Structural bioinformatics and machine learning for drug design
1. Target-centric drug discovery
2. Binding pockets
3. Geometric objects and their representation
1. Target-centric drug discovery
   a. Drug discovery challenges
   b. Drugs mechanism of action: target-centric

2. Binding pockets

3. Geometric objects and their representation
Drug discovery is a **Large** market experiencing **strong pain points**, with **disruptions** happening now.

### Massive R&D market with large budgets

- **190 $B**

### More and more difficult to find new drugs

<table>
<thead>
<tr>
<th>Year</th>
<th>R&amp;D return of top biopharma companies</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>10.1%</td>
</tr>
<tr>
<td>2019</td>
<td>1.8%</td>
</tr>
</tbody>
</table>

### Computational methods are on the rise

<table>
<thead>
<tr>
<th>Year</th>
<th>Investments in AI in Pharma</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>0.2 $B</td>
</tr>
<tr>
<td>2019</td>
<td>1.4 $B</td>
</tr>
</tbody>
</table>

---

1. Projected returns from late stage pipeline assets divided by total spend incurred bringing those assets to launch, for the 12 largest biopharma companies by 2009 R&D spend.

Sources: Deloitte “Measuring the return from pharmaceutical innovation 2019”; EvaluatePharma June 2020; Deep Pharma Intelligence “AI for drug discovery, biomarker development and advanced R&D landscape - Overview 2020”
Drug discovery is a **high stake bet** on the **right target and right molecule**

Research: Find the **right target** and the **right molecule**
Pharmas struggle to find the right molecules that will have a high affinity with the target.

There are too many possible molecules: Pharmas need to select a few only to test them...

... but current selection methods do not allow to make the right choices.

Choices from Chemists are biased and limited by individuals experience.

Choices from computational approaches are not satisfactory either.
Takeaways

- It is increasingly hard to find drugs (we found all the easy solutions)
- Computational methods play a key role there as the room for improvement there is still huge
Mechanism of action

- The underlying chemical process of drugs is almost always a binding event to enzymes or receptors

- Enzymes
  - Inhibition
  - Activation

- Receptors
  - Agonism (activators)
  - Antagonism (blockers)
Target-centric drug discovery

- Target-centric starts by finding the target: the biomolecule to be affected by the drug
- Then it uses the target structure to simulate the interaction with potential binders
- This is the focus of my research :)
Novel targets for drug discovery

- Traditional targets are exhausted
- Multi-target drugs
- PPI drugs
- RNA drugs
Target-centric challenges

- One needs to have the target and its structure
- Chemical space dilemma:
  - Classical space is exhausted
  - Most of the known space is hard to synthetize
Takeaways

- Drugs act by binding a molecule (traditionally a receptor or enzyme, now other things)
- Using the structure of the targeted molecule is known as target-centric drug discovery
1. Target-centric drug discovery

2. Binding pockets
   a. Finding the binding pockets
   b. Usage and representation of the binding pocket

3. Geometric objects and their representation
Finding the binding pockets

- A binding pocket is a part of the target where a compound might bind
  - Traditionally a cavity, ideally charged and able to create h-bonds
  - Lock-Key model

- Traditional way
  - Fpocket is kind of the reference
  - Based on insights and simulations from the physics

Fpocket: an open source platform for ligand pocket detection (2009)
Pocket finding: ML approach

- A concurrent approach is to use machine learning to find these pockets

- Several example of bound compounds are in the PDB
  - Pioneer method is DeepSite

Addressing more challenging targets

- RNA targets are difficult because the pockets are less stable
- PPI targets prediction need to have a prediction of the PPI site along with the druggable site
Takeaways

- Drugs usually bind to specific parts of the biomolecules: binding pockets
- Geometric deterministic and more recently ML methods were developed to detect these pockets
Docking on a binding pocket

- Given a pocket and a small molecule:
  - Minimize a binding energy score by randomly moving the compound
  - Get a list of poses and their energy scores

- This has been WIDELY used by pharmas to filter drug banks

*AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading (2010)*
RNA pockets representation

- A first step for machine learning alternatives is pocket representation (more later)
- RNA can be represented with special graphs that were shown to perform well

Augmented base pairing networks encode RNA-small molecule binding preferences (2020)
RNA pockets binder prediction

- We now have RNA pockets represented as graphs
- Can we directly try to predict the optimal binder?
  - We predict a pharmacophore: helps filtering the banks

Augmented base pairing networks encode RNA-small molecule binding preferences (2020)
Target-centric challenges

- One needs to have the target and its structure

- Chemical space dilemma:
  - Classical space is exhausted
  - Most of the known space is hard to synthetize

Remember this slide?
What are generative models?

- Samplers basically: they learn a data distribution by learning to transform a known one.

Learnable Function: f

Training

Table of Random Numbers

<table>
<thead>
<tr>
<th>Z</th>
<th>30006 3077</th>
<th>39813 50396</th>
<th>70655 36621</th>
<th>3187 62605</th>
<th>49428 1275</th>
<th>46352 16338</th>
<th>28157 73701</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>10402 16454</td>
<td>31491 4868</td>
<td>8602 5089</td>
<td>12907 20506</td>
<td>29836 74583</td>
<td>58005 51766</td>
<td>65809 4416</td>
</tr>
<tr>
<td>Z</td>
<td>46993 97405</td>
<td>6149 89507</td>
<td>63922 37358</td>
<td>70635 47547</td>
<td>71514 73303</td>
<td>7962 37373</td>
<td>80920 77771</td>
</tr>
<tr>
<td>Z</td>
<td>10387 15446</td>
<td>31873 42034</td>
<td>86035 12058</td>
<td>12908 20501</td>
<td>29835 74583</td>
<td>58005 51766</td>
<td>65809 4416</td>
</tr>
<tr>
<td>Z</td>
<td>46993 97405</td>
<td>6149 89507</td>
<td>63922 37358</td>
<td>70635 47547</td>
<td>71514 73303</td>
<td>7962 37373</td>
<td>80920 77771</td>
</tr>
<tr>
<td>Z</td>
<td>10387 15446</td>
<td>31873 42034</td>
<td>86035 12058</td>
<td>12908 20501</td>
<td>29835 74583</td>
<td>58005 51766</td>
<td>65809 4416</td>
</tr>
<tr>
<td>Z</td>
<td>46993 97405</td>
<td>6149 89507</td>
<td>63922 37358</td>
<td>70635 47547</td>
<td>71514 73303</td>
<td>7962 37373</td>
<td>80920 77771</td>
</tr>
</tbody>
</table>
Molecular VAEs

Sample conditioned by $z$:
$$\tilde{z} \sim \mathcal{N}(z)$$

Push $z$ to be close to $\mathcal{N}(0, 1)$
Docking score optimization

- First train a molecular generative model
- Fine-tune the generative model to change the distribution to optimize a property (here docking scores)
Pocket flexibility and molecular dynamics

- Biomolecules evolve through time, they are dynamic objects.
- Their properties depend on the whole conformational ensemble:
  - We should be using this ensemble as a pocket representation.
  - Frame-by-frame analysis is a first step!
Takeaways

- Once equipped with the pocket, we can try to find a set of candidate molecules to bind to it
  - Several settings exist

- The pocket representation should also include some flexibility although it is not well established yet
Agenda

1. Target-centric drug discovery
2. Binding pockets
3. Geometric objects and their representation
   a. Representation for machine learning
   b. Learning on graphs
   c. Equivariant networks
Representing objects

- Machine learning is about mining relationships between complex concepts.
- Machine learning methods mostly take vectors as inputs and outputs.
- These vectors are a representation of the complicated concepts we want to learn.
Experts descriptors

Hand-crafted Descriptors

- Fixed-sized vectors
- Crafted by experts
- More understandable

How much does this apartment cost?

Vector used for ML:

- 140 \textit{(size in square meters)}
- 20 \textit{(size of bedroom 1)}
- 10 \textit{(size of bedroom 2)}
- 15 \textit{(size of terrace)}
Learnt representations

- Input only raw data
- Let the algorithm find the patterns
- Need more data and better models

How much does this apartment cost?
Representing geometric objects

- How can we represent this complicated object into a computer?
  - A photograph in 3D?
  - Cloud of points? (x,y,z coordinates)

- What is the best representation?

- How to learn over stuff more complicated than vectors?
  - Graphs
  - Spaces acted upon by a group
Takeaways

- In machine learning, our models take inputs and outputs (they are functions)
  - These are often vector without structure, for instance expert descriptors (traditional ML)
- They could be more structured objects
- If we learn the representation, we need to enforce some properties in the model (prior information)
What are graphs?

- Mathematically speaking, graphs are:
  - Vertices or nodes
  - Edges

- The nodes can be isolated or linked to many others. The edges can include information and possibly a direction.

- It's a abstract object than can be adapted to many things: when objects can be related one to another.

- This flexibility makes it ubiquitous.
What are graph neural networks?

- Let us imagine that we want to learn something over graph data, leveraging graph structure.
- One way to see it is that the node initial information should be shared with its successive neighborhoods.
- The sharing happens through a parametric function that is learnable.
RNA as graphs

- RNA can be represented with special graphs that were shown to perform well.

- Using this graph representation is useful for several machine learning applications:
  - Structural motif mining
  - Drug design
  - Pockets detection

Augmented base pairing networks encode RNA-small molecule binding preferences (2020)
RNAGlib: a python package for RNA 2.5 D graphs (2022)
Takeaways

● Graphs are an example of structured data for machine learning systems.

● Graph neural networks are an example of model that accept this structured data as inputs

● GNNs can be used for structural biology and drug design research, for instance through the modelling of RNA structure
What is a group?

- Theoretical point of view:
  - A set along with an operation: \( \cdot : G \times G \rightarrow G \)
  - This operation must satisfy associativity, existence of identity and of inverse

- Examples:
  - Addition for \( \mathbb{Z} \) or \( \mathbb{R} \)
  - Multiplication for \( \mathbb{R}^* \) or \( \mathbb{Q}^* \)
  - \( \text{GL}(K) \) (invertible matrices) or subgroups (rotations or translations)
What is a representation?

Given a group and a set of finite dimension $V$, $G$ acts on $V$ with:

$$\rho : G \rightarrow \text{Bij}(V)$$

$$\forall g_i \in G, \rho(g_1 \cdot g_2) = \rho(g_1) \circ \rho(g_2)$$

$$R = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

$$R_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{bmatrix}$$
Equivariance property:

Given a group action on the input of the function psi, we want to have a predictable group action of the output.

\[ \pi_2^g(\phi(x)) = \phi(\pi_1^g(x)) \]  

Cohen et al (2016)
Equivariant networks

- If we stack equivariant functions, we get an equivariant network
- The learnt representation does not depend on the group action anymore!
Rotation equivariance

- An example problem: rotation invariance

- Gravity is not relevant for these objects
  - The nature of proteins does not depend on a top direction
  - Our hidden truth should not depend on it

*General E(2)-Equivariant Steerable CNNs (2019)*
Reverse complement symmetry:

- **Sequencing** gives access to substrings called *reads*
  - One strand
  - Follows biochemical direction

- From one strand, one knows the other one
RC equivariance

Framing the redundancies as group actions:
- Random strand sequenced: RC symmetry ($G = \mathbb{Z}_2$)
- Random cuts: translation symmetry ($G = \mathbb{Z}$)
Results

- Theoretical result:
  - We exhibit all linear RC-equivariant layers and all pointwise RC-equivariant nonlinearities

- State-of-the-art results
  - Larger space gives more freedom for the usual deep learning tinkering!
Takeaways and questions

- Some input data is acted upon by a group action, usually this means that the raw data has an arbitrary choice.
- Some models exist that ‘respect’ this group action, (equivariant networks), enforcing invariance to arbitrary choices.
- The most classical application is rotation equivariance, but other applications exist such as RC-equivariance for DNA.
Conclusion

- Drug design companies are massively investing in AI and in structural bioinformatics: this research shows great promise.

- Structure is key for the target-centric approach and several tools were already developed to help finding the right binders for a target.

- To account for this structure in the data, dedicated methods are needed. Some exist but there is still room for research there too.
Thanks for your attention! Any questions?

You can contact me by mail too: vincent.mallet96@gmail.com