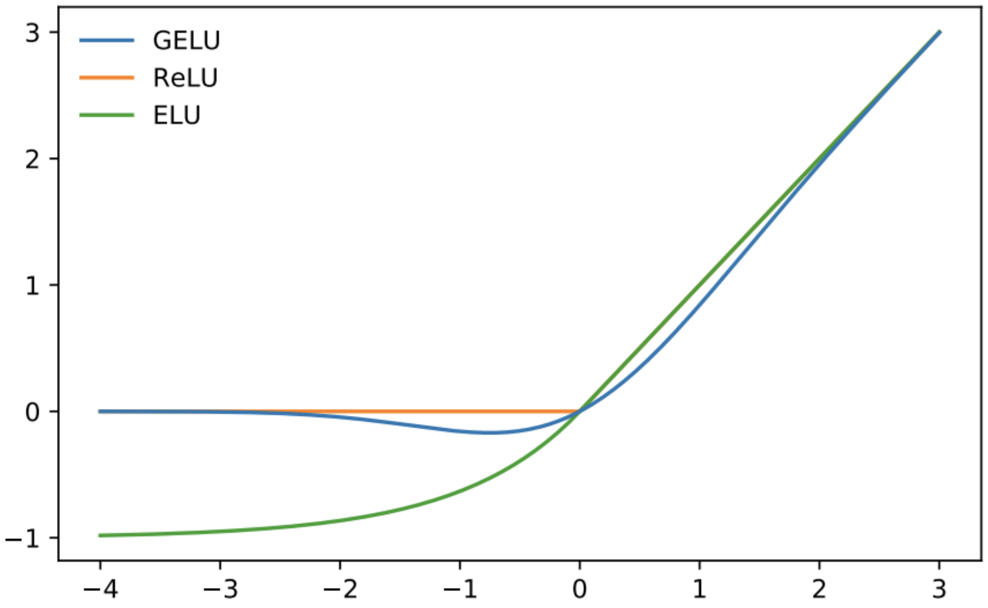


Figure 1: The GELU ($\mu = 0, \sigma = 1$), ReLU, and ELU ($\alpha = 1$).

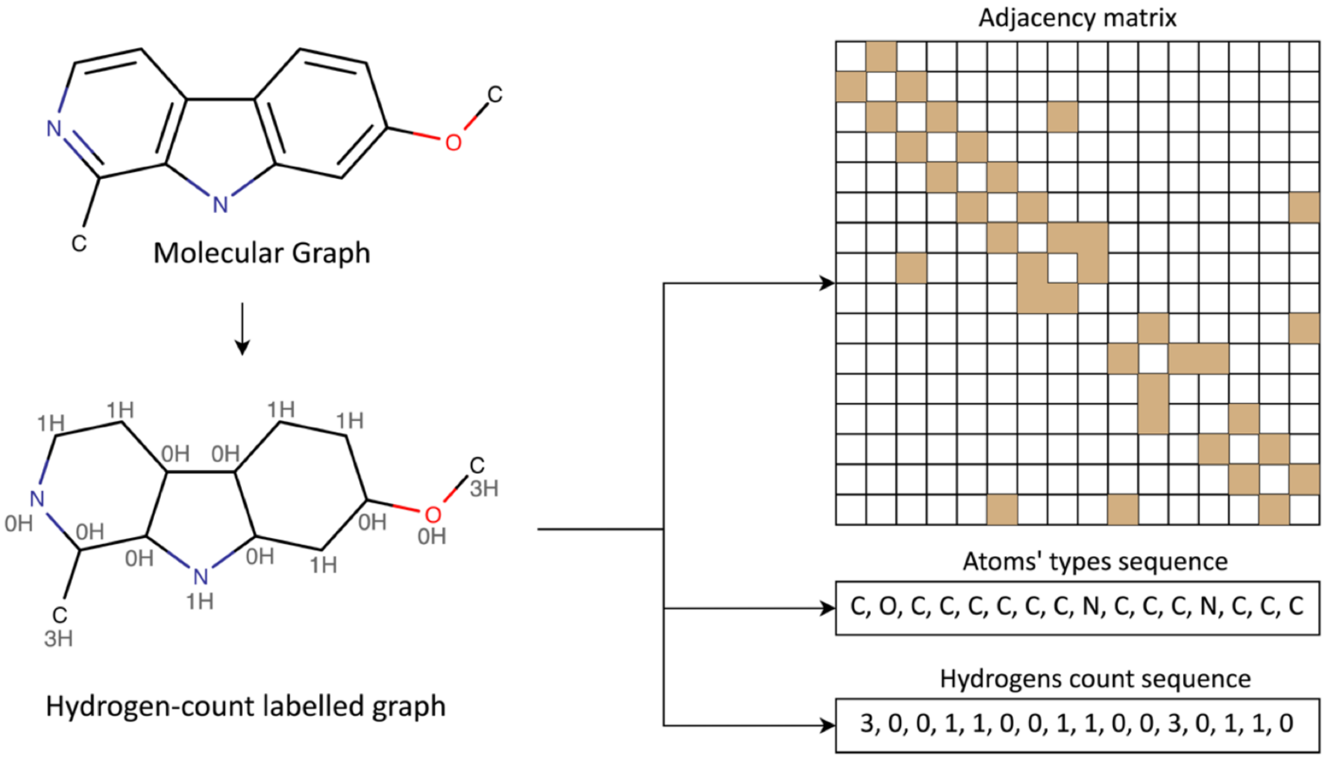
- CoLiNN is coupled with 5 Graph Convolution Network layers where GELU (Gaussian Linear Unit) is used as an activation function

- GELU is used instead of original ReLU since it is more smooth than ReLU and is differentiable at every point leading to the improved gradient flow during backpropagation and decrease in the number of dead neurons (that do not contribute to learning)

# Hydrogen-count labelled graph

- A molecular graph can be represented by three objects: 1) a vector of atom types, 2) a vector of the numbers of attached hydrogens, and 3) a single binary adjacency matrix.

- The number of hydrogens attached to each heavy atom is used instead of bond order.



Molecular Graph

Hydrogen-count labelled graph

Adjacency matrix

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ▨ | | | | | | | | | | | | | | |
| ▨ | | ▨ | | | | | | | | | | | | | |
| | ▨ | | ▨ | | | | ▨ | | | | | | | | |
| | | ▨ | | ▨ | | | | | | | | | | | |
| | | | ▨ | | ▨ | | | | | | | | | | |
| | | | | ▨ | | ▨ | | | | | | | | | ▨ |
| | | | | | ▨ | | ▨ | ▨ | | | | | | | |
| | | ▨ | | | | ▨ | | ▨ | | | | | | | |
| | | | | | | ▨ | ▨ | | | | | | | | |
| | | | | | | | | | | ▨ | | | | | ▨ |
| | | | | | | | | | ▨ | | ▨ | ▨ | | | |
| | | | | | | | | | | ▨ | | | | | |
| | | | | | | | | | | ▨ | | | ▨ | | |
| | | | | | | | | | | | | ▨ | | ▨ | |
| | | | | | | | | | | | | | ▨ | | ▨ |
| | | | | | ▨ | | | | ▨ | | | | | ▨ | |

Atoms' types sequence

C, O, C, C, C, C, C, C, N, C, C, C, N, C, C, C

Hydrogens count sequence

3, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 3, 0, 1, 1, 0

## Advantages:

- Reduces the amount of GPU memory required to store the model as well as training time without loss of accuracy.
- Instead of three or four bond- type-specific adjacency matrices and specific trainable weight matrices (necessary for relational GCNs that leads to too much memory + numerous math. operations) the number of hydrogens attached to each heavy atom is used instead of bond order.
- By using H atom numbers instead of bond orders, functional groups standardization and aromatization steps can be omitted.

Akhmetshin T, et al. HyFactor: Hydrogen-count labelled graph-based defactorization Autoencoder. ChemRxiv. 2021; doi:10.26434/chemrxiv-2021-18x0d

# Why Graph Representation for a molecule?

- **Intuitive and Natural Representation**: Molecules are inherently graph-like structures, where atoms are nodes and bonds are edges.

- **Independent of atom ordering:** Unlike some other methods, such as SMILES strings, which require a specific ordering of atoms, graph representations are independent of the atom's ordering.

- **Handling Cyclic Structures**: Graph representations naturally accommodate rings and cyclic structures without additional complexity, while some string-based methods (e.g., SMILES) require special notations to represent these features.

- **Scalability for Larger Molecules**: Graph representations scale well for larger, more complex molecules, like proteins or polymers, where other methods might become cumbersome.

- **Efficient Storage of Structural Data**: While not the most compact representation, molecular graphs balance between efficiency and detail, preserving essential structural features without requiring massive data storage.

Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des. 2016 Aug;30(8):595-608.

# Why GCN?

**Direct Application to Graph Structures**: GCNs work directly on molecular graphs, preserving the structure of atoms and bonds. Traditional neural networks expect vectorized inputs, which requires converting molecular structures into fixed-length feature vectors (like fingerprints), often leading to loss of structural information.

**Local Neighborhood Aggregation**: GCNs aggregate information from neighboring atoms, capturing chemical context about atom’s environment.

**Automatic Feature Learning**: GCNs automatically learn relevant molecular features without manual feature engineering.

**Scalability**: GCNs scale well to large molecular datasets and complex structures like proteins or polymers.

**Adaptability to Various Tasks**: GCNs can be applied to various tasks such as property prediction, reaction outcomes, and drug-target interaction.

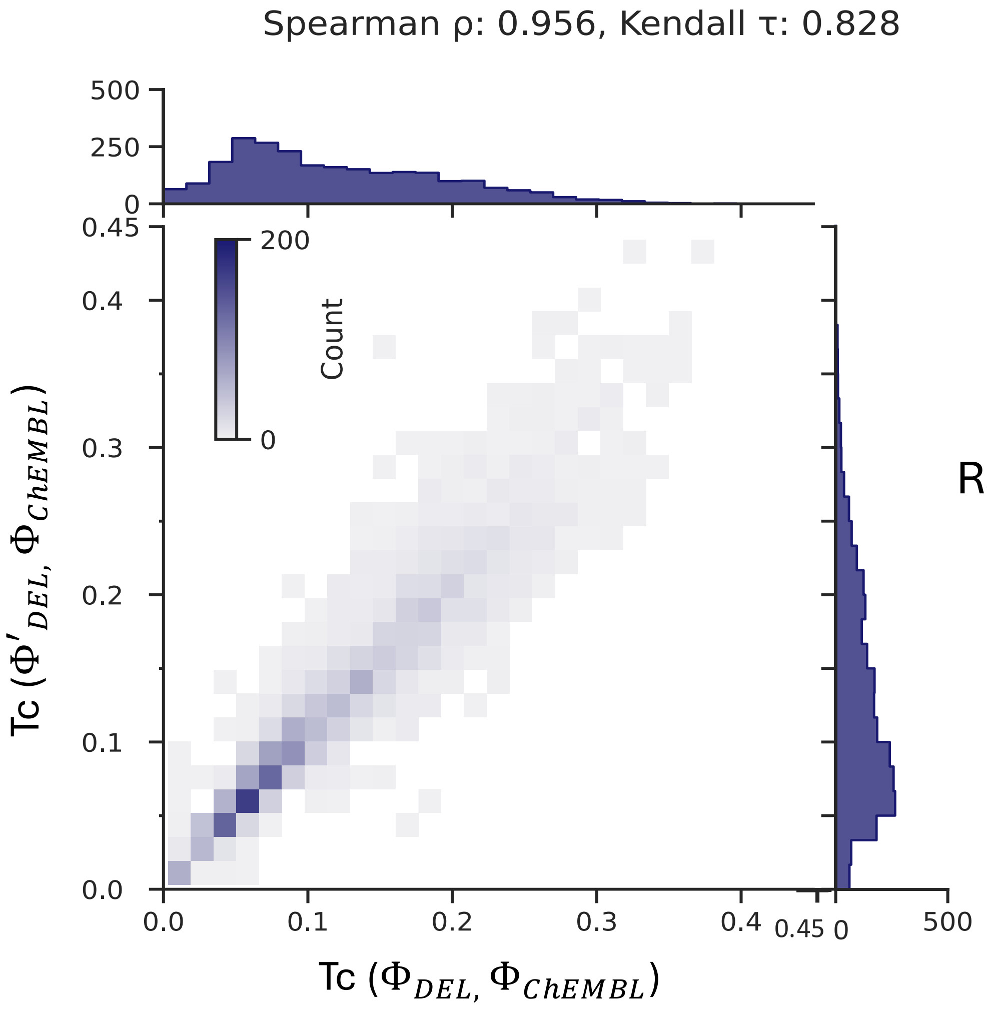
**Effective in Low-Data Regimes**: GCNs perform well even with limited data, making them useful in areas with sparse datasets.

**Improved Performance Over Traditional Methods**: GCNs often outperform traditional machine learning methods like random forests or SVMs.

Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des. 2016 Aug;30(8):595-608.

Fout, Alex, et al. "Protein interface prediction using graph convolutional networks." *Advances in neural information processing systems* 30 (2017).

# Results: Predicted maps allow to correctly rank DELs by similarity to ChEMBL

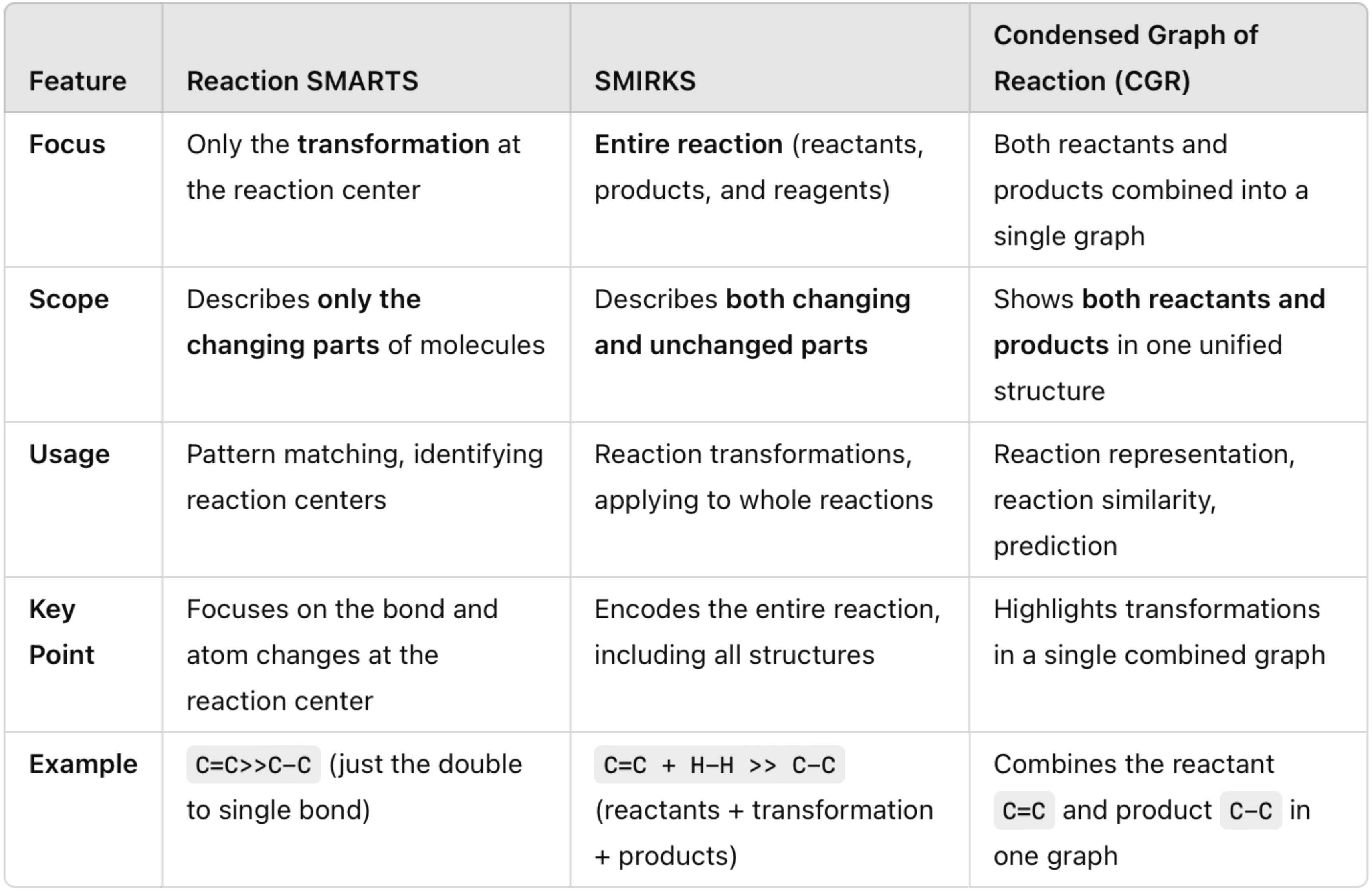


**Task:**

VS
DELs
ChEMBL

Ranking of DEL maps predicted by CoLiNN with respect to their similarity to ChEMBL correlate with the true ranking

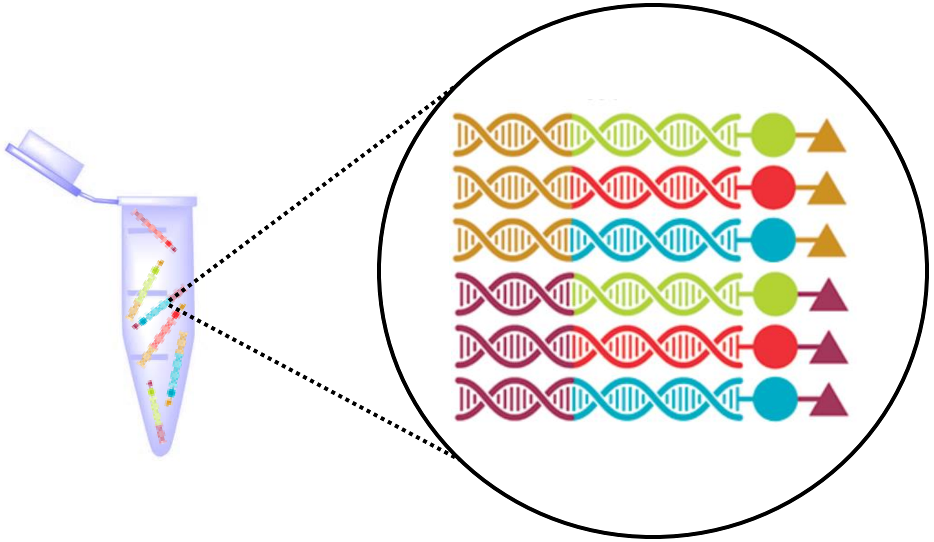
# Reaction SMARTS/SMIRKS/CGR



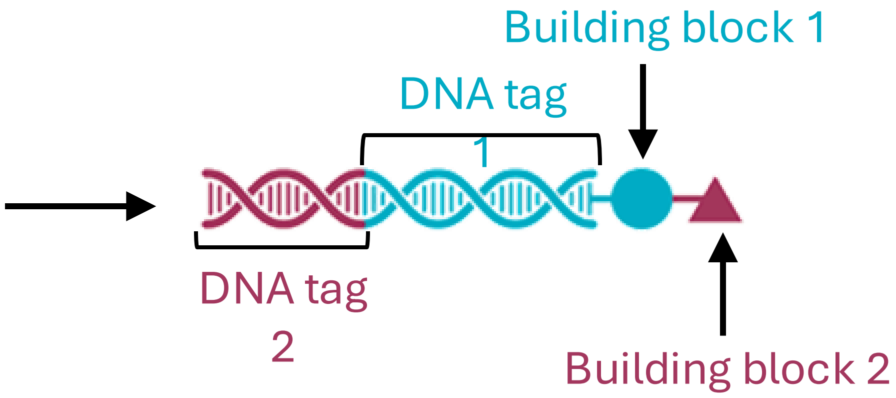
| Feature | Reaction SMARTS | SMIRKS | Condensed Graph of Reaction (CGR) |
|---|---|---|---|
| **Focus** | Only the **transformation** at the reaction center | **Entire reaction** (reactants, products, and reagents) | Both reactants and products combined into a single graph |
| **Scope** | Describes **only the changing parts** of molecules | Describes **both changing and unchanged parts** | Shows **both reactants and products** in one unified structure |
| **Usage** | Pattern matching, identifying reaction centers | Reaction transformations, applying to whole reactions | Reaction representation, reaction similarity, prediction |
| **Key Point** | Focuses on the bond and atom changes at the reaction center | Encodes the entire reaction, including all structures | Highlights transformations in a single combined graph |
| **Example** | `C=C>>C-C` (just the double to single bond) | `C=C + H-H >> C-C` (reactants + transformation + products) | Combines the reactant `C=C` and product `C-C` in one graph |

# DNA-Encoded Library (DEL)

**DNA-Encoded Library** is a combinatorial collection of small molecules covalently attached to the DNA tag



**DNA-Encoded Library**



Building block 1

DNA tag 1

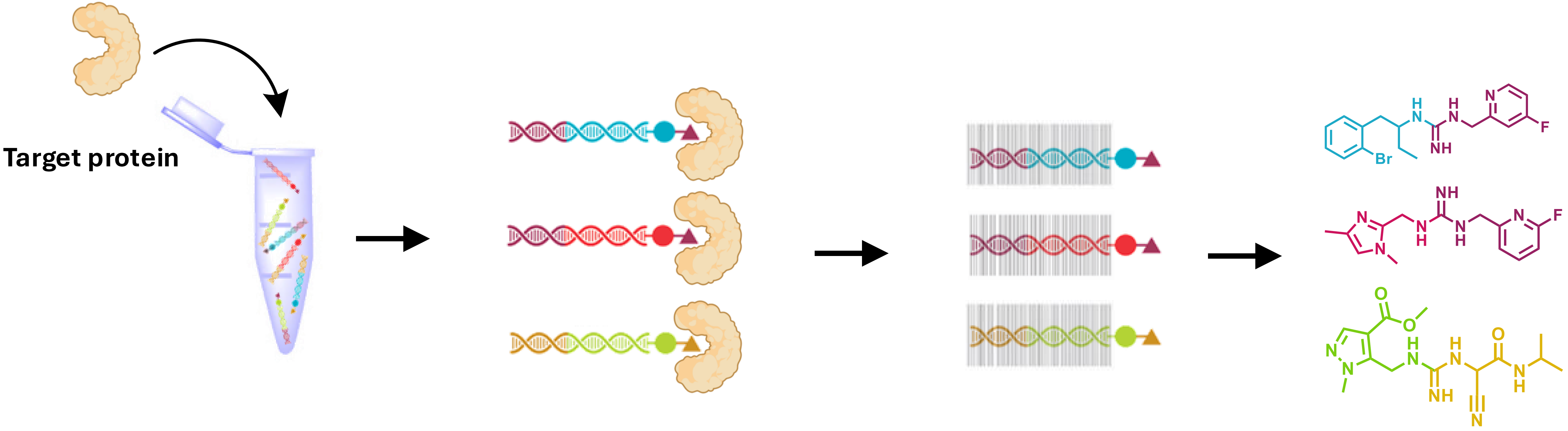
DNA tag 2

Building block 2

Library Size =(**Purple** BBs) x (**Blue** BBs)

Images: Halford, B. Chem. Eng. News 2017, 95, 28.
Images taken from biorender.com

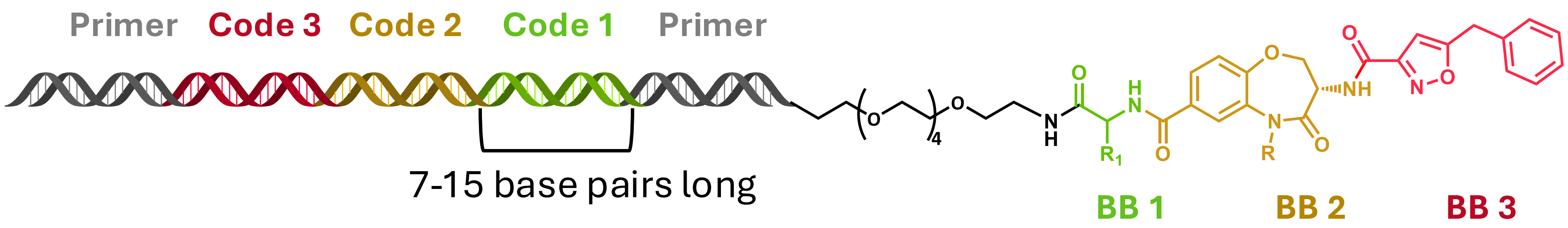
# DNA-Encoded Library (DEL) screening

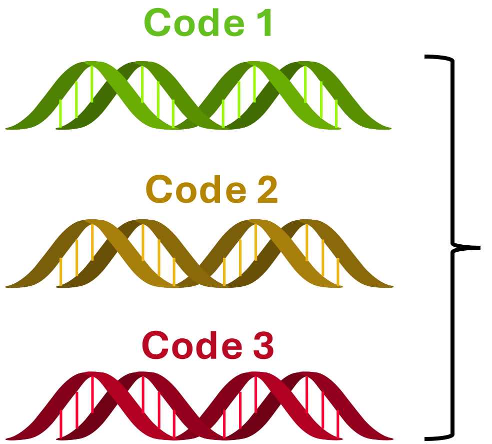


**DNA-Encoded Library** → **Isolation of compounds that bind to the target protein** → **Identification through DNA-sequencing** → **Structures of active compounds identified**

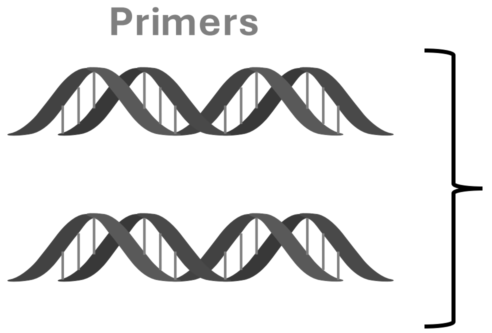
Images: Halford, B. Chem. Eng. News 2017, 95, 28.
Images taken from biorender.com

# DNA-tagging





DNA sequences identifying different building blocks (BBs) that make up a molecule



DNA sequences that initiate the Polymerase Chain Reaction essential for the compound identification step