Fondazione Ri.MED

SPAIS Scuola Permanente per l'Aggiornamento degli Insegnanti di Scienze Sperimentali

SPAIS- Scuola Permanente per l'Aggiornamento degli insegnanti di Scienze Sperimentali - 29 Luglio 2022 Cefalù

L'uso dell'intelligenza artificiale quale supporto nella ricerca di nuove terapie

Dr. Ugo Perricone
Group Leader in Molecular Informatics

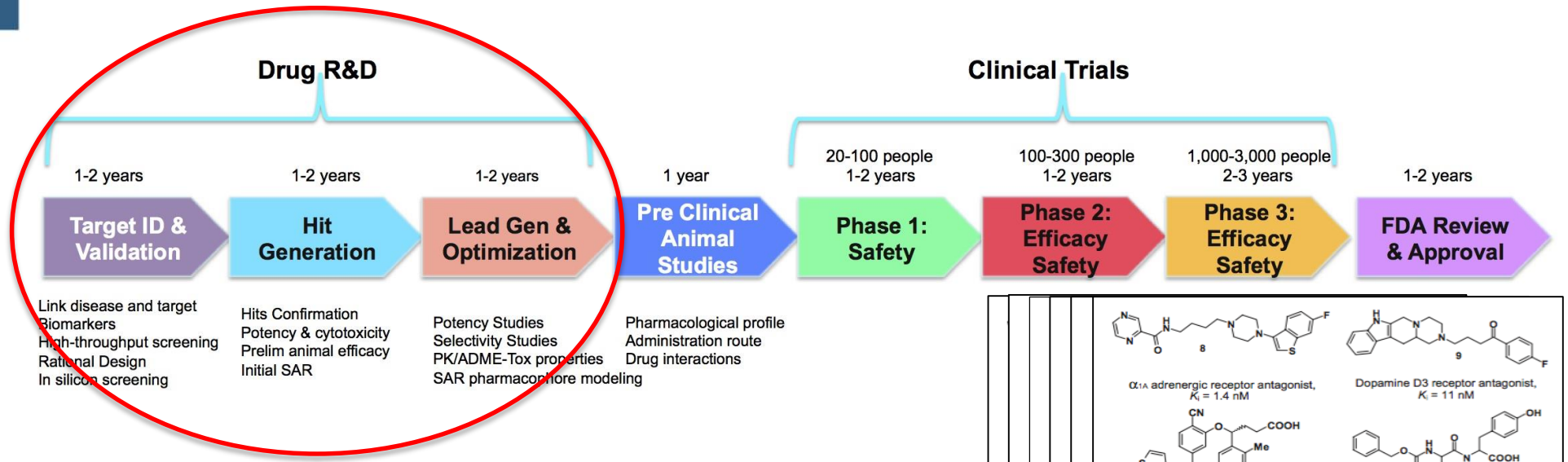
Chi Sono

- Laurea in Chimica e Tecnologia Farmaceutica (UNIPA)
- MBA (ISIDA-AFOR)-Specializzazione in Quality Management
- PhD in Scienze Molecolari e Biomolecolari (UNIPA-UNIVIE)

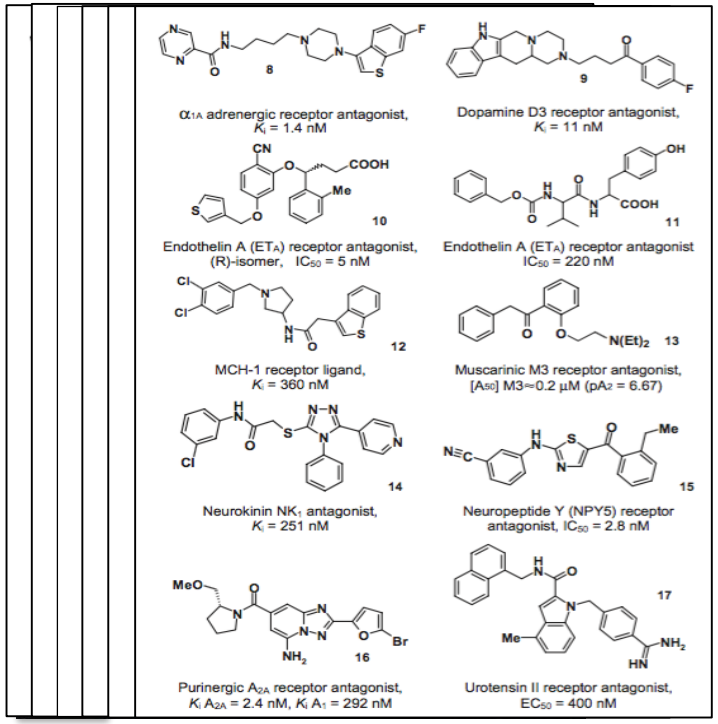
Agenda

- Metodologie Computazionali nella progettazione farmaceutica
- Introduzione a Big Data Artificial Intelligence (AI)
- Applicazioni di AI nella scoperta di Farmaci
- Applicazioni di AI in Diagnostica

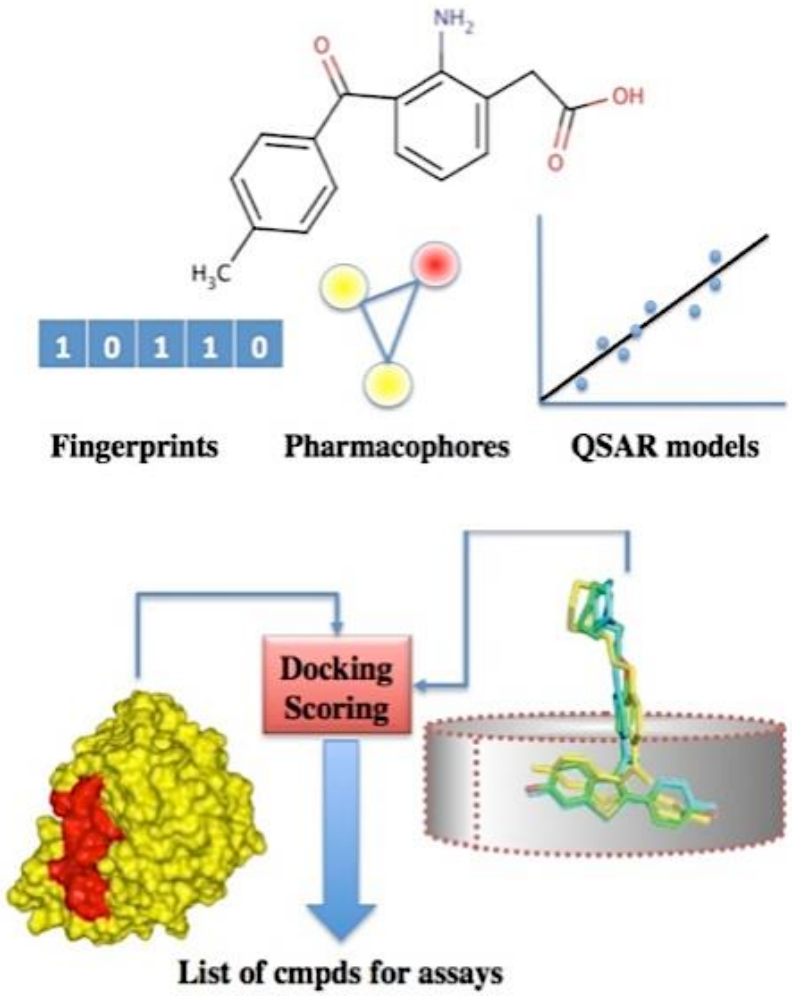
Drug Discovery Process



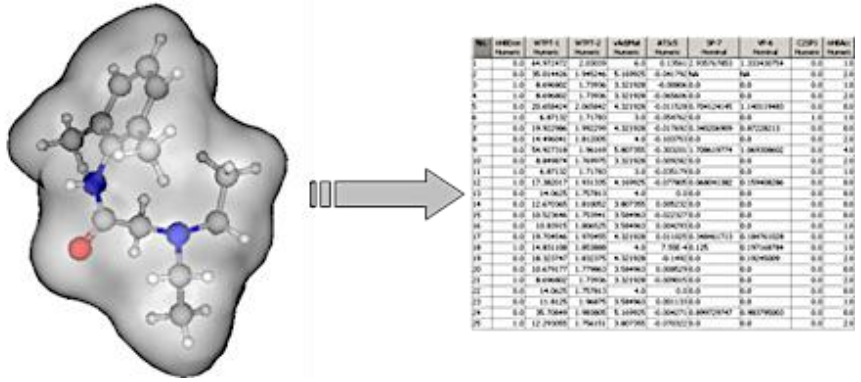
On October 5, 1981, *Fortune* magazine published a cover article entitled the **"Next Industrial Revolution: Designing Drugs by Computer at Merck".**



Approcci di Computer-Assisted Molecular Design



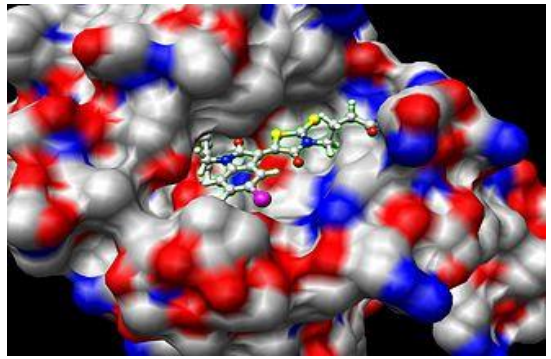
Ligand-based approach



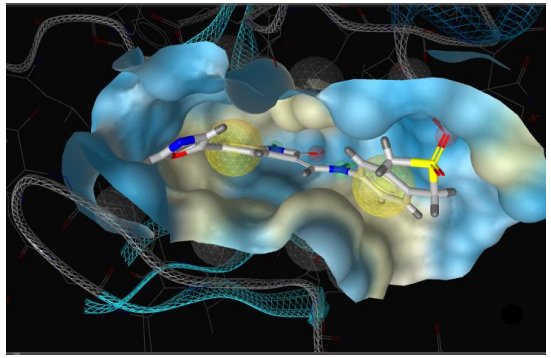
"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment." ***R. Todeschini***

- 0D Descriptors (i.e. constitutional, counting);
- 1D Descriptors(i.e. structural fragments, fingerprints);
- 2D Descriptors (i.e. molecular graphs and topological connectivity);
- 3D Descriptors (i.e. Molecular volume);
- 4D Descriptors (i.e. Interaction with a potential grid or a probe)

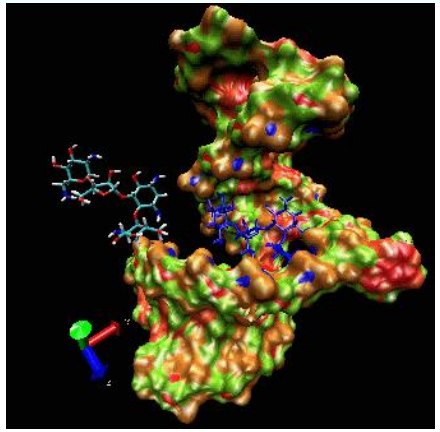
Structure-based Approach



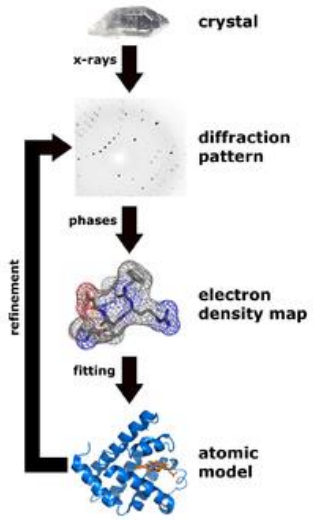
DOCKING



*PHARMACOPHORE



MOLECULAR DYNAMICS (MD)



- Missing loops ➔ lack of some receptor structural and geometrical information
- Mean atomic coordinates ➔ possible misinterpretation of ligand-receptor interaction pattern
- X-ray crystal resolution not always good (> 2 Å)

- Fill-in missing loops with homology modelling or threading
- MD-based structure refinement

INTEGRATING MOLECULAR DYNAMICS WITH VIRTUAL SCREENING TECHNIQUES ➔ DYNAMIC STUDY OF THE PROTEIN-LIGAND INTERACTION

***Pharmacophore** = Ensemble of electronic and steric features necessary for the supramolecular host-guest interaction and for the modulation of the biochemical response of a protein.

Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA (1998). "Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998)". Pure and Applied Chemistry. 70 (5): 1129–1143

Virtual screening workflow

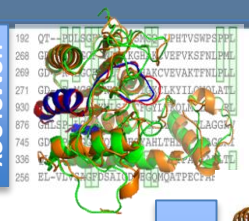
Fondazione Ri.MED

Scuola Permanente per l'Aggiornamento degli Insegnanti di Scienze Sperimentali

Società Chimica Italiana

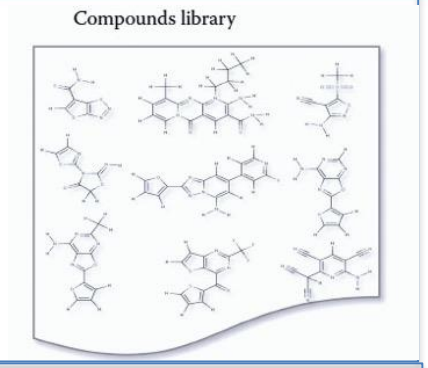
Input structure

HOMOLOGY MODELING



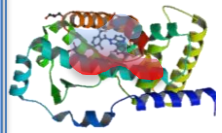
NMR

X-RAY

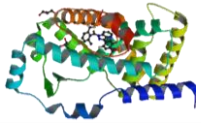


Preparation

- Hydrogen atoms assignement
- Histidine tautomers selection



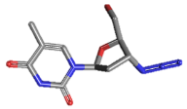
Binding site identification



Binding hypotesis generation

REOS
Rule of 5
PAINS

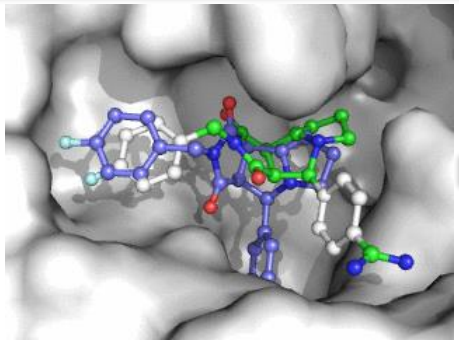
Filters application

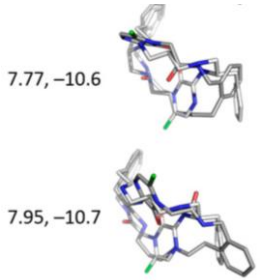


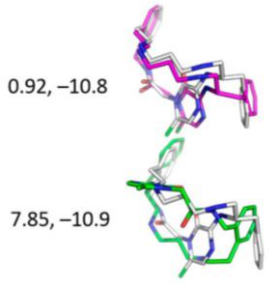
3D conformation

- Protonation states assignement
- Tautomers selection

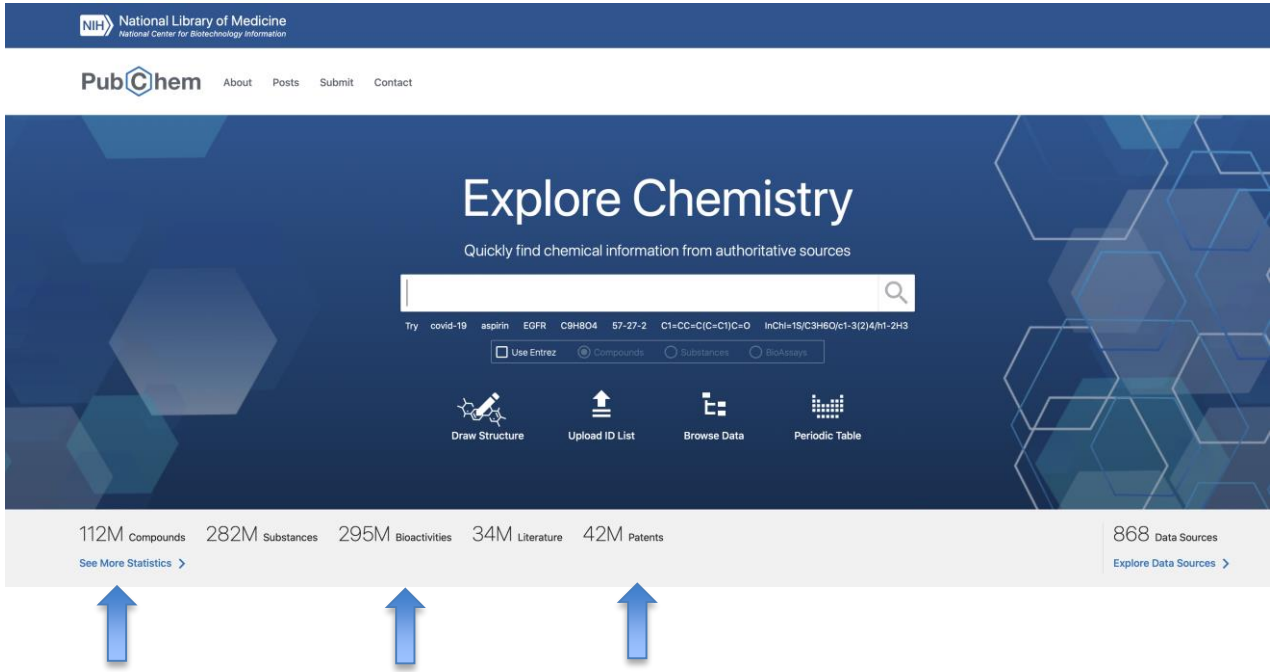
Docking & Scoring

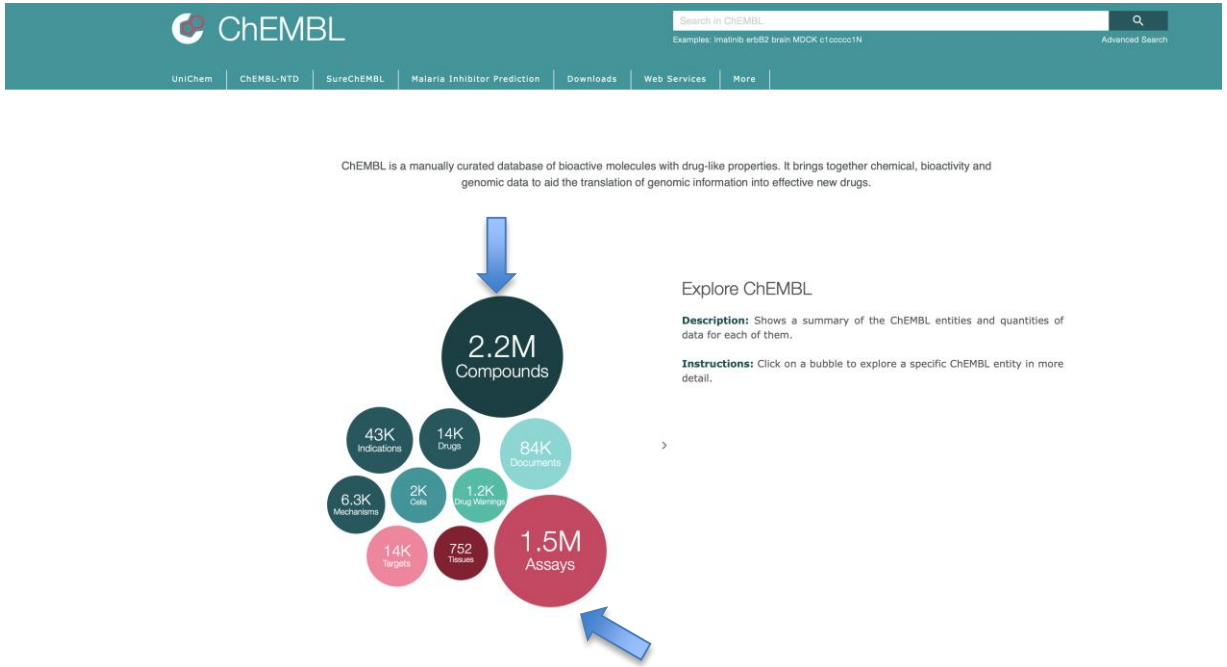




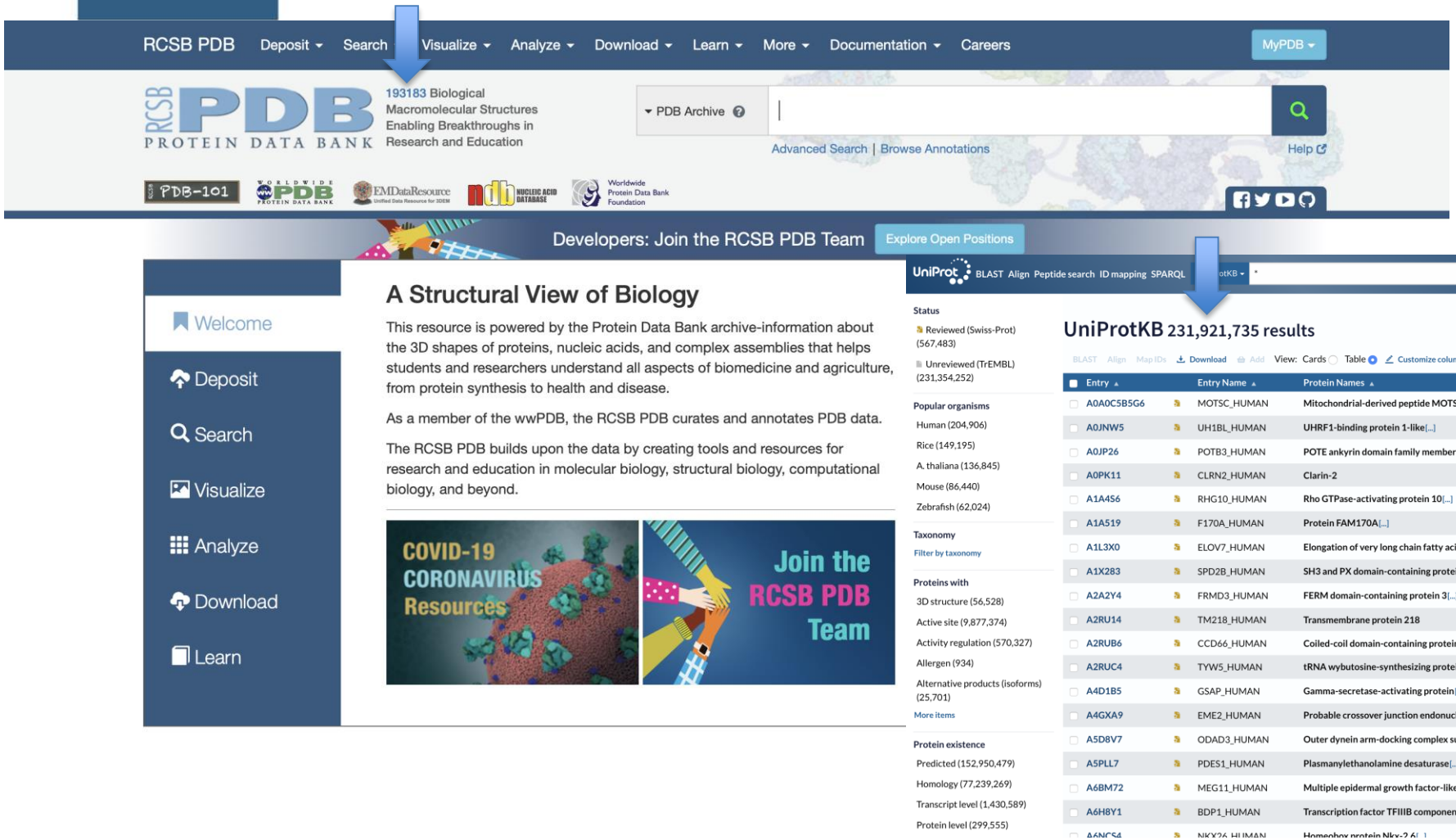


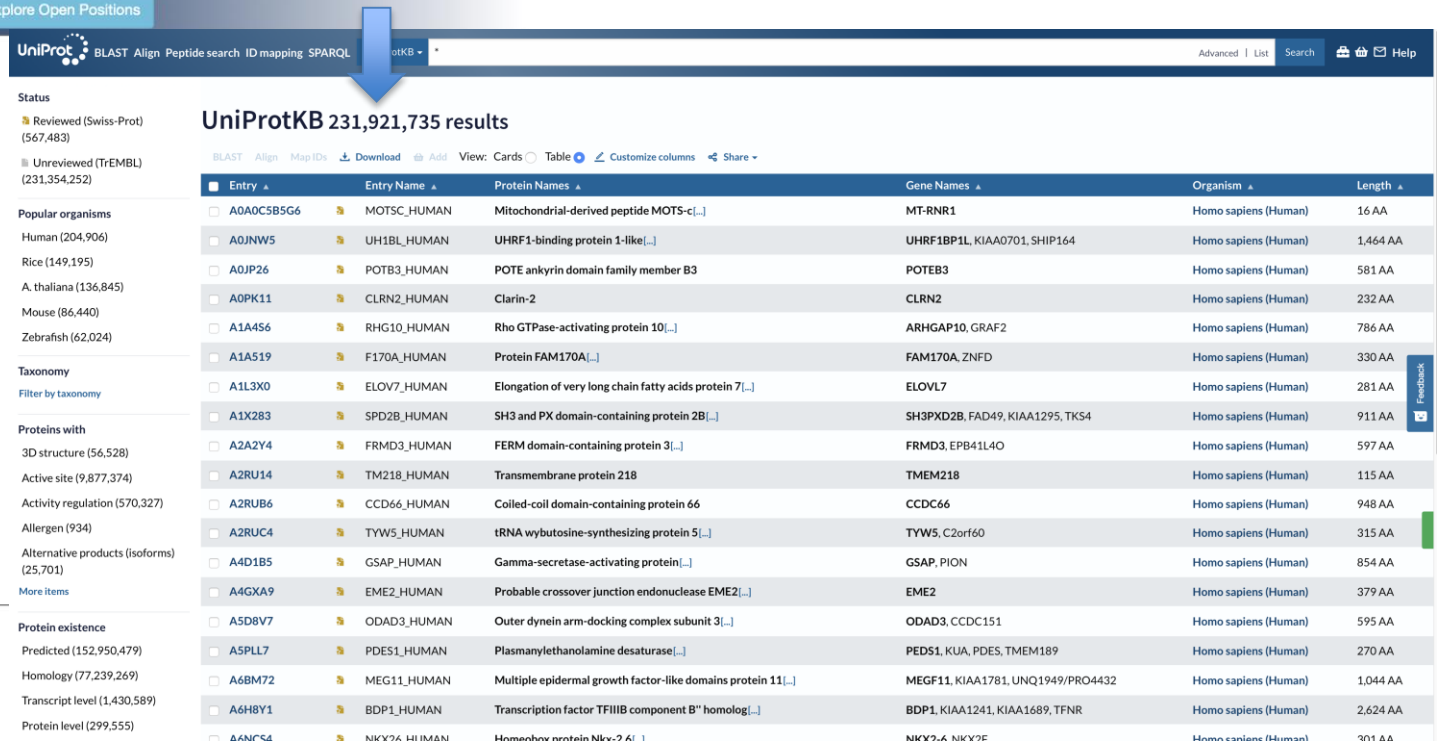
Database disponibili (molecole)





Database disponibili (proteine)





Big Data

Big Data: dati che contengono una grande varietà, che arrivano in volumi crescenti e con più velocità. Questo concetto è anche noto come le tre V.

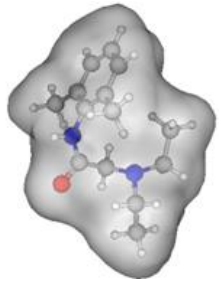
Volume	La quantità di dati è importante. Con i Big Data, dovrai elaborare volumi elevati di dati non strutturati a bassa densità. Può trattarsi di dati di valore sconosciuto, come feed di dati di Twitter, clickstream su una pagina Web o un'app mobile o apparecchiature abilitate per sensori. Per alcune organizzazioni, potrebbero essere decine di terabyte di dati. Per altre, potrebbero essere centinaia di petabyte.
Velocità	La velocità è la velocità con cui i dati vengono ricevuti e (forse) su cui si agisce. Normalmente, la velocità più elevata dei dati fluisce direttamente nella memoria invece di essere scritta sul disco. Alcuni prodotti intelligenti abilitati per Internet funzionano in tempo reale e richiedono valutazioni e azioni in tempo reale.
Varietà	La varietà si riferisce ai molti tipi di dati disponibili. I tipi di dati tradizionali erano strutturati e si adattavano perfettamente a un database relazionale. Con l'avvento dei Big Data, i dati arrivano come nuovi tipi di dati non strutturati. I tipi di dati non strutturati e semistrutturati, come testo, audio e video, richiedono un'ulteriore elaborazione preliminare per ricavare significato e supportare i metadati.

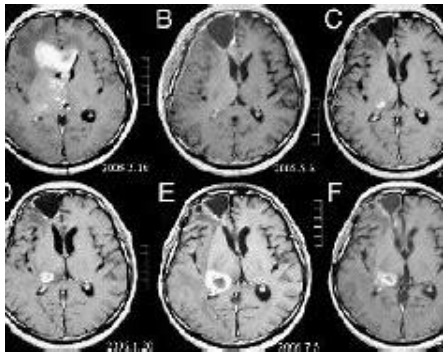
In parole povere, i big data sono set di dati più grandi e complessi, provenienti soprattutto da nuove origini dati.

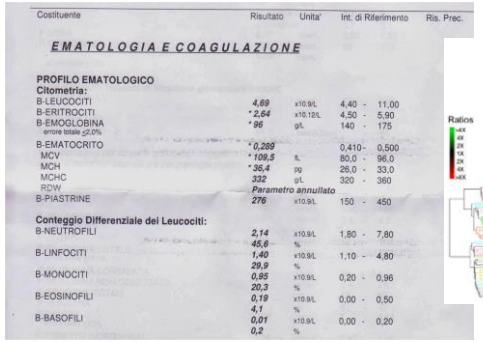
Questi set di dati sono così voluminosi che i sistemi di elaborazione dati tradizionale non sono in grado di gestirli.

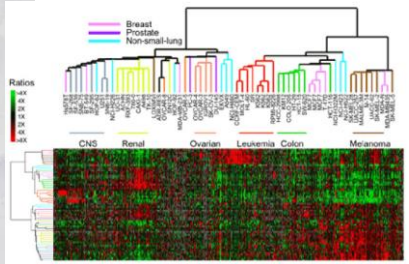
Big Data

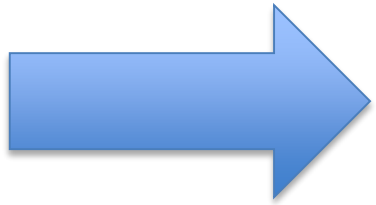
Negli ultimi anni sono emerse altre due V: **Valore** e **Veridicità**. I dati hanno un valore intrinseco. Ma sono inutili fino a quando non viene scoperto quel valore. Ugualmente importante: quanto sono veritieri i tuoi dati—e quanto puoi fare affidamento su di loro?

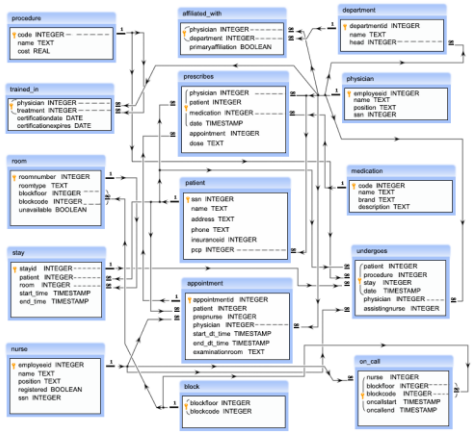


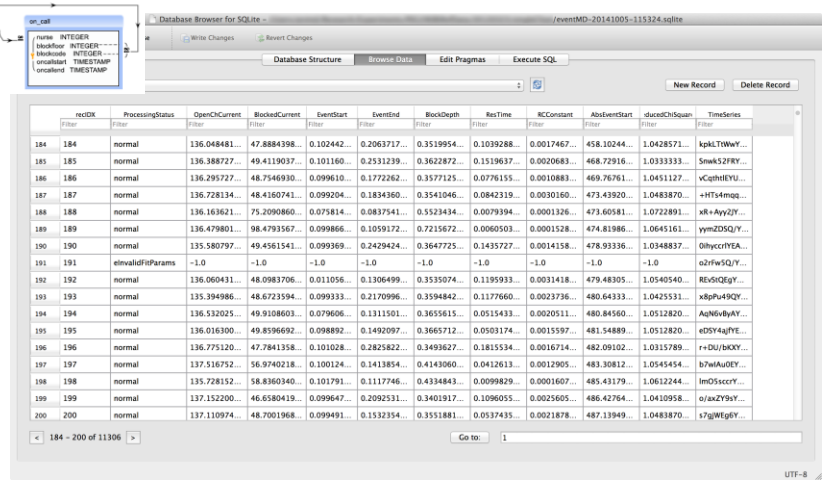












Big Data

Come funzionano i Big Data

I Big Data offrono nuovi insight che aprono nuove opportunità e modelli di lavoro. Lavorare con i big data prevede tre azioni chiave:

1. **Integrare**

dati provenienti da molte origini e applicazioni disparate. I tradizionali meccanismi di integrazione dei dati; necessarie nuove strategie e tecnologie per analizzare i set di Big Data su scala terabyte o addirittura petabyte. Durante l'integrazione e l'elaborazione dati formattati e disponibili in una forma con cui si possa lavorare in base all'applicazione prevista.

2. **Gestire**

Necessario un grande spazio di archiviazione. Molte persone scelgono la loro soluzione di archiviazione in base a dove risiedono attualmente i loro dati. Il cloud sta gradualmente guadagnando popolarità perché supporta i requisiti di calcolo correnti e consente di aumentare le risorse secondo necessità.

3. **Analizzare**

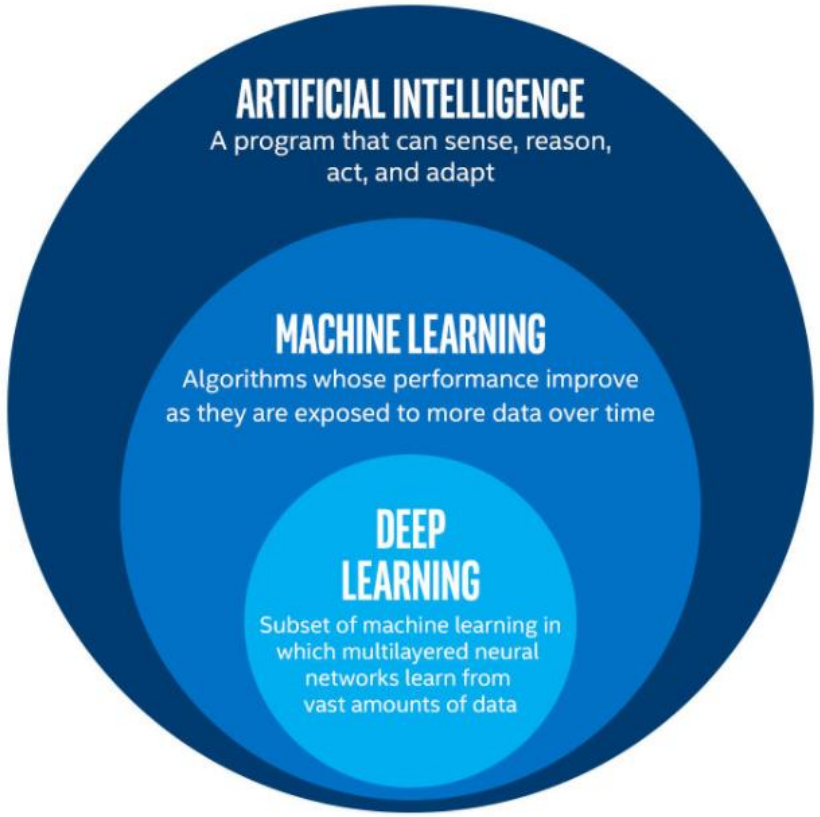
Creazione di modelli machine learning e intelligenza artificiale per sfruttare al massimo il potenziale dei dati archiviati e dare vita continua al dato.

A cosa serve l'artificial intelligence?

E' uno strumento a servizio dell'uomo. Una delle applicazioni più utili è la realizzazione di agenti decisionali autonomi, in grado di pianificare da sé le proprie scelte ma per perseguire degli obiettivi ben precisi prefissati dall'uomo. Le applicazioni sono comunque molteplici. Nel campo della ricerca un sistema AI può sperimentare in modo molto più veloce e razionale rispetto all'uomo, prendendo in considerazione un volume di dati enorme (big data) altrimenti impossibile da analizzare per l'uomo.

Gli agenti robotici, inoltre, possono svolgere operazioni ripetitive o pericolose, consentendo all'uomo di svolgere lavori di supervisione.

Artificial Intelligence (AI)



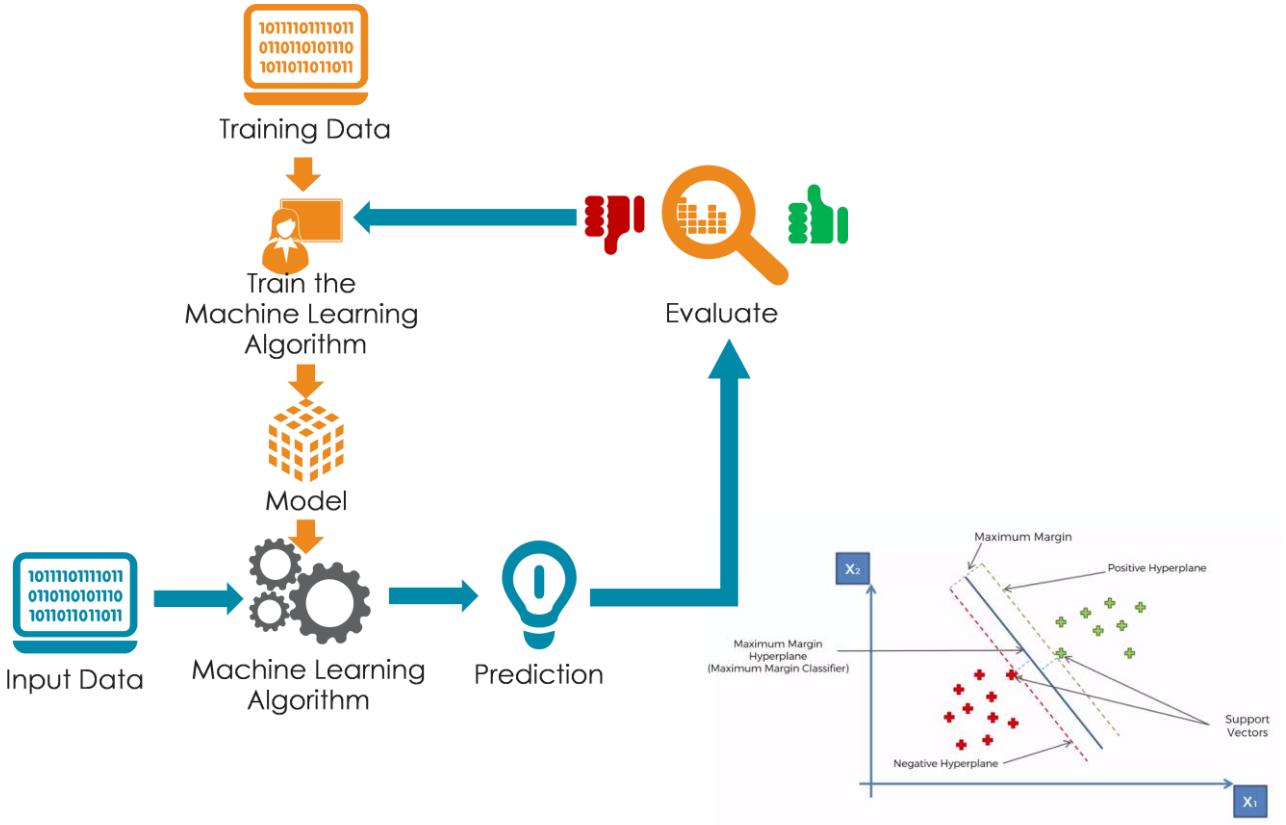
L'**intelligenza artificiale (**o **AI)** è una disciplina che studia se e in che modo si possano realizzare sistemi informatici intelligenti in grado di simulare la capacità e il comportamento del pensiero umano

Il **Machine Learning (ML)** o **apprendimento automatico** è una branca dell'intelligenza artificiale che raccoglie metodi sviluppati negli ultimi decenni e che utilizza metodi statistici per migliorare la performance di un algoritmo nell'identificare pattern nei dati. Nell'ambito dell'informatica, l'apprendimento automatico è una variante alla programmazione tradizionale nella quale in una macchina si predispone l'abilità di apprendere qualcosa dai dati in maniera autonoma, senza istruzioni esplicite[2,3]

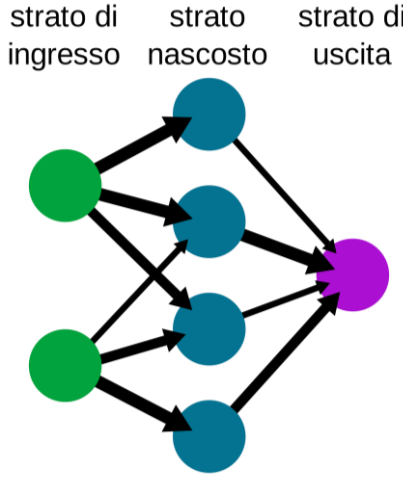
Il **Deep Learning**, la cui traduzione letterale significa apprendimento profondo, è una sottocategoria del Machine Learning (che letteralmente viene tradotto come apprendimento automatico) e indica quella branca dell'intelligenza artificiale che fa riferimento agli algoritmi ispirati alla struttura e alla funzione del cervello chiamate reti neurali artificiali.

Artificial Intelligence (AI)

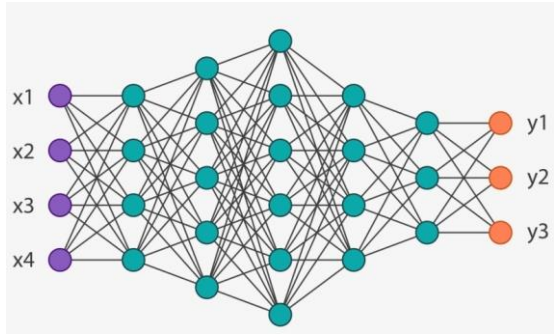
Gli algoritmi di apprendimento automatico sono utilizzati in un'ampia varietà di branche del sapere, come la medicina, il filtraggio delle e-mail, il riconoscimento vocale e la visione artificiale, dove è difficile o non fattibile sviluppare algoritmi convenzionali per eseguire i compiti richiesti



Rete semplice



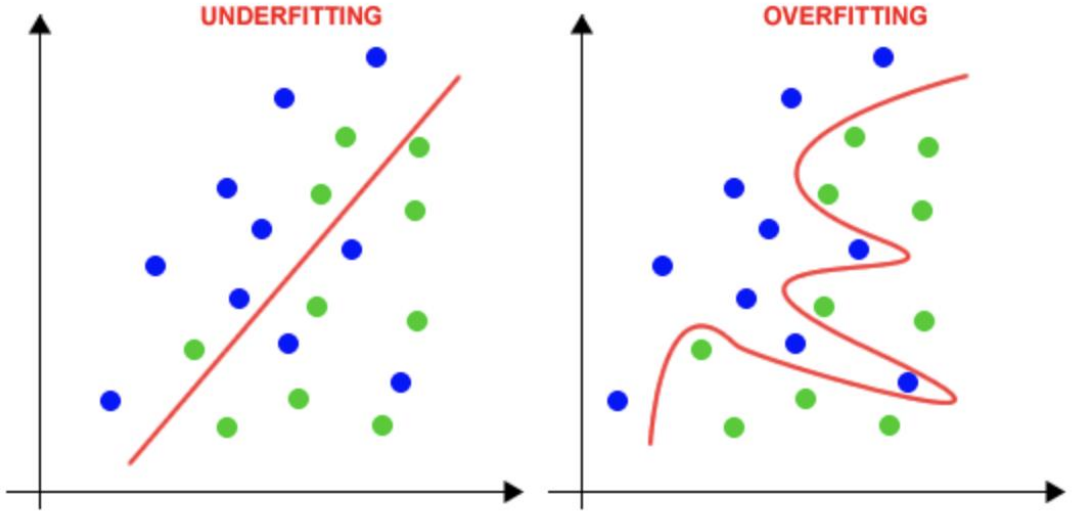
Rete complessa



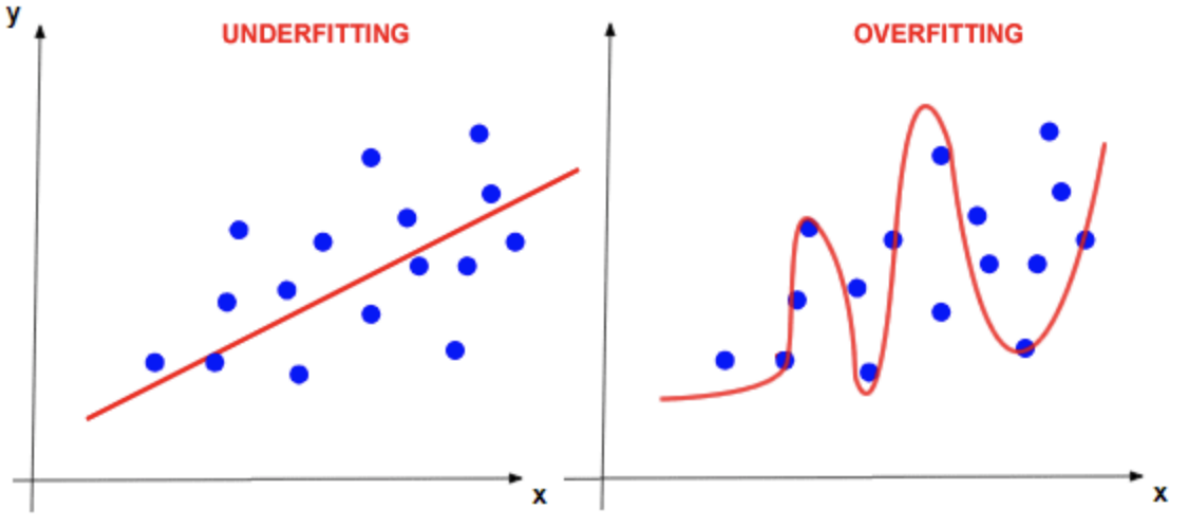
Artificial Intelligence (AI) –machine learning classico

I classificatori (classification) separano i dati in classi mentre i regressori (regression) interpolano i dati. Quindi, l'output di un modello classificatore è una classe mentre l'output di un modello regressore è un dato numerico.

La classificazione può essere lineare o non lineare, a due o più dimensioni, ma la logica è sempre la stessa.
.

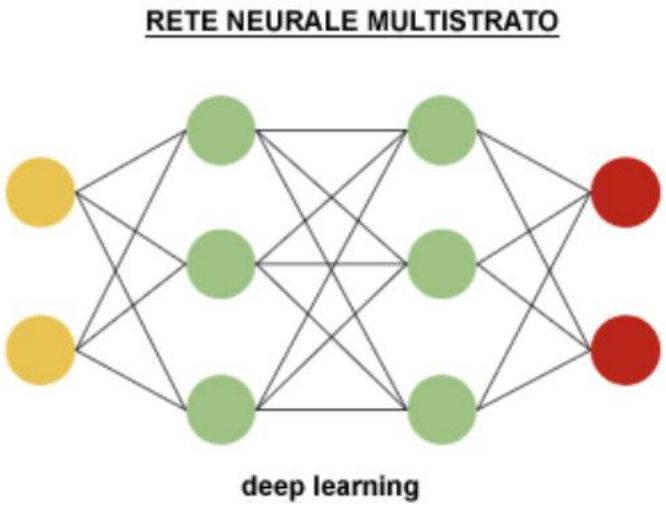


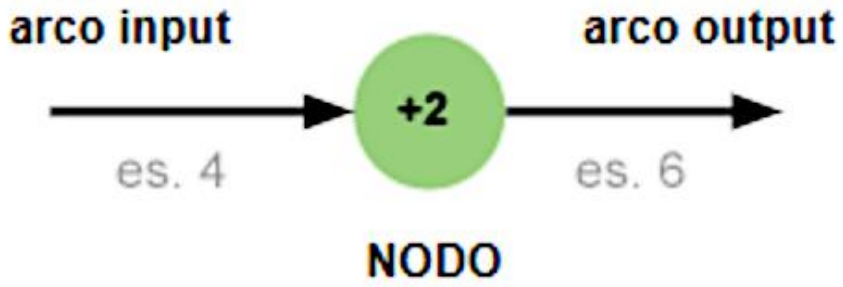
La regressione può essere lineare o non lineare, in uno spazio a due o più dimensioni, ma la logica è sempre la stessa.



Come funziona la rete neurale

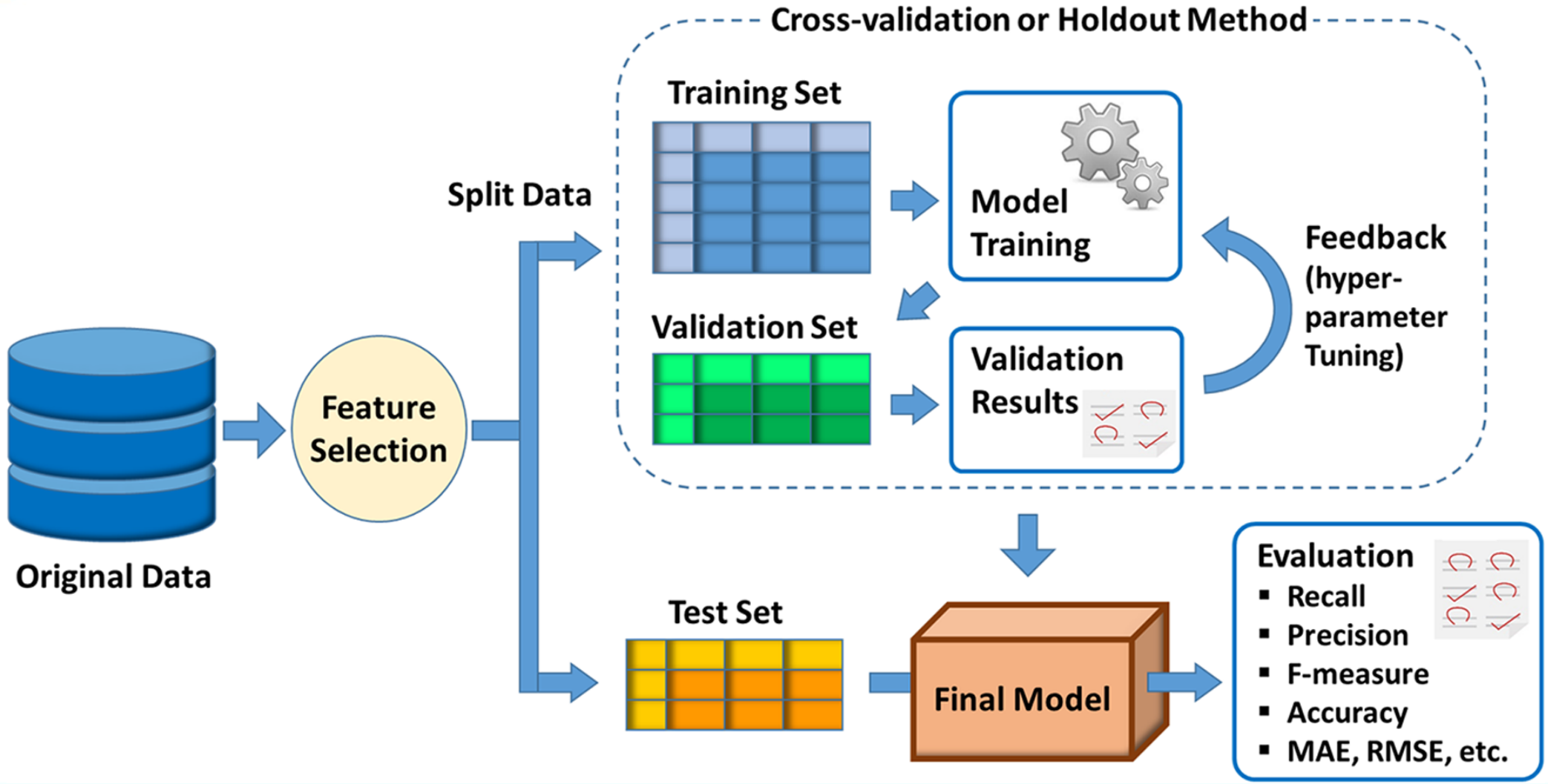
Gli archi trasmettono i dati dell'elaborazione nella neural network da un nodo a un altro.
L'informazione si propaga da sinistra verso destra (forward).



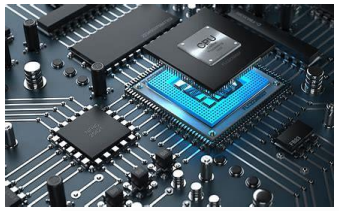


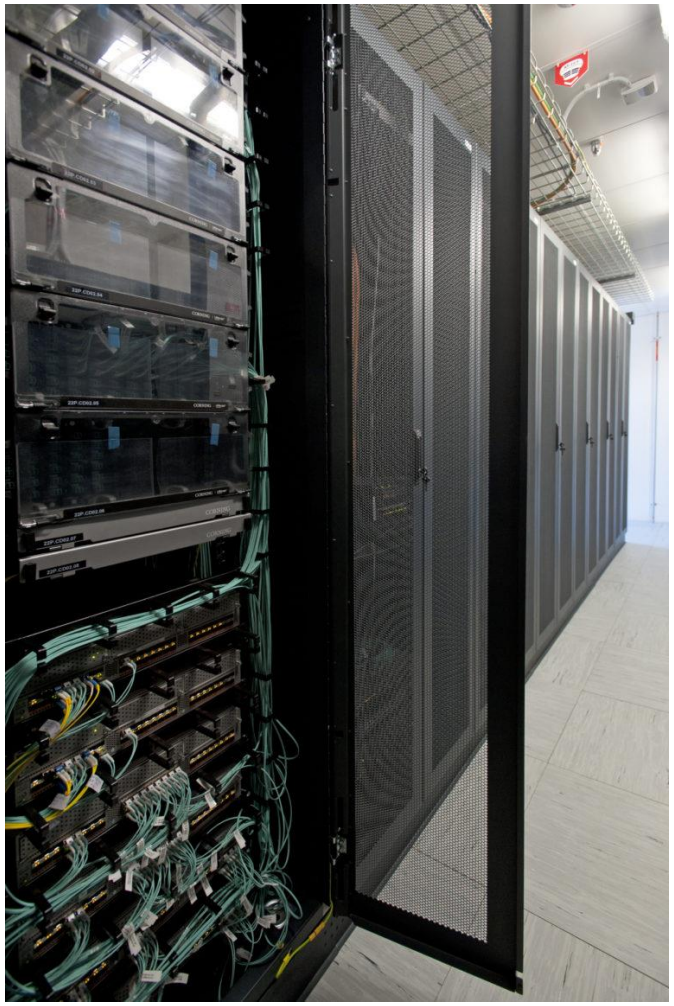
Gli archi in entrata sui nodi veicolano i valori di input (es. 4) mentre gli archi in uscita trasmettono il risultato dell'operazione in avanti (es. 4+2 = 6).
Ogni nodo può essere collegato tramite più archi sia in entrata che in uscita.

Artificial Intelligence (AI)



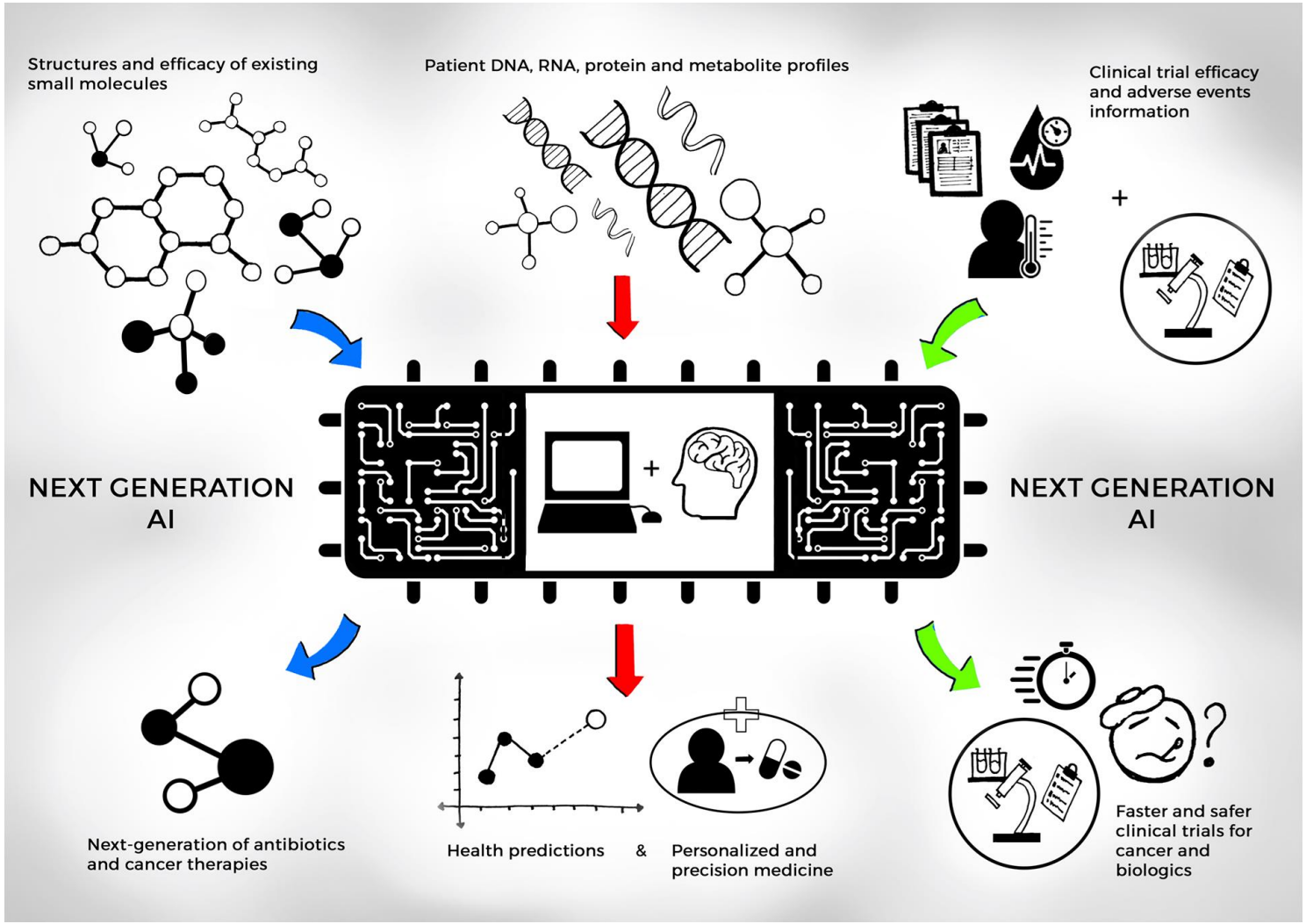
Strumenti per gestire Big Data e AI





Le GPU sono processori che supportano un tipo di calcolo parallelo altamente specializzato. **Ogni server accelerato da una GPU sostituisce decine di server CPU generici, fornendo un aumento sostanziale del throughput applicativo a costi più contenuti**.

AI in Biomedical Sciences



AI nella progettazione di nuovi antitumorali

Problema:

- Progettare molecole selettive su specifici target biologici (Kinasi)
- Metodologie classiche non sono in grado di guidare la progettazione selettiva → Il problema va risolto ad un livello superiore

Soluzione:

- ✓ Sfruttare dati strutturali disponibili
- ✓ Reti neurali in grado di cogliere caratteristiche fondamentali e discriminanti la selettività su un target o su un altro considerando contemporaneamente caratteristiche multiparametriche

BMC Bioinformatics

RESEARCH Open Access

Convolutional architectures for virtual screening

Isabella Mendolia, Salvatore Contino, Ugo Perricone, Edoardo Ardizzone and Roberto Pirrone

From Annual Meeting of the Bioinformatics Italian Society (BITS 2019)
Palermo, Italy. 26-28 June 2019

Abstract

Background: A Virtual Screening algorithm has to adapt to the different stages of this process. Early screening needs to ensure that all bioactive compounds are ranked in the first positions despite of the number of false positives, while a second screening round is aimed at increasing the prediction accuracy.

Results: A novel CNN architecture is presented to this aim, which predicts bioactivity of candidate compounds on CDK1 using a combination of molecular fingerprints as their vector representation, and has been trained suitably to achieve good results as regards both enrichment factor and accuracy in different screening modes (98.55% accuracy in active-only selection, and 98.88% in high precision discrimination).

Conclusion: The proposed architecture outperforms state-of-the-art ML approaches, and some interesting insights on molecular fingerprints are devised.

Keywords: Deep learning, Drug design, Molecular fingerprints, Bioactivity prediction, Virtual screening

Background

Virtual Screening (VS) is a routinely applied computational technique useful for drug design. However, some issues remain uncertain due to the complexity of the algorithms used behind the screening campaign, and this leads to generate models with different prediction reliability. Clinical candidate molecules selected by drug detection must have a profile responding to different criteria, that are based not only on the effect potency but also on the selectivity, safety as well as the so called *ADMET* properties (Absorption, Distribution, Metabolism, Excretion and Toxicity). Therefore, the design of the optimal compound is a multidimensional cha
Biology, which can be faced using Mac
One key aspect for ML approach
possibility to access and mining large
Until recent years, the best performi

BMC

International Journal of *Molecular Sciences*

MDPI

Article

EMBER—Embedding Multiple Molecular Fingerprints for Virtual Screening

Isabella Mendolia, Salvatore Contino, Giada De Simone, Ugo Perricone and Roberto Pirrone

Abstract: In recent years, the debate in the field of applications of Deep Learning to Virtual Screening has focused on the use of neural embeddings with respect to classical descriptors in order to encode both structural and physical properties of ligands and/or targets. The attention on embeddings with the increasing use of Graph Neural Networks aimed at overcoming molecular fingerprints that are short range embeddings for atomic neighborhoods. Here, we present EMBER, a novel molecular embedding made by seven molecular fingerprints arranged as different "spectra" to describe the same molecule, and we prove its effectiveness by using deep convolutional architecture that assesses ligands' bioactivity on a data set containing twenty protein kinases with similar binding sites to CDK1. The data set itself is presented, and the architecture is explained in detail along with its training procedure. We report experimental results and an explainability analysis to assess the contribution of each fingerprint to different targets.

Keywords: deep learning; drug design; virtual screening; embedding

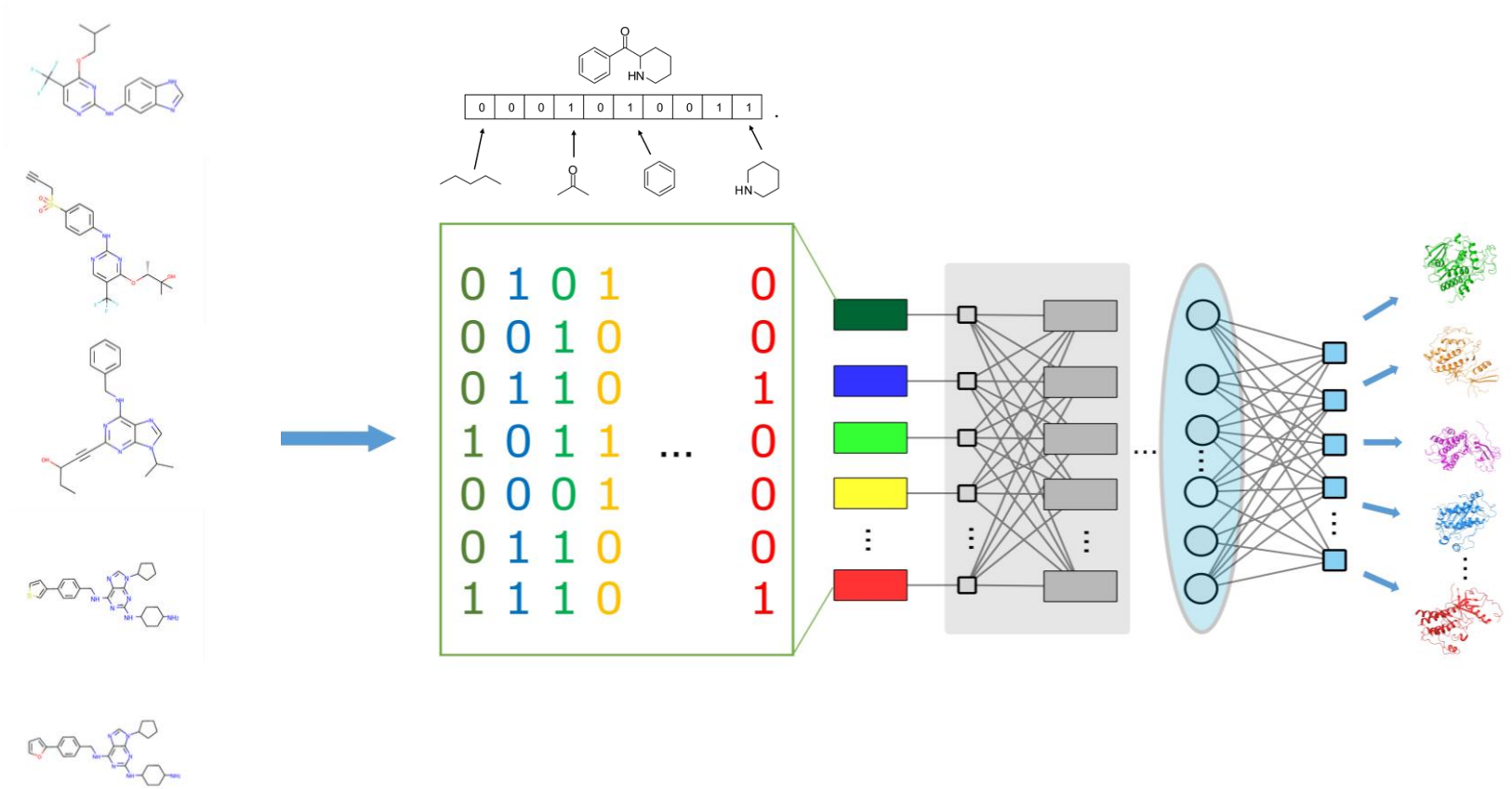
1. Introduction

Drug discovery is a very long and expensive process that includes many stages such as drug target identification, target validation, virtual screening (VS), hit-to-lead generation, lead optimization, and so on [1] . Moreover, developing a new drug has a mean pretax expenditure above 2 billion USD and takes about 10–15 years [2,3]. Despite the huge investment of time and money, the estimated clinical approval success rate of innovative small molecules during the drug discovery process is about 13%; thus, the overall risk of failure is very high. Drug design is supported by computational methods in almost every stage. Yu and MacKerell [4] report a review that describes the drug discovery process and the corresponding computer-aided drug design methods. Computational methods do not guarantee a systematical assessment of molecular characteristics (e.g., bioactivity, *ADMET* properties, selectivity, and physicochemical properties) but generate lead molecules with favourable properties in silico.

In particular, Virtual Screening (VS) is an often discussed topic in Chemoinformatics and Medicinal Chemistry and is widely applied in pharmaceutical research. VS consists of screening large small-molecule databases searching for bioactive molecules with respect to the target under investigation. This enables the researcher to cut the cost of experimentally testing thousands of compounds through a severe reduction in the number of candidate molecules. Research in the field of VS gained increasing importance in the last decade when Deep Learning (DL) became a mature discipline [5] . In this field, the scientific debate is very rich with respect to the proper method for representing molecular structures that are learned by the network. The very first architectures used classical representations such as molecular fingerprints [6] and SMILES notation [7] . Recently, molecular graphs have been investigated along with neural embeddings, essentially a learned low-dimension vector

BMC Bioinformatics. 2020 Sep 16;21(Suppl 8):310.doi: 10.1186/s12859-020-03645-9 - Int. J. Mol. Sci. 2022, 23(4), 2156; https://doi.org/10.3390/ijms23042156

AI nella progettazione di nuovi antitumorali




BMC Bioinformatics

RESEARCH Open Access

Convolutional architectures for virtual screening

International Journal of Molecular Sciences

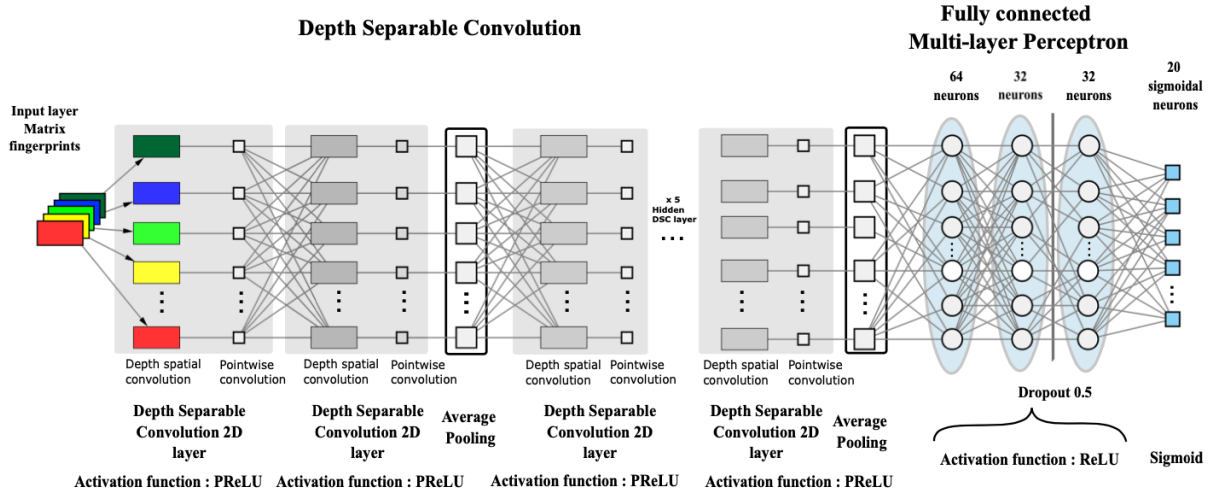
EMBER—Embedding Multiple Molecular Fingerprints for Virtual Screening

BMC Bioinformatics. 2020 Sep 16;21(Suppl 8):310.doi: 10.1186/s12859-020-03645-9 - Int. J. Mol. Sci. 2022, 23(4), 2156; https://doi.org/10.3390/ijms23042156

AI nella progettazione di nuovi antitumorali

Target	PDB ID	Ligand Code *	Actives	Inactives
ACK	5ZXB	9KO	746	159,775
ALK	6E0R	HKJ	1665	227,247
CDK1	6GU2	F9Z	1241	124,473
CDK2	6INL	AJR	1924	225,087
CDK6	5L2S	6ZV	646	256,561
INSR	5E1S	5JA	1423	195,990
ITK	4RFM	3P6	1001	135,007
JAK2	6M9H	J9D	5526	577,409
JNK3	2B1P	AIZ	658	95,252
MELK	6GVX	TAK	1215	246,662
CHK1	6FC8	D4Q	2175	21,763
CK2a1	6JWA	5ID	1053	10,534
CLK2	6FYL	3NG	671	6800
DYRK1A	4YLK	4E2	1126	11,274
EGFR	5GNK	80U	4757	47,541
ERK2	6OPH	6QB	3525	35,237
GSK3B	5F94	3UO		
IRAK4	6EG9	OLI		
MAPK2K1	4AN9	ACP; 2P7		
PDK1	3NAX	MP7		

Target	Acc.	Loss	Sensitivity	MCC	AUC	F1-Score
ACK	0.9957	0.0226	0.5000	0.6742	0.9834	0.6463
ALK	0.9930	0.0402	0.6575	0.7913	0.9904	0.7804
CDK1	0.9910	0.0314	0.4537	0.6397	0.9850	0.6059
CDK2	0.9859	0.0431	0.5281	0.6338	0.9845	0.6287
CDK6	0.9966	0.0210	0.5865	0.7523	0.9895	0.7305
INSR	0.9893	0.0329	0.3779	0.5830	0.9858	0.5342
ITK	0.9945	0.0232	0.5886	0.7302	0.9905	0.7154
JAK2	0.9898	0.0472	**0.8474**	**0.9090**	**0.9950**	**0.9114**
JNK3	**0.9967**	**0.0154**	0.5905	0.7610	0.9901	0.7381
MELK	0.9957	0.0229	0.7081	0.8270	0.9897	0.8188
CHK1	0.9895	0.0512	0.6385	0.7650	0.9846	0.7565
CK2A1	0.9942	0.0253	0.5166	0.6944	0.9857	0.6667
CLK2	0.9936	0.0259	*0.2255*	*0.4137*	0.9771	*0.3485*
DYRK1A	0.9916	0.0321	0.4080	0.5987	0.9776	0.5591
EGFR	0.9845	*0.0604*	0.7536	0.8331	0.9874	0.8357
ERK2	0.9881	0.0563	0.7295	0.8292	0.9886	0.8272
GSK3	*0.9843*	0.0554	0.5827	0.6892	*0.9762*	0.6856
IRAK4	0.9936	0.0287	0.7611	0.8611	0.9938	0.8571
MAP2K1	0.9931	0.0319	0.5497	0.7184	0.9795	0.6954
PDK1	0.9945	0.0271	0.6310	0.7757	0.9875	0.7613



AI nella progettazione di antivirali (anti Sars-COV-2)

Problema:

- Progettare molecole selettive sulla proteasi virale di Sars-COV-2 (Mpro)
- Le proteasi hanno similarità dei siti di binding
- Prioritizzazione molecole con metodiche MM classiche insufficiente se usata singolarmente

Soluzione:

- ✓ Sfruttare dati strutturali disponibili
- ✓ Creazione di un modello di classificazione binaria (Attivo/Inattivo) per 'blindare' la predizione di attività

International Journal of Molecular Sciences

MDPI

Article

Support Vector Machine as a Supervised Learning for the Prioritization of Novel Potential SARS-CoV-2 Main Protease Inhibitors

Nedra Mekni [1,2,*], Claudia Coronnello [2], Thierry Langer [1], Maria De Rosa [2,†] and Ugo Perricone [2,*,†]

[1] Department of Pharmaceutical Chemistry, University of Vienna, 1090 Vienna, Austria; thierry.langer@univie.ac.at
[2] Drug Discovery Unit, Fondazione Ri.MED, 90128 Palermo, Italy; ccoronnello@fondazionerimed.com (C.C.); mderosa@fondazionerimed.com (M.D.R.)
* Correspondence: nmekni@fondazionerimed.com (N.M.); uperricone@fondazionerimed.com (U.P.)
† These authors contributed equally to this work.

Abstract: In the last year, the COVID-19 pandemic has highly affected the lifestyle of the world population, encouraging the scientific community towards a great effort on studying the infection molecular mechanisms. Several vaccine formulations are nowadays available and helping to reach immunity. Nevertheless, there is a growing interest towards the development of novel anti-covid drugs. In this scenario, the main protease (Mpro) represents an appealing target, being the enzyme responsible for the cleavage of polypeptides during the viral genome transcription. With the aim of sharing new insights for the design of novel Mpro inhibitors, our research group developed a machine learning approach using the support vector machine (SVM) classification. Starting from a dataset of two million commercially available compounds, the model was able to classify two hundred novel chemo-types as potentially active against the viral protease. The compounds labelled as actives by SVM were next evaluated through consensus docking studies on two PDB structures and their binding mode was compared to well-known protease inhibitors. The best five compounds selected by consensus docking were then submitted to molecular dynamics to deepen binding interactions stability. Of note, the compounds selected via SVM retrieved all the most important interactions known in the literature.

Keywords: machine learning; classification; main protease; COVID-19; molecular docking

check for updates

Citation: Mekni, N.; Coronnello, C.; Langer, T.; Rosa, M.D.; Perricone, U. Support Vector Machine as a Supervised Learning for the Prioritization of Novel Potential SARS-CoV-2 Main Protease Inhibitors. *Int. J. Mol. Sci.* **2021**, *22*, 7714. https://doi.org/10.3390/ijms22147714

Academic Editor: Daeui Park

Received: 2 July 2021
Accepted: 19 July 2021
Published: 19 July 2021

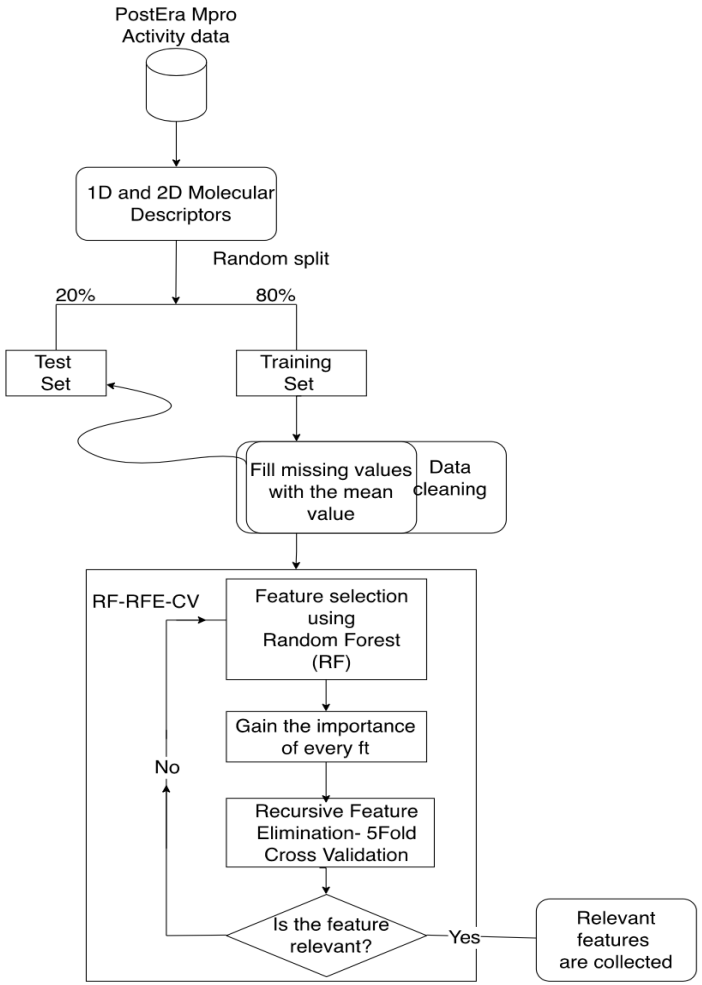
1. Introduction

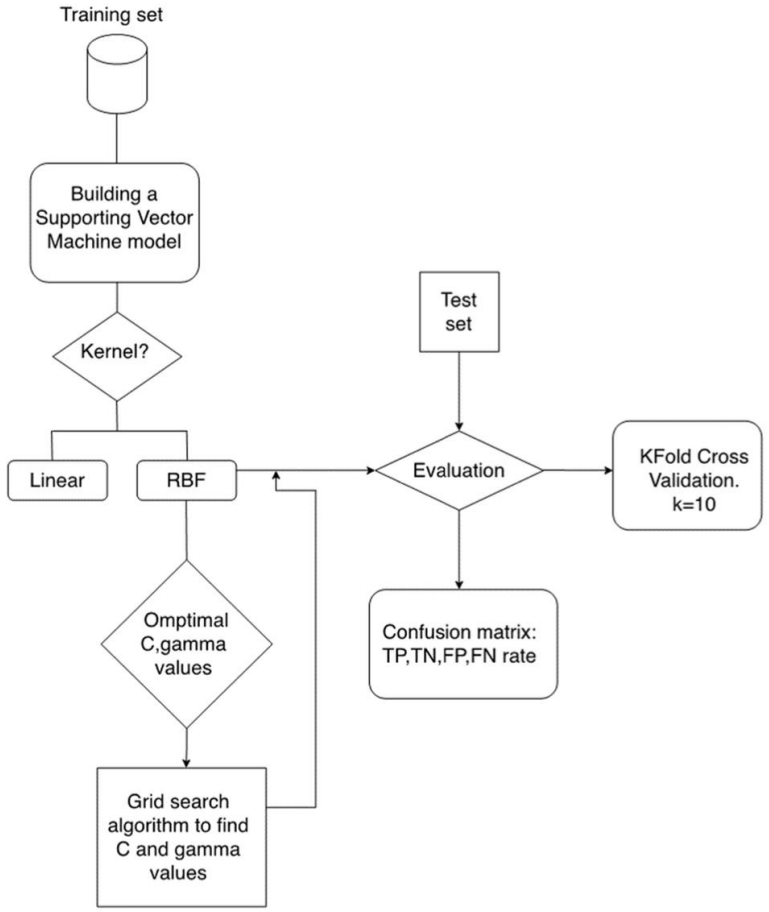
The COVID-19 pandemic, also known as Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is afflicting the health and routines of billions of people worldwide.

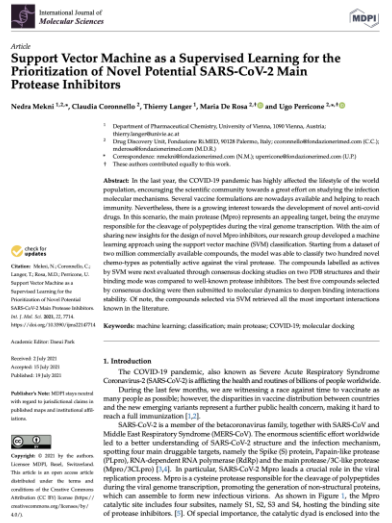
During the last few months, we are witnessing a race against time to vaccinate as many people as possible; however, the disparities in vaccine distribution between countries and the new emerging variants represent a further public health concern, making it hard to reach a full immunization [1,2].

SARS-CoV-2 is a member of the betacoronavirus family, together with SARS-CoV and Middle East Respiratory Syndrome (MERS-CoV). The enormous scientific effort worldwide led to a better understanding of SARS-CoV-2 structure and the infection mechanism, spotting four main druggable targets, namely the Spike (S) protein, Papain-like protease (PLpro), RNA-dependent RNA polymerase (RdRp) and the main protease/3C-like protease (Mpro/3CLpro) [3,4]. In particular, SARS-CoV-2 Mpro leads a crucial role in the viral replication process. Mpro is a cysteine protease responsible for the cleavage of polypeptides during the viral genome transcription, promoting the generation of non-structural proteins, which can assemble to form new infectious virions. As shown in Figure 1, the Mpro catalytic site includes four subsites, namely S1, S2, S3 and S4, hosting the binding site of protease inhibitors. [5]. Of special importance, the catalytic dyad is enclosed into the

AI nella progettazione di antivirali (anti Sars-COV-2)







International Journal of *Molecular Sciences*

MDPI

Article

Support Vector Machine as a Supervised Learning for the Prioritization of Novel Potential SARS-CoV-2 Main Protease Inhibitors

Nedra Mekni [1,2,*], Claudia Coronnello [3], Thierry Langer [1], Maria De Rosa [2,†] and Ugo Perricone [2,*,†]

[1] Department of Pharmaceutical Chemistry, University of Vienna, 1090 Vienna, Austria; thierry.langer@univie.ac.at
[2] Drug Discovery Unit, Fondazione Ri.MED, 90128 Palermo, Italy; coronnello@fondazionerimed.com (C.C.); mderosa@fondazionerimed.com (M.D.R.)
* Correspondence: nmekni@fondazionerimed.com (N.M.); uperricone@fondazionerimed.com (U.P.)
† These authors contributed equally to this work.

Abstract: In the last year, the COVID-19 pandemic has highly affected the lifestyle of the world population, encouraging the scientific community towards a great effort on studying the infection molecular mechanisms. Several vaccine formulations are nowadays available and helping to reach immunity. Nevertheless, there is a growing interest towards the development of novel anti-covid drugs. In this scenario, the main protease (Mpro) represents an appealing target, being the enzyme responsible for the cleavage of polypeptides during the viral genome transcription. With the aim of sharing new insights for the design of novel Mpro inhibitors, our research group developed a machine learning approach using the support vector machine (SVM) classification. Starting from a dataset of two million commercially available compounds, the model was able to classify two hundred novel chemo-types as potentially active against the viral protease. The compounds labelled as actives by SVM were next evaluated through consensus docking studies on two PDB structures and their binding mode was compared to well-known protease inhibitors. The best five compounds selected by consensus docking were then submitted to molecular dynamics to deepen binding interactions stability. Of note, the compounds selected via SVM retrieved all the most important interactions known in the literature.

Keywords: machine learning; classification; main protease; COVID-19; molecular docking

Citation: Mekni, N.; Coronnello, C.; Langer, T.; Rosa, M.D.; Perricone, U. Support Vector Machine as a Supervised Learning for the Prioritization of Novel Potential SARS-CoV-2 Main Protease Inhibitors. Int. J. Mol. Sci. 2021, 22, 7714. https://doi.org/10.3390/ijms22147714

Academic Editor: Dansi Park

Received: 2 July 2021
Accepted: 15 July 2021
Published: 19 July 2021

1. Introduction

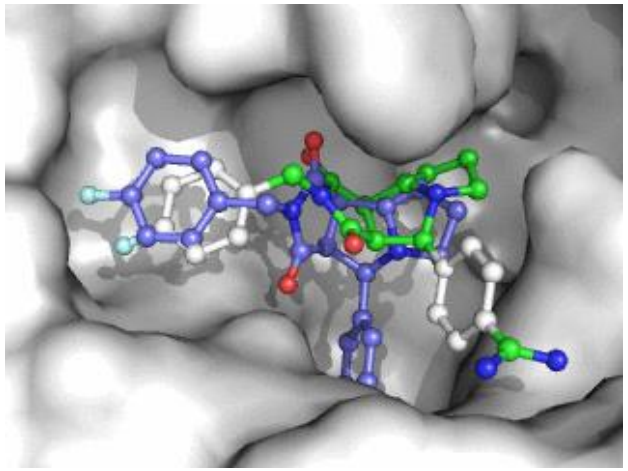
The COVID-19 pandemic, also known as Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is afflicting the health and routines of billions of people worldwide.

During the last few months, we are witnessing a race against time to vaccinate as many people as possible; however, the disparities in vaccine distribution between countries and the new emerging variants represent a further public health concern, making it hard to reach a full immunization [1,2].

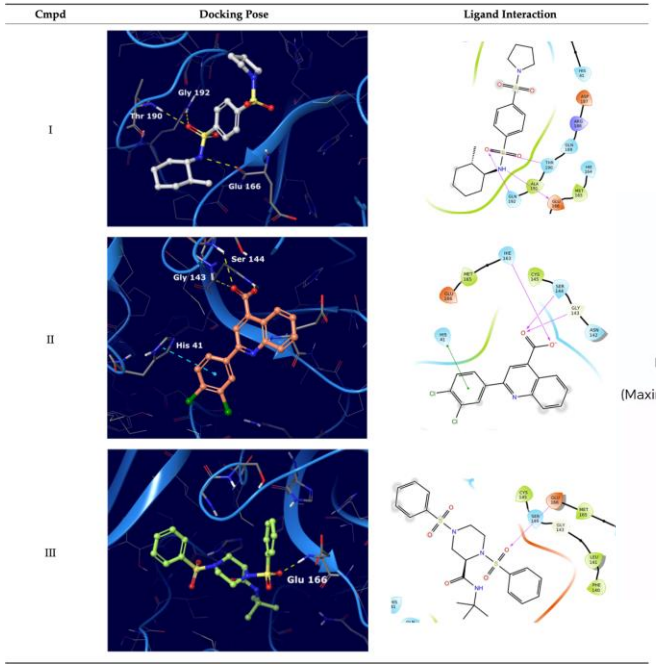
SARS-CoV-2 is a member of the betacoronavirus family, together with SARS-CoV and Middle East Respiratory Syndrome (MERS-CoV). The enormous scientific effort worldwide led to a better understanding of SARS-CoV-2 structure and the infection mechanism, spotting four main druggable targets, namely the Spike (S) protein, Papain-like protease (PLpro), RNA-dependent RNA polymerase (RdRp) and the main protease/3C-like protease (Mpro/3CLpro) [3,4]. In particular, SARS-CoV-2 Mpro leads a crucial role in the viral replication process. Mpro is a cysteine protease responsible for the cleavage of polypeptides during the viral genome transcription, promoting the generation of non-structural proteins, which can assemble to form new infectious virions. As shown in Figure 1, the Mpro catalytic site includes four subsites, namely S1, S2, S3 and S4, hosting the binding site of protease inhibitors. [5]. Of special importance, the catalytic dyad is enclosed into the

AI nella progettazione di antivirali (anti Sars-COV-2)

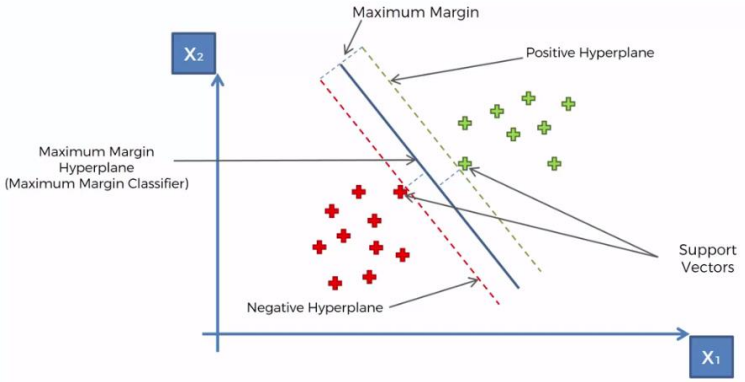
Molecular Docking

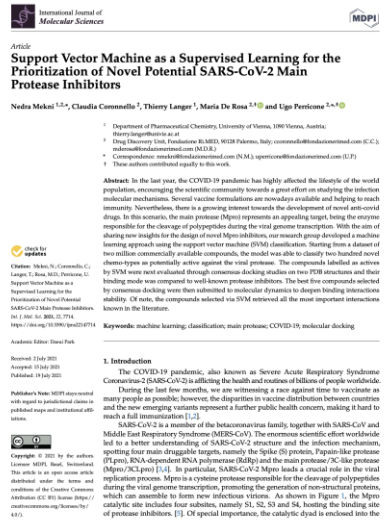


Ranking /Prioritisation



Modello ML per la classificazione





International Journal of Molecular Sciences

Article

Support Vector Machine as a Supervised Learning for the Prioritization of Novel Potential SARS-CoV-2 Main Protease Inhibitors

Nedra Mekni [1,2,*], Claudia Coronnello [3], Thierry Langer [1], Maria De Rosa [2,†] and Ugo Perricone [2,*,†]

Abstract: In the last year, the COVID-19 pandemic has highly affected the lifestyle of the world population, encouraging the scientific community towards a great effort on studying the infection molecular mechanisms. Several vaccine formulations are nowadays available and helping to reach immunity. Nevertheless, there is a growing interest towards the development of novel anti-covid drugs. In this scenario, the main protease (Mpro) represents an appealing target, being the enzyme responsible for the cleavage of polypeptides during the viral genome transcription. With the aim of sharing new insights for the design of novel Mpro inhibitors, our research group developed a machine learning approach using the support vector machine (SVM) classification. Starting from a dataset of two million commercially available compounds, the model was able to classify two hundred novel chemo-types as potentially active against the viral protease. The compounds labelled as actives by SVM were next evaluated through consensus docking studies on two PDB structures and their binding mode was compared to well-known protease inhibitors. The best five compounds selected by consensus docking were then submitted to molecular dynamics to deepen binding interactions stability. Of note, the compounds selected via SVM retrieved all the most important interactions known in the literature.

Keywords: machine learning; classification; main protease; COVID-19; molecular docking

1. Introduction

The COVID-19 pandemic, also known as Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is afflicting the health and routines of billions of people worldwide.

During the last few months, we are witnessing a race against time to vaccinate as many people as possible; however, the disparities in vaccine distribution between countries and the new emerging variants represent a further public health concern, making it hard to reach a full immunization [1,2].

SARS-CoV-2 is a member of the betacoronavirus family, together with SARS-CoV and Middle East Respiratory Syndrome (MERS-CoV). The enormous scientific effort worldwide led to a better understanding of SARS-CoV-2 structure and the infection mechanism, spotting four main druggable targets, namely the Spike (S) protein, Papain-like protease (PLpro), RNA-dependent RNA polymerase (RdRp) and the main protease/3C-like protease (Mpro/3CLpro) [3,4]. In particular, SARS-CoV-2 Mpro leads a crucial role in the viral replication process. Mpro is a cysteine protease responsible for the cleavage of polypeptides during the viral genome transcription, promoting the generation of non-structural proteins, which can assemble to form new infectious virions. As shown in Figure 1, the Mpro catalytic site includes four subsites, namely S1, S2, S3 and S4, hosting the binding site of protease inhibitors. [5]. Of special importance, the catalytic dyad is enclosed into the

Int. J. Mol. Sci. 2021, 22, 7714. https://doi.org/10.3390/ijms22147714

AI nella predizione di Tossicità

Problema:

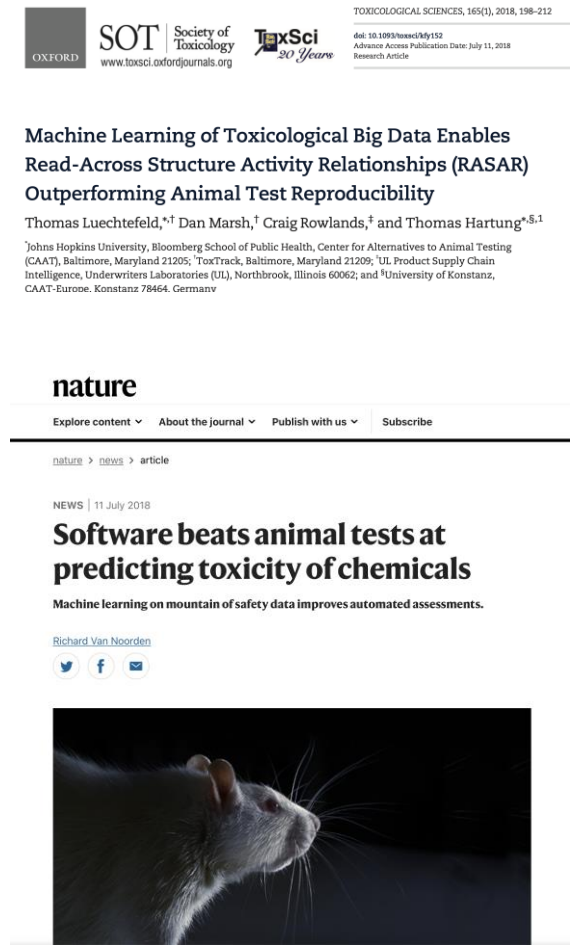
- La tossicità di un farmaco è un fattore multiparametrico
- Testare la tossicità comporta il sacrificio di molti animali e costi molto elevati
- Trattandosi di fattori multiparmaetrici, molti test in vivo sono poco affidabili perché il «sistema» animale è troppo differente da quello umano

Soluzione:

- ✓ Applicare multiclassificatori che considerino contemporanemente tutti i parametri disponibili
- ✓ Applicazione di modelli quali-quantitativi per permettere l'ottimizzazione molecolare e la riduzione della tossicità

Vantaggi:

- ➢ Predire per molecole simili a quelle studiate la possibile tossicità e anticipare fallimenti nelle campagne di drug discovery
- ➢ Lavorare con dati umani opportunamente modellati e possibilità di trasferire i modelli da un comparto ad un altro



TOXICOLOGICAL SCIENCES, 165(1), 2018, 198–212

Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility

Thomas Luechtefeld,*,† Dan Marsh,† Craig Rowlands,‡ and Thomas Hartung*,§,1

nature

Explore content · About the journal · Publish with us · Subscribe

nature › news › article

NEWS | 11 July 2018

Software beats animal tests at predicting toxicity of chemicals

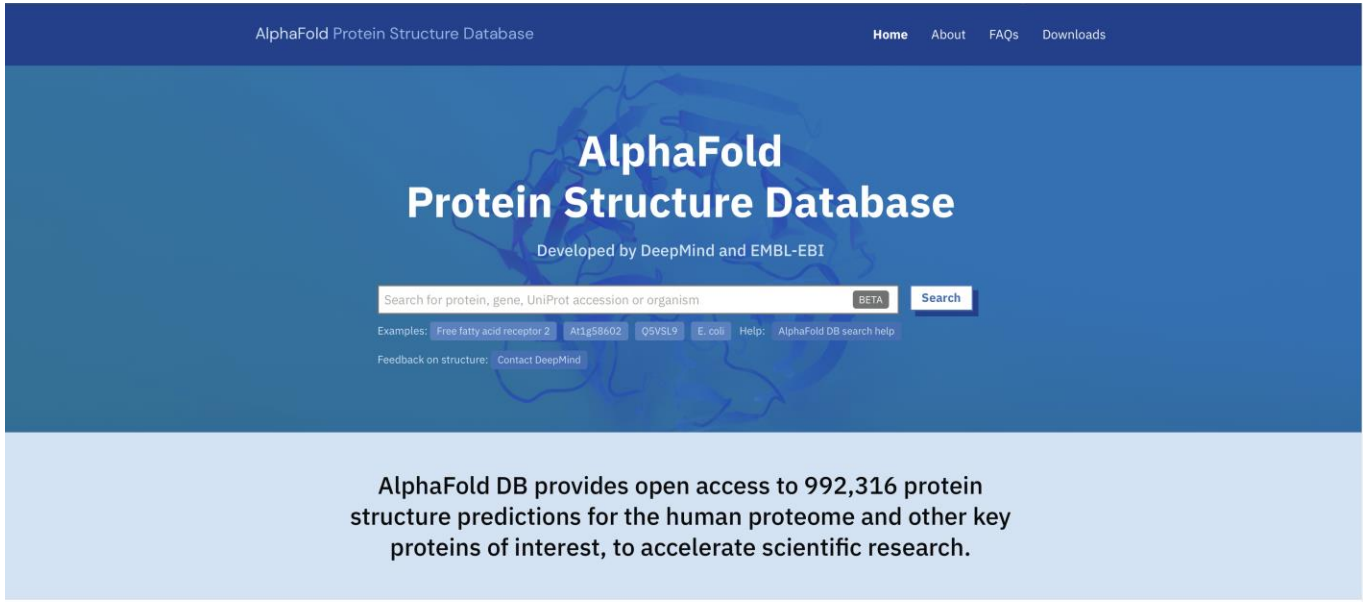
Machine learning on mountain of safety data improves automated assessments.

Richard Van Noorden

AI nella predizione di Strutture 3D…il caso Alphafold

Problema:

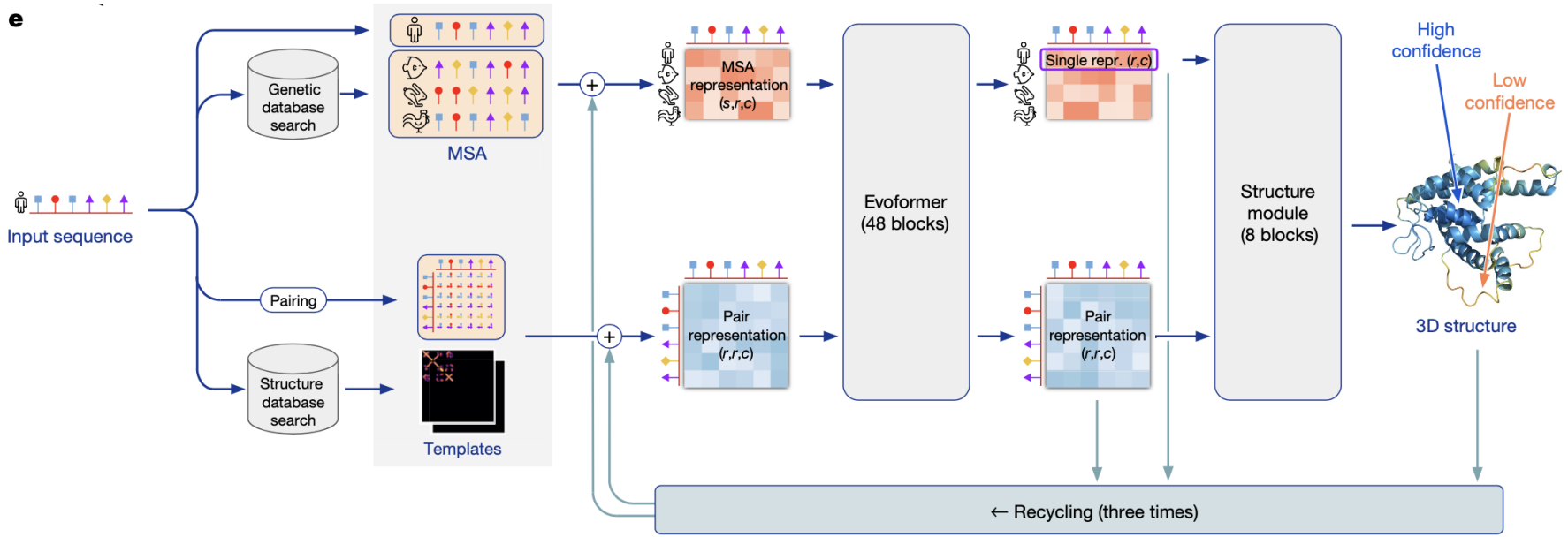
- Non tutte le strutture tridimensionali delle proteine sono disponibili (es. difficoltà legate ad isolamento e purificazione proteina)
- Gli studi strutturali stanno alla base della comprensione di processi biologici e patofisiologici

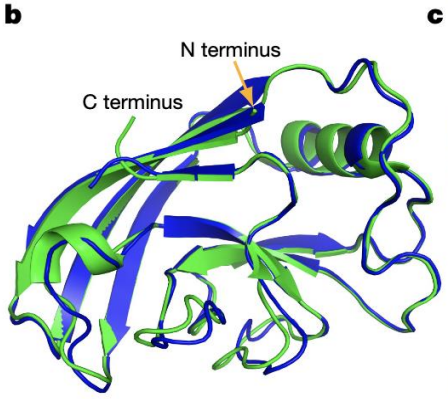


Soluzione:

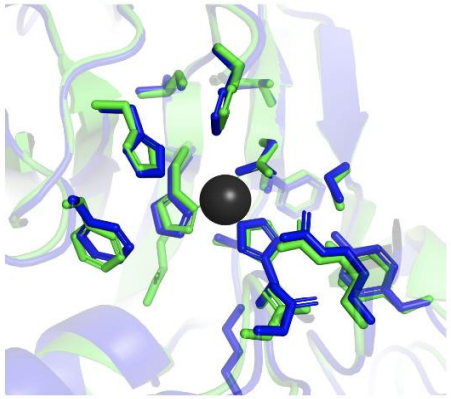
- ✓ Sfruttare tutti i dati disponibili in letteratura su sequenze e folding proteine (input : training/validation/test sets)
- ✓ Addestrare reti in grado di predire la struttura tridimensionale partendo dalla sequenza aminoacidica
- ✓ Predire strutture non disponibili

AI nella predizione di Strutture 3D…il caso Alphafold

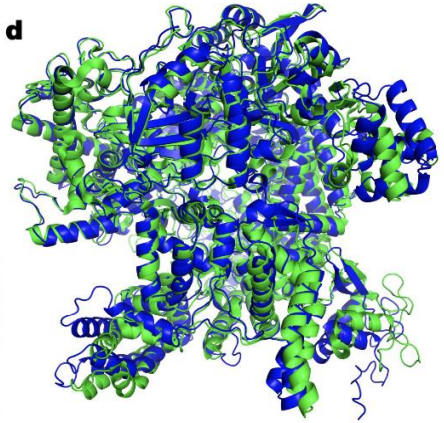




AlphaFold Experiment
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93



AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn



AlphaFold Experiment
r.m.s.d.$_{95}$ = 2.2 Å; TM-score = 0.96

AI nella diagnostica (Imaging)

La radiomica rappresenta una nuova frontiera della medicina basata sull'estrazione di dati quantitativi dalle immagini radiologiche che non possono essere rilevati dall'occhio umano e sull'uso di questi dati per la creazione di sistemi di supporto alle decisioni cliniche.

L'obiettivo a lungo termine della radiomica è quello di migliorare la diagnosi non invasiva delle patologie focali e diffuse di vari organi attraverso la comprensione di collegamenti tra i dati di imaging quantitativo e le caratteristiche molecolari e patologiche delle lesioni.

Studi pubblicati nell'ultimo decennio hanno dimostrato l'enorme potenziale della radiomica nella patologia tumorale e non, nell'ambito di diversi sistemi e apparati inclusi encefalo, polmone, mammella, apparato gastrointestinale e genitourinario.

AI nella diagnostica (Imaging)

Il processo di creazione di un database di caratteristiche quantitative correlative, che può essere utilizzato per analizzare casi successivi (sconosciuti), comprende i seguenti passaggi:

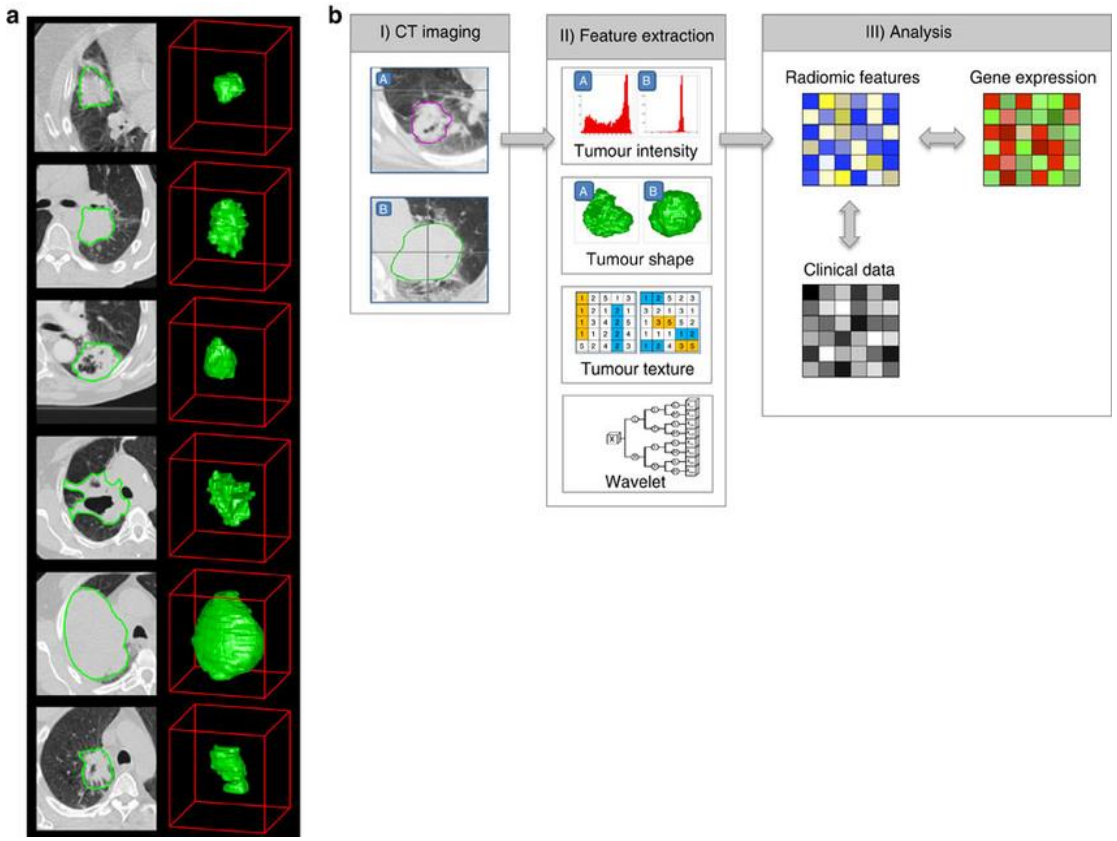
Elaborazione iniziale dell'immagine
Utilizzando una varietà di algoritmi di ricostruzione come contrasto, miglioramento dei bordi, ecc. → qualità e usabilità delle immagini → Accuratezza informazioni dedotte

Segmentazione dell'immagine
Identifica/crea aree (immagini 2D) o volumi di interesse (immagini 3D). Può essere eseguito manualmente, semiautomatico o completamente automatizzato utilizzando l'intelligenza artificiale (usato in big data)

Estrazione e classificazione
Le caratteristiche includono volume, forma, superficie, densità e intensità, consistenza, posizione e relazioni con i tessuti circostanti.

Le caratteristiche semantiche sono quelle comunemente usate nel lessico radiologico per descrivere le regioni di interesse.

Le caratteristiche agnostiche sono quelle che tentano di catturare l'eterogeneità della lesione attraverso descrittori matematici quantitativi (kurtosis or skewness (of the image histogram), Haralick textures, Laws textures).



Case courtesy of Dr Mohammed Sultan, Radiopaedia.org, rID: 65398

Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, et al Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, NatCommun, 2014 Jun 3;5:4006, doi: 10.1038/ncomms5006

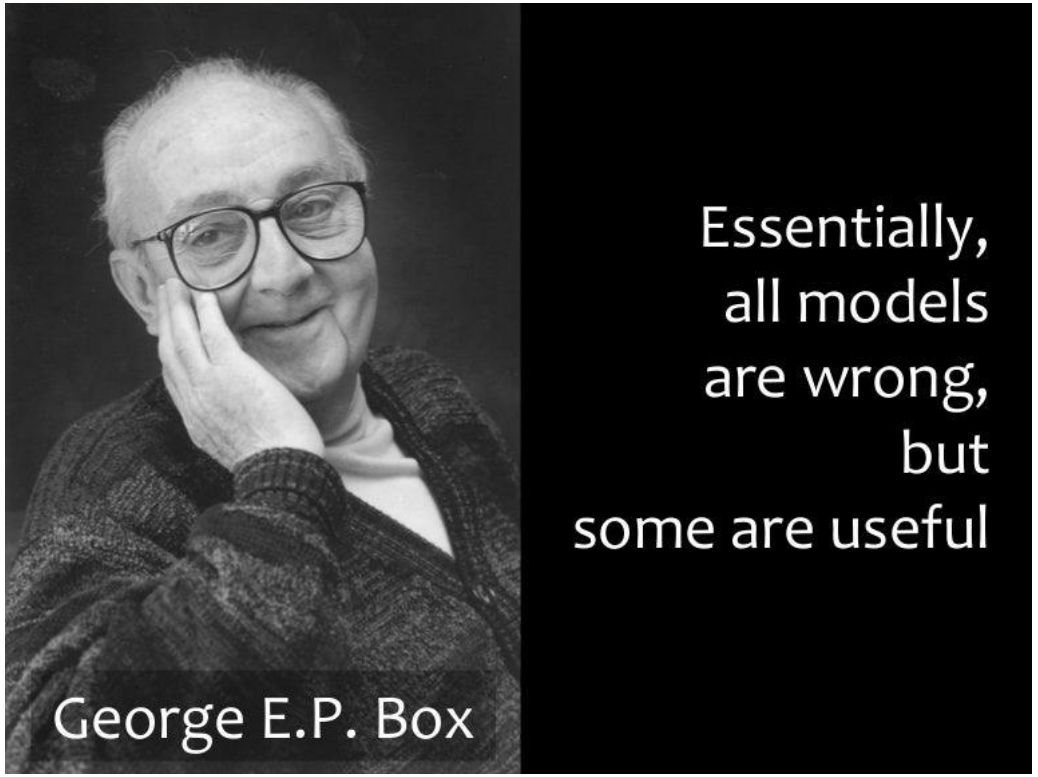
Conclusioni

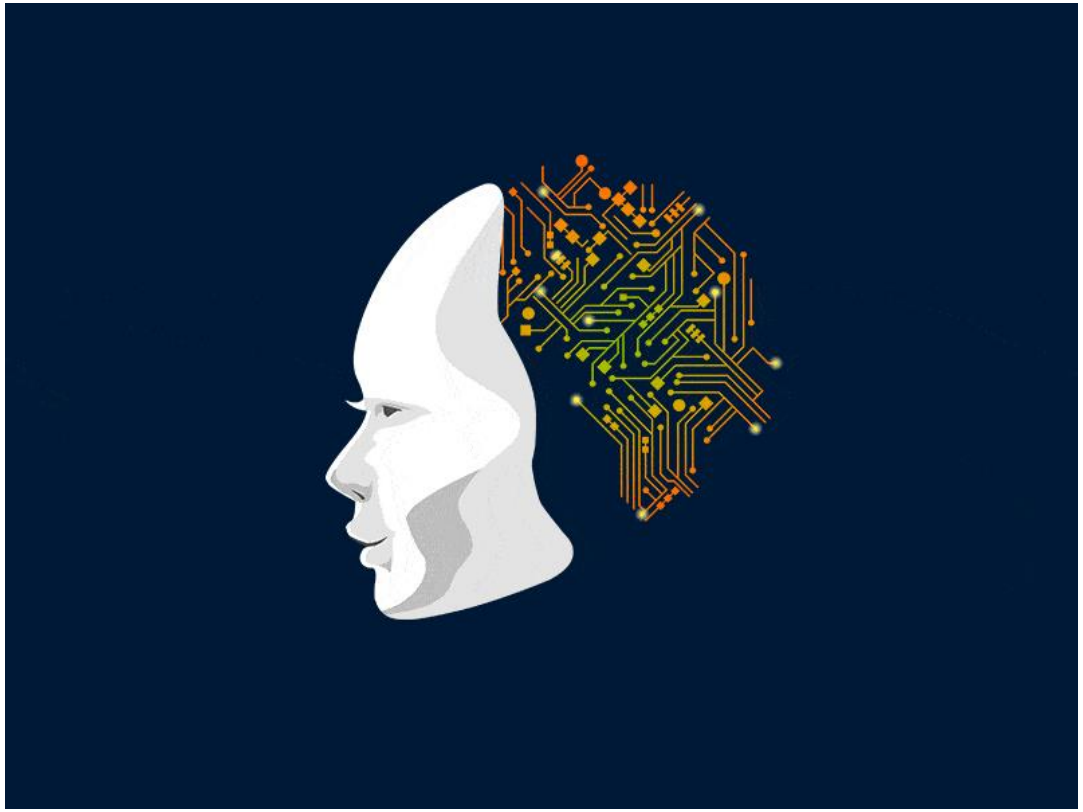
Punti di forza:

- Predire per molecole simili a quelle studiate la possibile tossicità e anticipare fallimenti nelle campagne di drug discovery
- Lavorare con dati umani opportunamente modellati e possibilità di trasferire i modelli da un comparto ad un altro
- Capacità di predire in largo anticipo situazioni che potrebbero sfuggire all'indagine visiva dell'operatore
- Possibilità di clusterizzare i pazienti su dati multiparametrici

Punti di debolezza:

- Reperibilità di dati in quantità sufficiente per creare modelli robusti
- Validazione molto complessa (non sempre disponibili sufficienti dati sperimentali)
- Riproducibilità non ottimale cambiando macchine (es. due modelli di GPU non è detto che rispondano allo stesso modo)
- Necessario alto livello di specializzazione degli operatori





GRAZIE PER LA VOSTRA ATTENZIONE