Multimodal Training

DL and GNNs for supervised and generative tasks in the bio/medical domain

M. Vazirgiannis

LIX @ Ecole Polytechnique
http://www.lix.polytechnique.fr/dascim
Google Scholar: https://bit.ly/2rwmvQU
Twitter: @mvazirg

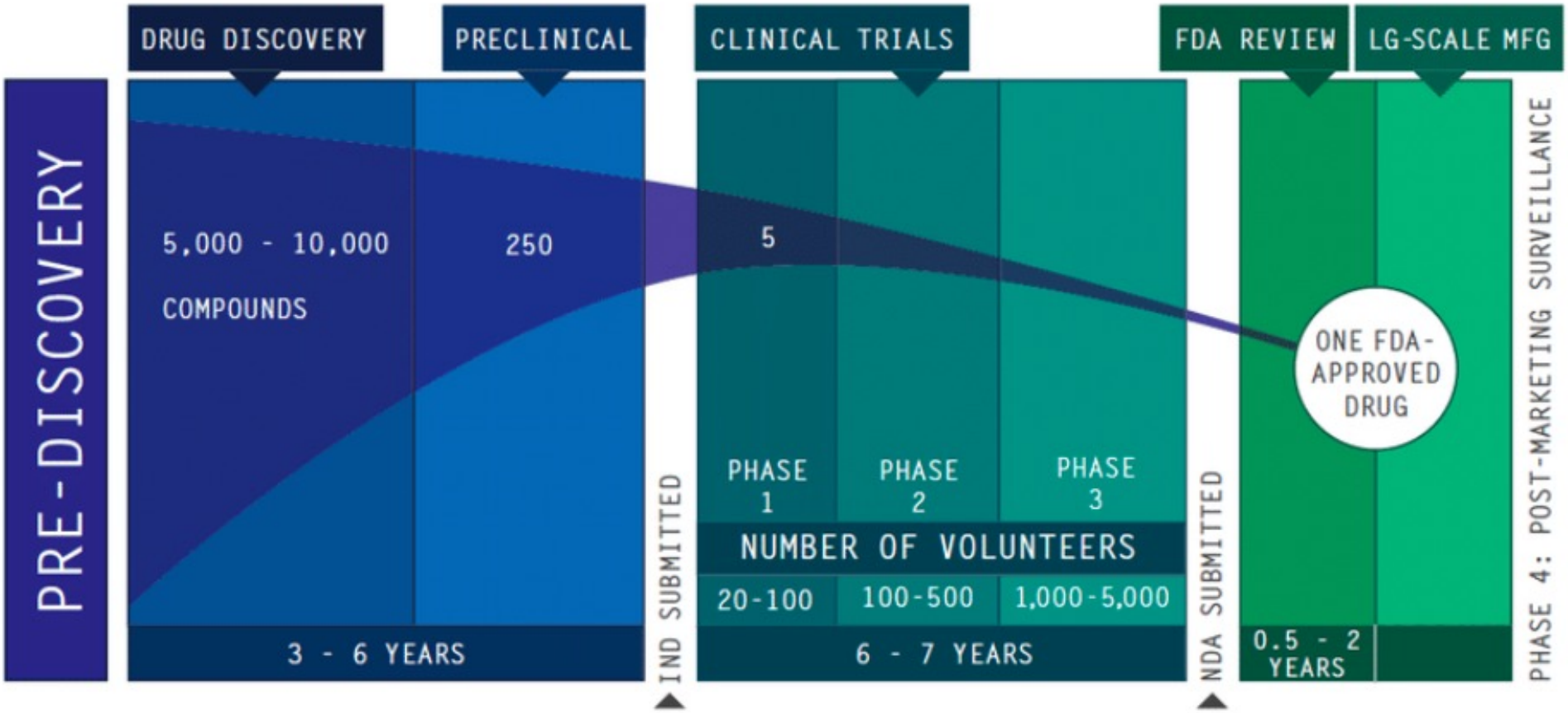
November 2022

Overview

- AI & Drugs design
- GNNs for Antibiotic Resistance prediction
- GVAE for medical graph generation
- Our vision

Drug Discovery – Long process

In average : **10 years** and **$2.6B**

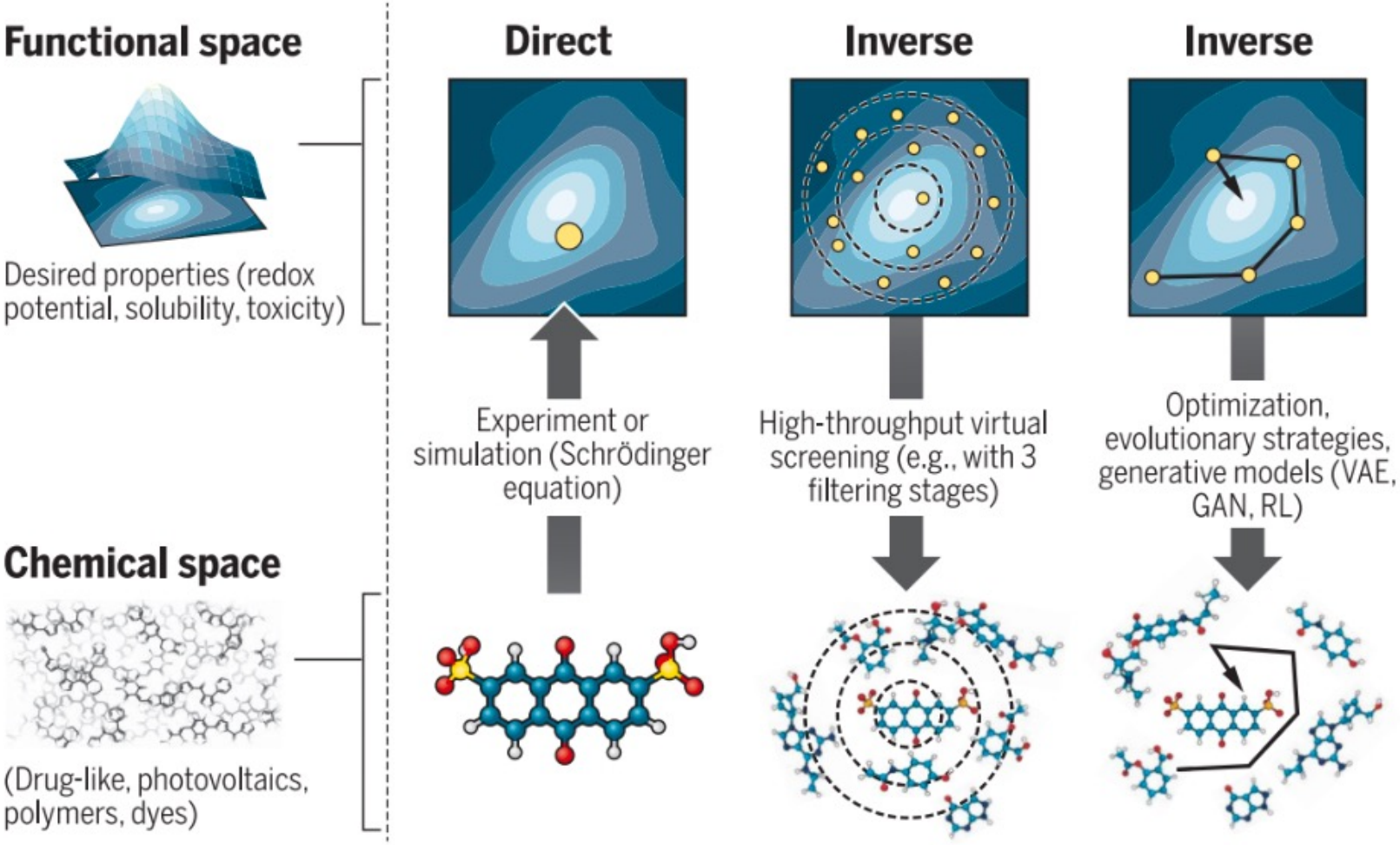


Source : *Pharmaceutical Research and Manufacturers of America*

Drug discovery is a challenging search problem

- The total number of possible small organic molecules that populate **chemical space** has been estimated to exceed 10^{60}.
- Experimental facilities in industry can only test 10^5 compounds/day.
  We need efficient computational methods to design drugs automatically and
- Artificial Intelligence can help !

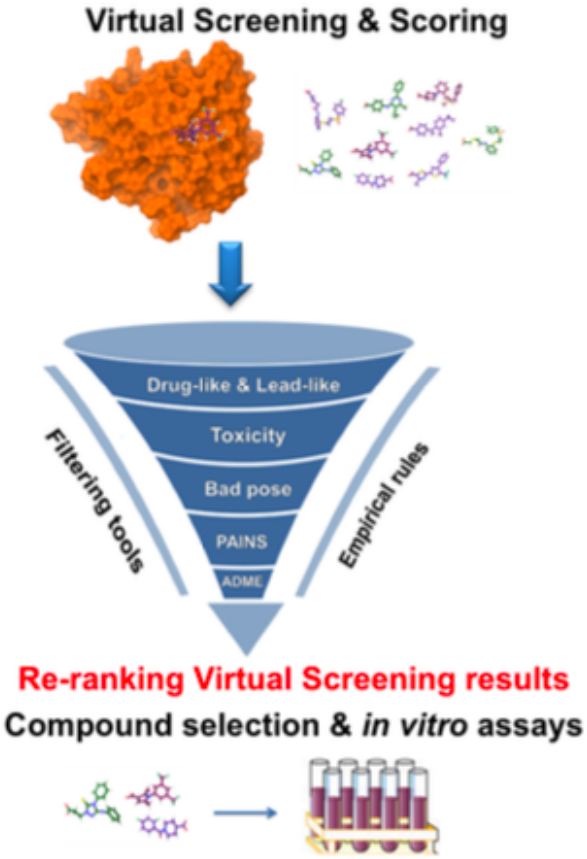
Approaches towards molecular design



Source : *Sanchez-Lengeling et al., Science 361, 360–365 (2018)*

Virtual Screening

- **Virtual screening** : assess whether a compound is a good drug using computation models (Walters et al., 1998 ; McGregor et al., 2007 ; . . .)

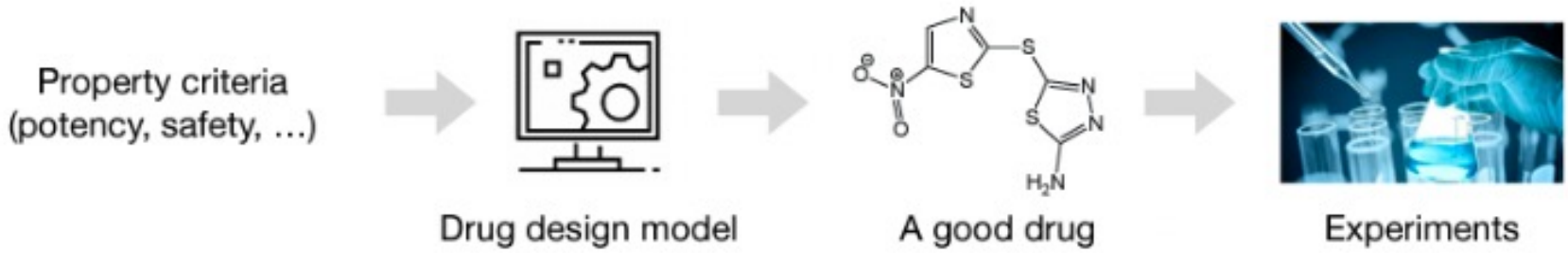


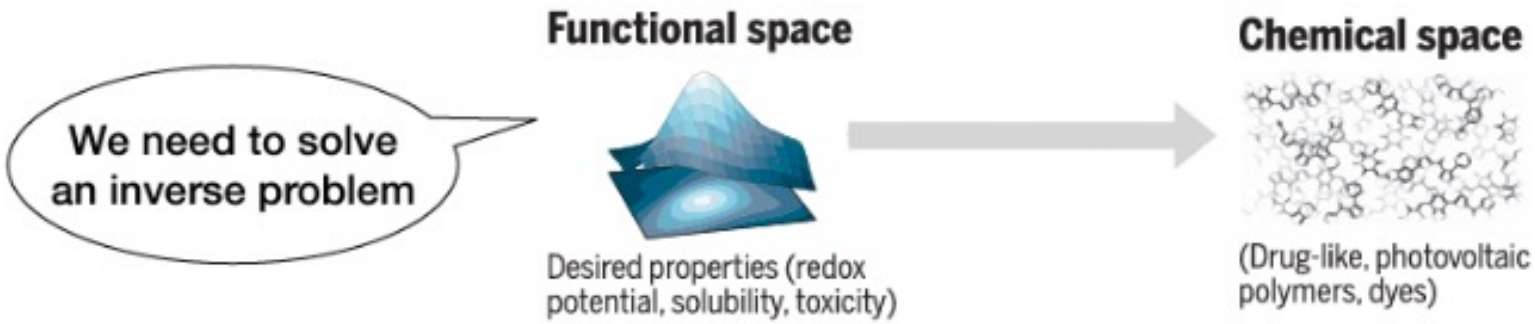
- Virtual screening is much faster than experimental screening in web labs.
- It can test 10^8 compounds within a day, while experimental screening takes years.
- It is also much cheaper than experimental screening.

Source : *Surabhi, Surabhi Singh, BK. (2018).*

De Novo Drug Design

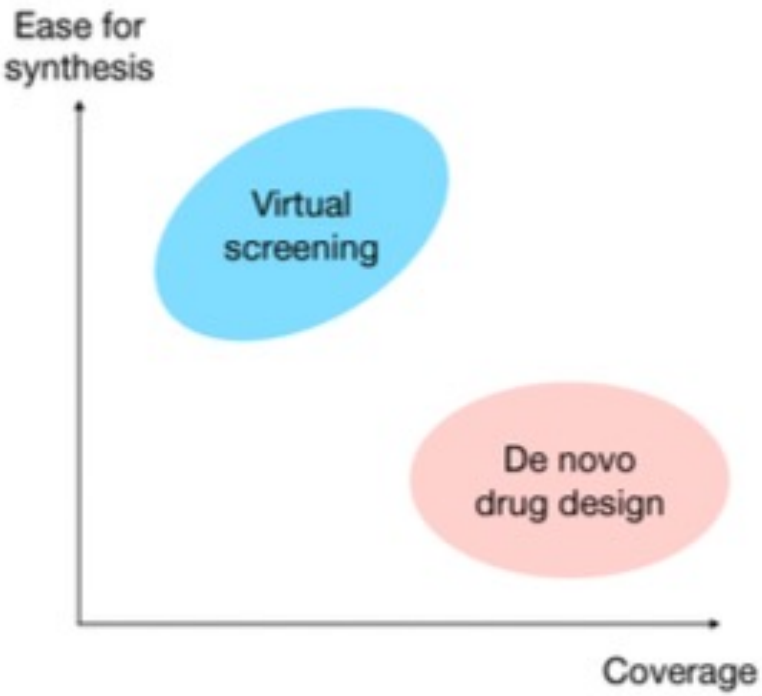
- **De novo drug design** : directly generate a compound with desired properties (Moon et al., 1991 ; Clark et al., 1995 ; Schneider Fechner, 2005 ; . . .)





Source : *Wengong Jin, Deep Learning for drug discovery (2021)*

Virtual Screening vs De Novo Drug Design

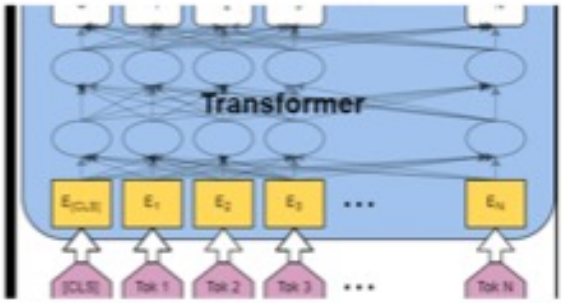


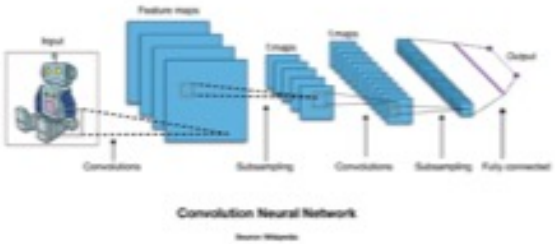
- Virtual screening is restricted to commercially available compounds (e.g., ZINC library - containing $2 * 10^8$ compounds).
- De novo drug design can explore the entire chemical space efficiently.
- Issue : feasibility of the new compounds synthesis.
- Traditional Techniques for both based on
- hand-crafted features and hand-designed rules

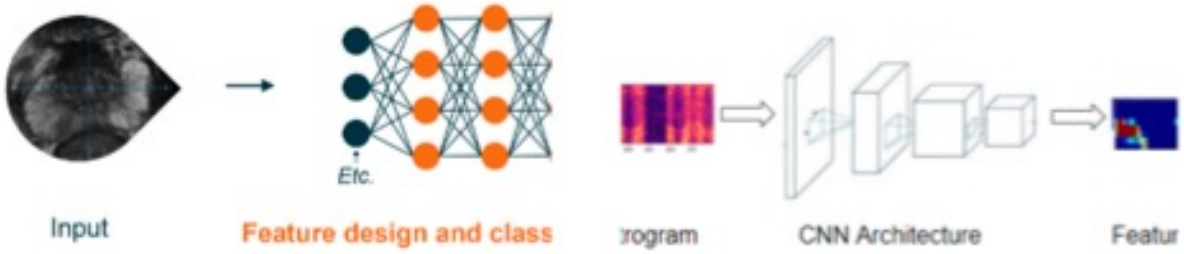
- Source : *Wengong Jin, Deep Learningrufloers.drug discovery (2021)*

Deep Learning – a promising direction

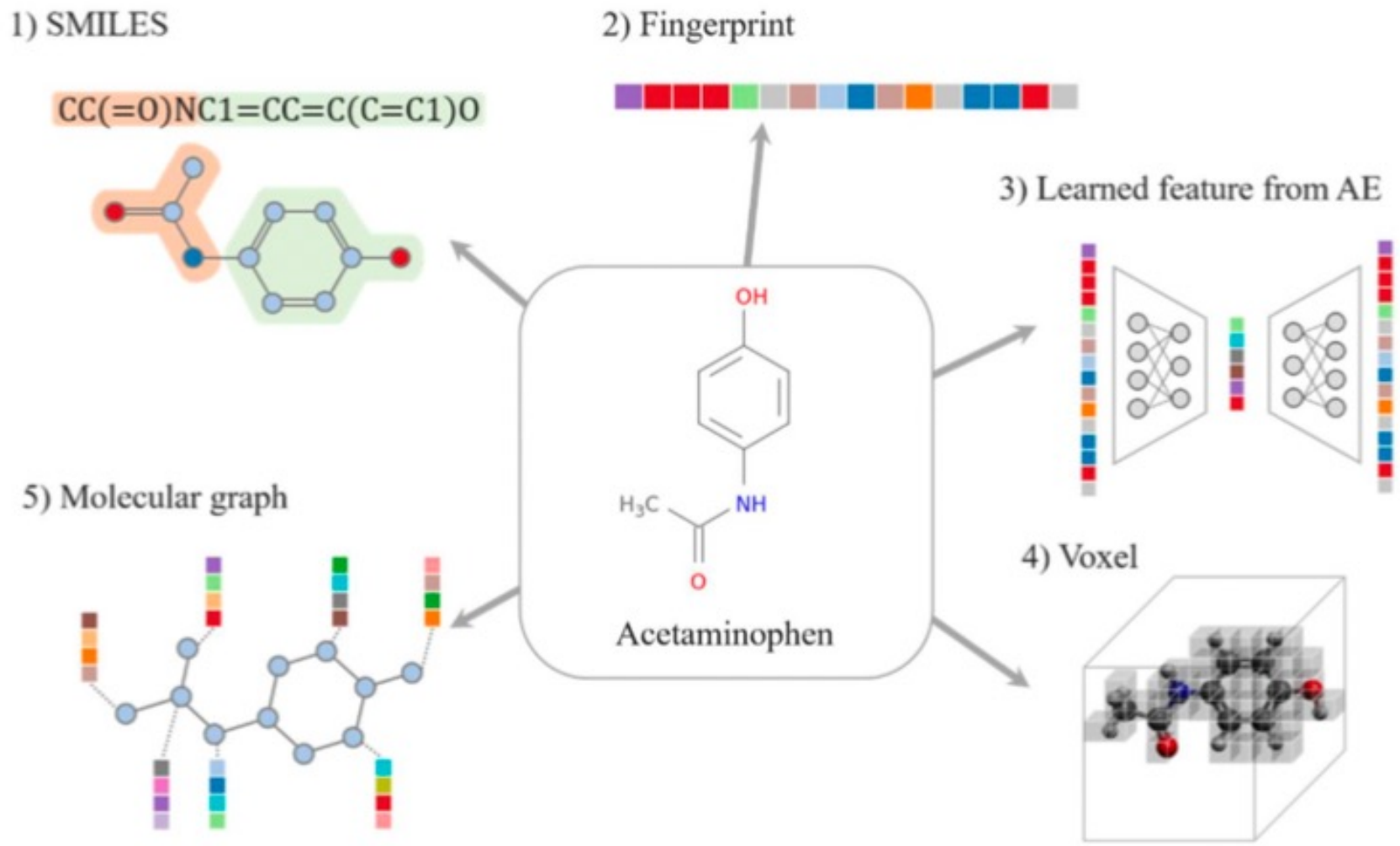
- Deep Learning has achieved many advances in recent years, with human-level accuracy in many tasks (Image Processing, Natural Language Processing, Recommendation Systems, Protein Folding, ...)







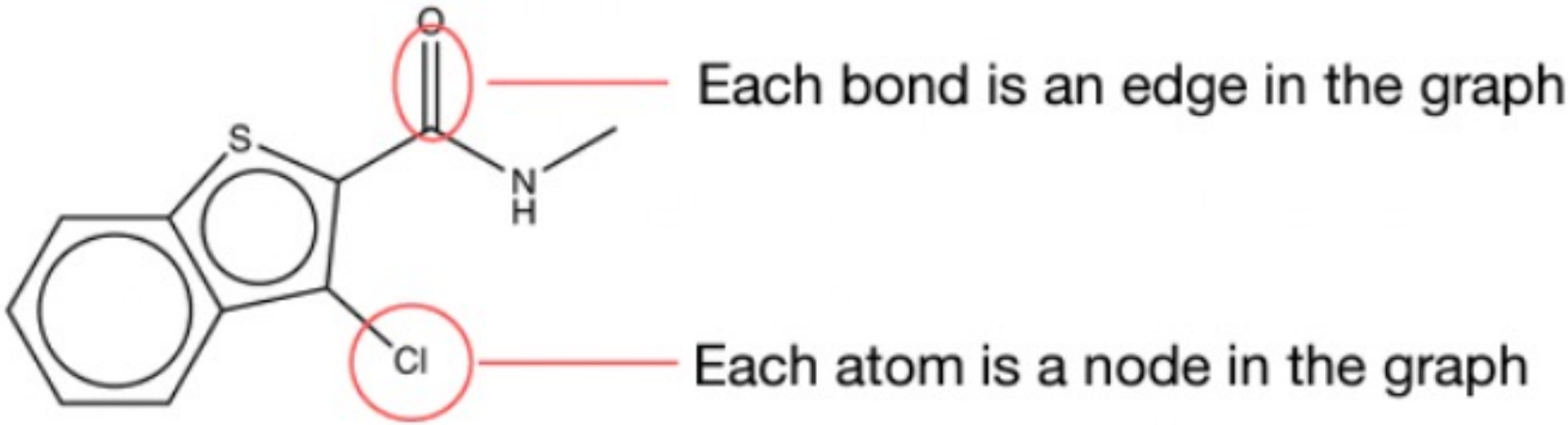
Molecules representations



Source : *Kim, Jintae et al.(2021)*

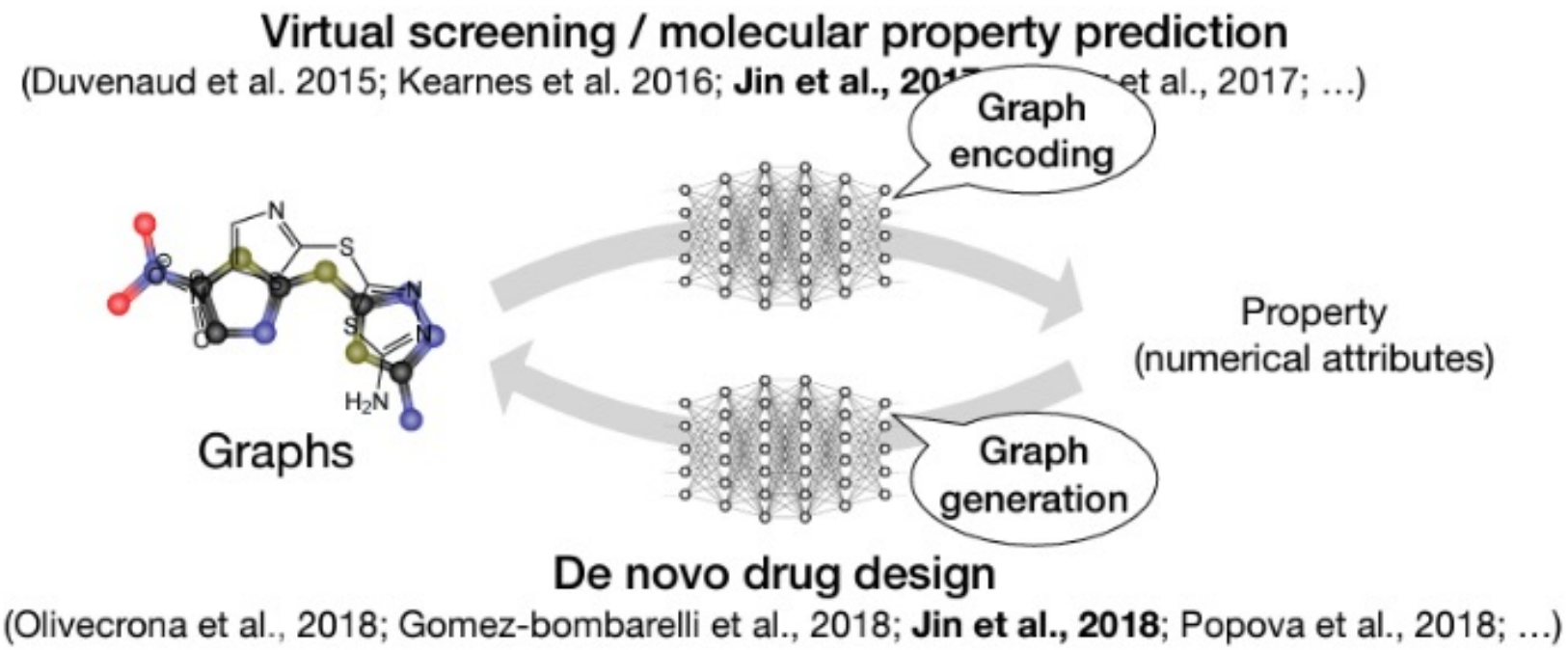
Molecules as Graphs

- Molecules can be represented as graphs.
- Node features : Atom type, Formal charge, etc.
- Edge features : Bond type, etc.



Deep Learning for drug discovery

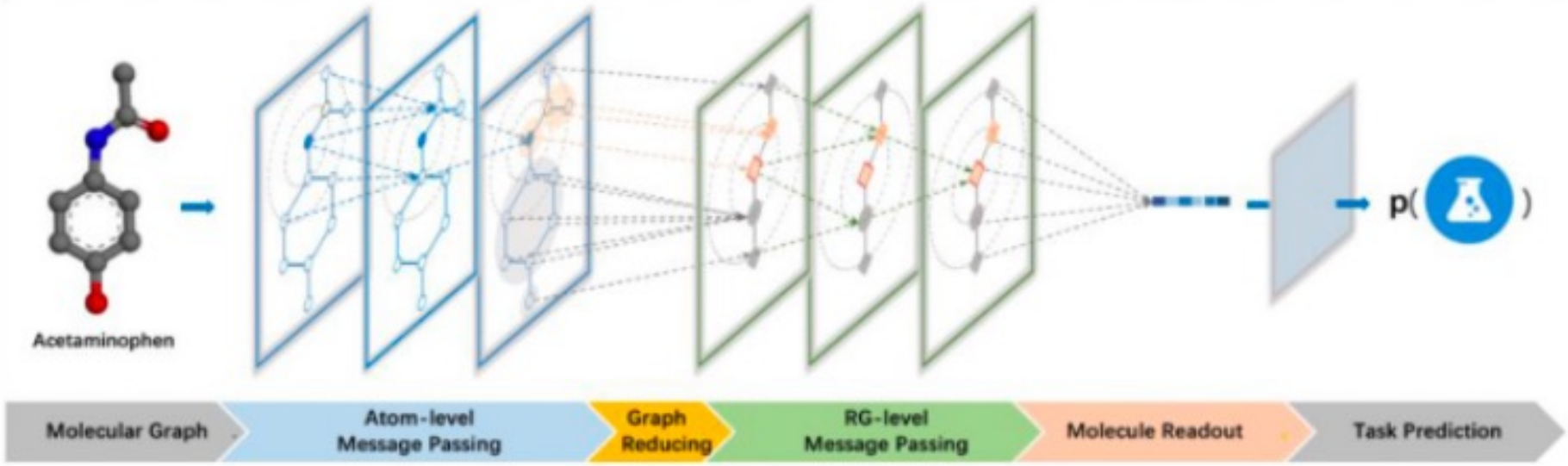
- **Virtual Screening** : Can be seen as a classification problem of molecules.
- **De Novo Drug Design** : Can be seen as a generative problem of molecules, with constraints.



Source : *Wengong Jin, Deep Learning for drug discovery (2021)*

Graph Neural Networks for molecules

- Rich history of GNNs (Gori et al., 2005, Scarselli et al., 2009, Duvenaud et al. 2015, Kearnes et al. 2016, Jin et al., 2017, Gilmer et al., 2017, Zitnik et al., 2018, etc.)
- Adapted to leverage the structure and properties of graphs.
- Used in Chemistry, Natural Language Processing, Recommendation systems, Computer Vision, Graph Generation ,etc.



Source : *Kong, Y., Zhao, X., Liu, R. et al., 2022*

Overview

- AI & Drugs design
- GNNs for Antibiotic Resistance prediction
- GVAE for medical graph generation
- Our vision

Structure-Aware Antibiotic Resistance Classification Using Graph Neural Networks [1]

Definition : Antimicrobial resistance (AR) happens when germs like bacteria and fungi develop the ability to defeat the drugs designed to kill them.

Why ?

- Major Health Challenge
- Natural Process, but enhanced by overuse and misuse in humans and animals
- Increased mortality (up to 10 million people by 2050)[5]
- Few antibiotics are developed each year, due to expensive process and low profit.

Antibiotics : Many antibiotics work by inactivating an essential bacterial protein.

How ? It can happen for many reasons :

- Spontaneous mutation in the bacterium's DNA, which change the essential bacterial protein.
- Transfer of antibiotic-resistance genes

[1] A. Qabel et. al, "Structure-Aware Antibiotic Resistance Classification Using Graph Neural Networks", https://www.biorxiv.org/content/10.1101/2022.10.06.511103v1, accepted in the AI4Science WS@Neurips2022

ARG prediction - Current baselines

Sequence based Algorithms :

- BLAST: Finds regions of local similarity between sequences.
- DIAMOND : Another sequence alignment tool, faster than BLAST.

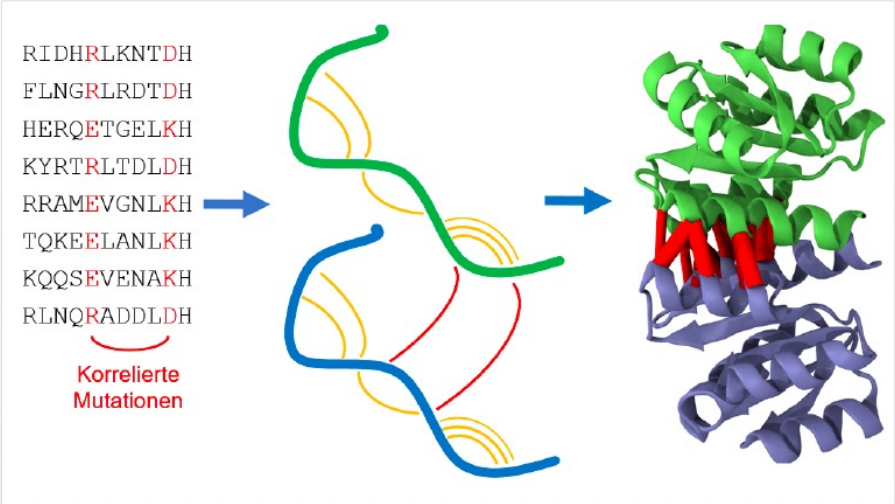
Machine Learning :

- DeepARG: Features = BLAST best hit score with some known ARs.
- TRAC: Pretraining of an LSTM-based LM + finetune on AR prediction task.
- ARG-SHINE: an ensemble method using BLAST resutls and deep convolutional network on the sequences.

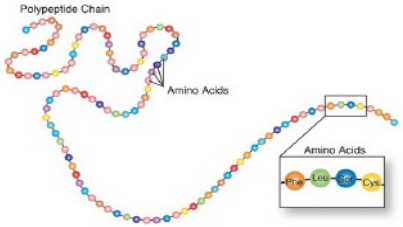
All these methods use Protein Sequence !

Protein folding

- Protein Folding: physical process - protein chain translated to its native three-dimensional structure – *a very tough problem*
- computational methods using AI to predict the protein structure:
  - DeepMind developed *AlphaFold2*, a model that achieved high quality *structure prediction.*
  - Facebook released *ESM2*, a Large Language model pretrained on *sequences* that achieves similar results to AlphaFold when using the structure module of AlphaFold2.

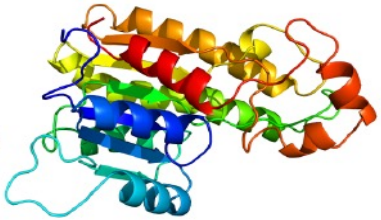


How to use protein structure



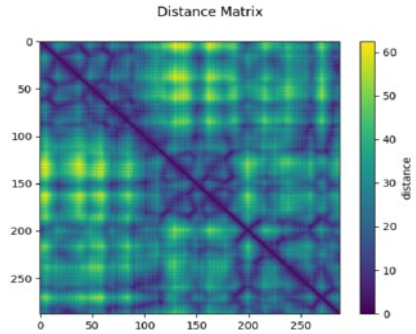
Protein Sequence

Protein Structure

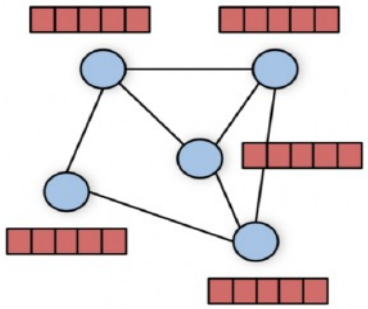


PDB Files with 3D coordinates of atoms

```
MODEL     1
ATOM      1  N   MET A   1      15.514  -1.247 -21.632  1.00 58.66           N
ATOM      2  CA  MET A   1      16.451  -2.365 -21.608  1.00 58.66           C
ATOM      3  C   MET A   1      15.712  -3.697 -21.686  1.00 58.66           C
ATOM      4  CB  MET A   1      17.313  -2.318 -20.345  1.00 58.66           C
ATOM      5  O   MET A   1      14.719  -3.902 -20.985  1.00 58.66           O
ATOM      6  CG  MET A   1      18.564  -1.468 -20.487  1.00 58.66           C
ATOM      7  SD  MET A   1      19.660  -1.577 -19.019  1.00 58.66           S
ATOM      8  CE  MET A   1      20.901  -0.327 -19.452  1.00 58.66           C
ATOM      9  N   THR A   2      15.923  -4.426 -22.730  1.00 68.58           N
ATOM     10  CA  THR A   2      15.337  -5.751 -22.893  1.00 68.58           C
ATOM     11  C   THR A   2      16.252  -6.823 -22.307  1.00 68.58           C
ATOM     12  CB  THR A   2      15.061  -6.062 -24.375  1.00 68.58           C
ATOM     13  O   THR A   2      17.449  -6.853 -22.602  1.00 68.58           O
ATOM     14  CG2 THR A   2      14.073  -7.215 -24.521  1.00 68.58           C
ATOM     15  OG1 THR A   2      14.516  -4.890 -25.008  1.00 68.58           O
ATOM     16  N   PHE A   3      15.828  -7.481 -21.207  1.00 76.05           N
ATOM     17  CA  PHE A   3      16.592  -8.525 -20.534  1.00 76.05           C
ATOM     18  C   PHE A   3      16.394  -9.870 -21.223  1.00 76.05           C
ATOM     19  CB  PHE A   3      16.186  -8.625 -19.058  1.00 76.05           C
ATOM     20  O   PHE A   3      15.282 -10.204 -21.636  1.00 76.05           O
ATOM     21  CG  PHE A   3      16.507  -7.394 -18.257  1.00 76.05           C
```



Protein Distance heatmap



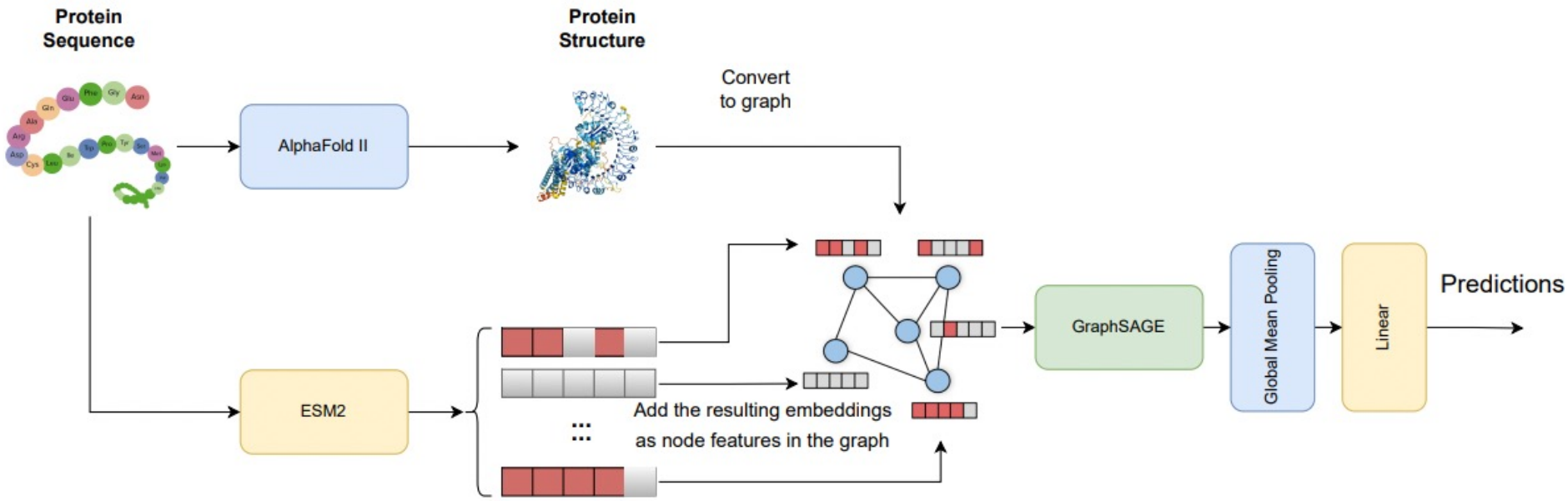
Protein Graph

Protein 3D structure - Features

Proteins represented as Graphs, where :

- **Nodes** : Amino-Acids
- **Node Features** :
  - **Amino-Acid type** : One-hot-encoding of the type (20 different types).
  - **Phi** angle : phi-angles of each residue.
  - **Psi** angle : psi-angles of each residue.
  - **RSA** : relative solvent accessibility of each residue.
- **Edges** : defined based on distance threshold ($<$12 Å) + Peptide bonds.
- **Edge Features :**
  - **Distance** : distance between residues.
  - **Bond type** : Peptide bond or distance bond

LM + GNN for AR prediction



The framework consists of two main components : (1) The AlphaFold model predicting the 3D structure of an input protein; (2) A classifier, consisting of a GraphSAGE model applied to the graph constructed from the structure and with node features from a pretrained protein language model (ESM2) applied to the sequence.

ARG Data set

- **COALA** : COllection of ALl Antibiotic resistance gene databases
- 70% max identity between sequence pairs.
- 16 AR classes.
- **Splitting :**
  - train set : 80% of the dataset
  - test set : 20% of the dataset
  - Stratified Splitting between train and test set.
  - Ensure test sequences have at most 40% identity with the training set sequences : To do so, we use CD-HIT, which produces global identity score between sequence pairs and cluster them and iteratively add clusters, making sure we have a stratified split.
  - –> 5 different splits of data

GNN for AR prediction

Results on test set				
	Accuracy	Accuracy (<50% id)	Accuracy (>50% id)	Accuracy (Homologue not found)
BLAST	0.6542±0.0057	0.7511±0.0178	**0.9529±0.0118**	0
Diamond	0.5886±0.0062	0.6469±0.0172	**0.9511±0.0153**	0.0005±0.0010
TF-IDF	0.5419±0.0162	0.5518% (± 0.172%)	0.8405% (± 0.207%)	0.3543% (± 0.125%)
TRAC	0.6980±0.0066	0.7278±0.0177	0.9092±0.0216	0.3683±0.0436
ARG-SHINE-CNN	0.6834±0.0127	0.6960±0.0178	0.9288±0.0070	0.3800±0.0388
Efficient-Net	0.6983±0.0089	0.7310±0.0217	0.9167±0.0236	0.3507±0.0288
GraphSAGE	0.6666±0.0098	0.6938±0.0168	0.8972±0.0148	0.3257±0.0318
ESM2_8M	0.7086±0.0133	0.7437±0.0170	0.9260±0.0050	0.3535±0.0493
ESM2_650M	**0.7310±0.0138**	**0.7652±0.0174**	0.9094±0.0080	**0.4216±0.0357**
ESM2_8M_GraphSAGE	**0.7290±0.0065**	**0.7649±0.0130**	0.9300±0.011	**0.3890±0.0398**

Conclusions

- Overall improvement of accuracy : More than 3%.
- Transformer Language models perform very well.
- Combining Sequence + Structure information leads to better performances.
- Deep Learning methods showed good performance detecting Ars even without similar homologues in the training set.

Overview

- Graph Neural Networks introduction
- GNNs for Antibiotic Resistance prediction
- GVAE for medical graph generation
- Our vision

GVAE for Generating Synthetic Patient Trajectories [1]

[1] Synthetic electronic health records generated with variational graph autoencoders,
Giannis Nikolentzos, Michalis Vazirgiannis, Christos Xypolopoulos, Markus Lingman, Erik G. Brandt
doi: https://doi.org/10.1101/2022.10.17.22281145

Health data - constraints

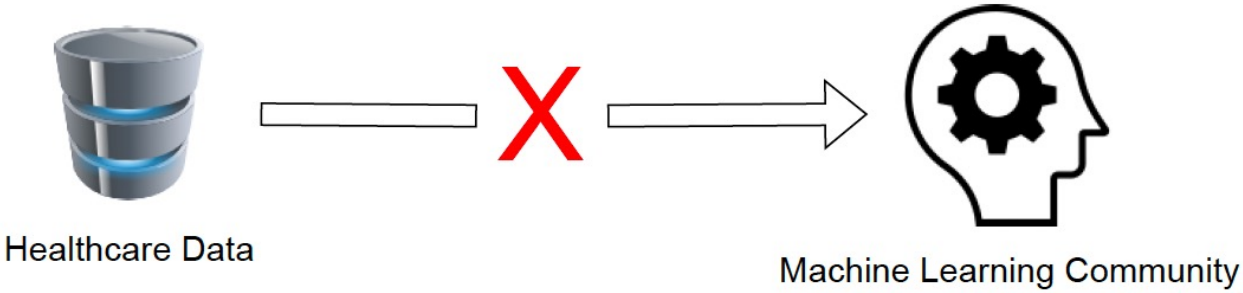
In many fields, open access data sets lead to significant progress, e. g.,

- Computer vision → Imagenet
- Natural language processing → Wordnet

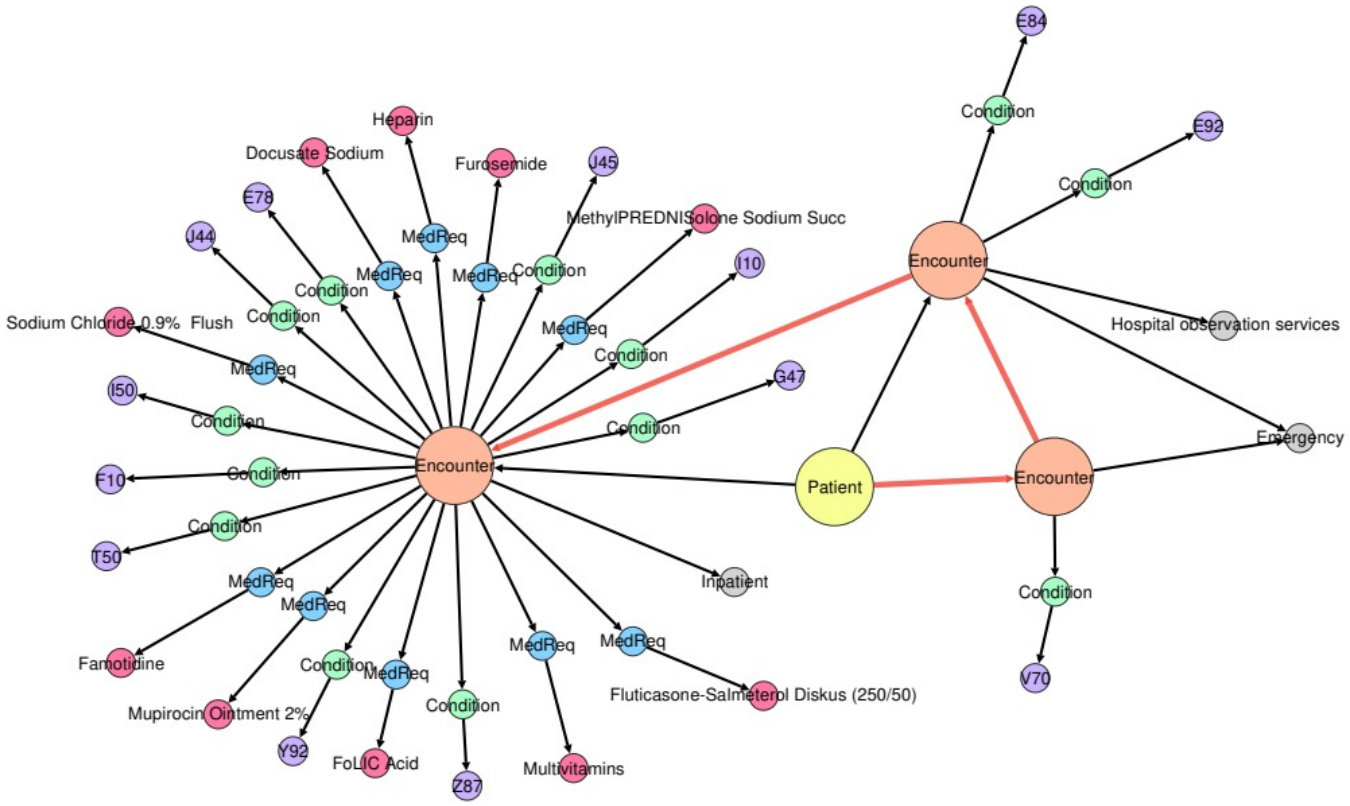
However, in the case of healthcare data:

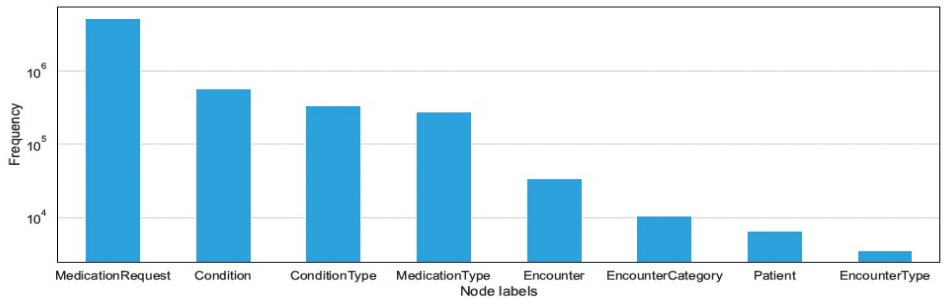
- Lack of high-quality healthcare data
- Strict regulations for data access:
  - The use of patient data leads to privacy concerns
  - Regulations like HIPAA prohibit the unauthorized use and disclosure of protected health information
  - Patient data cannot be freely shared

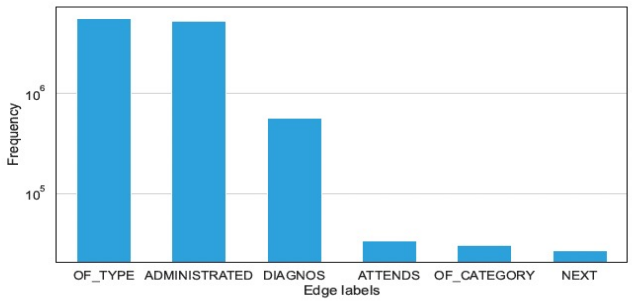
The above impedes machine learning research in healthcare!



Shaarpeec graph model for patient trajectories



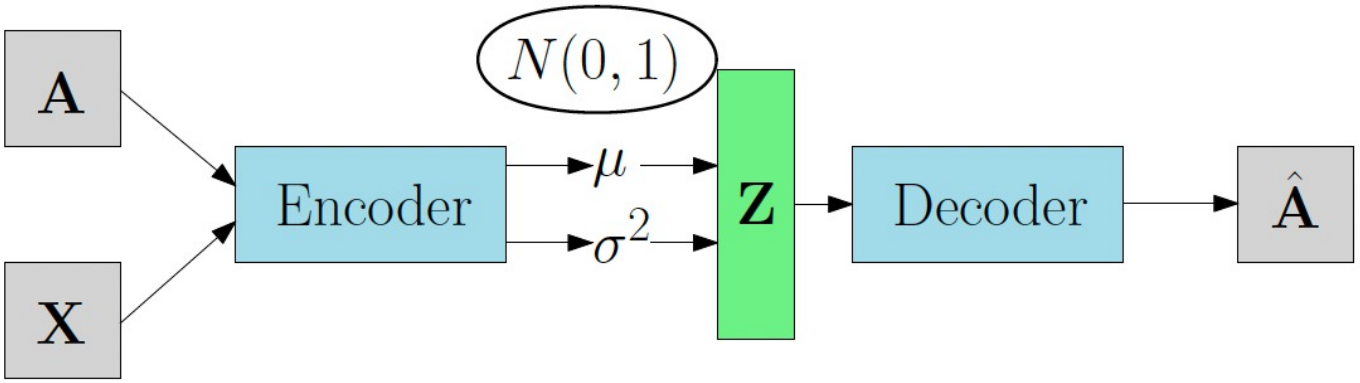




GVAE architecture

We use a variational autoencoder:

- The encoder maps input DAGs into d-dimensional gaussian distributions
- The decoder reconstructs the input DAGs given vectors sampled from the gaussian distributions



GVAE

Technical Details:

- Gaussian distribution represented by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ (encoded as a diagonal matrix $\rightarrow$ no correlation between elements of latent vector)
- Encoder actually generates mean vector and logarithm of the variance vector (to ensure encoder's output is unbounded)
- Minimize Kullback–Leibler (KL) divergence between produced distribution and the standard normal distribution
- Input to decoder is a vector sampled from the produced gaussian distribution, i. e., $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is sampled from a multivariate standard normal distribution
- Since KL divergence ensures that the encoder maps as close to a standard normal distribution as possible, to generate new graphs, we can sample vectors from a d-dimensional standard normal distribution and feed them to the decoder

GVAE Encoder

The encoder corresponds to a message passing graph neural network (GNN)

- First, an embedding layer maps nodes to vectors
- Then, an asynchronous message passing scheme is employed to update those representations (based on the topological sorting of the DAGs)
- The representation of each node is updated as follows:

$$\mathbf{m}_v = \sum_{u \in \mathcal{N}_{in}(v)} f(\mathbf{h}_u)$$

$$\mathbf{h}_v = \mathrm{GRU}(\mathbf{x}_v, \mathbf{m}_v)$$

  where $\mathcal{N}_{in}(v)$ is the set of incoming neighbors of v, f is some neural network (e. g., MLP), $\mathbf{x}_v$ is the embedding of node v and GRU is a gated recurrent unit

- After computing the representations of all the nodes of the DAG, we use the representation of the *end node* (i. e., the node without any successors) as the output of the encoder
- Then, we pass that vector on to two fully-connected layers to get the mean and variance parameters

GVAE Decoder

The decoder reconstructs the input graph:

- Given the latent vector $\mathbf{z}$ to decode, initial hidden state $\mathbf{h}_0$ is produced using a fully-connected layer
- Then, a DAG is constructed node by node
  - For the first node that is added to the graph (i.e., patient), we obtain its new representation as follows: $\mathbf{h}_{v_1} = \text{GRU}(\mathbf{x}_{v_1}, \mathbf{h}_0)$
  - For the i-th generated node v_i where $i > 1$, the following steps are performed:
    1. Compute node v_i's label distribution using an MLP based on the current graph state $\mathbf{h}_G = \mathbf{h}_{v_{i-1}}$
    2. Sample node v_i's label. If the sampled label is the end label, stop the decoding, connect all nodes without successors to v_i, and return the DAG; otherwise, continue the generation
    3. For $j = i-1, \ldots, 1$: (i) compute the edge probability of (v_j, v_i) using an MLP based on the representations of the two nodes $\mathbf{h}_{v_j}$ and $\mathbf{h}_{v_i}$; (ii) sample the edge; and (iii) if a new edge is added, update $\mathbf{h}_{v_i}$ as follows:

$$\mathbf{m}_{v_i} = \sum_{u \in \mathcal{N}_{in}(v_i)} f(\mathbf{h}_u)$$

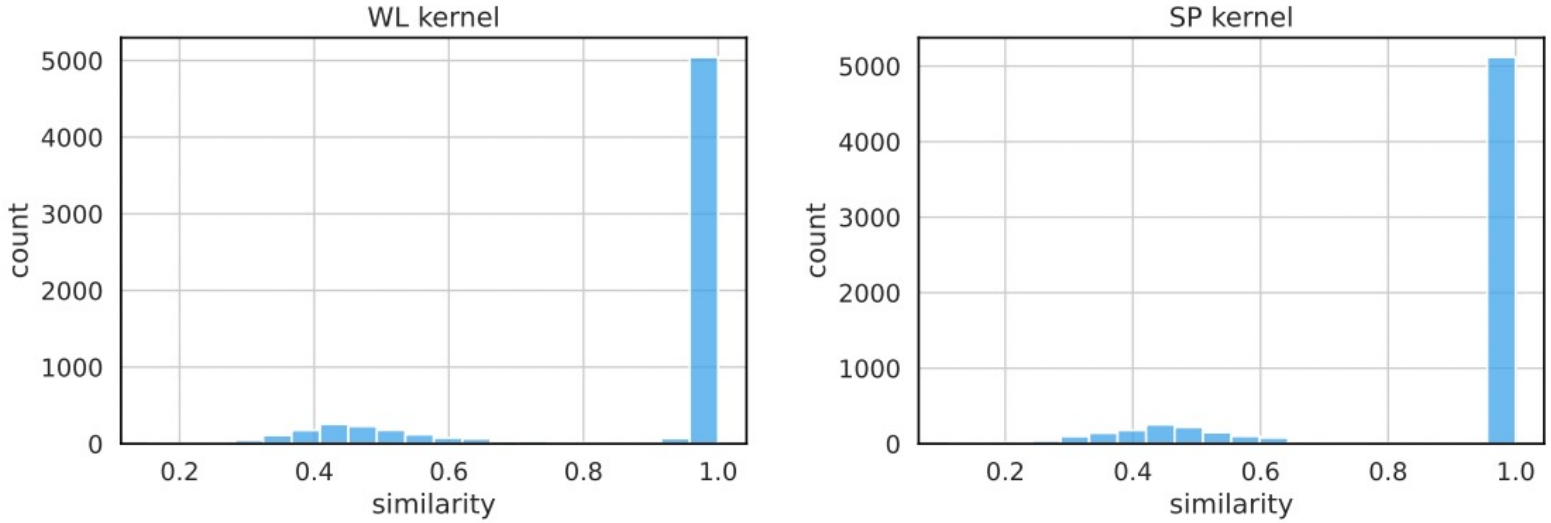
$$\mathbf{h}_{v_i} = \text{GRU}(\mathbf{x}_{v_i}, \mathbf{m}_{v_i})$$

EXPERIMENTS - MIMIC dataset

- Statistics on the patient trajectories calculated from the atrial fibrillation cohort extracted from the MIMIC-IV database.

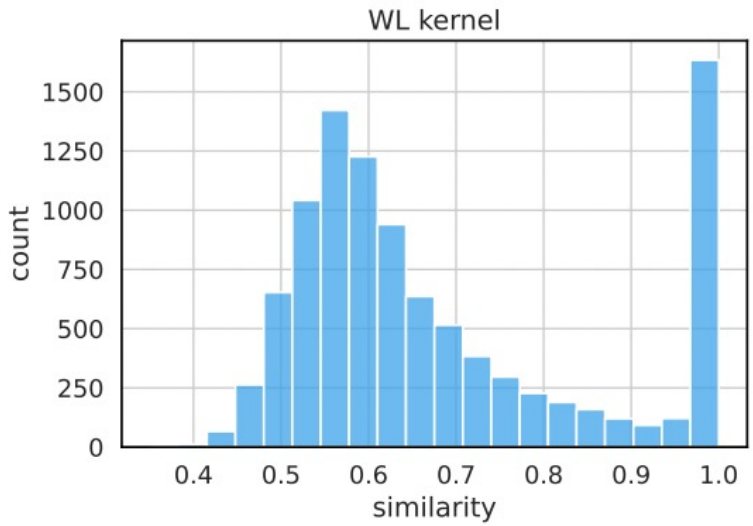
	Raw	**After Pre-processing**
Max # nodes	18,947	2,772
Min # nodes	10	10
Average # nodes	1,044.1	221.2
Max # edges	36,811	5,162
Min # edges	9	9
Average # edges	1,867.3	294.7
# node labels ($\lvert\Sigma_V\rvert$)	13,980	944
# edge labels ($\lvert\Sigma_E\rvert$)	6	6
# graphs	6,535	6,535

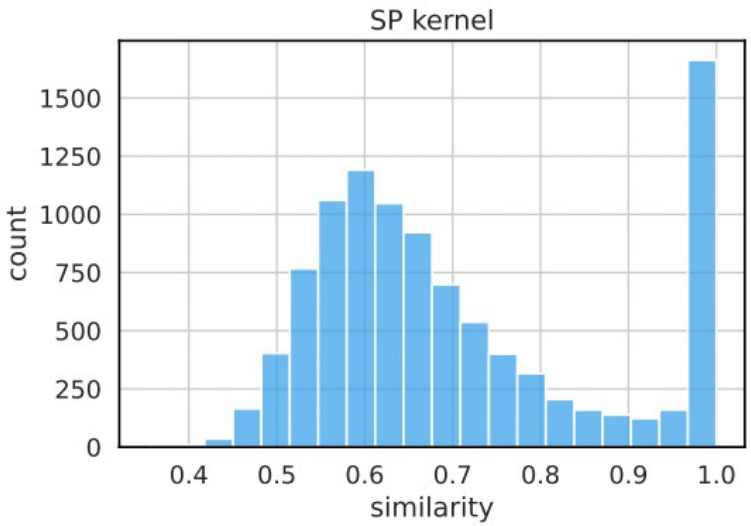
Quality of graph reconstruction



- Real graph input to encoder - decoder produces a reconstructed version
- Histogram of similarities between input graphs and reconstructed graphs using the Weisfeiler Lehman subtree (WL) kernel and the shortest path (SP) kernel.

Graph generation – similarity to real graphs





- Generated graph: Sample from the learned distribution and feed the decoder
- Similarity histogram between input and generated graphs using the Weisfeiler-Lehman subtree (WL) kernel and the shortest path (SP) kernel.
- low similarity => privacy

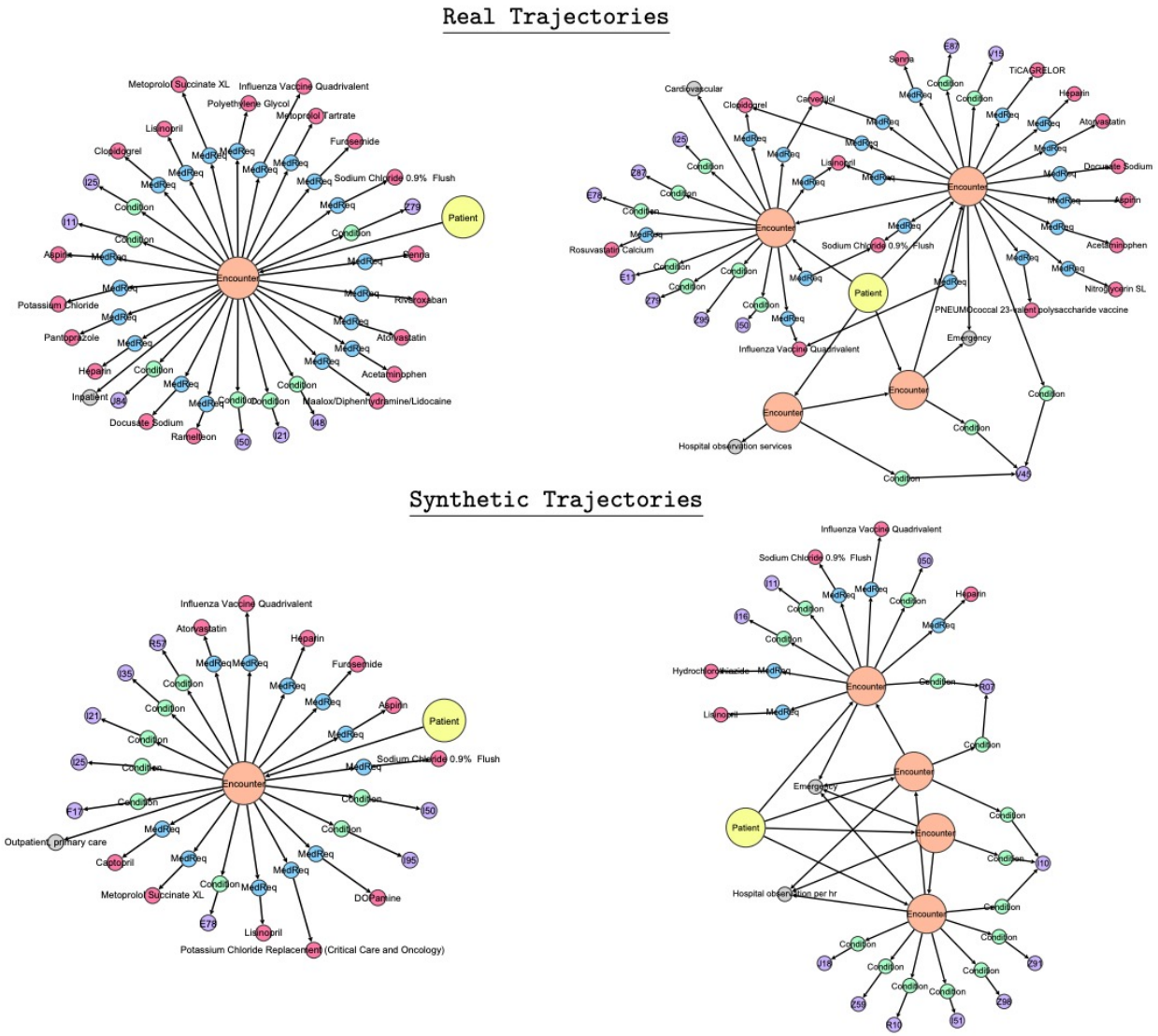
Graph generation – similarity to real graphs

- A mixture of real and generated graphs
- Classifier
  - cannot distinguish between real and generated ones when structure only
  - Less performant when labels are involved.

	Without Labels	With Labels
WL	58%	72%
SP	56%	73%

GVAEs for medical graph generation - Conclusions

- the model generates novel synthetic patient trajectories,
- sufficiently different to preserve patient privacy,
- yet retains the characteristics of the real-world data. Thus data exchange and ML is feasible
- most significant feature: capacity to learn long-range correlations between trajectory nodes.



Overview

- Graph Neural Networks introduction
- GNNs for Antibiotic Resistance prediction
- GVAE for medical graph generation
- Our vision

Our Vision – structure assisted protein pretrained models

- NLP pretrained models
  - GPT3, BARTHEZ
  - Very successful for downstream and generative tasks (i.e. translation, summarization, text generation…)
- Pretrain protein models based on structure as well
  - Capitalise on Alphafold2 database (~200M proteins) and large LM for proteins models (ESM2)
  - Downstream tasks (i.e. protein function prediction)
  - Generative tasks (new proteins …)

Our vision – multimodal pretrained models

- For proteins with textual description NOT "unknown" in UNIPROT
- multimodal pretraining of protein structures and textual descriptions (functions)
  - available resources for French: http://nlp.polytechnique.fr/

- Objectives would be to
  - given a description the decoder *produces a protein structure* that could be evaluated based on its k-nn neighbors.
  - given a protein structure *generate a textual description* that predicts its function
- **Tasks for evaluation..?**

Thank you!

Google Scholar: https://bit.ly/2rwmvQU

Submitted papers

[1] *Structure-Aware Antibiotic Resistance Classification Using Graph Neural Networks*, Aymen Qabel, **Sofiane Ennadir**, Giannis Nikolentzos, Johannes F Lutzeyer, Michail Chatzianastasis, **Henrik Bostrom**, Michalis Vazirgiannis. 2022/1/1, Journal, bioRxiv, https://www.biorxiv.org/content/10.1101/2022.10.06.511103v1

[2] *Synthetic electronic health records generated with variational graph autoencoders*, Giannis Nikolentzos, Michalis Vazirgiannis, Christos Xypolopoulos, Markus, Lingman, Erik G. Brandt, doi: https://doi.org/10.1101/2022.10.17.22281145

References [graph autoencoders]

[1] Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia. *Learning deep generative models of graphs*. https://openreview.net/forum?id=Hy1d-ebAb. 2018.

[2] Jiaxuan You, Rex Ying, Xiang Ren, William Hamilton, and Jure Leskovec. *GraphRNN: Generating Realistic graphs with Deep Auto-Regressive models.* International Conference on Machine Learning. PMLR, 2018.

[3] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. *Junction tree variational autoencoder for molecular graph generation.* International Conference on Machine Learning. PMLR, 2018.

[4] Matt Kusner, Brooks Paige, and Jose Miguel Hernandez-Lobato. *Grammar variational autoencoder*. International Conference on Machine Learning. PMLR, 2017.

[5] Teague Sterling, and John Irwin. *ZINC 15–ligand discovery for everyone*. Journal of Chemical Information and Modeling, 2015.

[6] Martin Simonovsky, and Nikos Komodakis. *GraphVAE: Towards generation of small graphs using variational autoencoders.* International Conference on Artificial Neural Networks, pp. 412-422. Springer, Cham, 2018.

[7] Hanjun Dai, Bo Dai, and Le Song. *Discriminative embeddings of latent variable models for structured data.* International Conference on Machine Learning, pp. 2702-2711. PMLR, 2016.

[8] David Alvarez-Melis and Tommi S. Jaakkola. *Tree-structured decoding with doubly-recurrent neural networks*. International Conference on Learning Representations. 2017.

[9] Manfred Droste and Paul Gastin. *Weighted automata and weighted logics*. Theoretical Computer Science, 380(1–2):69–86, 2007.

References [ARG]

1. Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia. *Learning deep generative models of graphs*. https://openreview.net/forum?id=Hy1d-ebAb. 2018.
2. S. F. Altschul, W. Gish, W. Miller, E. W. Myers & D. J. Lipman – « Basic local alignment search tool », J Mol Biol 215 (1990), no. 3, p. 403–410 (en).
3. F. AS. – « Complexity of protein folding », Bull Math Biol (1993).
4. B. Buchfink, K. Reuter & H.-G. Drost – « Sensitive protein alignments at tree-of-life scale using DIAMOND », Nat Methods 18 (2021), no. 4, p. 366–368 .
5. M.-N. Hamid & I. Friedberg – « Transfer learning improves antibiotic resistance class prediction », bioRxiv (2020).
6. J. R. L. W. R. M. e. a. J O'Neill, S Davies – « Review on antimicrobial resistance, tackling drug-resistant infections globally : final report and recommendations. », Welcome Trust and UK Government (2016).
7. E. R. P. A. e. a. Jumper, J. – « Highly accurate protein structure prediction with alphafold. », Nature (2021) (en).
8. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. d. Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido & A. Rives – « Language models of protein sequences at the scale of evolution enable accurate structure prediction », bioRxiv (2022).
9. Z. Wang, S. Li, R. You, S. Zhu, X. J. Zhou & F. Sun – « ARG-SHINE : improve antibiotic resistance class prediction by integrating sequence homology, functional information and deep convolutional neural network », NAR Genomics and Bioinformatics 3 (2021), no. 3, lqab066.