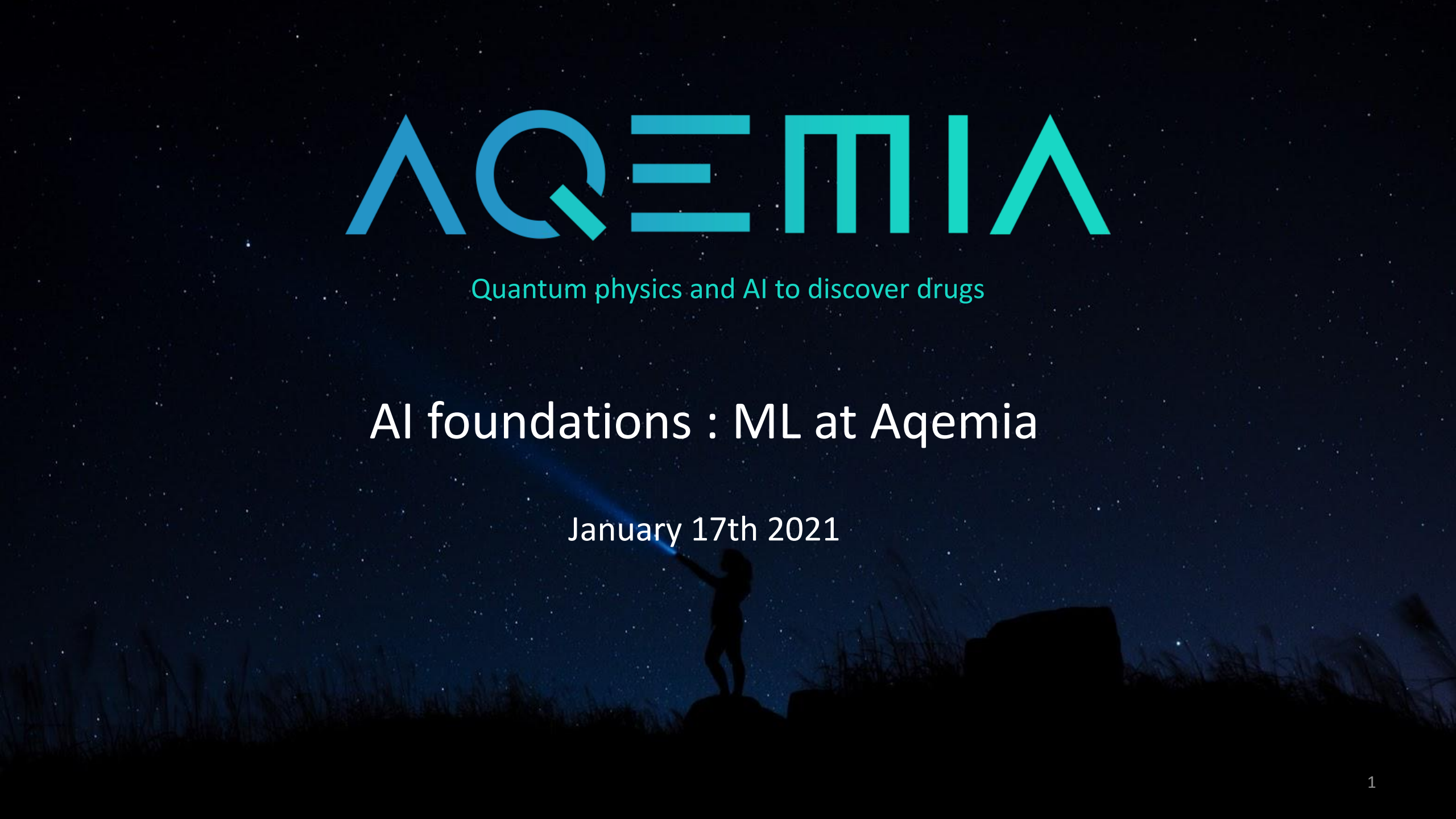


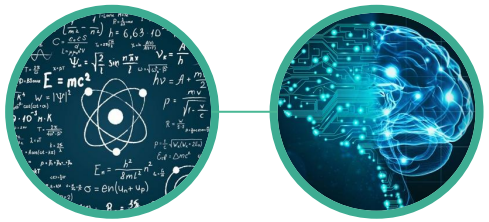
AI foundations : ML at Aqemia

January 17th 2021

Aqemia at a glance

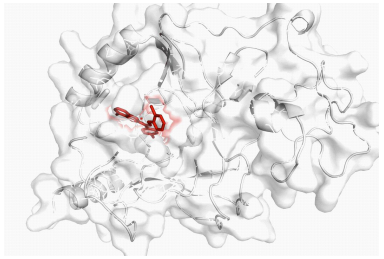
AQEMIA

Digital platform to invent new molecules



Quantum inspired Physics

Artificial intelligence



Applied to drug discovery

ENS ÉCOLE NORMALE SUPÉRIEURE

UNIVERSITY OF CAMBRIDGE

UNIVERSITY OF OXFORD

10+ years academic research

Product up & running

Team of 25

Collab with top pharmas

$12M funding

Internal pipeline

24 months

AGENDA

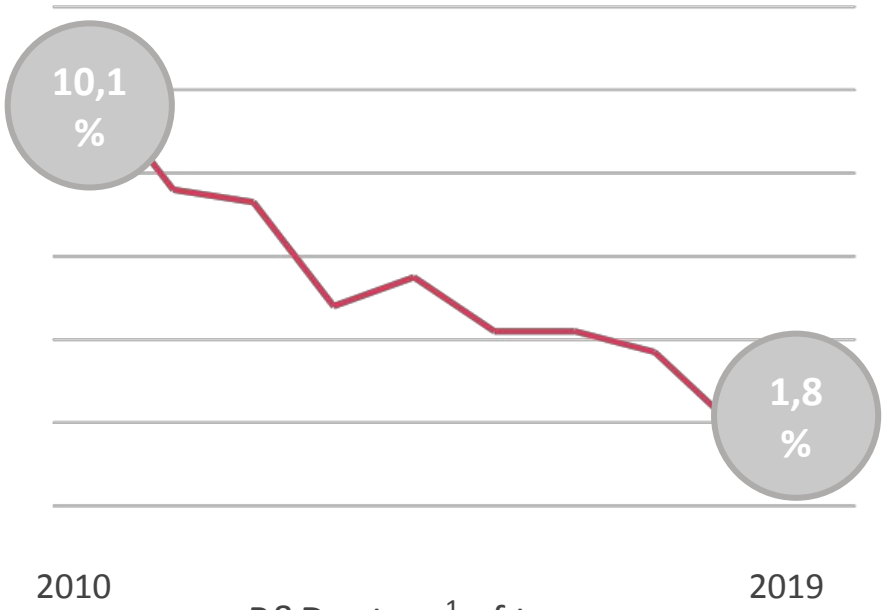
- Introduction to drug discovery
- Challenges in ML models design
- Everyday-use of ML models

Drug discovery is a **Large** market experiencing **strong pain points,** with **disruptions** happening now

Massive R&D market with large budgets

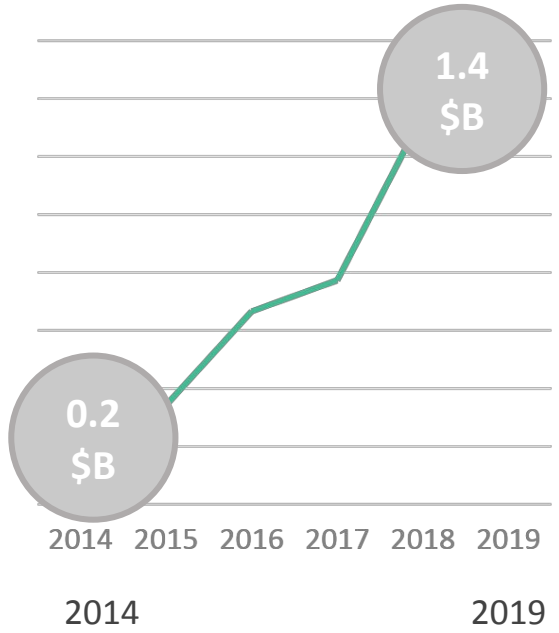
Annual pharmaceutical R&D expenses

More and more difficult to find new drugs



R&D return[1] of top biopharma companies

Computational methods are on the rise



Investments in AI in Pharma

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**

10-15 years

1.5B$ to put a drug on market

Target identification

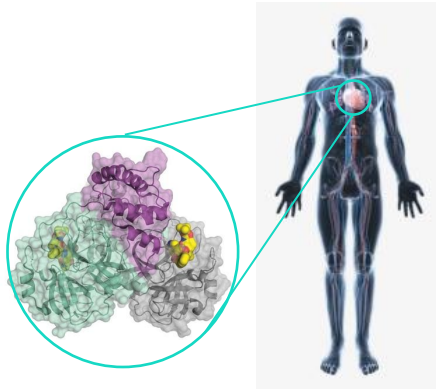
Candidate invention

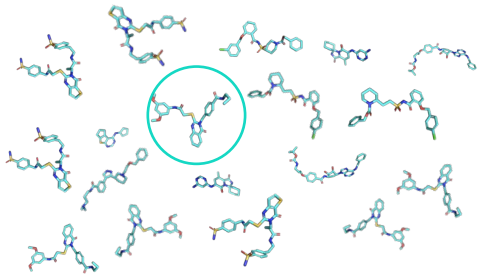
Pre-clinical trials

Phase I trials

Phase II trials

Phase III trials



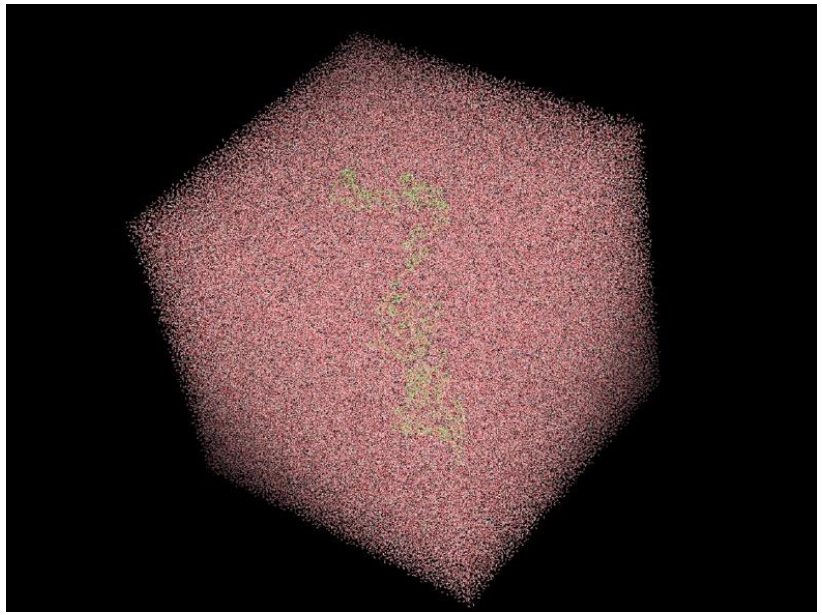


⚠ Research: Find the right target and the right molecule

Aqemia unique technology: Our secret sauce to compute affinity 10 000x faster

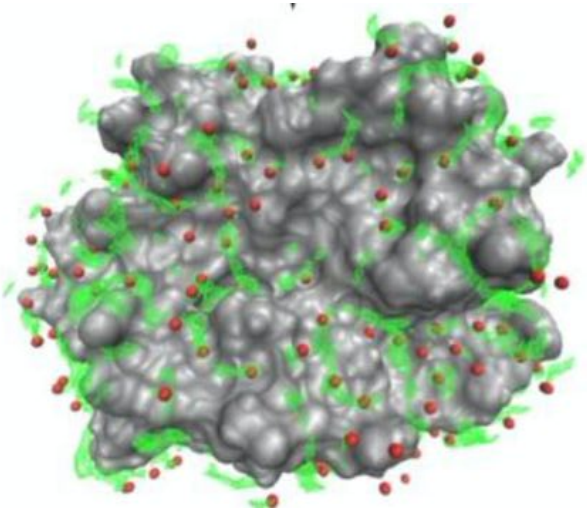
(hint: it is not about 10 000x more computers in parallel)

The **historical statistical mechanics** accurate prediction of affinity computes **averages of all interactions** between protein, small molecule and **100 000's of water molecules**



Vs.

We use **novel quantum-inspired theory** to compute very fast **the water probability density distribution** around the protein and the small molecule



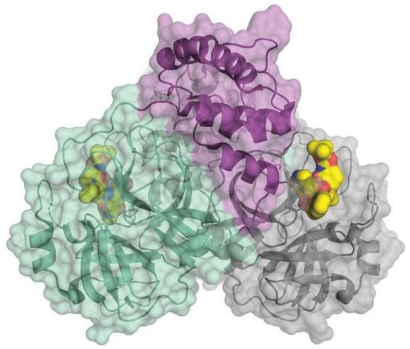
$$\hat{\gamma}^{m}_{\mu;\chi}(\mathbf{q}) = \sum_{n\nu} (-1)^{\chi+\nu}\, \hat{c}^{mn}_{\mu\nu;\chi}(q)\Delta\hat{\rho}^{n}_{\underline{\nu};\chi}(\mathbf{q})$$

Mamilien Levesque
Cofounder & CEO

- ***10+ years*** *of research and* ***40+*** *publications*
- *Published theory selected as 2017* ***Editor's choice by American Institute of Physics***
- *Aqemia has an* ***exclusive licence*** *on technology from Paris Sciences Lettres*

ENS ÉCOLE NORMALE SUPÉRIEURE · PSL UNIVERSITÉ PARIS · UNIVERSITY OF OXFORD · UNIVERSITY OF CAMBRIDGE · AIP American Institute of Physics

Our unique fast & accurate affinity prediction **unlocks de novo drug design**



Target protein structure

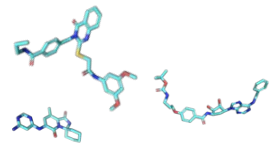
AQEMIA

Generation of molecules

With AI

Scoring of molecules

With Physics

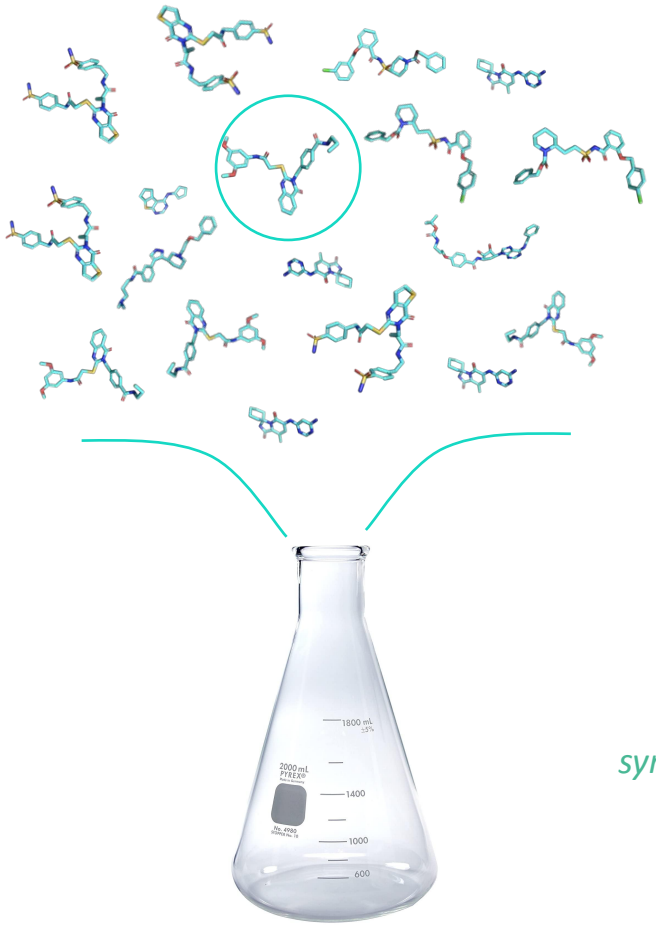


Drug candidates

Aqemia **unique fast & accurate affinity scoring**
+ Other property predictions based on AI models
+ Synthetic accessibility of generated molecules

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…



10^{60}

potential molecules

10^{3}

Molecules that can be synthetized & tested in vitro in a project

… 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

see next slide

AGENDA

- Introduction to drug discovery
- Challenges in ML models design
- Everyday-use of ML models

Designing ML models from a usecase

Preliminary questions can impact a lot the success of a ML project :

What is the target we want to predict ?

- Categorical vs continuous
- Error bars
- Comparable across experiments ?

What data is available ?

- Comparable across datasets ?
- Acquisition biases ?

What is the metric we care about ?

- False positives vs false negatives
- Correlation vs mean squared error

How should we split the data ?

- Keep a separate dataset for test ?
- Stratified split, time-split ?

Designing interpretable models

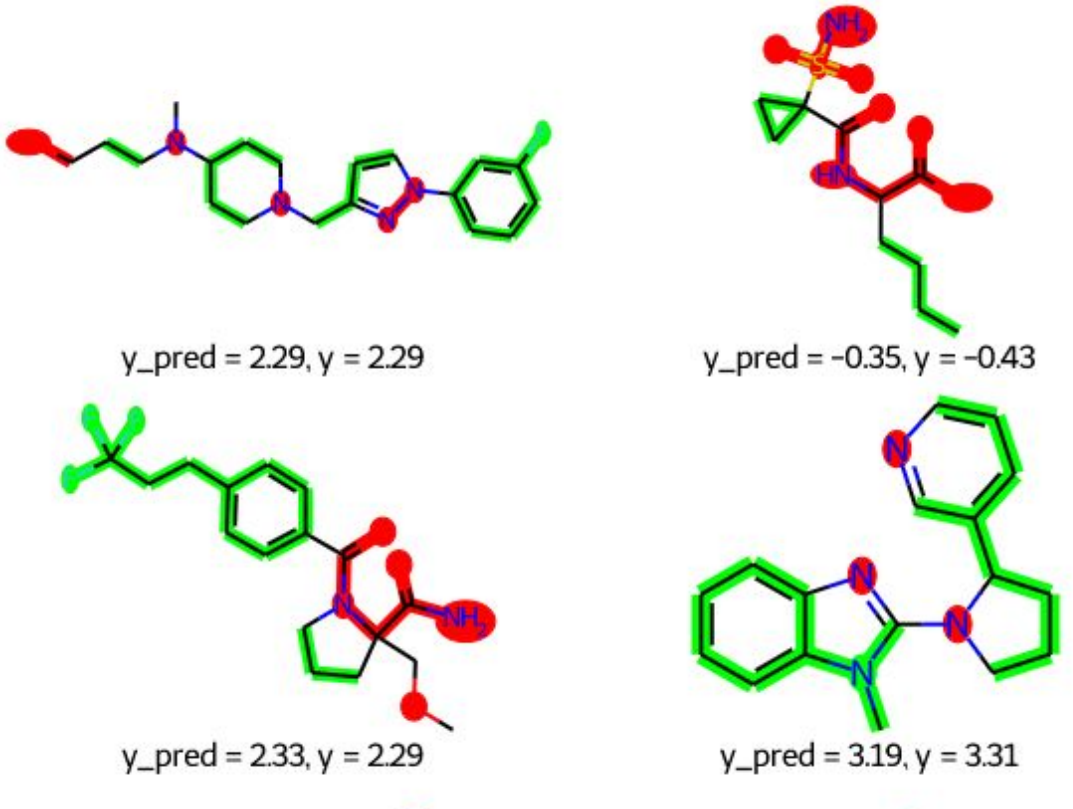
Interpretability methods **attribute** the prediction to the features of the input
-> *Integrated Gradients (Sundararajan 2017)*

Model debugging :

- Can we explain the error on a validation set by looking at model attributions ?

Focused optimization :

- Can chemists propose modifications of the substructures that get negative attributions ?



Dealing with data constraints

Biological data often falls in two settings :

1) Abundant data, mostly unlabeled and noisy →
   - data processing pipelines affect model performance
   - self-supervised learning

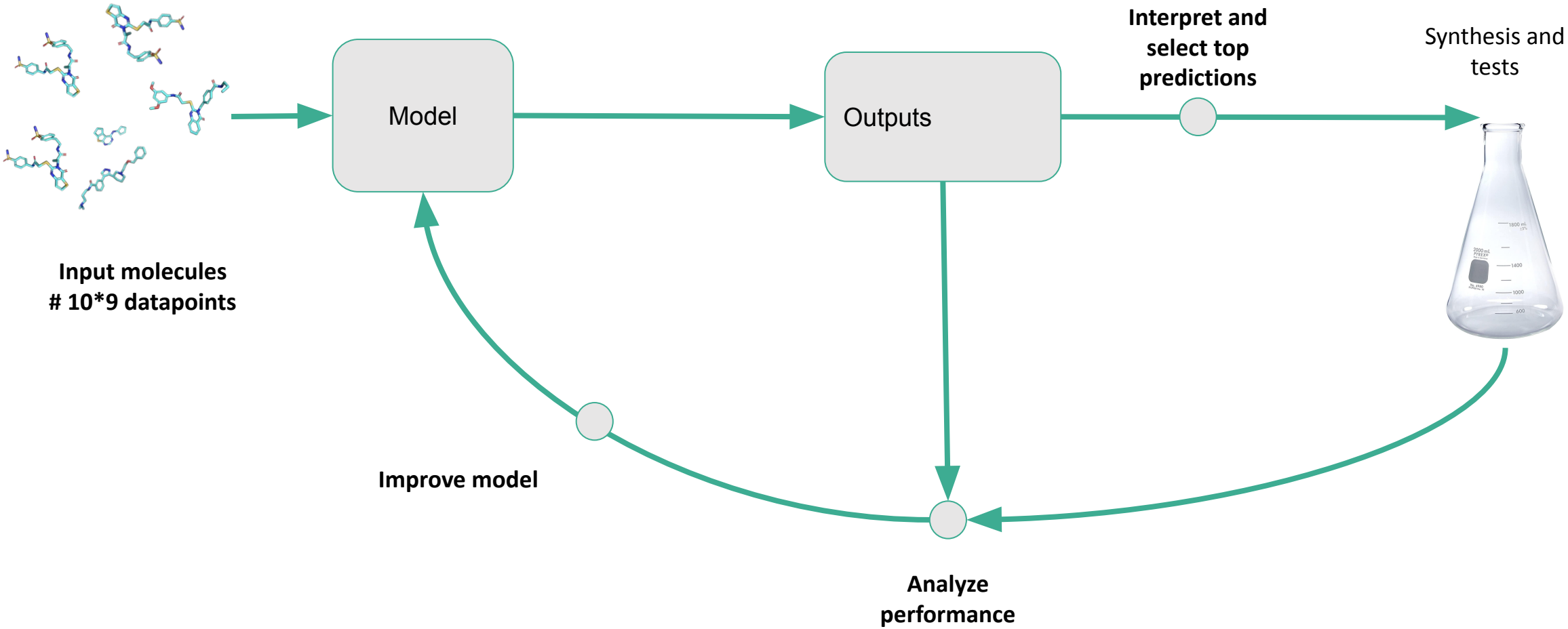
2) Scarce, labeled data obtained from costly experiments
   - Focus on model ability to generalize to new datasets
   - Iteratively re-train models on new datapoints

AGENDA

- Introduction to drug discovery
- Challenges in ML models design
- Everyday-use of ML models

Designing a ML model is only the beginning …

- Models are integrated in an experiment loop to design better and better molecules.
- Automation is key to cut the loop time.

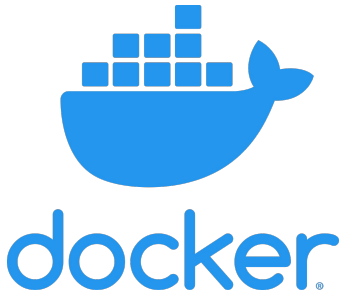


Sharing models between researchers and across teams

- Pytorch Lightning **organizes pytorch code** so that models are interchangeable and easy to maintain

```
class LitModel(LightningModule):
    def training_step(self, batch, batch_idx):
        x, y = batch
        y_hat = self(x)
        loss = F.cross_entropy(y_hat, y)
        return loss
```

- Docker embarks all dependencies and softwares needed for your model into a **docker image** that can run on any machine



From model design to predictions on billions of molecules :

Development

Make model

Lightning

Push code to github

Push weights to cloud storage

Automatic deployment

Test model performance
Build docker image

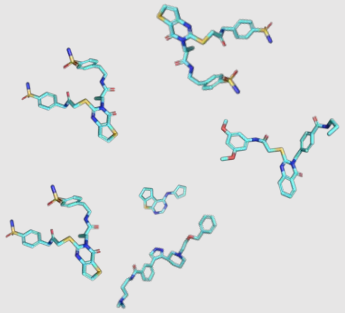
GitHub Actions

Image pushed to the cloud, ready to run

docker

Serving

Inputs cloud storage



Computing cluster

Predictions dataframes **cloud storage**

Recommended reading to build "ML-ops" pipelines : https://github.com/graviraja/MLOps-Basics

If you want to revolutionize drug discovery

- Developing cutting-edge algorithms
- Based on DL, chemistry & physics
- With an amazing & friendly multidisciplinary team

careers@aqemia.com

AQEMIA