netherlands eScience center

DeepRank:
finding the true love between proteins

Cunliang Geng
eScience Research Engineer
Netherlands eScience Center

SUPER DAY
17/12/2019 | EYE A'DAM

WHAT

AI, Deep Learning & CNN

HOW

PDaMED work model for ML projects

EXAMPLES

DeepRank for classification

DeepRank for ranking

AI: a powerful and versatile tool

Tools' mission is to solve problems

AI is a computer-based tool

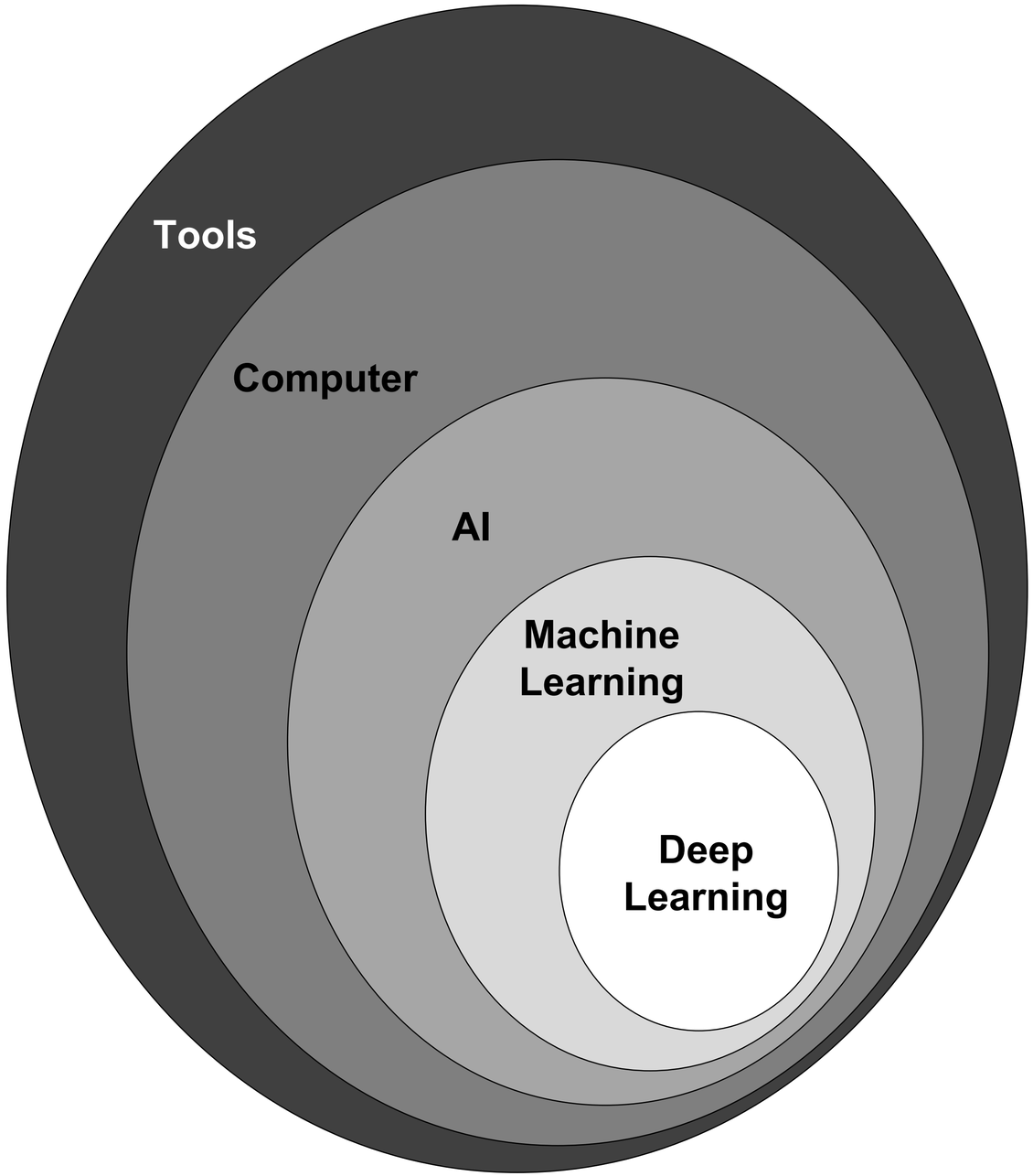
AI cannot do everything as computer has limits

Machine learning (ML):

learn experience from data to improve performance on specific tasks

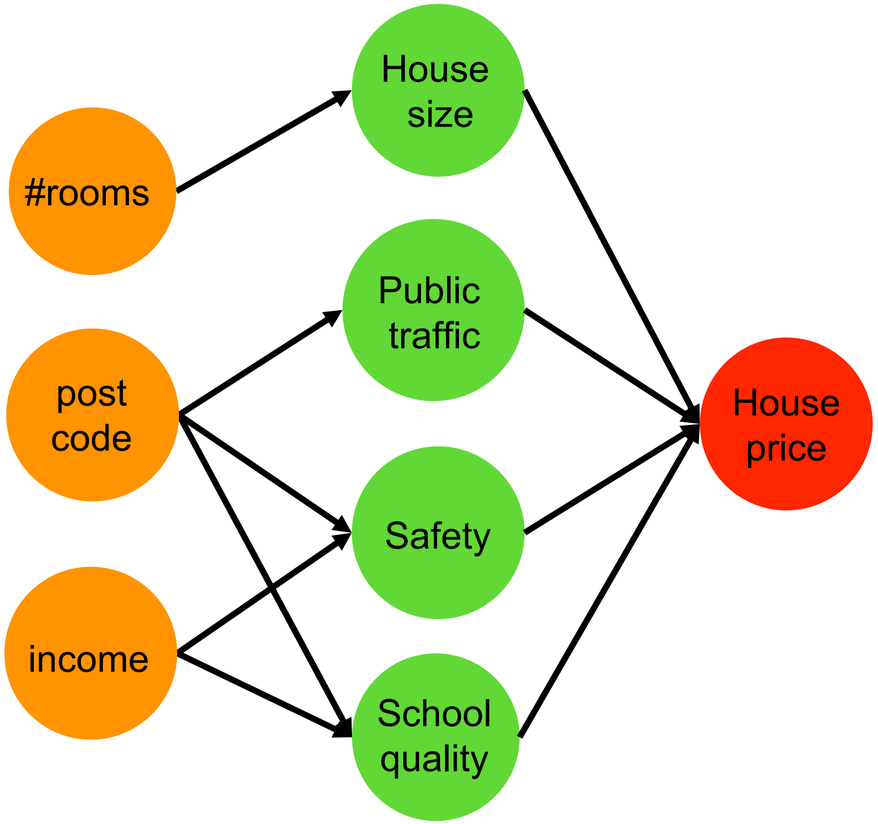
$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X})$$

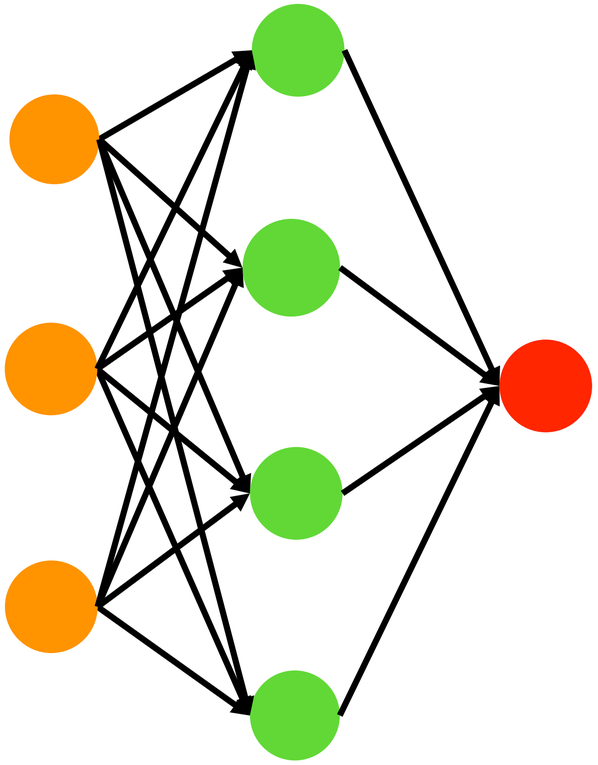
Deep learning can learn more complex functions to solve more complex problems



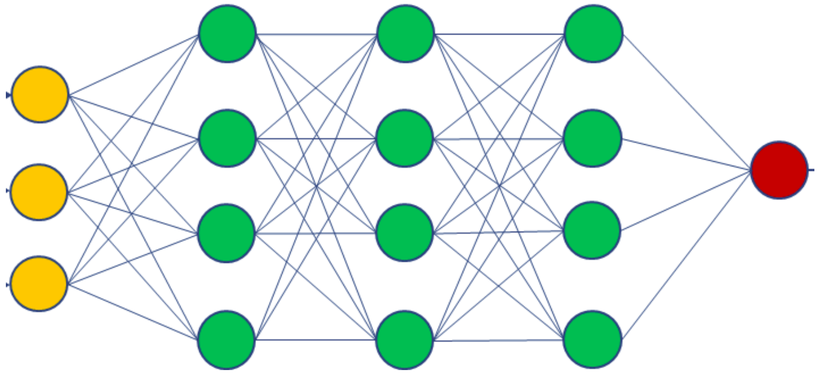
Deep Learning (DL)

- DL is a method based on artificial neural networks (NN)
- In essence, an artificial NN is a very complex (non-linear) function $\boldsymbol{Y = f(X)}$





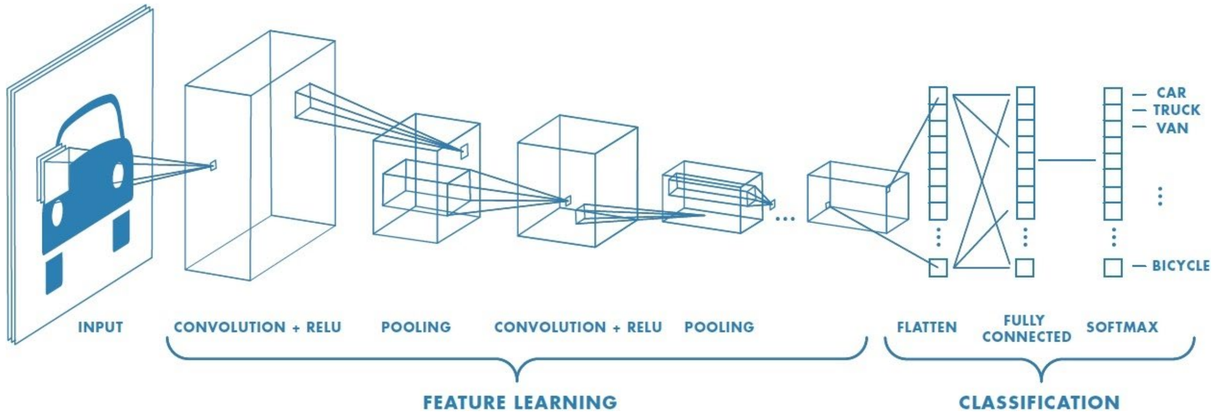
input hidden output



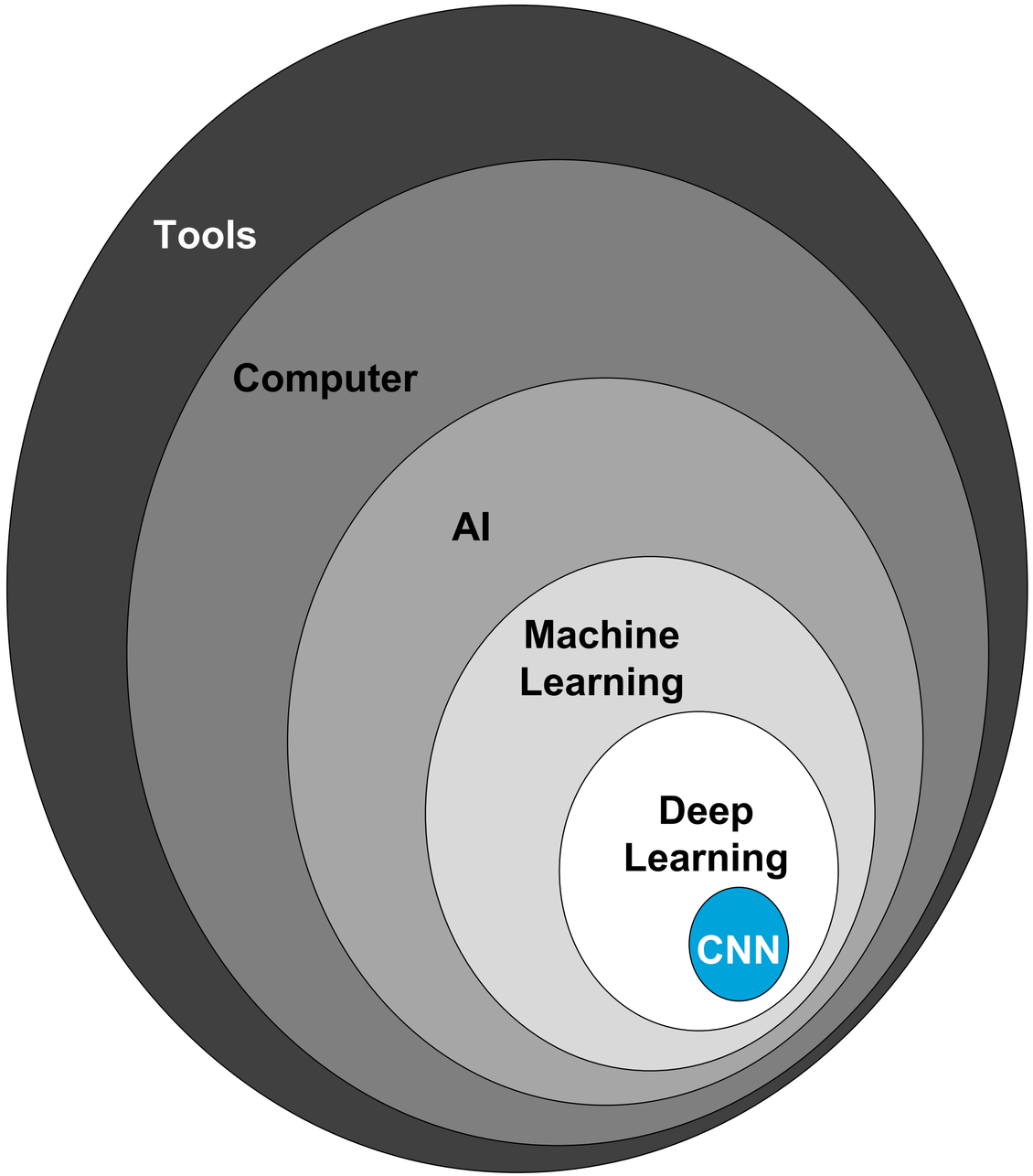
Go deeper: more hidden layers

Convolutional Neural Network (CNN)

- CNN is one of the most successful DL tools
- CNN is good at learning from image-like data



https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html



WHAT

AI, Deep Learning & CNN

HOW

PDaMED work model for ML projects

EXAMPLES

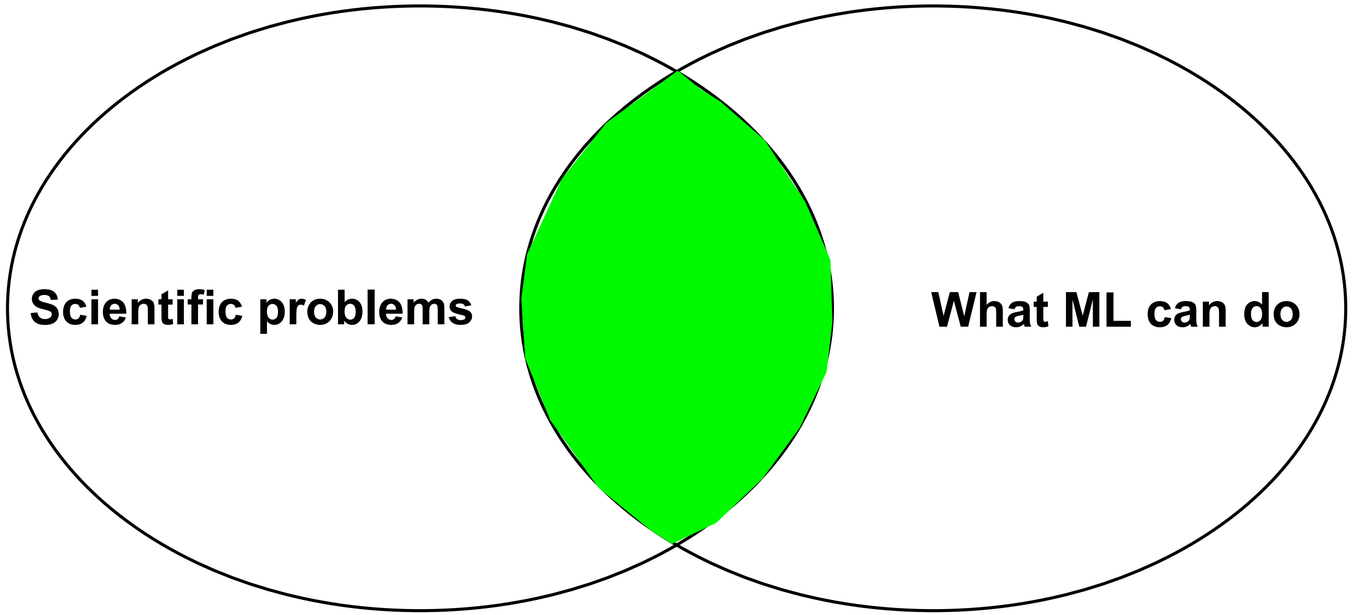
DeepRank for classification

DeepRank for ranking

PDaMED: simple & useful work model to speed up ML projects

- **P**roblem
- **Da**ta
- **M**odel
- **E**valuation
- **D**eployment

- **Problem**
  - Define your scientific problem and check if ML can solve it or not
    - Keep the problem simple; if not, decompose it
    - Choose proper tool, DL might not be a must
  - Transform the scientific problem to a ML problem
    - e.g. classification, regression…
  - Set a clear target



PDaMED *speeding up ML projects*

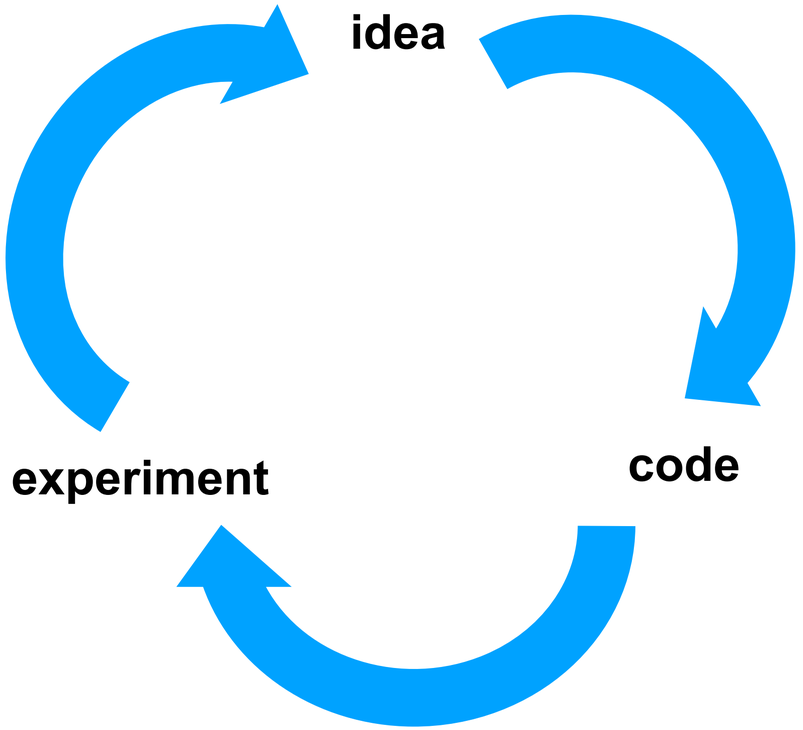
- **Problem**
- **Data**
  - Do you have enough data?
  - Can you collect or generate enough data with reasonable effort?
  - *“Enough” data:* volume, labels, variety, quality…
  - Data engineering
    - e.g. transformation, scaling, augmentation…

If DL is a rocket, then data is the fuel.

--- Andrew Ng

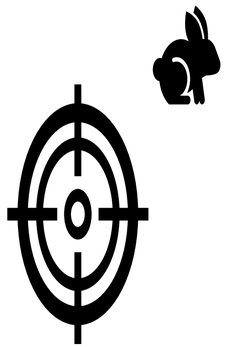
PDaMED *speeding up ML projects*

- **Problem**
- **Data**
- **Model**
  - Training a model is an iterative cycle
  - Versioning is important
    - e.g. data, code, architectures, hyperparameters…

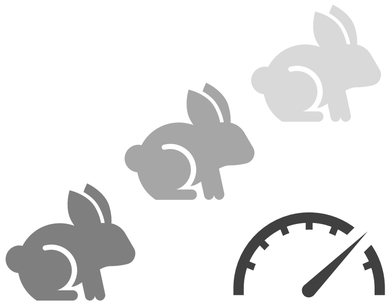


PDaMED *speeding up ML projects*

- **Problem**
- **Data**
- **Model**
- **Evaluation**
  - Metric selection depends on the target
  - Multiple targets: ONE optimizing metric and others using satisficing metrics
    - e.g. accuracy as high as possible, and speed lower than 1sec/case



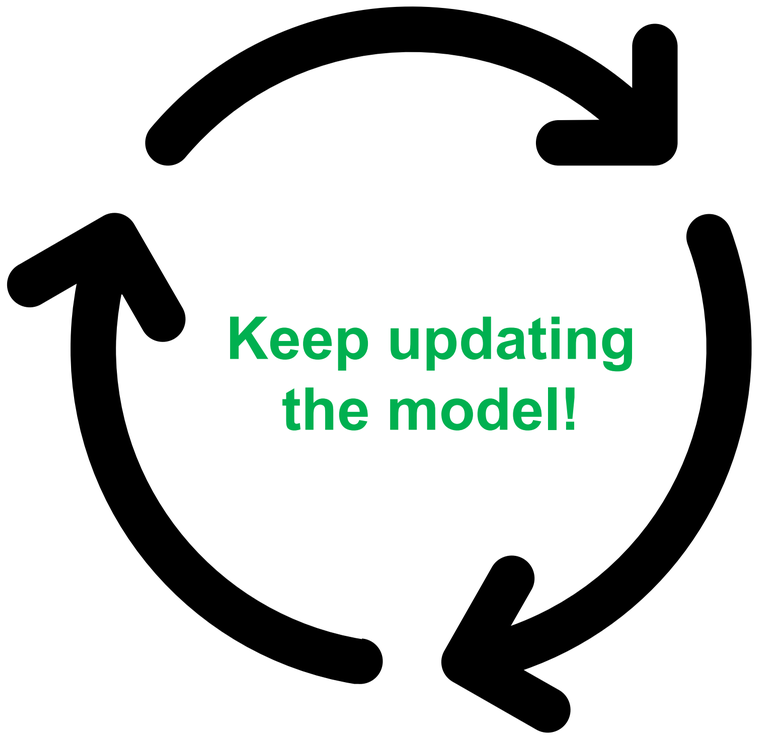
optimizing metric



satisficing metrics

PDaMED *speeding up ML projects*

- **Problem**
- **Data**
- **Model**
- **Evaluation**
- **Deployment**
  - software, webserver, docker, cloud…



WHAT

AI, Deep Learning & CNN

HOW

PDaMED work model for ML projects

EXAMPLES

DeepRank for classification

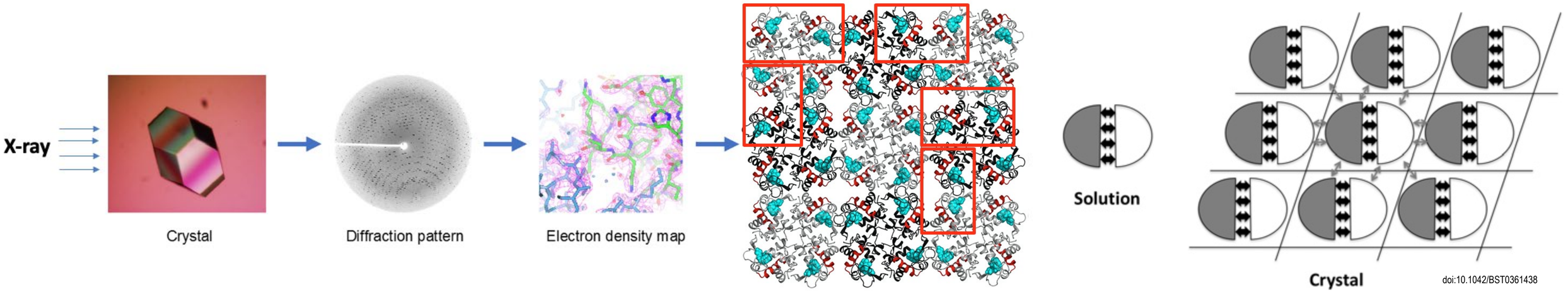
DeepRank for ranking

DeepRank for classifying protein-protein interactions

SURF Open Innovation Lab
ML4HPC project

PDaMED

- **Problem**



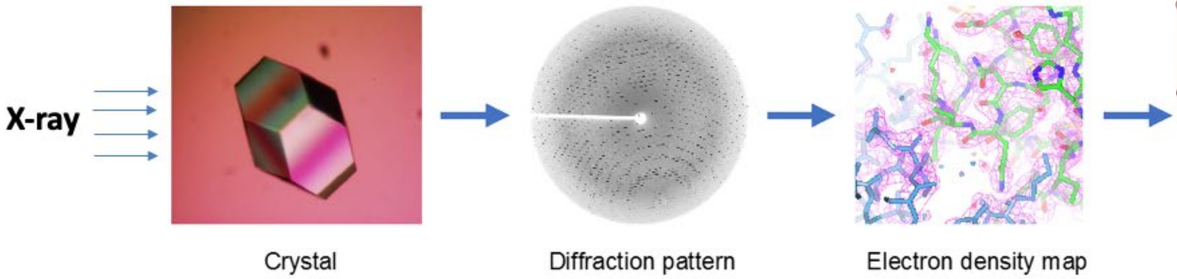
Scientific problem:

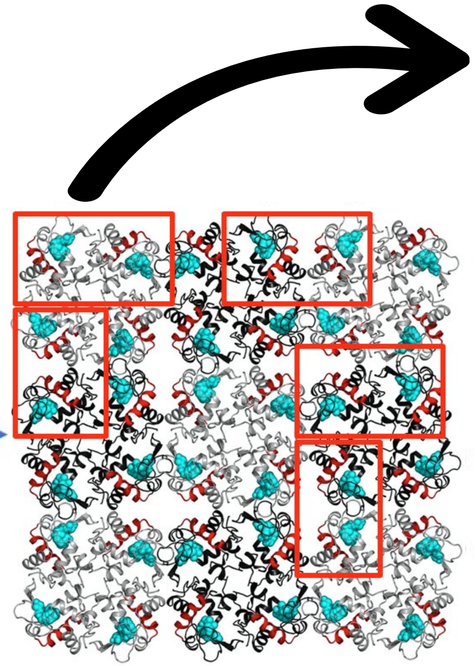
how to distinguish 3D biological interface from crystal contacts

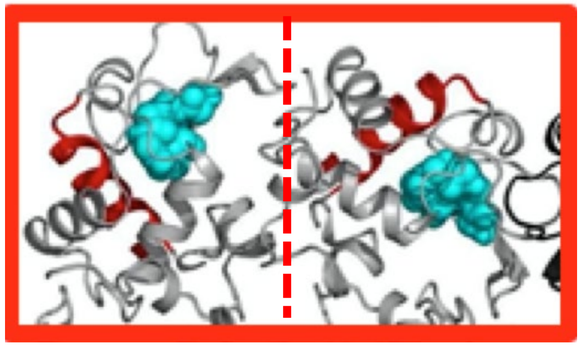
DeepRank for classifying protein-protein interactions

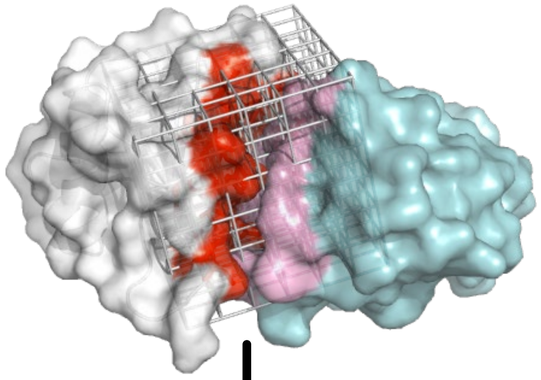
PDaMED

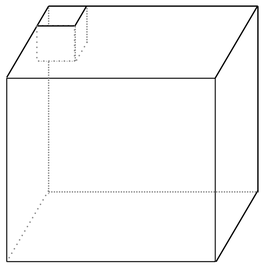
- **Problem**



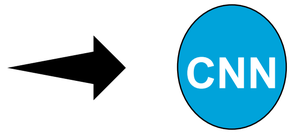








3D image



Scientific problem:

how to distinguish 3D biological interface from crystal contacts

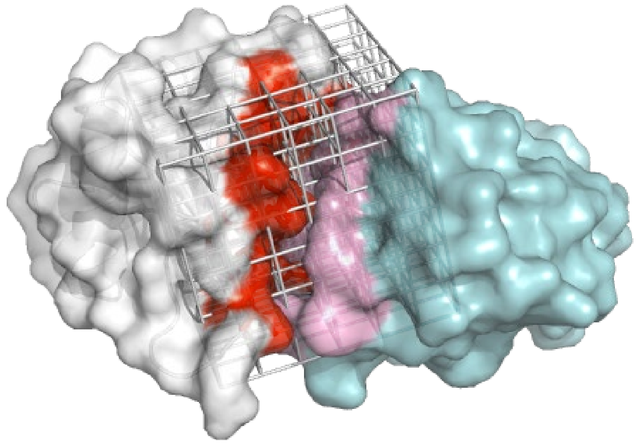
ML problem:

Given a 3D structure of protein-protein complex,
to classify its interface is biologically relevant or not

DeepRank for classifying protein-protein interactions

PDaMED

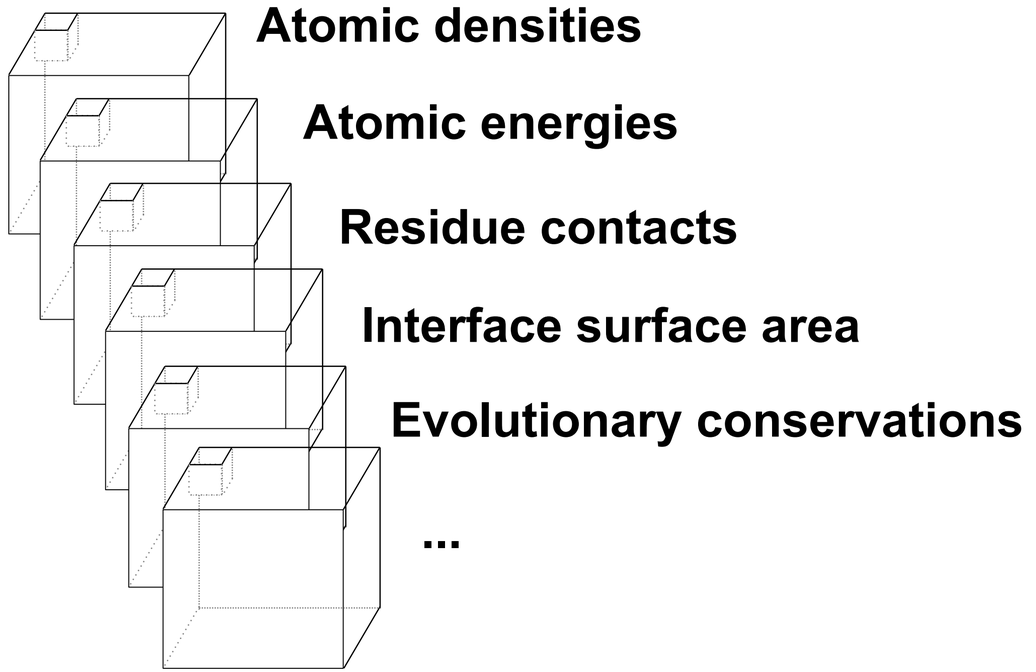
- Problem
- Data



Labeled 3D structures of protein-protein complexes

Training:
positive (biological) 2829
negative (crystal) 2911

Test:
pos 81, neg 81



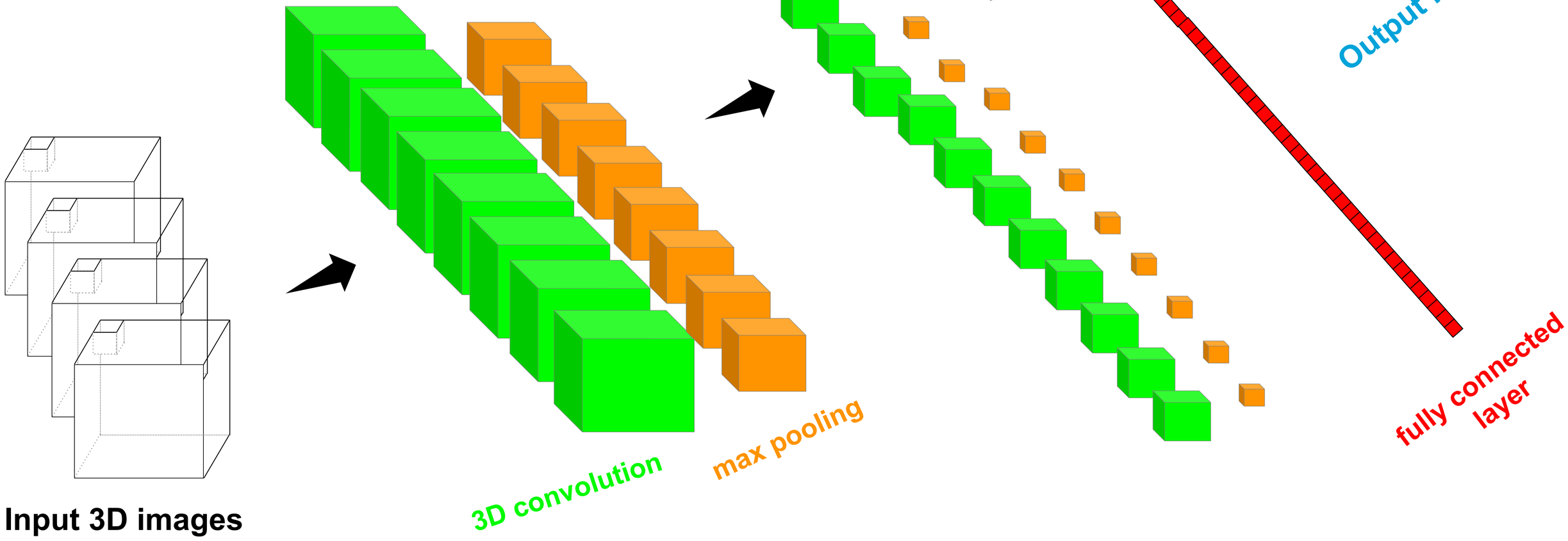
Input data:

Various structural properties,
each type of property is a 3D image

DeepRank for classifying protein-protein interactions

PDaMED

- **Problem**
- **Data**
- **Model**
  - 3D CNN architecture

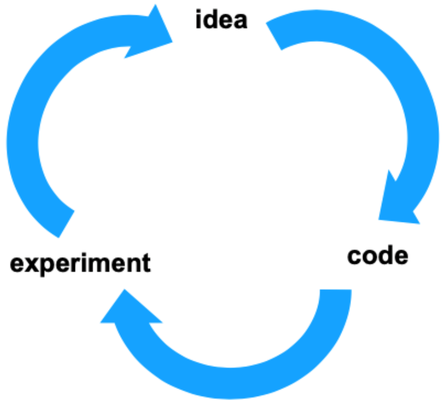


DeepRank for classifying protein-protein interactions

PDaMED

- Problem
- Data
- Model

Computational support from SURF Cartesius & Lisa



One example of our experiments

3D CNN architecture

layer	function	activation	#channel	channelSize	kernelSize	stride	padding
input	-	-	8	10x10x10	-	-	-
conv1	conv3d	relu	80	8x8x8	2	1	0
	maxpool3d	-	80	4x4x4	2	1	0
conv2	conv3d	relu	120	2x2x2	2	1	0
	maxpool3d	-	120	1	2	1	0
fc1	linear	relu	-	120	-	-	-
fc2	-	logSoftMax	-	2	-	-	-

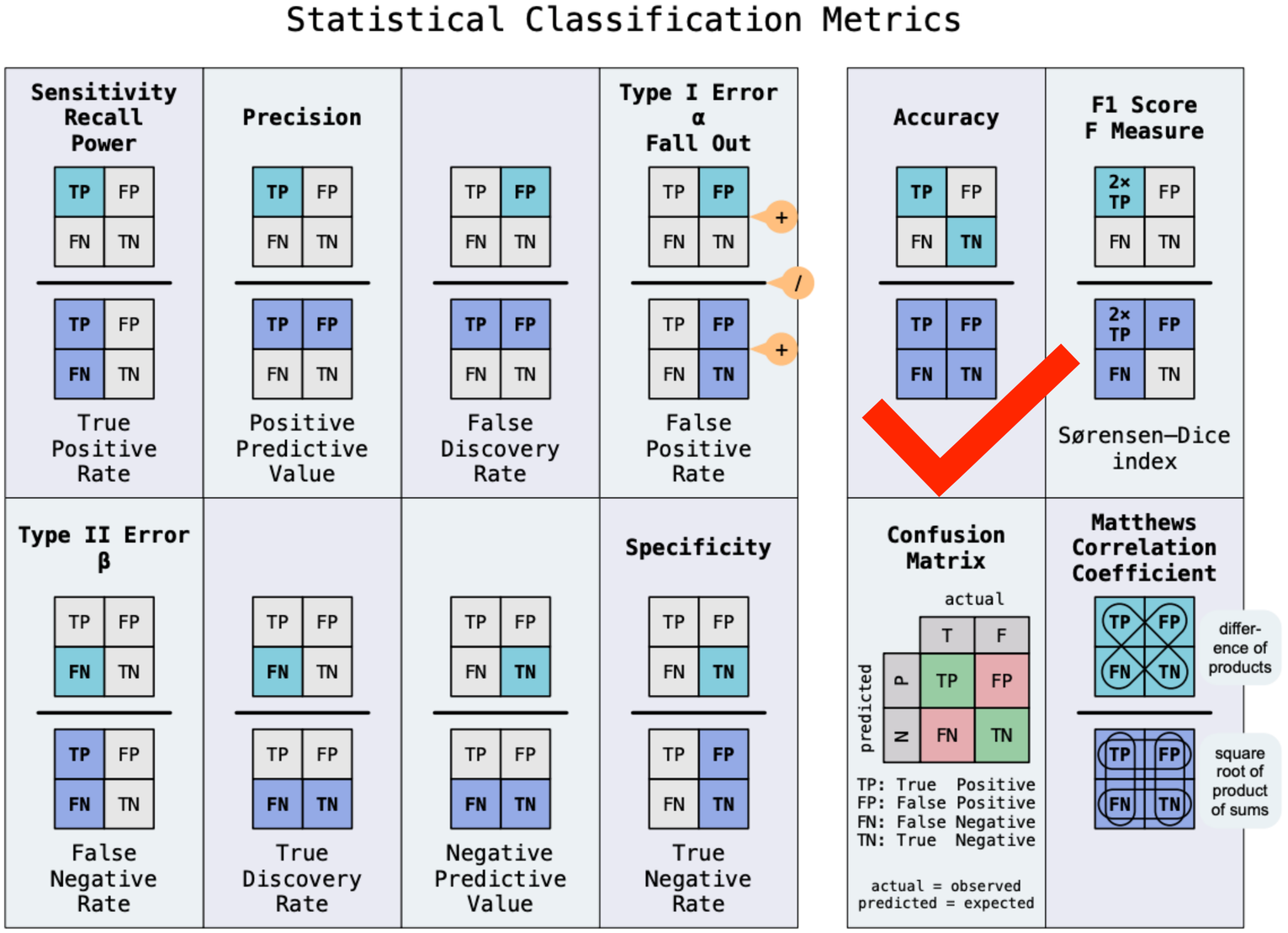
Training/validation/test data, optimizer, CNN hyper-parameters, GPU/CPU…

Training dataset	augmentation	datasetSplit	Test dataset	Feature	Architecture	Optimiser				epoch	batchSize	workers	Node			
						optimType	learningRate	momentum	weighDecay				GPUnode	#GPU	CPUnode	#CPU
MANY	0	8:2	DC	pssm+pssmic	arch001-02	SGD	e-4	0.9	e-4	30	2	16	1	2	0	

DeepRank for classifying protein-protein interactions

PDaMED

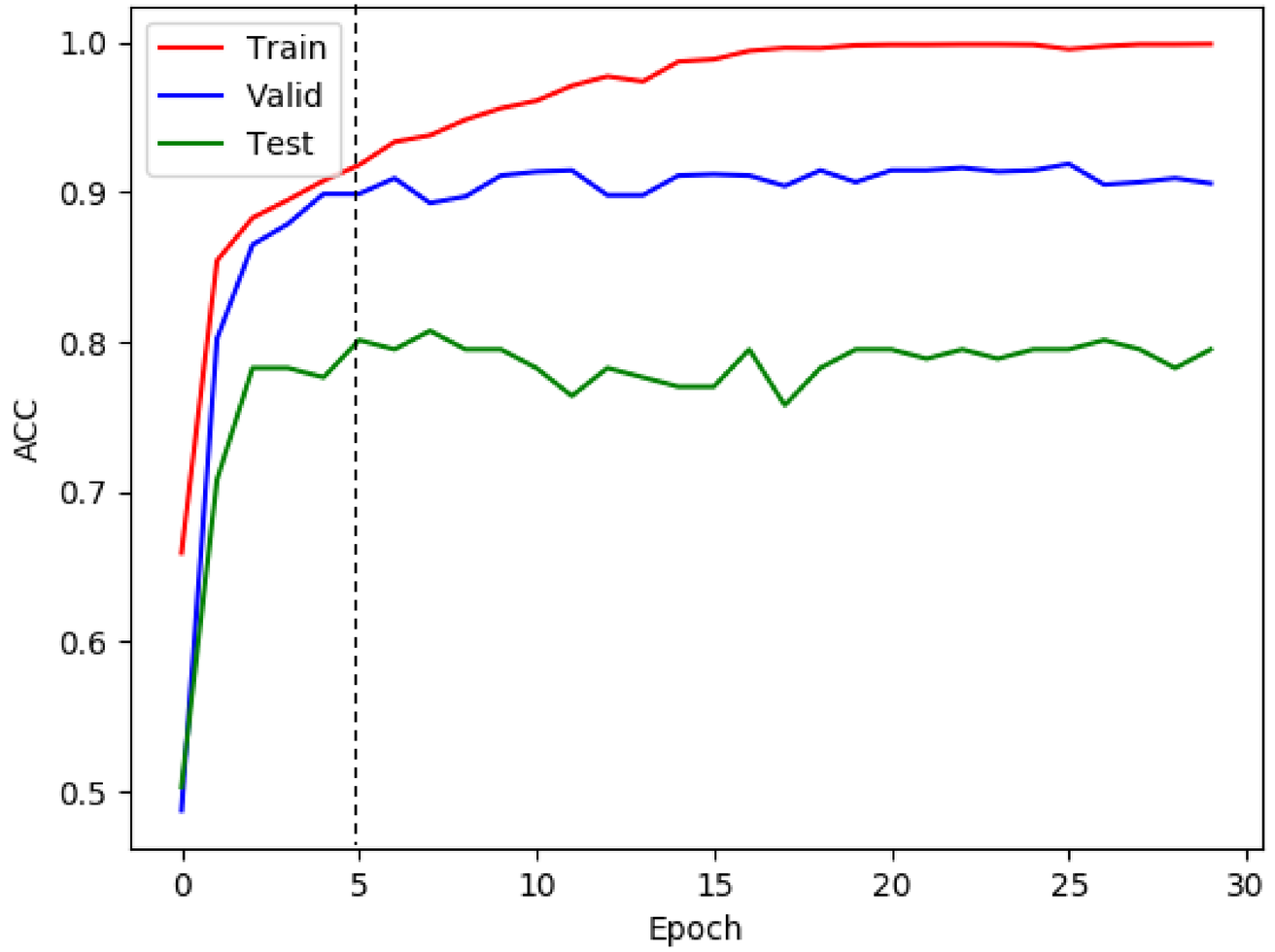
- **Problem**
- **Data**
- **Model**
- **Evaluation**



DeepRank for classifying protein-protein interactions

PDaMED

- **Problem**
- **Data**
- **Model**
- **Evaluation**



Test accuracy ≈ 0.8

Can correctly select 80% biologically relevant and non-relevant ones from all given complexes

DeepRank for classifying protein-protein interactions

PDaMED

- **P**roblem
- **D**ata
- **M**odel
- **E**valuation
- **D**eployment

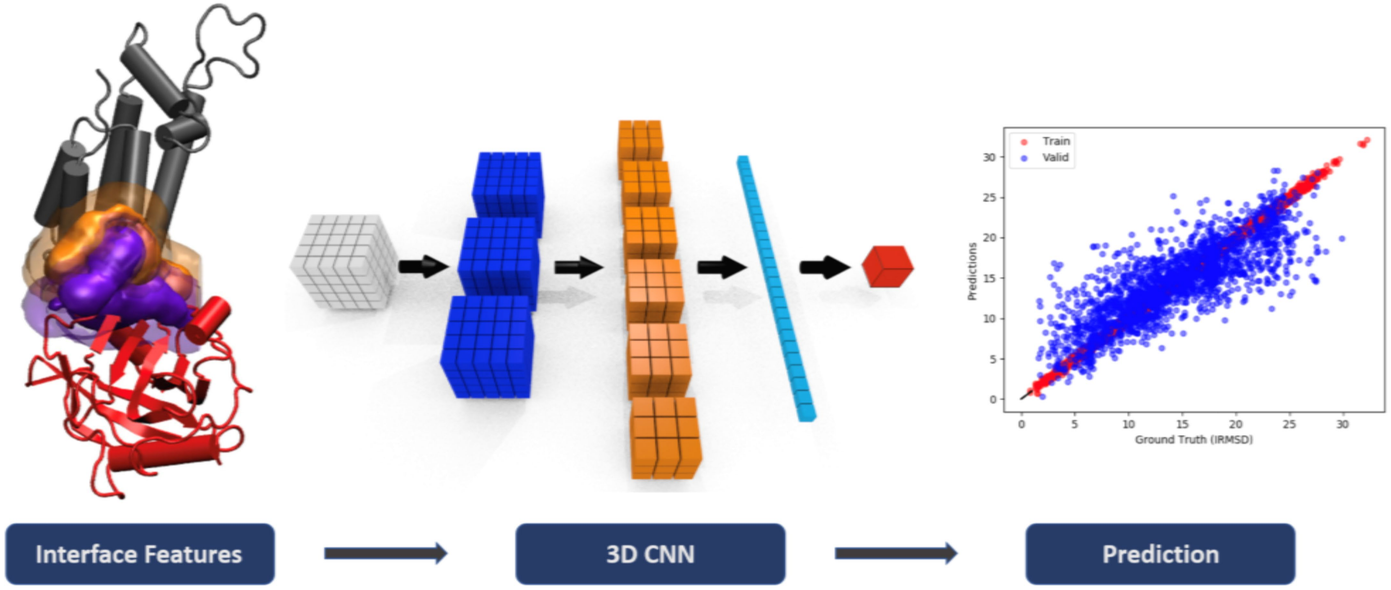
https://github.com/DeepRank/deeprank

DeepRank

Deep Learning for ranking protein-protein conformations

build failing | code quality B | docs passing | coverage 77%

The documentation of the module can be found on readthedocs : http://deeprank.readthedocs.io/en/latest/



PYTORCH

EXAMPLES

DeepRank for classification

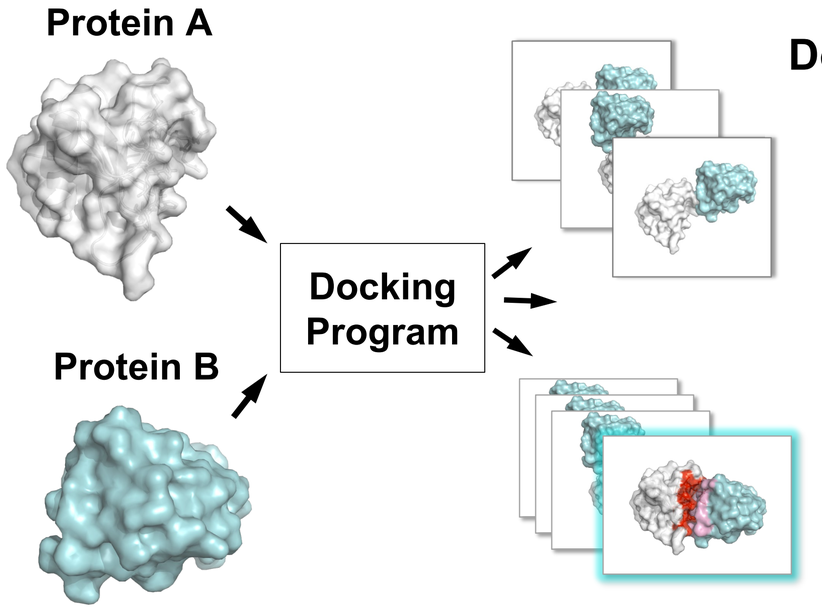
DeepRank for ranking

DeepRank for ranking protein-protein interactions

eScience Center
DeepRank project

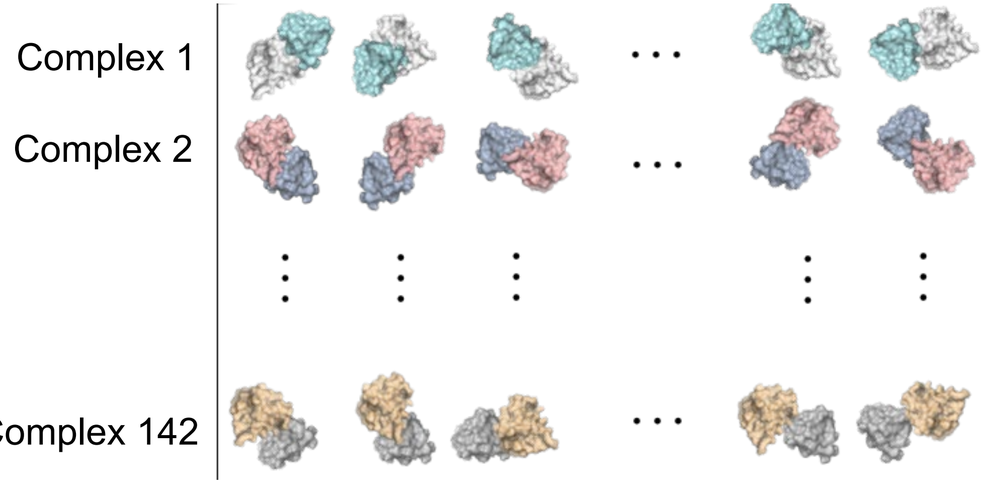
PDaMED

- **P**roblem
- **D**ata
- **M**odel
- **E**valuation
- **D**eployment



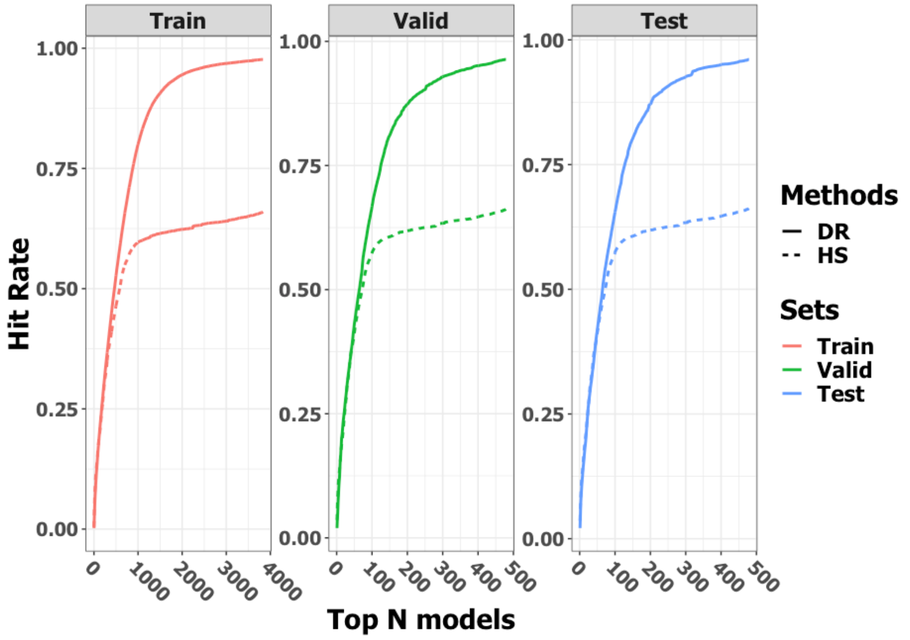
Docking generates *vast numbers* of structure models

The challenge:
How to *select* the most native-like structure models



DeepRank

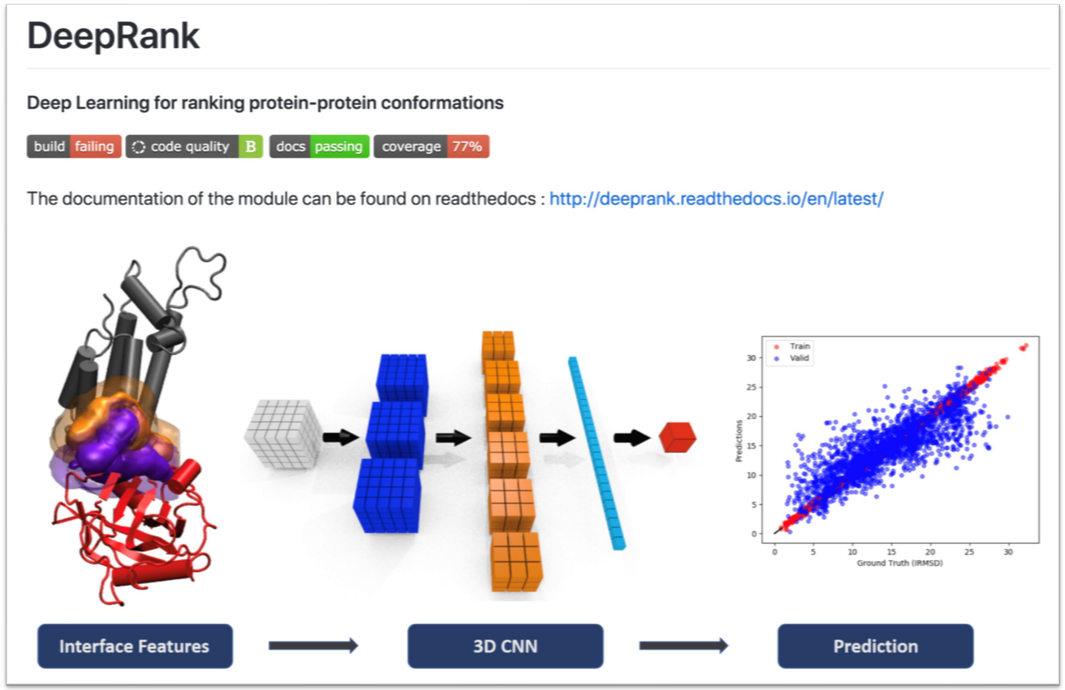
CNN



Take home message

- AI is a tool
- ML can empower scientific research, enabling a new paradigm
- PDaMED, a simple & useful work model to speed up ML projects
- DeepRank, a rich DL framework for studying biomolecular interactions

https://github.com/DeepRank/deeprank

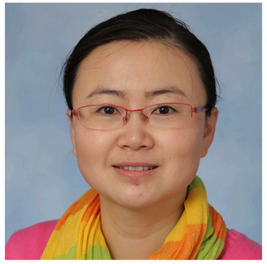


Contributors and support

Universiteit Utrecht

Alexandre Bonvin

Radboud Universiteit

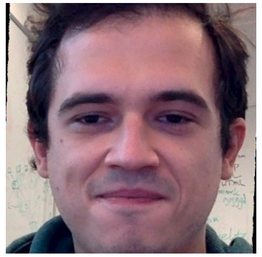


Li Xue

SURF SARA

Valeriu Codreanu

Caspar van Leeuwen



Damian Podareanu

SURF Open Innovation Lab
ML4HPC project

SURF Cartesius & Lisa

netherlands eScience center

Nicolas Renaud

Sonja Georgievska

Lars Ridder

Elena Ranguelova

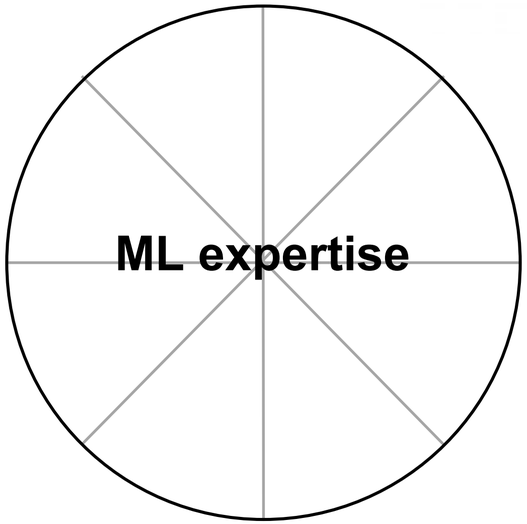
DeepRank project

Thank you for your attention!

Cunliang Geng

c.geng@esciencecenter.nl

netherlands eScience center



SURF SARA