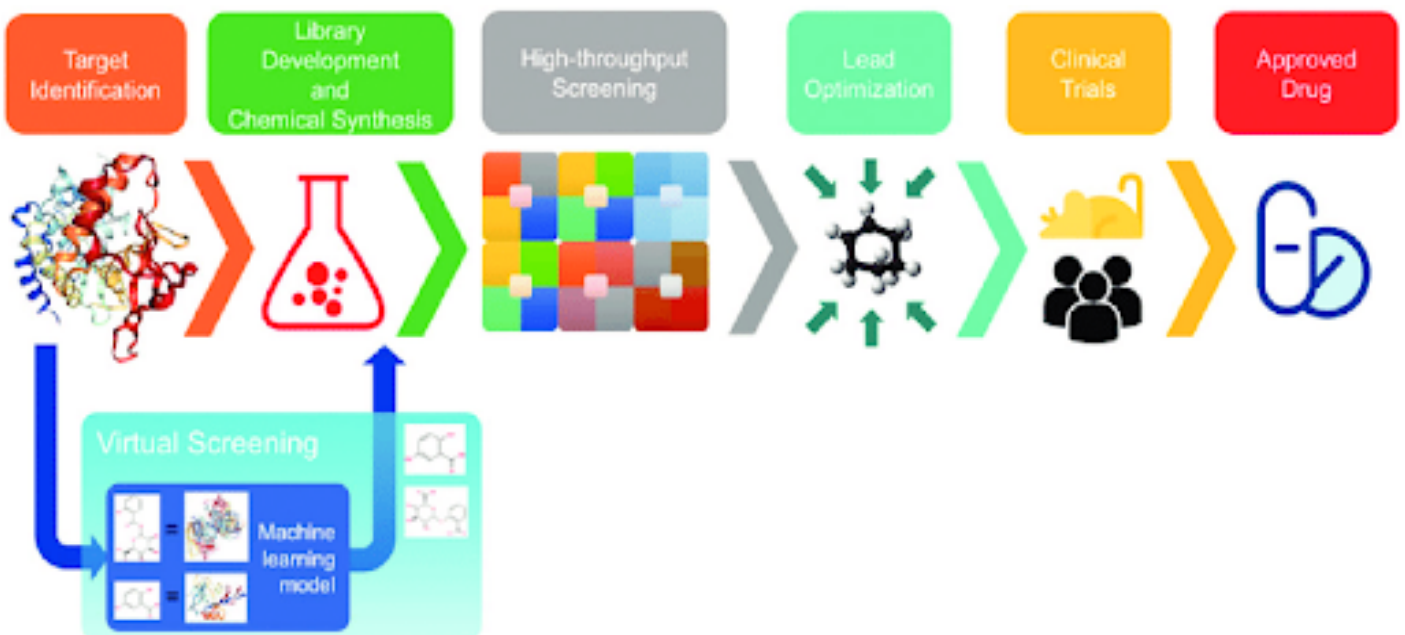
ML for Molecular Properties Prediction

Roi Naveiro (CUNEF) Simón Rodríguez Santana (ICMAT)

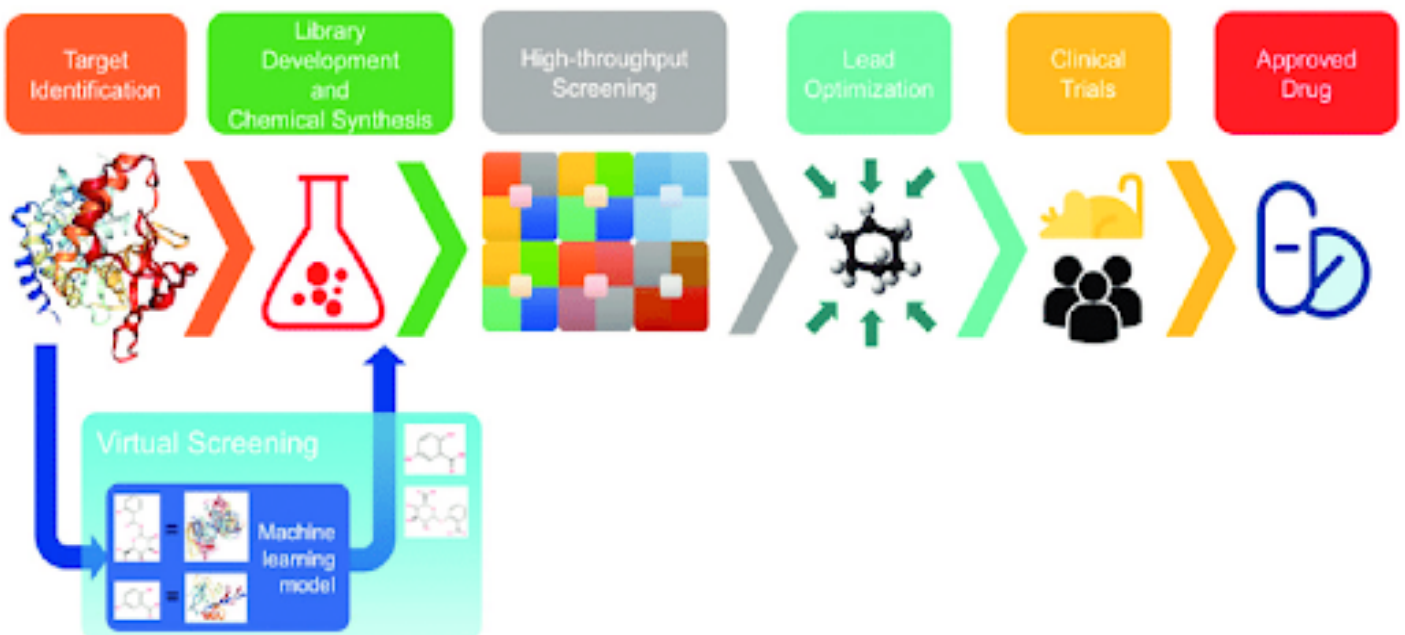
Discovering new molecules - Process

- Design of new molecule: countless applications in various sectors, e.g. pharmaceuticals and materials.
- Pharma: average time discovery starts - market, 13 years. Outside pharma: 25 years



Discovering new molecules - Process

- Crucial 1st step: generate pool of promising candidates
- Daunting task (chemical space is huge and has complex structural constraints molecules)



The old and soon-to-be-old ways

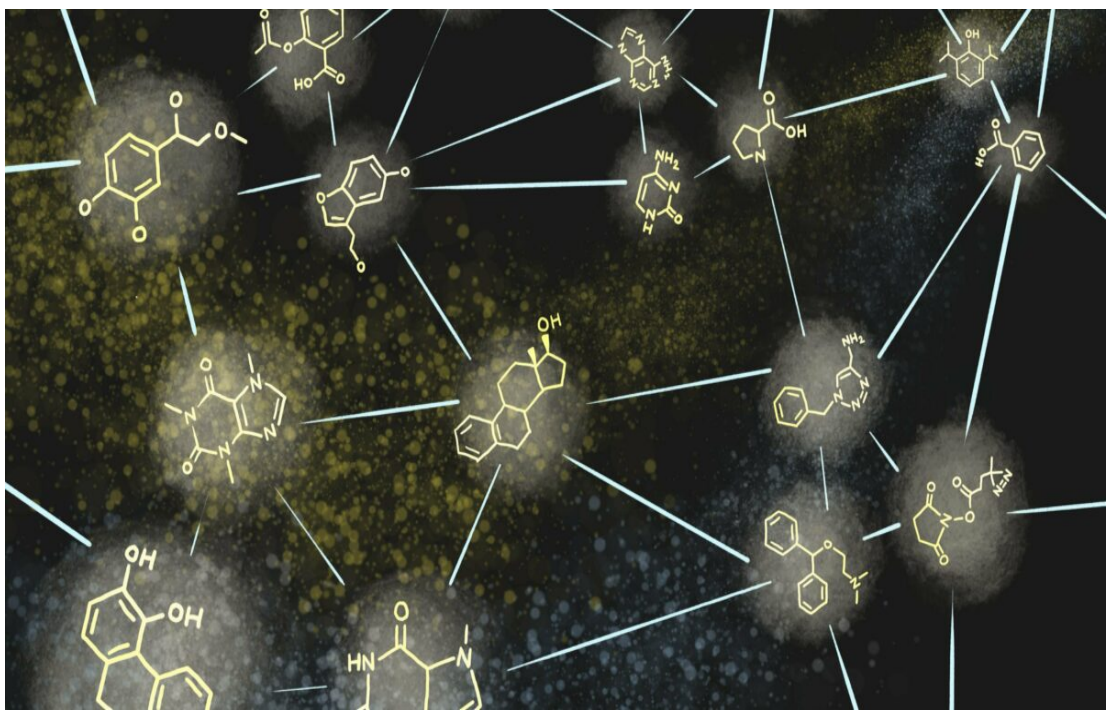
- Old way
  - Human experts propose, synthesize and test (in vitro)
- Soon-to-be-old way: high throughput virtual screening (HTVS)
  - Predict properties through computational chemistry…
  - …leverage rapid ML-based property predictions

Problems with previous approaches

- Just existing molecules are explored
- Much time lost evaluating bad leads
- **Goal**: traverse chemical space more “effectively”: reach optimal molecules with less evaluations than brute-force screening

De novo design

The process of **automatically proposing novel chemical structures** that **optimally satisfy desired properties**

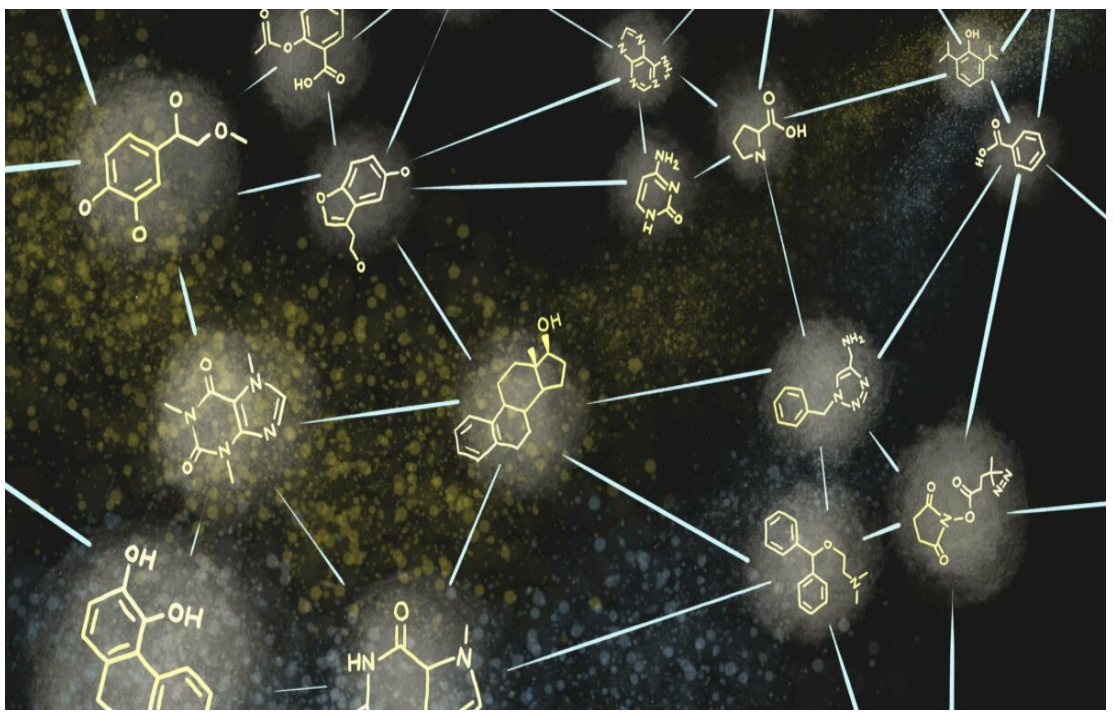


Mathematically speaking

- Combinatorial optimization problem
- Often stochastic and multi-objective
- Black-box objective functions
- Black-box constraints

De novo design

The process of **automatically proposing novel chemical structures** that **optimally satisfy desired properties**



Two interrelated steps

1. **Optimally satisfy desired properties:**
   Predictive models to forecast/approximate properties/ objective functions from chemical structure
2. **Automatically proposing novel chemical structures**
   Automatic generation of molecules that optimize properties (predictions from first stage)

This workshop

- **Session 1:** Predictive (QSAR) Models, with focus in low data regime
- **Session 2:** Generative Models
- **Session 3:** The Tailor's Drawer (+ Case Study)

Predictive Models

Predictive models to forecast properties of molecules given structure, wiht Focus on small data regime

1. Computational representations of molecules
2. An overview of predictive models for molecular properties
3. Evaluating model performance

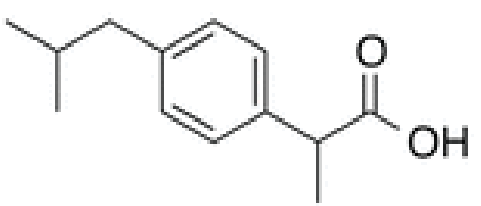
Representating molecules

Molecules are **3D QM objects** with: nuclei with defined positions surrounded by electrons described by complex wave-functions

- Digital encoding that serves as input to model
- Uniqueness and invertibility
- Trade-off: information lost vs complexity
  - 3D coord. representation (symmetries?)
  - More compact 2D (graph) representation
- 1D, 2D and 3D Representations

1D Representations

- **Simplified Molecular Input Line Entry System** (SMILES)
- Molecule as graph (bond length and conformational info lost)
- Traverse graph
- Generate Sequence of ASCII characters

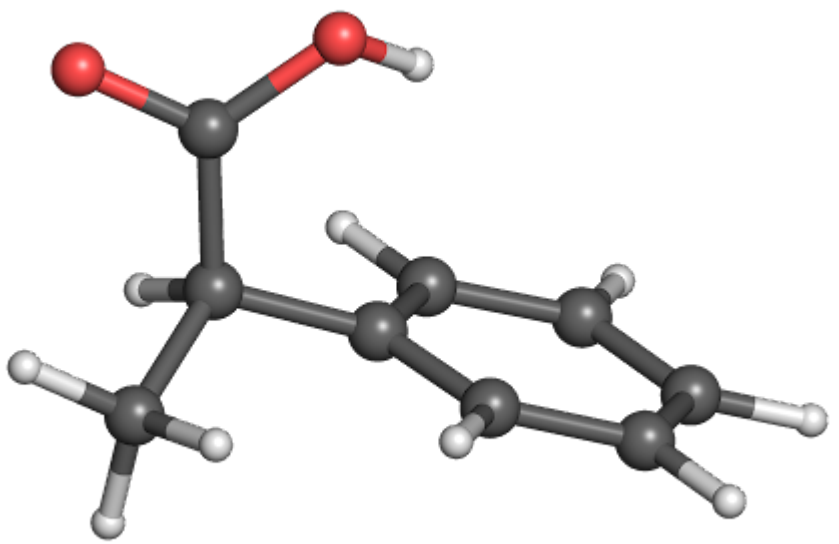


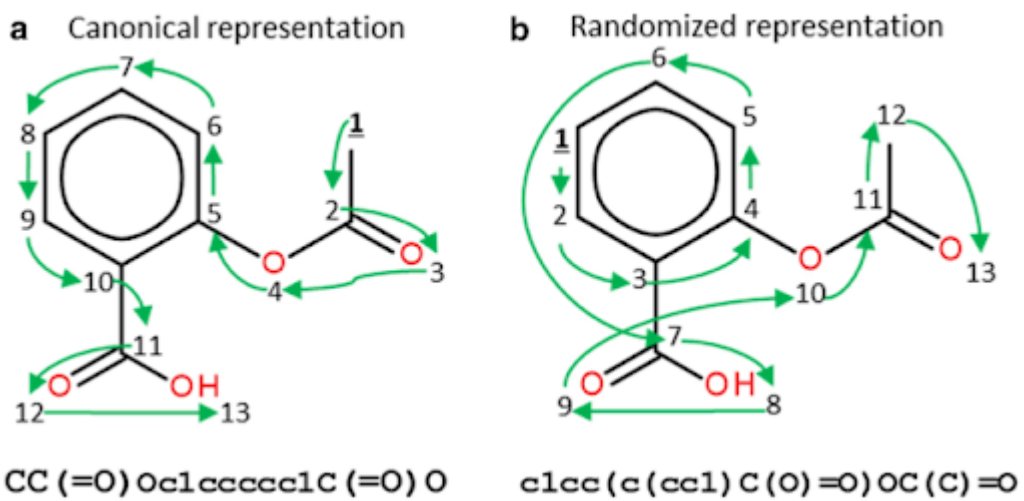
Ibuprofen

`CC(C)Cc1ccc(cc1)C(C)C(0)=0`

1D Representations

- Non-Unique! **Canonical SMILES**
- Tabular data:
  - One-Hot Encoding (NLP)
  - Molecular Descriptors (usual ML models)



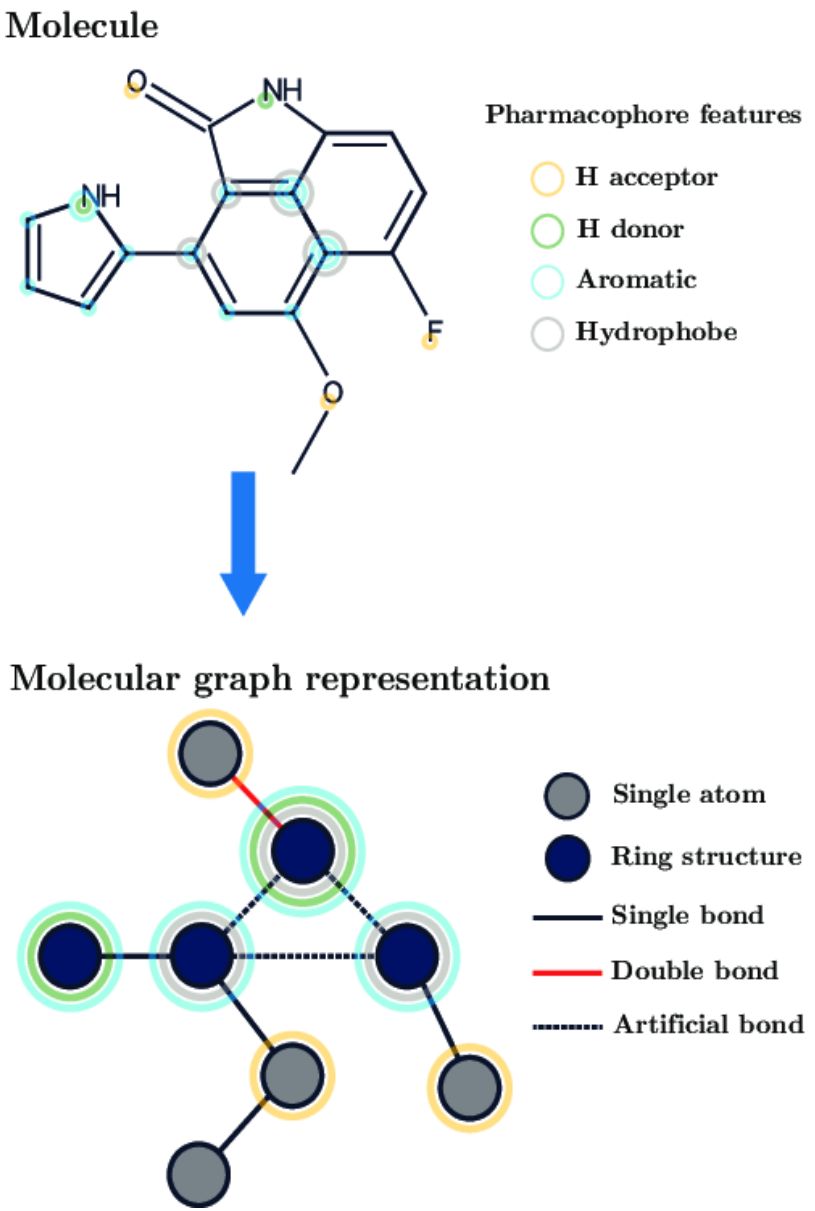


Molecular Descriptors

- Morgan Fingerprints Capecci et. al. (2020)
- Mordred Descriptors Moriwaki et. al. (2018)
- More… e.g. **molecular embeddings**

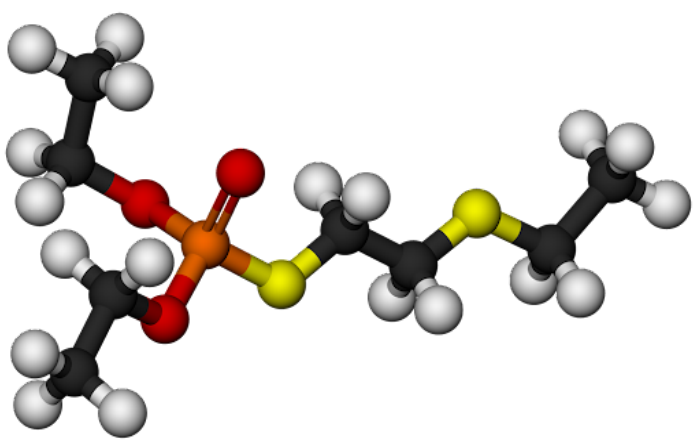
2D Representations

- Nodes represent atoms
- Edges represent bonds
- Nodes/Edges have associated features (atom number, bond type, etc.)
- Capture connectivity!
- Respect symmetries
- Tailored algorithms (GNNs!)



3D Representations

- 3D point clouds: , where are features and are coordinates
- Minimal information lost (conformational preferences, bond lengths, etc.)
- Tailored predictive algorithms that respect translational and rotational invariance



An overview of predictive models for molecular properties

- Molecular representation and property
- Given training data …
- … predictive **regression** model of given .
- Deterministic models - **Point Forecasts**
- Probabilistic (Bayesian) models - **Probabilistic Forecasts**

Models for 1D representations - Descriptors

- Usual desterministic models: linear regression, RF, XGBoost, SVR…
- Low-data regime:
  - : need for regularization
  - Uncertainty is key probabilistic (Bayesian) models

Models for 1D representations - Strings

- One-hot encoding of SMILES representations
- Deep Neural Nets: RNN, 1D Conv, Transformers
- BNNs
  - Computationally expensive to train
  - Variational Inference: uncertainty underestimation Blei et. al. (2018)

Models for 2D molecular representations

- Graph Neural Networks
- Sequence of graph-to-graph blocks + output layer
- (Infinitely) many architectures: Graph Networks Battaglia et. al. (2018)

GNNs (on a nutshell)

- Functions on graph-structured data
- GN block (graph-to-graph map): primary computational unit in GNN
- Graph nodes and edges: tuple
  - : global attribute
  - : set of node attribute vectors
  - : set of edges. edge attribute, index of receiving node, and is index of sending node.

GN Block

- Edge update function
- Node update function
- Global update function .
- : aggregates edge attributes per node
- : aggregates edge attributes globally
- : aggregates node attributes globally.

GN Block - Computations

Algorithm 1 Steps of computation in a full GN block.

function GRAPHNETWORK(E, V, $\mathbf{u}$)

  for $k \in \{1 \dots N^e\}$ **do**

    $\mathbf{e}'_k \leftarrow \phi^e\left(\mathbf{e}_k, \mathbf{v}_{r_k}, \mathbf{v}_{s_k}, \mathbf{u}\right)$ ▷ 1. Compute updated edge attributes

  end for

  for $i \in \{1 \dots N^n\}$ **do**

    let $E'_i = \left\{\left(\mathbf{e}'_k, r_k, s_k\right)\right\}_{r_k=i,\ k=1:N^e}$

    $\bar{\mathbf{e}}'_i \leftarrow \rho^{e\to v}\left(E'_i\right)$ ▷ 2. Aggregate edge attributes per node

    $\mathbf{v}'_i \leftarrow \phi^v\left(\bar{\mathbf{e}}'_i, \mathbf{v}_i, \mathbf{u}\right)$ ▷ 3. Compute updated node attributes

  end for

  let $V' = \left\{\mathbf{v}'\right\}_{i=1:N^v}$

  let $E' = \left\{\left(\mathbf{e}'_k, r_k, s_k\right)\right\}_{k=1:N^e}$

  $\bar{\mathbf{e}}' \leftarrow \rho^{e\to u}\left(E'\right)$ ▷ 4. Aggregate edge attributes globally

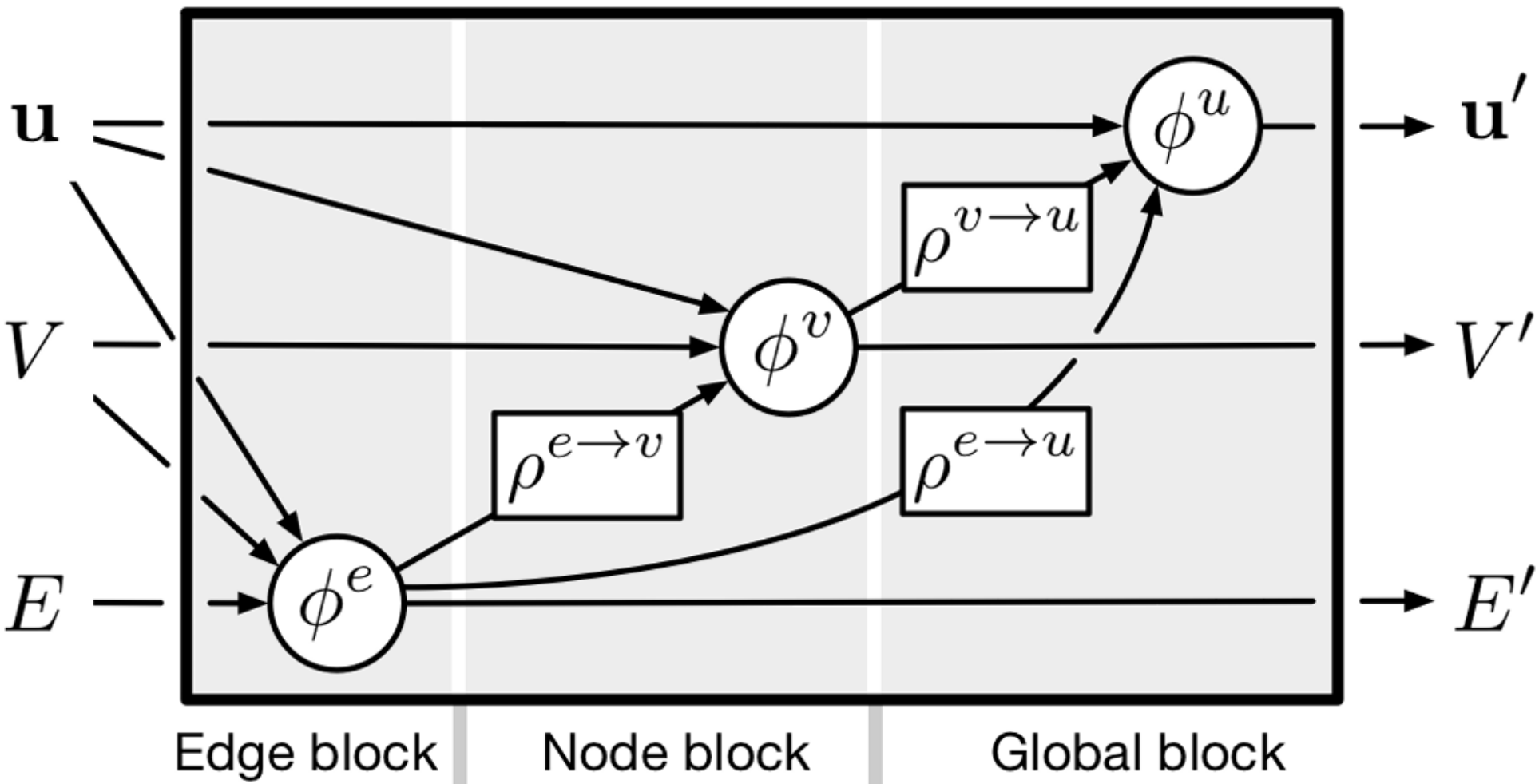
  $\bar{\mathbf{v}}' \leftarrow \rho^{v\to u}\left(V'\right)$ ▷ 5. Aggregate node attributes globally

  $\mathbf{u}' \leftarrow \phi^u\left(\bar{\mathbf{e}}', \bar{\mathbf{v}}', \mathbf{u}\right)$ ▷ 6. Compute updated global attribute

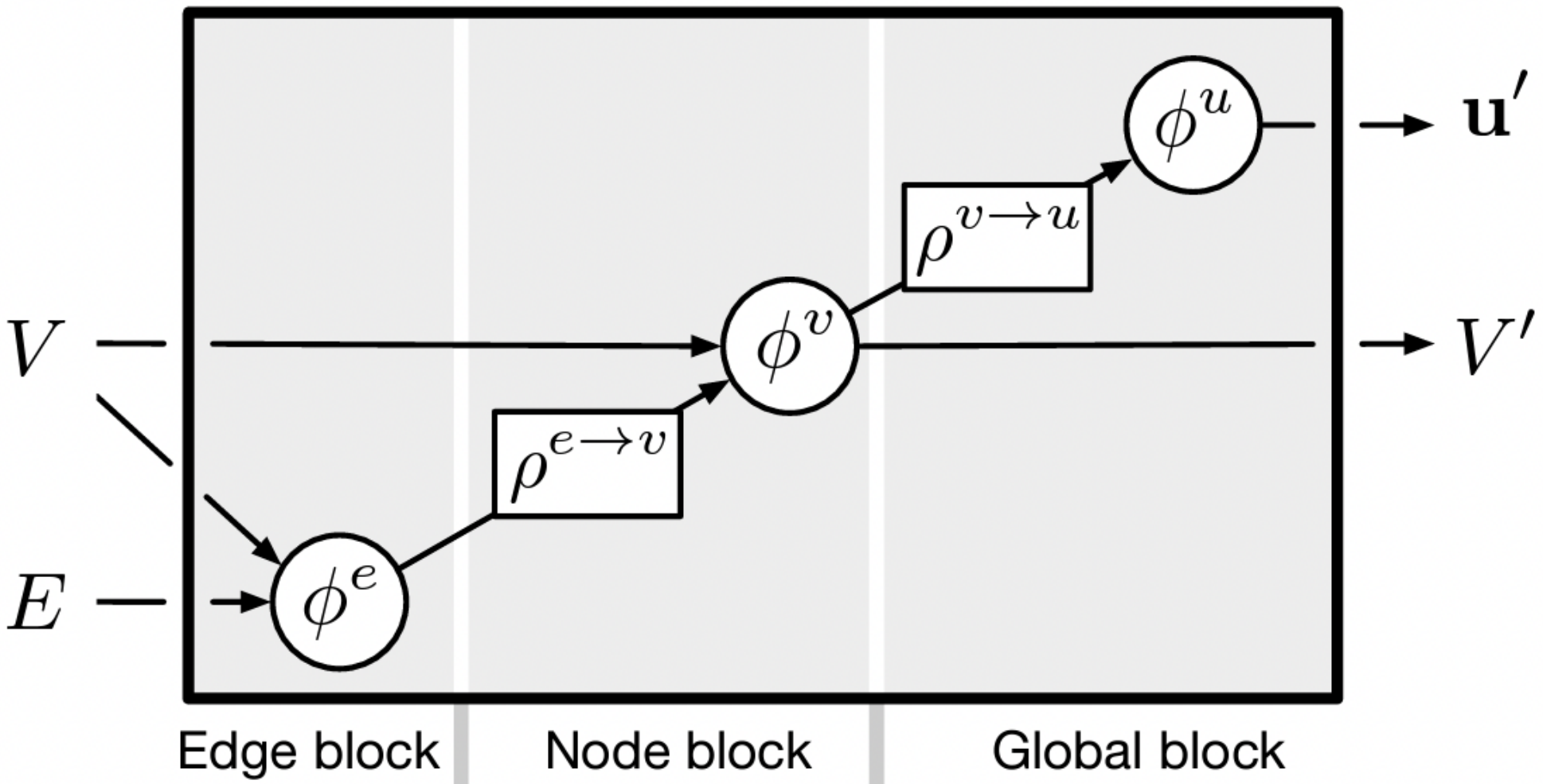
  return $\left(E', V', \mathbf{u}'\right)$

end function

GN Block - Computations



MPNN Block - Computations



GNN

- Various parametric forms for functions
- Multilayer perceptrons for the update functions and sums for the aggregate functions
- GN blocks can be concatenated
- Output layer of GNN depends on the task The entire architecture can be summarized as follows:

GNN Workflow

1. **Encode** the input graph using independent node and edge update functions to match the internal node and edge feature sizes
2. Apply **multiple GN blocks**
3. Use an **output** layer to map the updated global features to a property prediction

Once the architecture is defined, the parameters can be optimized using **standard optimizers and loss functions**.

Models for 3D molecular representations

- Geometric Neural Networks
- (Again) many architectures
- In a Geometric Net Block we update::
  - Node features, s.t. updated features are **invariant** to 3D translations and rotations
  - Node coordinates, s.t. updated coordinates are **equivariant** to 3D translations and rotations
- equivariant graph neural nets Satorras et. al. (2022)

E(n) equivariante GNNs

- Refinement of MPNN
-
- In addition to node features, coordinates: .

In a MPNN

1. edges ,
2. nodes
    -
    -
    -
3.
4.
5. .

E(n) equivariante GNNs

1. edges ,
2. nodes
    -
    -
    -
    -
3.
4.
5. .

Evaluating model performance - Point Predictions

Usual metrics for regression

- RMSE
- MAE
- MAPE
-

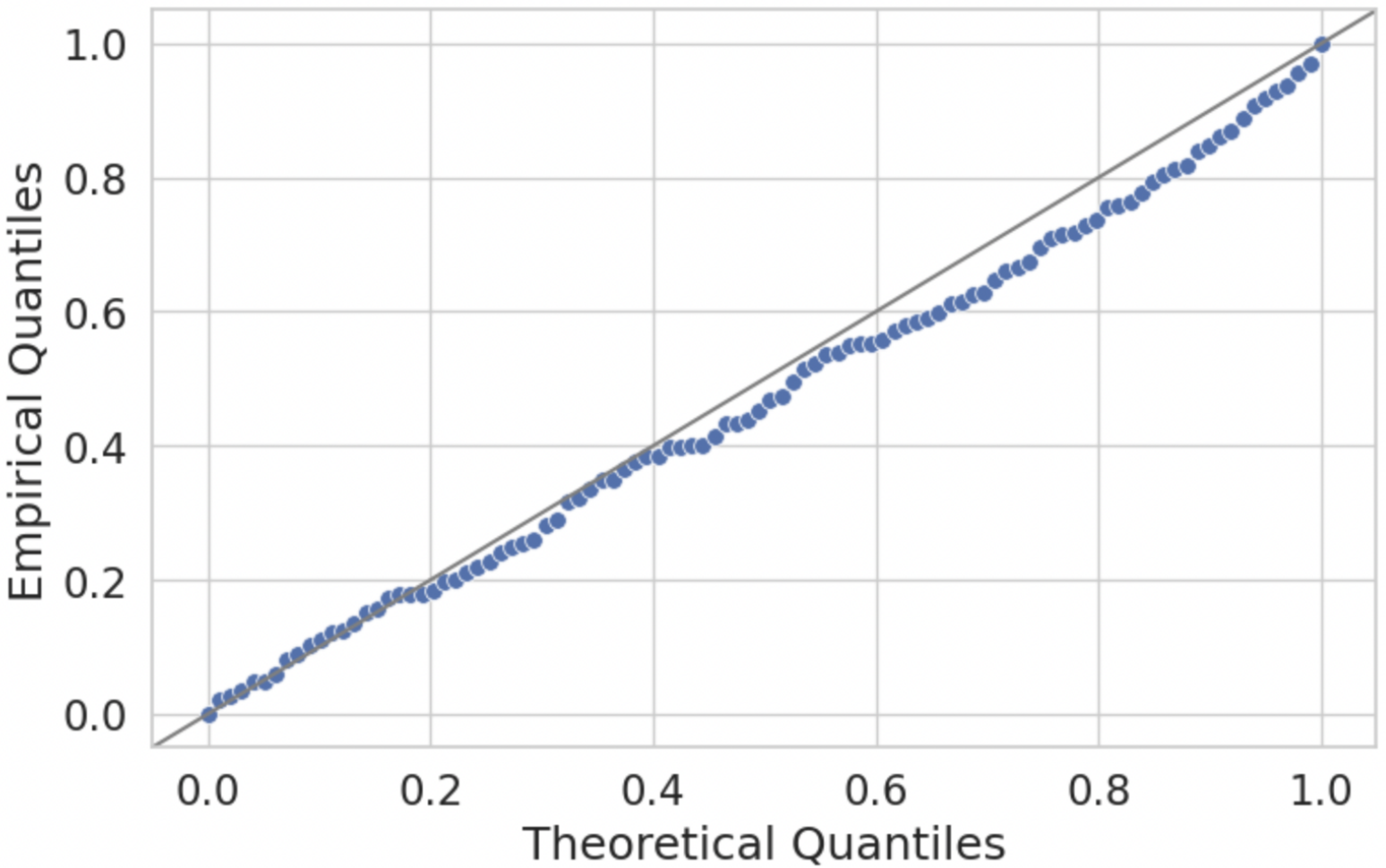
Evaluating quality of probabilistic predictions

- Multiple ways, research area itself! Gneiting and Raftery (2007)
- Calibration measures

Evaluating quality of probabilistic predictions

- Idea: create % prediction intervals for the property prediction of every molecules in a test set.
- is the proportion of the molecules in the test set whose property value is in the interval calculated for such molecule.
  - If we say that the model is well calibrated.
  - If we say that the model is overconfident.
  - If we say that the model is underconfident.

Evaluating quality of probabilistic predictions



Hands-on!