



Fondazione
Ri.MED



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

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



Molecular Informatics



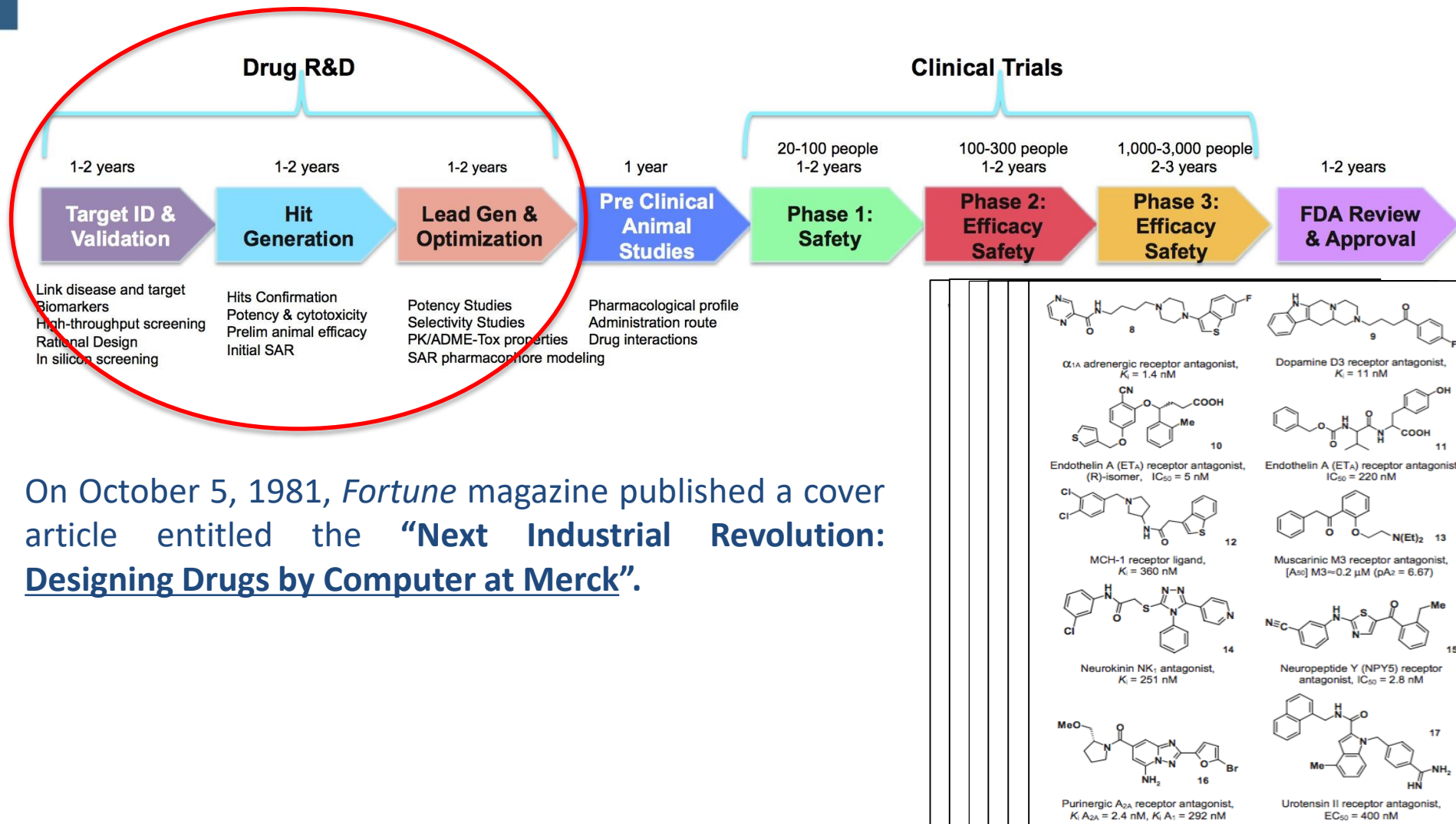
Ugo Perricone
Group Leader
in Molecular
Informatics

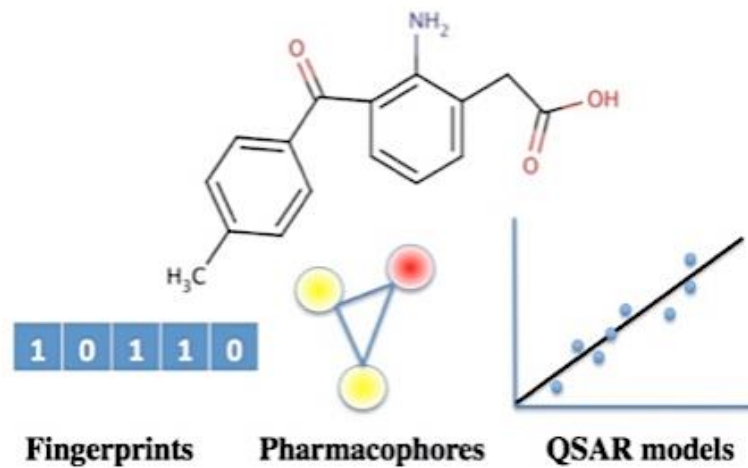


4 FELLOWSHIPS

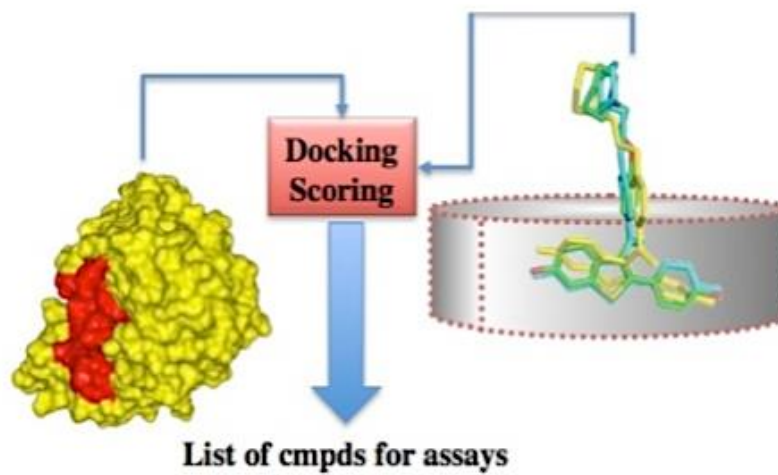
4 PhD STUDENTS

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

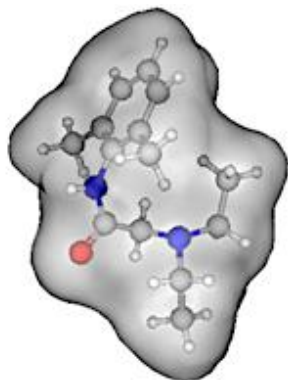




Ligand-based



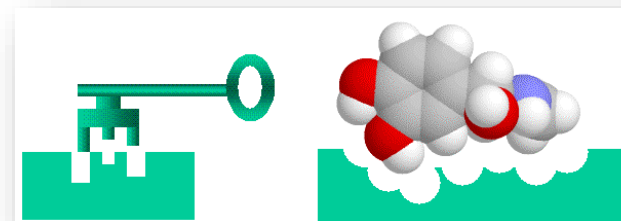
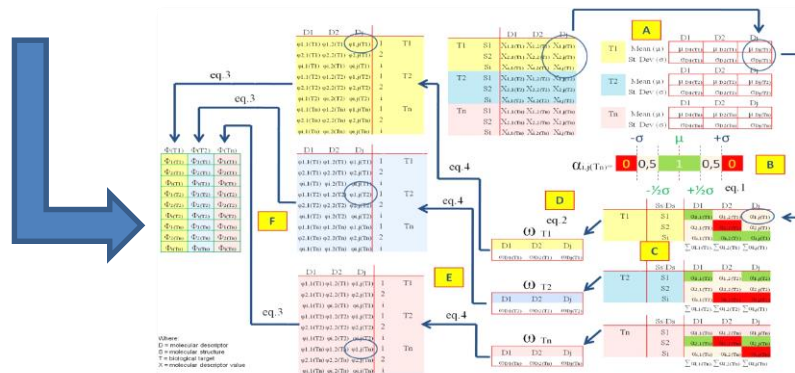
Structure-based

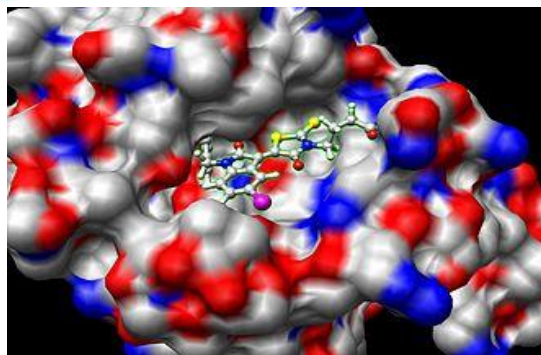


NO	INDICE	INDICE	INDICE	INDICE	INDICE	INDICE	INDICE	INDICE	INDICE
Numero	Numero	Numero	Numero	Numero	Numero	Numero	Numero	Numero	Numero
1	0.0	44.912412	2.09109	4.0	0.17961	1.932761903	3.332430794	0.0	0.0
2	0.0	35.014526	1.945246	5.149926	0.241792746	0.0	0.0	0.0	0.0
3	0.0	8.696802	1.77976	3.321926	-0.3999810.0	0.0	0.0	0.0	0.0
4	0.0	8.696802	1.77976	3.321926	-0.3999810.0	0.0	0.0	0.0	0.0
5	0.0	25.658424	2.058442	4.321926	-0.0115285.0	0.0	0.0	0.0	0.0
6	0.0	8.67152	1.71782	3.0	-0.3999810.0	0.0	0.0	0.0	0.0
7	0.0	19.84296	1.962296	4.321926	-0.0115285.0	0.0	0.0	0.0	0.0
8	0.0	14.94091	1.812005	4.0	-0.1037618.0	0.0	0.0	0.0	0.0
9	0.0	34.927198	2.36249	5.807395	-0.3020210.0	0.0	0.0	0.0	0.0
10	0.0	8.696802	1.77976	3.321926	-0.3999810.0	0.0	0.0	0.0	0.0
11	0.0	8.67152	1.71782	3.0	-0.3999810.0	0.0	0.0	0.0	0.0
12	0.0	17.362417	1.911325	4.321926	-0.0779895.0	0.0	0.0	0.0	0.0
13	0.0	19.84296	1.962296	4.0	0.0.0.0	0.0	0.0	0.0	0.0
14	0.0	12.476585	1.818892	3.807395	0.3020210.0	0.0	0.0	0.0	0.0
15	0.0	10.323466	1.751991	3.599967	-0.3020210.0	0.0	0.0	0.0	0.0
16	0.0	10.323466	1.751991	3.599967	-0.3020210.0	0.0	0.0	0.0	0.0
17	0.0	16.704546	1.876955	4.321926	-0.0115285.0	0.0	0.0	0.0	0.0
18	0.0	14.811268	1.823889	4.0	7.702.48.125	0.197088784	0.0	0.0	0.0
19	0.0	16.810747	1.893375	4.321926	-0.3999810.0	0.0	0.0	0.0	0.0
20	0.0	10.479177	1.779662	3.599967	0.3020210.0	0.0	0.0	0.0	0.0
21	0.0	8.696802	1.77976	3.321926	-0.3999810.0	0.0	0.0	0.0	0.0
22	0.0	16.84296	1.962296	4.0	0.0.0.0	0.0	0.0	0.0	0.0
23	0.0	15.84296	1.962296	3.599967	0.3020210.0	0.0	0.0	0.0	0.0
24	0.0	35.704546	1.962296	5.149926	-0.3020210.0	0.0	0.0	0.0	0.0
25	0.0	12.299967	1.766171	3.807395	-0.3020210.0	0.0	0.0	0.0	0.0

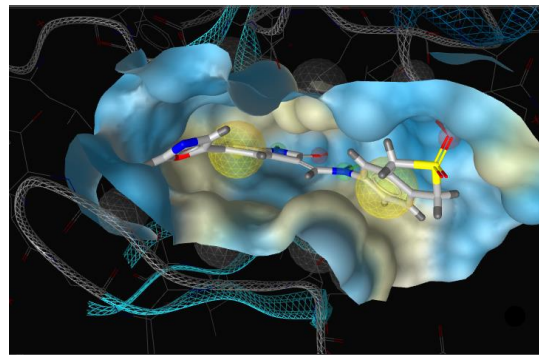
"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment." **R. Todeschini**

- 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)

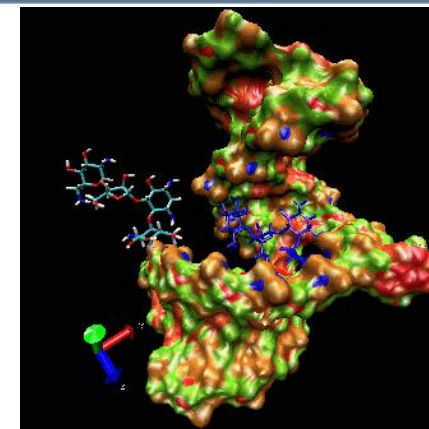




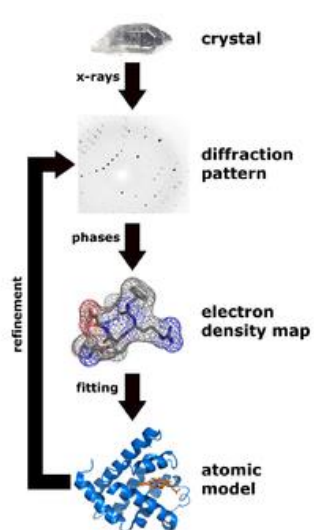
DOCKING



*PHARMACOPHORE



MOLECULAR DYNAMICS (MD)



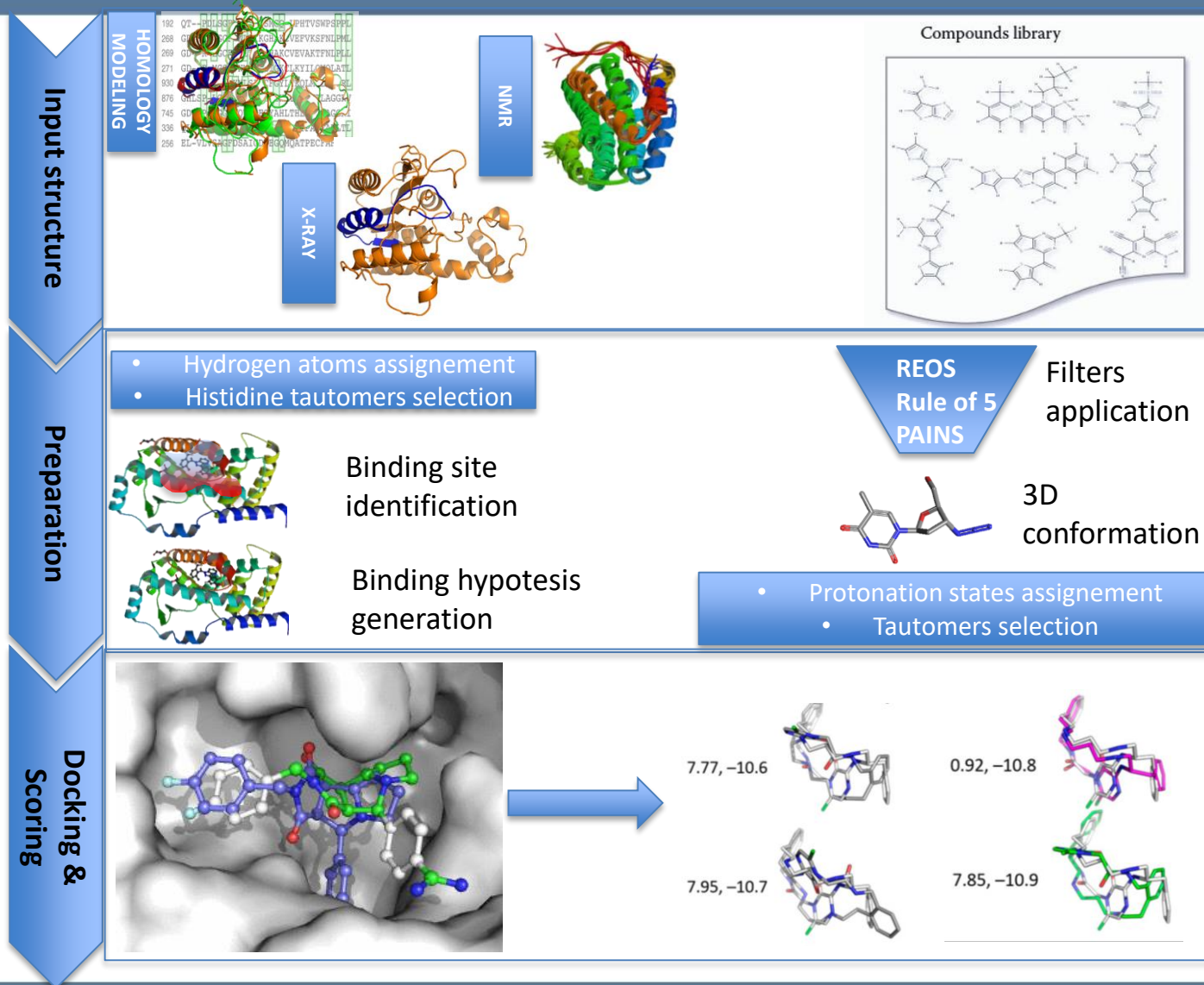
- Missing loops → lack of some receptor structural and geometrical information
- Mean atomic coordinates → possible misinterpretation of ligand-receptor interaction pattern
- X-ray crystal resolution not always good ($> 2 \text{ \AA}$)

- 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



Explore Chemistry

Quickly find chemical information from authoritative sources

Try covid-19 aspirin EGFR C9H8O4 57-27-2 C1=CC=C(C=C1)C=O InChI=1S/C3H6O/c1-3(2)4/h1-2H3

Use Entrez Compounds Substances BioAssays



Draw Structure



Upload ID List



Browse Data



Periodic Table

112M Compounds 282M Substances 295M Bioactivities 34M Literature 42M Patents

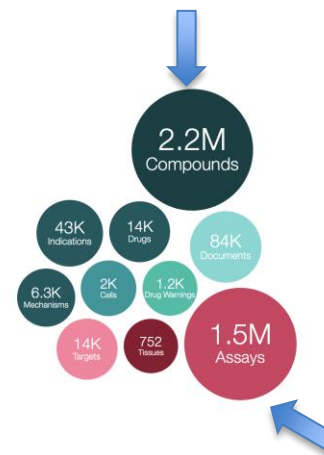
See More Statistics >

868 Data Sources

Explore Data Sources >



ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs.



Explore ChEMBL

Description: Shows a summary of the ChEMBL entities and quantities of data for each of them.

Instructions: Click on a bubble to explore a specific ChEMBL entity in more detail.

RCSB PDB Deposit Search Visualize Analyze Download Learn More Documentation Careers MyPDB

RCSB PDB
PROTEIN DATA BANK

193183 Biological
Macromolecular Structures
Enabling Breakthroughs in
Research and Education

PDB Archive

Advanced Search | Browse Annotations

Help

PDB-101

WORLDWIDE
PDB
PROTEIN DATA BANK

EMBL
Data Resource
United Data Resource for SDSC

NUCLEIC ACID
DATABASE

Worldwide
Protein Data Bank
Foundation

Developers: Join the RCSB PDB Team

Explore Open Positions

Welcome

Deposit

Search

Visualize

Analyze

Download

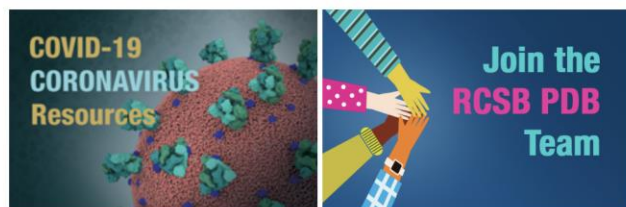
Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.



UniProt BLAST Align Peptide search ID mapping SPARQL

Status

Reviewed (Swiss-Prot)
(567,483)

Unreviewed (TrEMBL)
(231,354,252)

Popular organisms

Human (204,906)

Rice (149,195)

A. thaliana (136,845)

Mouse (86,440)

Zebrafish (62,024)

Taxonomy

Filter by taxonomy

Proteins with

3D structure (56,528)

Active site (9,877,374)

Activity regulation (570,327)

Allergen (934)

Alternative products (isoforms)
(25,701)

More items

Protein existence

Predicted (152,950,479)

Homology (77,239,269)

Transcript level (1,430,589)

Protein level (299,555)

UniProtKB 231,921,735 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
A0A0C5B5G6	MOTSC_HUMAN	Mitochondrial-derived peptide MOTS-c[...]	MT-RNR1	Homo sapiens (Human)	16 AA
A0JNW5	UH1BL_HUMAN	UHRF1-binding protein 1-like[...]	UHRF1BP1L, KIAA0701, SHIP164	Homo sapiens (Human)	1,464 AA
A0JP26	POTB3_HUMAN	POTE ankyrin domain family member B3	POTEB3	Homo sapiens (Human)	581 AA
A0PK11	CLRN2_HUMAN	Clarin-2	CLRN2	Homo sapiens (Human)	232 AA
A1A4S6	RHG10_HUMAN	Rho GTPase-activating protein 10[...]	ARHGAP10, GRAF2	Homo sapiens (Human)	786 AA
A1A519	F170A_HUMAN	Protein FAM170A[...]	FAM170A, ZNFD	Homo sapiens (Human)	330 AA
A1L3X0	ELOV7_HUMAN	Elongation of very long chain fatty acids protein 7[...]	ELOVL7	Homo sapiens (Human)	281 AA
A1X283	SPD2B_HUMAN	SH3 and PX domain-containing protein 2B[...]	SH3PXD2B, FAD49, KIAA1295, TKS4	Homo sapiens (Human)	911 AA
A2A2Y4	FRMD3_HUMAN	FERM domain-containing protein 3[...]	FRMD3, EPB41L4O	Homo sapiens (Human)	597 AA
A2RU14	TM218_HUMAN	Transmembrane protein 218	TMEM218	Homo sapiens (Human)	115 AA
A2RUB6	CCD66_HUMAN	Coiled-coil domain-containing protein 66	CCDC66	Homo sapiens (Human)	948 AA
A2RUC4	TYW5_HUMAN	tRNA wybutosine-synthesizing protein 5[...]	TYW5, C2orf60	Homo sapiens (Human)	315 AA
A4D1B5	GSAP_HUMAN	Gamma-secretase-activating protein[...]	GSAP, PION	Homo sapiens (Human)	854 AA
A4GXA9	EME2_HUMAN	Probable crossover junction endonuclease EME2[...]	EME2	Homo sapiens (Human)	379 AA
A5D8V7	ODAD3_HUMAN	Outer dynein arm-docking complex subunit 3[...]	ODAD3, CCDC151	Homo sapiens (Human)	595 AA
A5PLL7	PDES1_HUMAN	Plasmamylethanolamine desaturase[...]	PDES1, KUA, PDES, TMEM189	Homo sapiens (Human)	270 AA
A6BM72	MEG11_HUMAN	Multiple epidermal growth factor-like domains protein 11[...]	MEGF11, KIAA1781, UNQ1949/PRO4432	Homo sapiens (Human)	1,044 AA
A6H8Y1	BDP1_HUMAN	Transcription factor TFIIB component B'' homolog[...]	BDP1, KIAA1241, KIAA1689, TFNR	Homo sapiens (Human)	2,624 AA
A6ANC4	NKY2A_HUMAN	Humanov protein Nky-2 A1	NKY2A, NKY2E	Homo sapiens (Human)	901 AA

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

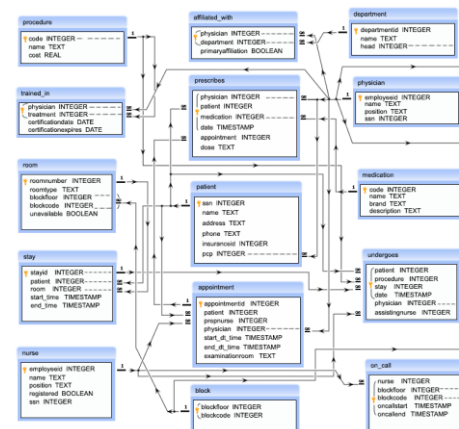
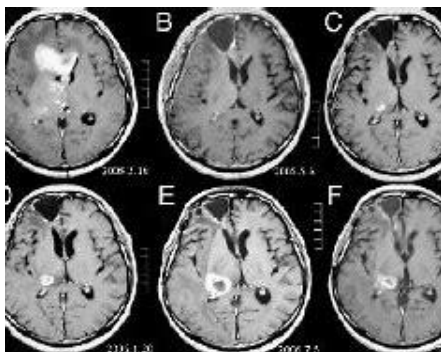
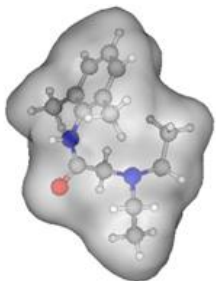
La varietà si riferisce ai molti tipi di dati disponibili. I tipi di dati tradizionali erano strutturati e si adattavano perfettamente a un [database relazionale](#). Con l'avvento dei Big Data, i dati arrivano come nuovi tipi di dati non strutturati. I tipi di dati non strutturati e semistrutturati, 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.

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?



Database Browser for SQLite

Database Structure | Browse Data | Edit Pragma | Execute SQL

recID	ProcessingStatus	OpenCurrent	BlockedCurrent	EventStart	EventEnd	BlockDepth	ResTime	RCCount	AbsEventStart	AutoChSpan	TimeSeries
184	normal	136.048481...	47.8884398...	0.102442...	0.2063717...	0.3519954...	0.1039288...	0.0017467...	458.10244...	1.0428571...	kgkL7WwY...
185	normal	136.388727...	49.4119037...	0.101160...	0.2531239...	0.3622872...	0.1519637...	0.0020683...	468.72916...	1.0333333...	Smk5ZFwY...
186	normal	136.295727...	48.7546930...	0.099610...	0.1772262...	0.3577125...	0.0776155...	0.0010883...	469.76761...	1.0451127...	HT54mq...
187	normal	136.728134...	48.6160741...	0.099204...	0.1834360...	0.3541046...	0.0842319...	0.0001860...	473.43920...	1.0483870...	+Hf54mq...
188	normal	136.163621...	75.2090860...	0.075814...	0.0837541...	0.5523434...	0.0079394...	0.0001326...	473.60581...	1.0722891...	xR+Ay2Z...
189	normal	136.479801...	98.4793567...	0.099866...	0.1059172...	0.7215672...	0.0060503...	0.0001528...	474.81886...	1.0645161...	ymZD5QY...
190	normal	135.580797...	49.4561541...	0.099369...	0.2429424...	0.3647725...	0.1435727...	0.0014158...	478.93336...	1.0348837...	0MyccrFA...
191	normal	136.060431...	48.0983706...	0.111056...	0.1306499...	0.3535074...	0.1195933...	0.0031418...	479.48305...	1.0540540...	6z7fw5QY...
192	normal	135.394986...	48.6723594...	0.099333...	0.2170996...	0.3594842...	0.1177660...	0.0023736...	480.64333...	1.0425531...	x8Pu49QY...
194	normal	136.532025...	49.9108603...	0.079606...	0.1311501...	0.3655615...	0.0515433...	0.0020511...	480.84560...	1.0512820...	AqNvByAY...
195	normal	136.016300...	49.8596692...	0.099892...	0.1492097...	0.3665712...	0.0503174...	0.0015597...	481.54889...	1.0512820...	eDY4qfYE...
196	normal	136.775120...	47.7841358...	0.101028...	0.2825822...	0.3493627...	0.1815534...	0.0016714...	482.09102...	1.0315789...	rrDU4bQY...
197	normal	137.516752...	56.9740218...	0.100124...	0.1413854...	0.4143060...	0.0412613...	0.0012905...	483.30812...	1.0545454...	b7MwAuQY...
198	normal	135.728152...	58.8360340...	0.101791...	0.1117746...	0.4334843...	0.0099829...	0.0001607...	485.43179...	1.0612244...	ImD5ccrY...
199	normal	137.152200...	46.6580419...	0.099647...	0.2092531...	0.3401917...	0.1096055...	0.0025605...	486.42764...	1.0410958...	o7axZY9Y...
200	normal	137.110974...	48.7001968...	0.099491...	0.1532354...	0.3551881...	0.0374315...	0.0021878...	487.13949...	1.0483870...	v7qW6QY...

Costituente Risultato Unità Int. di Riferimento Ris. Prec.

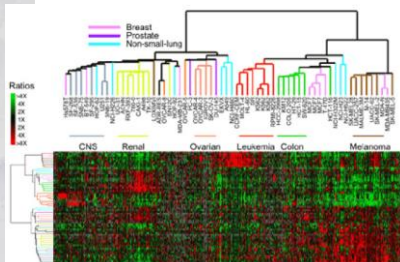
EMATOLOGIA E COAGULAZIONE

PROFilo EMATOLOGICO

Citometria: 4.89 x10.9L 4.40 - 11.00
 B-LEUCOCITI 2.264 x10.9L 4.50 - 5.90
 B-ERITROCITI *96 gt. 140 - 175
 B-EMOGLOBINA 0.289 % 0.410 - 0.500
 B-EMATOCRITO *109.5 % 80.0 - 96.0
 MCH *36.4 pg 28.0 - 33.0
 MCHC 332 gL 320 - 380
 RDW Parametro annullato
 B-PIASTRINE 276 x10.9L 150 - 450

Conteggio Differenziale dei Leucociti:

B-NEUTROFILI 2.14 x10.9L 1.80 - 7.80
 45.6 %
 B-LINFOCITI 1.40 x10.9L 1.10 - 4.80
 29.9 %
 B-MONOCITI 0.85 x10.9L 0.20 - 0.96
 4.1 %
 B-EOSINOFILI 0.19 x10.9L 0.00 - 0.50
 4.1 %
 B-BASOFILI 0.01 x10.9L 0.00 - 0.20
 0.2 %





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

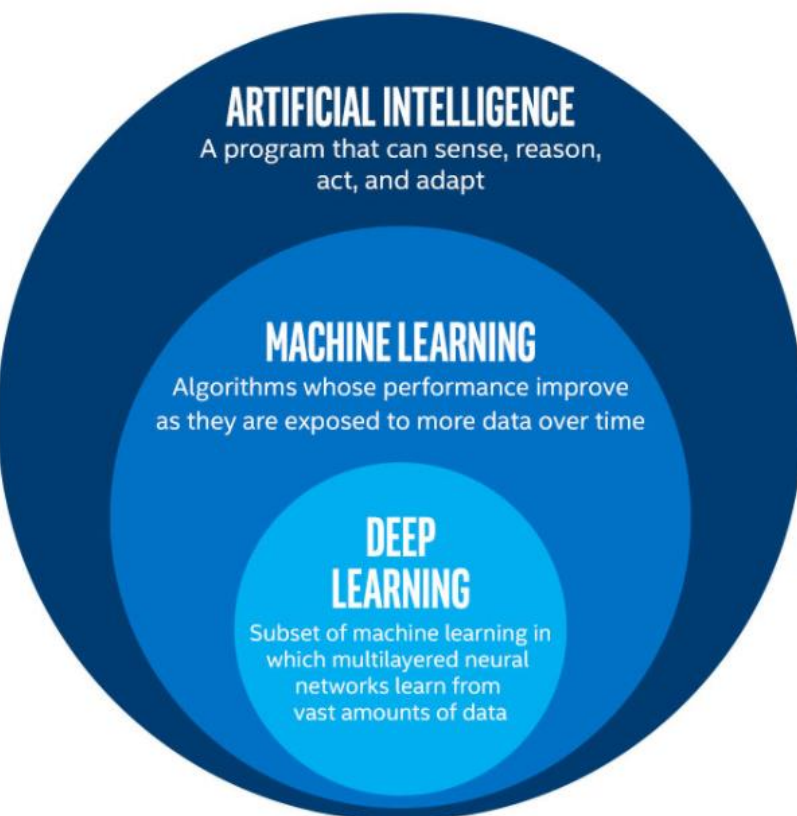
Creazione di modelli machine learning e intelligenza artificiale per sfruttare al massimo il potenziale dei dati archiviati e dare vita continua al dato.

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.



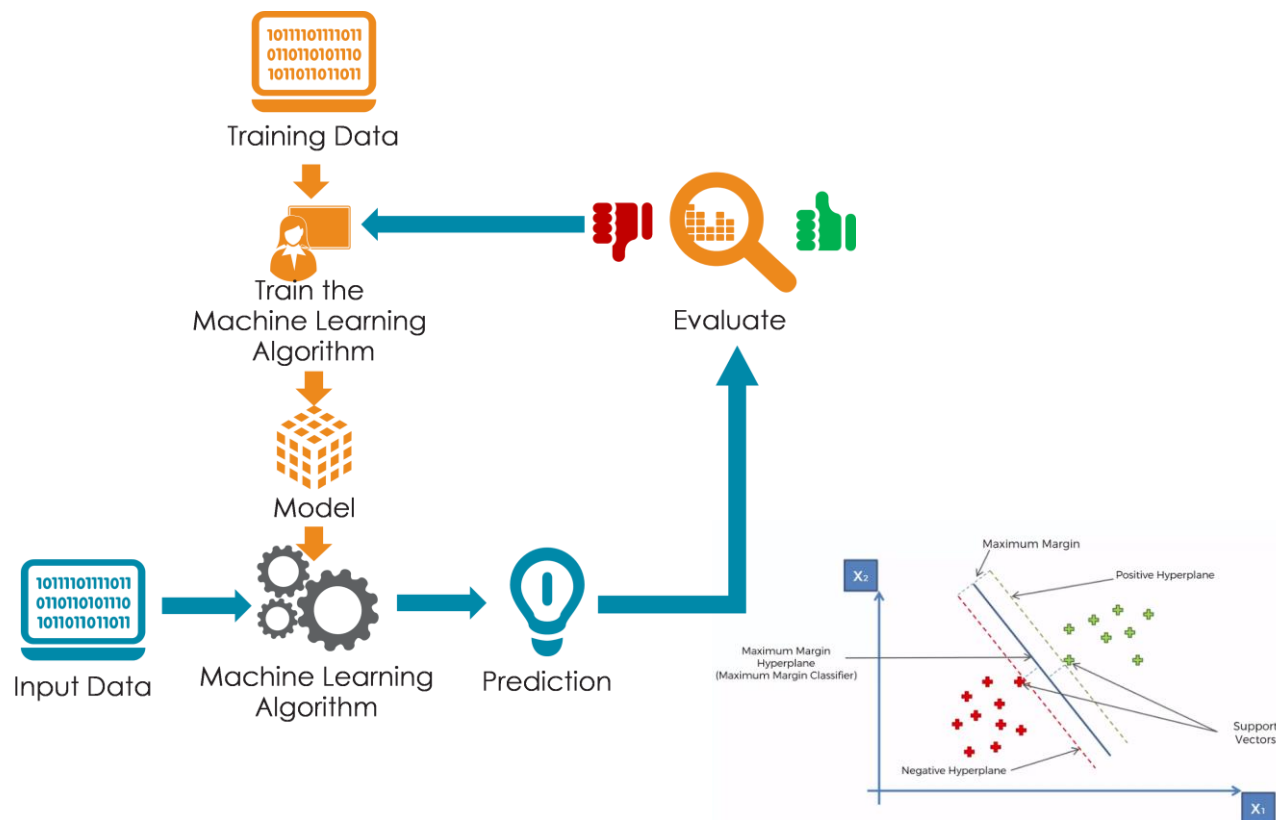


L'**intelligenza artificiale** (o **AI**) è una disciplina che studia se e in che modo si possano realizzare sistemi informatici intelligenti in grado di simulare la capacità e il comportamento del pensiero umano

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 predispongono l'abilità di apprendere qualcosa dai dati in maniera autonoma, senza istruzioni esplicite^{2,3}

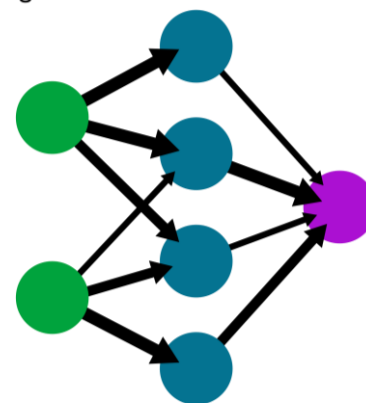
Il **Deep Learning**, la cui traduzione letterale significa apprendimento profondo, è una sottocategoria del Machine Learning (che letteralmente viene tradotto come apprendimento automatico) e indica quella branca dell'intelligenza artificiale che fa riferimento agli algoritmi ispirati alla struttura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

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

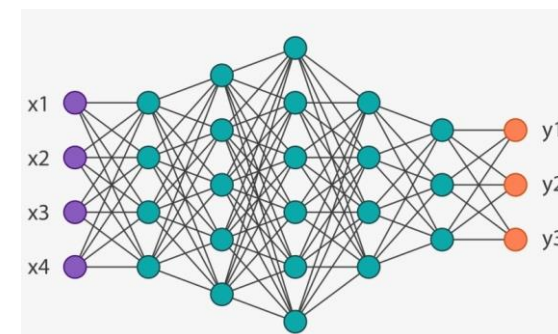


Rete semplice

strato di ingresso strato nascosto strato di uscita



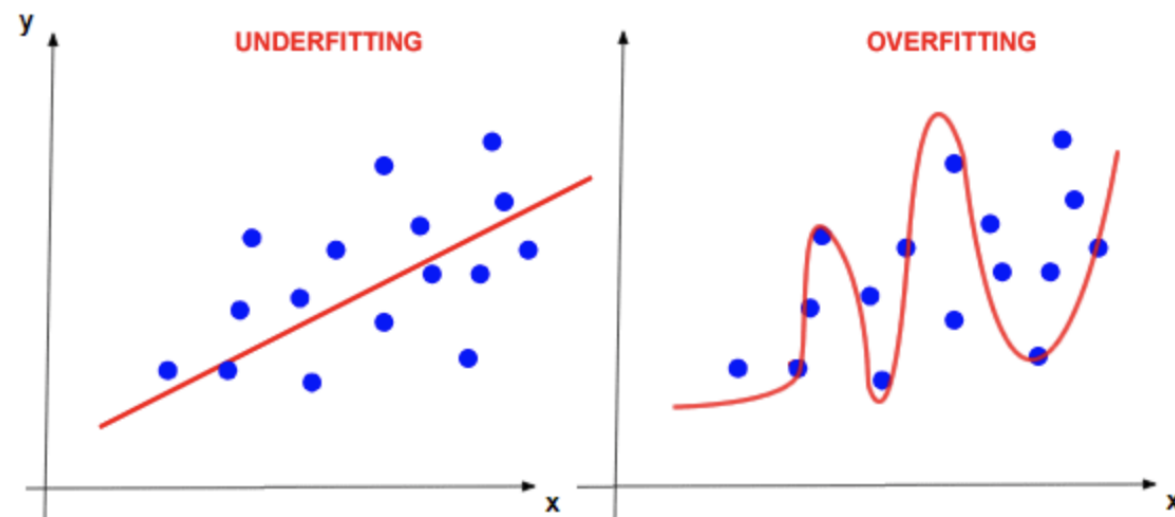
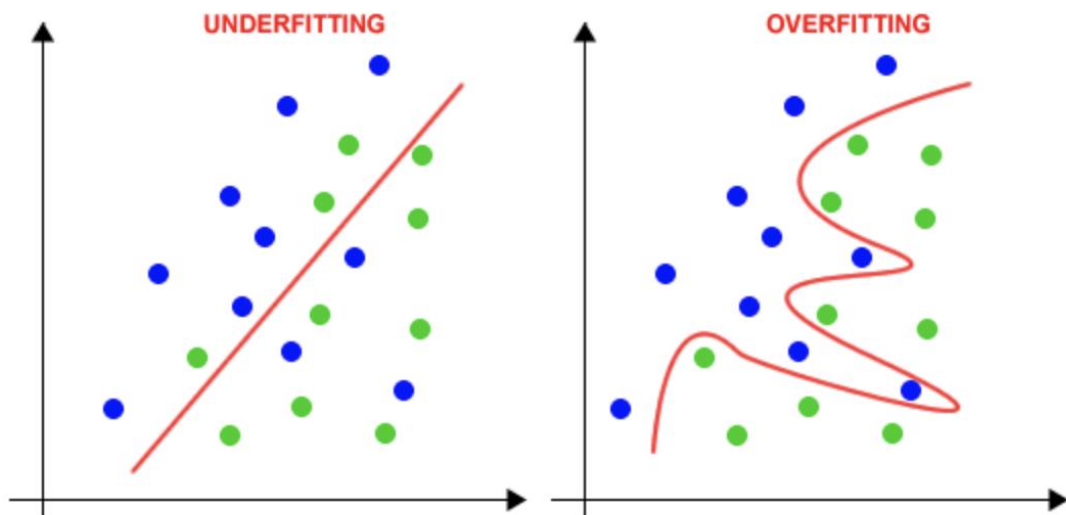
Rete complessa



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.

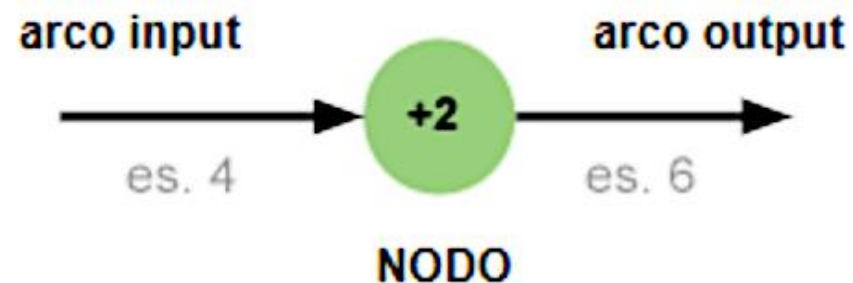
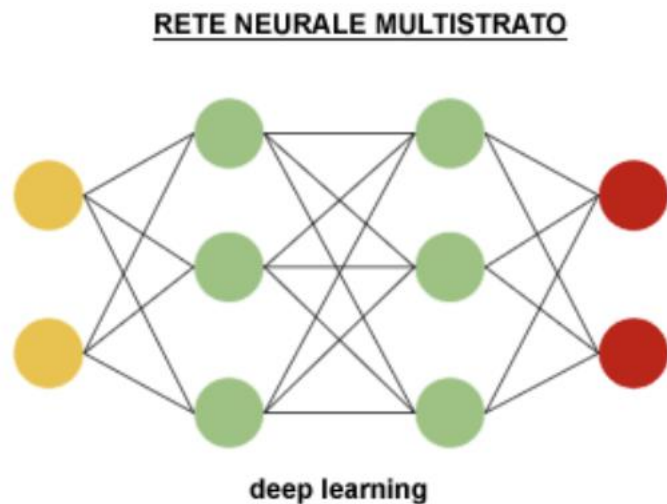
La classificazione può essere lineare o non lineare, a due o più dimensioni, ma la logica è sempre la stessa.

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



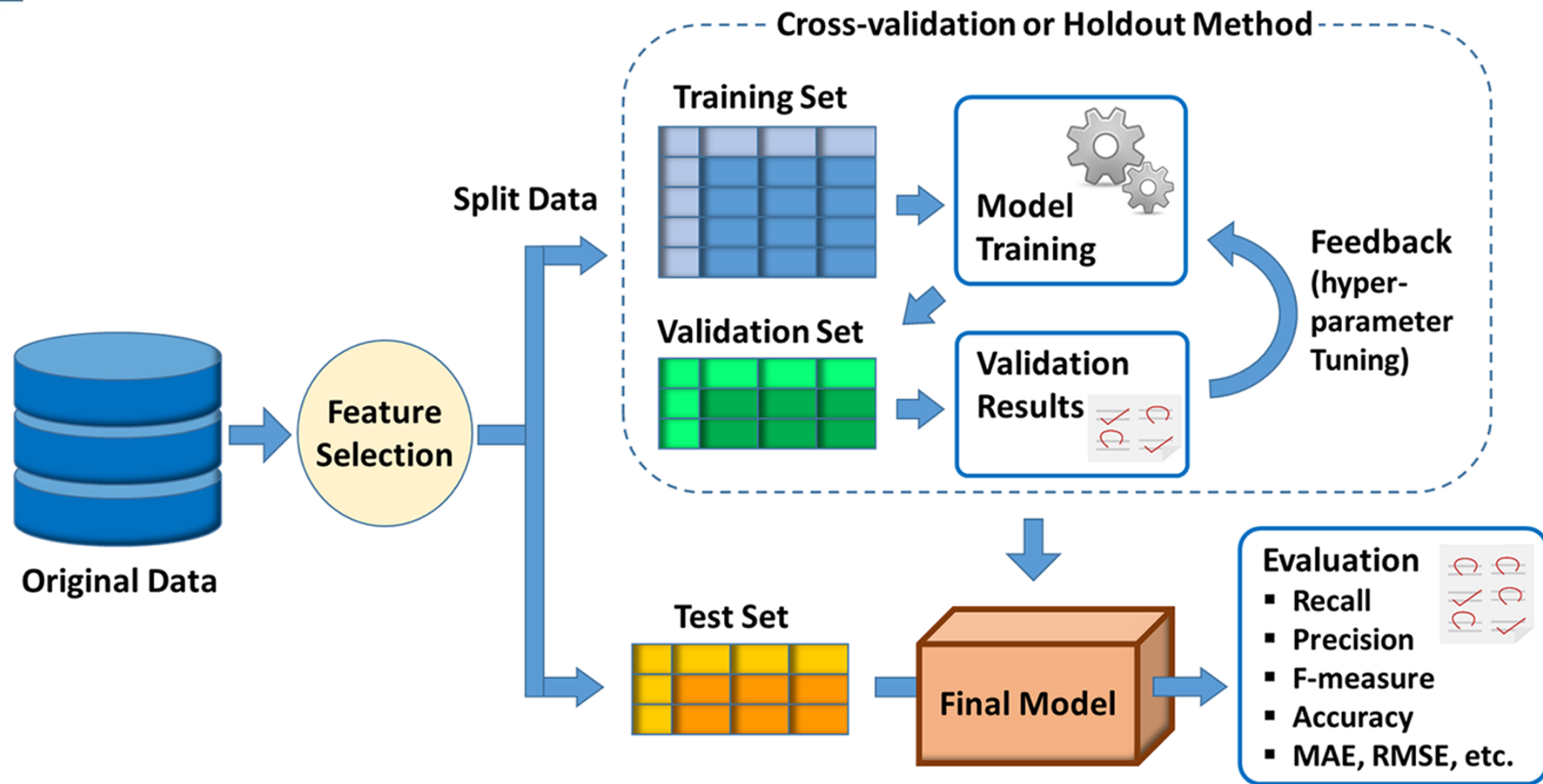
Come funziona la rete neurale

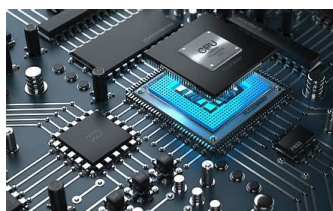
Gli archi trasmettono i dati dell'elaborazione nella neural network da un nodo a un altro. L'informazione si propaga da sinistra verso destra (forward).



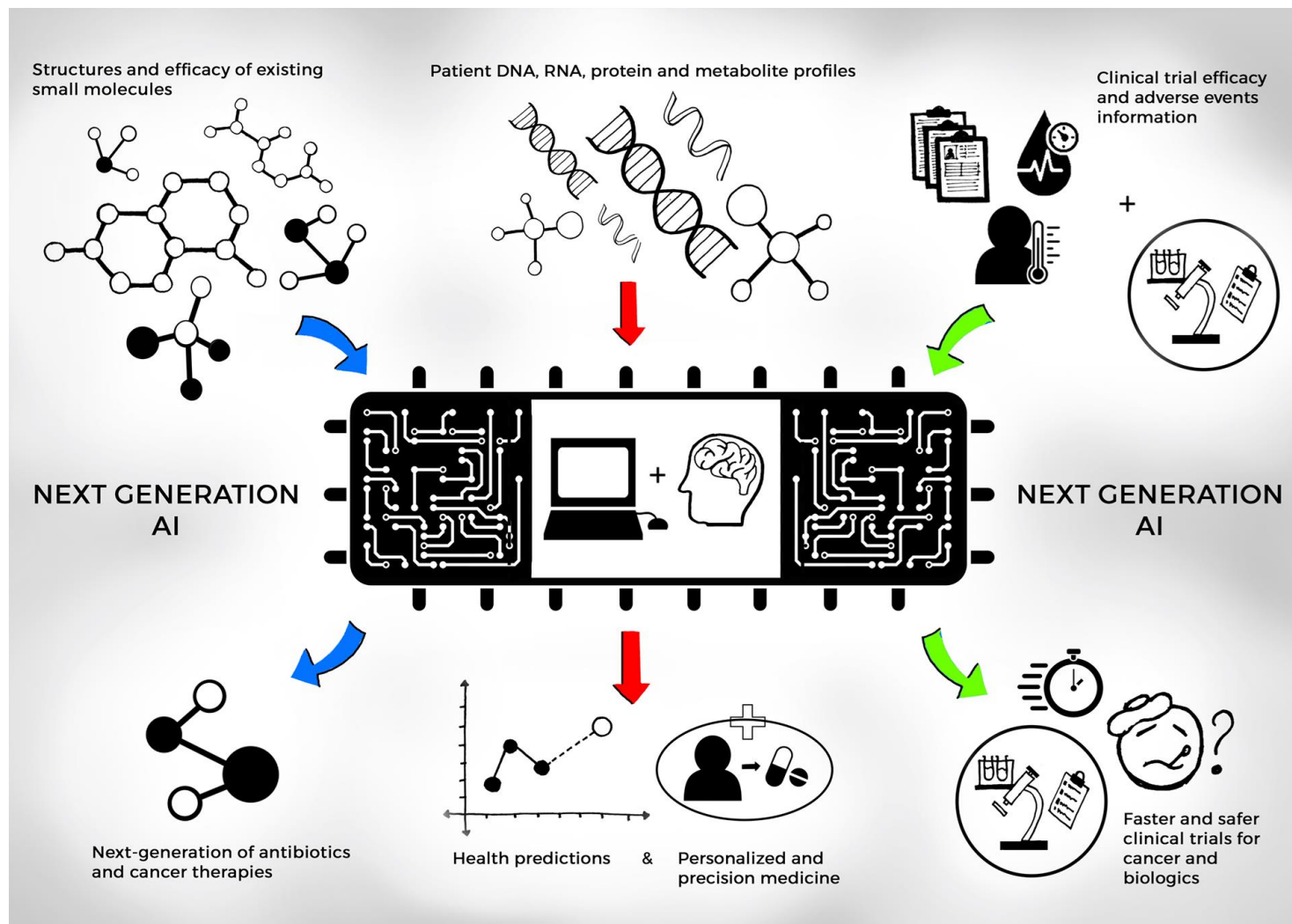
Gli archi in entrata sui nodi veicolano i valori di input (es. 4) mentre gli archi in uscita trasmettono il risultato dell'operazione in avanti (es. $4+2 = 6$).

Ogni nodo può essere collegato tramite più archi sia in entrata che in uscita.





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.



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

Mendola et al. BMC Bioinformatics 2022, 23(Suppl 8):310
https://doi.org/10.1186/s12859-020-03645-9

BMC Bioinformatics

RESEARCH Open Access

Convolutional architectures for virtual screening

Isabella Mendola^{1,2}, Salvatore Contino^{1,2,4}, Ugo Perricone^{2,4}, Edwards Ardizzone¹ and Roberto Pirrone¹
Palermo, Italy, 26-28 June 2019

* Correspondence: isabella.mendola@unipa.it
isabella.mendola@unipa.it
1 Dipartimento di Ingegneria, Università degli Studi di Palermo, Viale delle Scienze, Edificio 6, 90132 Palermo, Italy
2 Sezione Drug Design, Fondazione Ri.MED, 90132 Palermo, Italy

Abstract
Background: A Virtual Screening algorithm has to adapt to the different stages of drug discovery. 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 learned suitably to achieve good results as regards both enrichment factor and accuracy in different screening modes (88.52% 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 derived.
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 campaigns, 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 challenge.
One key aspect for ML approach possibility to access and mining large data sets in recent years, the best performed

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

BMC International Journal of Molecular Sciences **MDPI**

Article

EMBER—Embedding Multiple Molecular Fingerprints for Virtual Screening

Isabella Mendola^{1,2}, Salvatore Contino^{1,2,4}, Ugo Perricone^{2,4} and Roberto Pirrone¹

¹ Dipartimento di Ingegneria, Università degli Studi di Palermo, 90132 Palermo, Italy; isabella.mendola@unipa.it (I.M.); salvatore.contino@unipa.it (S.C.)
² Molecular Informatics Group, Fondazione Ri.MED, 90132 Palermo, Italy; ugo.perricone@fondazionerimed.com (U.P.)
⁴ Correspondence: salvatore.contino@fondazionerimed.com (S.C.); ugo.perricone@fondazionerimed.com (U.P.)
† These authors contributed equally to this work.

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 architectures that assesses ligands' bioactivity on a data set containing twenty protein kinases with similar binding sites to CDK1. The data set first 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 pre-tax expenditure above 2 billion USD and takes about 10–15 years [2–7]. 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 Mackenell [1] report a review that describes the drug discovery process and the corresponding computer-aided drug design methods. Computational methods do not guarantee a systematic assessment of molecular characteristics (e.g., bioactivity, ADMET properties, selectivity, and physicochemical properties) but generate lead molecules with favorable 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 [8]. 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 [9] and SMILES notation [7]. Recently, molecular graphs have been investigated along with neural embeddings, essentially a learned low-dimension vector

Manzi et al. BMC Bioinformatics 2020, 21(Suppl 8):10
https://doi.org/10.1186/s12859-020-03645-9

BMC Bioinformatics

RESEARCH Open Access

Convolutional architectures for virtual screening

Isabella Mandoli¹, Sabatino Costantini¹, Ligo Perricone², Edoardo Ardizzone¹ and Roberto Pirodda^{1*}

*Correspondence: rpirodda@uniba.it
¹Department of Chemistry, University of Bari, Via G. Cesare, 4, 70125 Bari, Italy
 Full list of author information is available at the end of the article

© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Background: Virtual screening (VS) is a routinely applied computer-aided drug discovery strategy. However, some issues remain unsolved that hinder its performance. One of the main challenges is the high number of molecules to be screened, which requires efficient and accurate methods. In this paper, we propose a novel convolutional architecture for VS, which is able to handle large datasets and to extract relevant features from molecular fingerprints. The proposed architecture outperforms state-of-the-art VS approaches, and can be used for drug discovery, molecular fingerprinting, and virtual screening.

Keywords: Deep learning, Drug design, Molecular fingerprinting, Reactivity prediction, Virtual screening

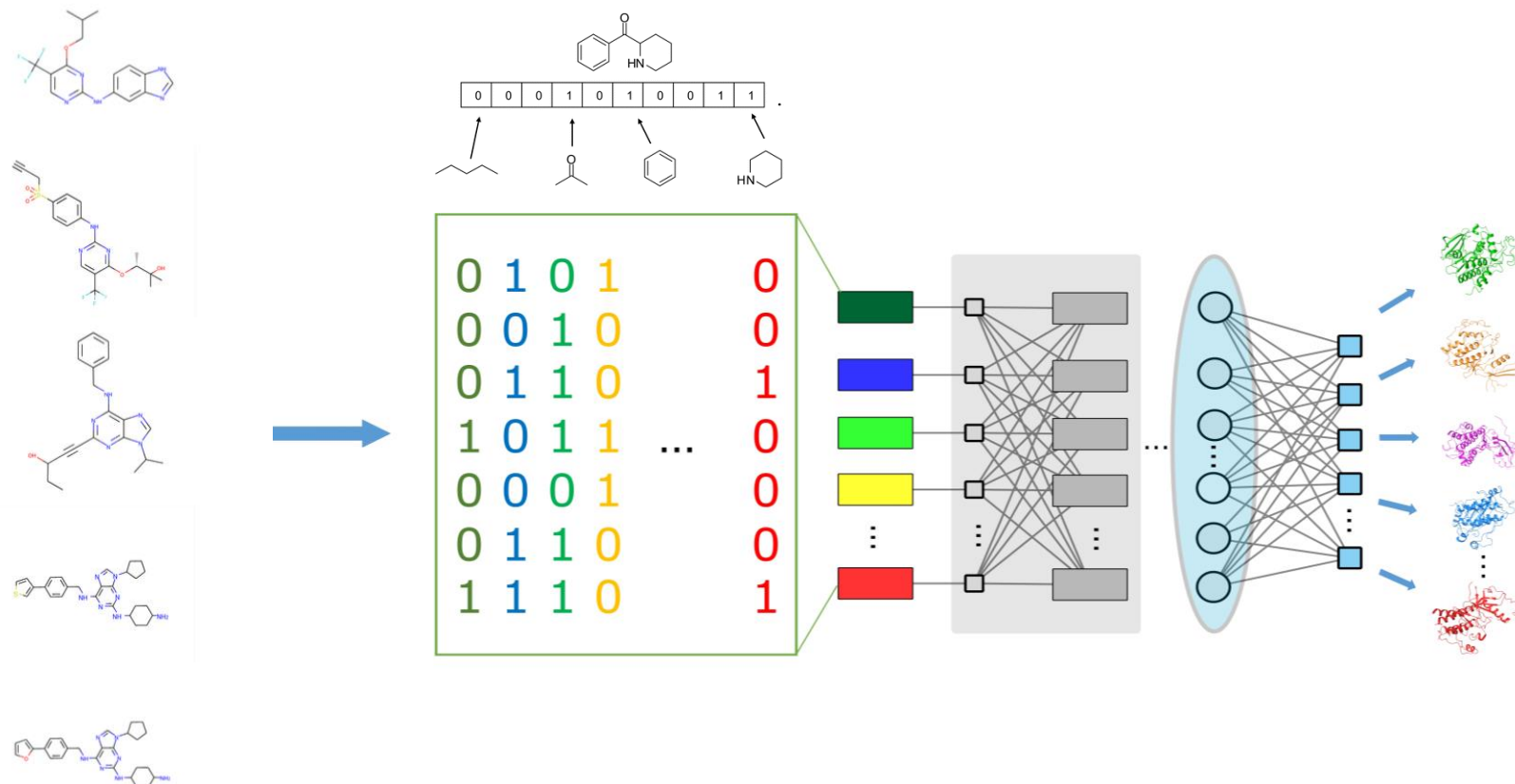
Background: Virtual screening (VS) is a routinely applied computer-aided drug discovery strategy. However, some issues remain unsolved that hinder its performance. One of the main challenges is the high number of molecules to be screened, which requires efficient and accurate methods. In this paper, we propose a novel convolutional architecture for VS, which is able to handle large datasets and to extract relevant features from molecular fingerprints. The proposed architecture outperforms state-of-the-art VS approaches, and can be used for drug discovery, molecular fingerprinting, and virtual screening.

Keywords: Deep learning, Drug design, Molecular fingerprinting, Reactivity prediction, Virtual screening

Introduction

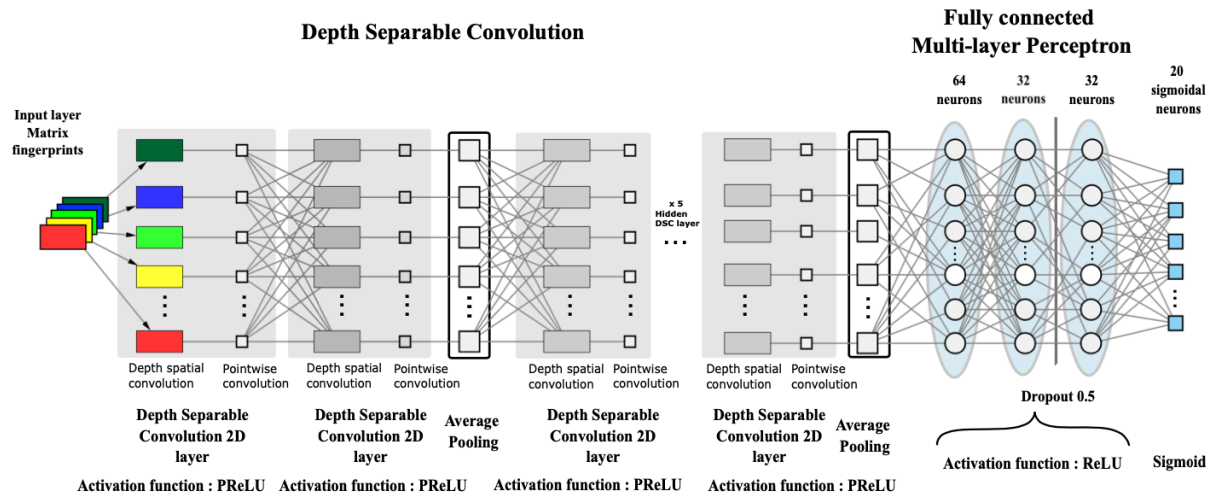
Drug discovery is a very long and expensive process that includes many stages such as target identification, target validation, virtual screening (VS), hit-to-lead generation, lead optimization, and so on [1]. Moreover, developing a new drug has a mean price expenditure above 2 billion USD and takes about 10-15 years [2]. Despite the huge investment of time and money, the experimental clinical approach remains the most conventional method during the drug discovery process. In about 10% of cases, the overall risk of failure is very high. This delay is supported by computational methods in almost every stage. Yu and MacKerell [3] report a review that describes the drug discovery process and the corresponding computer-aided drug design methods. Computational methods do not guarantee a systematic assessment of molecular characteristics (e.g., bioactivity, ADMET properties, solubility, and physicochemical properties) for general lead molecules with reasonable priorities in silico.

In particular, Virtual Screening (VS) is an often-discussed topic in Chemoinformatics and Molecular Chemistry and is widely applied in pharmaceutical research. VS consists of screening large small-molecule databases searched 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 more reduced list of molecules of candidate molecules. However, in the field of VS general screening importance is the last decade when Deep Learning (DL) became more popular [4]. In this field, the research data are very rich and complex and they require the use of sophisticated machine learning techniques by the network. The very first architectures used classical representations such as molecular fingerprints [5] and SMILES molecules [6]. Recently, molecular graphs have been investigated along with neural embeddings, especially a so-called low-dimensional vector



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	35,227
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



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

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

Nedra Mekki ^{1,2,*}, Claudia Coronello 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, 90129 Palermo, Italy; coronello@fondazioneirmed.com (C.C.); mendera@fondazioneirmed.com (M.D.R.)
 * Correspondence: nmekki@fondazioneirmed.com (N.M.); uperricone@fondazioneirmed.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: Mekki, N.; Coronello, 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: Daniel Park

Received: 2 July 2021
 Accepted: 11 July 2021
 Published: 19 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

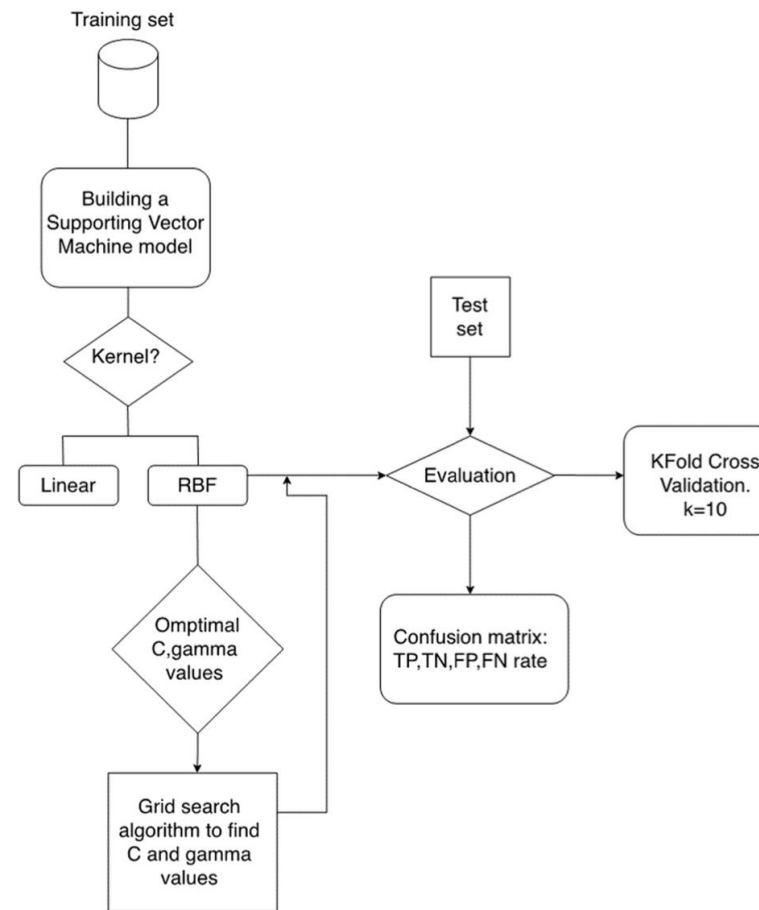
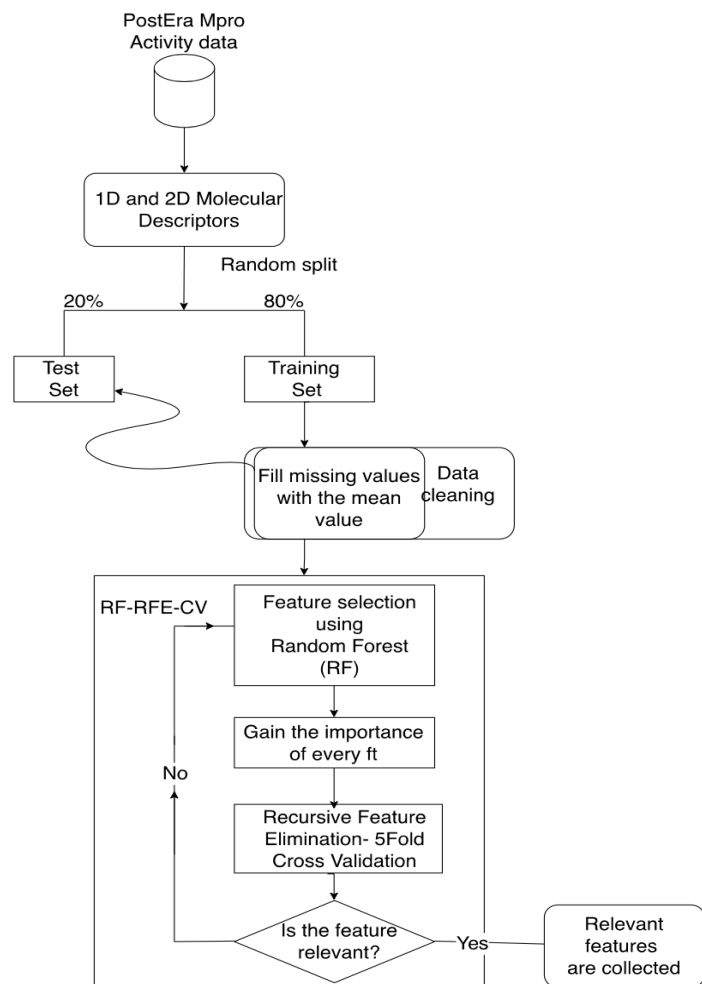


Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

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



International Journal of
Molecular Sciences

MDPI

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

Nedra Mekou^{1,2,*}, Claudia Ciommiello², Thierry Langer¹, Maria De Rosa^{2,3,4} and Ugo Perrone^{2,4,5}

¹ Department of Pharmaceutical Chemistry, University of Naples, 80131 Naples, Italy; thierri@unina.it (T.L.); mariade.rosa@unina.it (M.D.R.); ugo.perrone@unina.it (U.P.)
² Drug Discovery Unit, Fondazione Ri.MED, 80138 Palermo, Italy; ciommiello@fondazionerimed.com (C.C.); nedra@fondazionerimed.com (N.M.)
³ Correspondence: mekou@fondazionerimed.com (N.M.); ugo.perrone@fondazionerimed.com (U.P.)
⁴ These authors contributed equally to this work.

Abstract: In the last year, the COVID-19 pandemic has highly affected the identity of the world population, encouraging the scientific community towards a great effort in 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-viral drugs. In this scenario, the main protease (Mpro) represents an appealing target, being the enzyme responsible for the cleavage of polyproteins 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 chemical types as potentially active against the viral protease. The compounds labeled as active by SVM were next evaluated through consensus docking studies on two PDB structures and their binding mode were 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 revealed all the most important interactions known in the literature.

Check for updates

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

Received: 21 July 2021

Accepted: 17 July 2021

Published: 19 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction
The COVID-19 pandemic, also known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is afflicting the health and economies 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 represents a further public health concern, making it hard to reach a full immunization [1].

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/3C-like protease) [1]. In particular, SARS-CoV-2 Mpro has a crucial role in the viral replication process. Mpro is a cysteine protease responsible for the cleavage of polyproteins during the viral genome transcription, promoting the generation of non-structural proteins, which can assemble to form new infectious viruses. As shown in Figure 1, the Mpro catalytic site includes four subsites, namely S1, S2, S3 and S4, housing the binding site of protease inhibitors. [1]. Of special importance, the catalytic dyad is enclosed into the

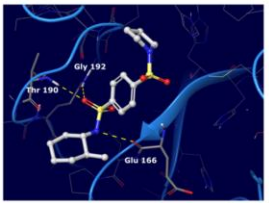

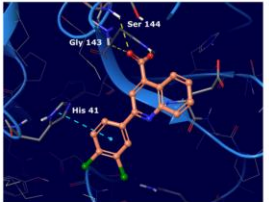
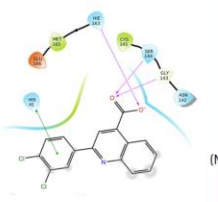
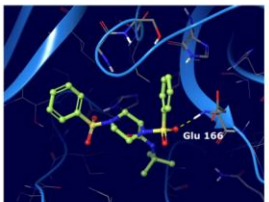
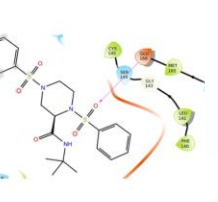
Abstract: In the last year, the COVID-19 pandemic has highly affected the identity of the world population, encouraging the scientific community towards a great effort in studying the infection molecular mechanisms. Several vaccine formulations are nowadays available and helping to reach immunity. Nevertheless, there is a growing interest towards the development of novel anti-viral drugs. In this scenario, the main protease (Mpro) represents an appealing target, being the enzyme responsible for the cleavage of poly-peptides 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 chemical types as potentially active against the viral protease. The compounds labeled as active by SVM were next evaluated through consensus docking studies on two PDB structures and their binding mode were 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

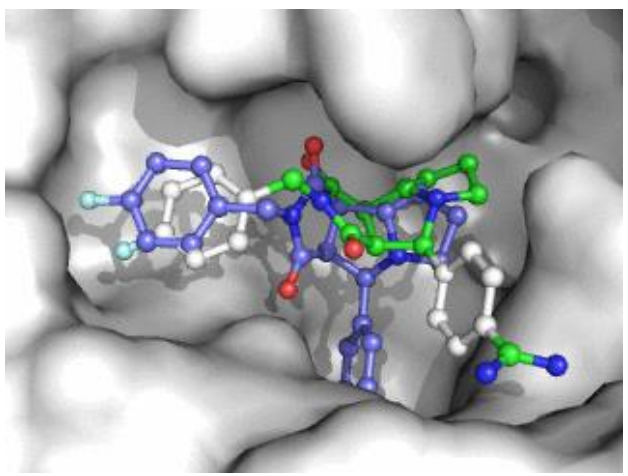
1. Introduction
The COVID-19 pandemic, also known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is afflicting the health and existence of billions of people worldwide. During the last few months, we are witnessing a race against time to vaccinate as many people as possible; however, the disparities in vaccine distribution between countries and the new emerging variants represents 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) [3,4]. In particular, SARS-CoV-2 Mpro holds a crucial role in the viral replication process. Mpro is a cysteine protease responsible for the cleavage of poly-peptides during the viral genome transcription, promoting the generation of non-structural proteins, which can assemble to form new infectious viruses. As shown in Figure 1, the Mpro catalytic site includes four subsites, namely S1, S2, S3 and S4, housing the binding site of protease inhibitors. [5]. Of special importance, the catalytic dyad is enclosed into the

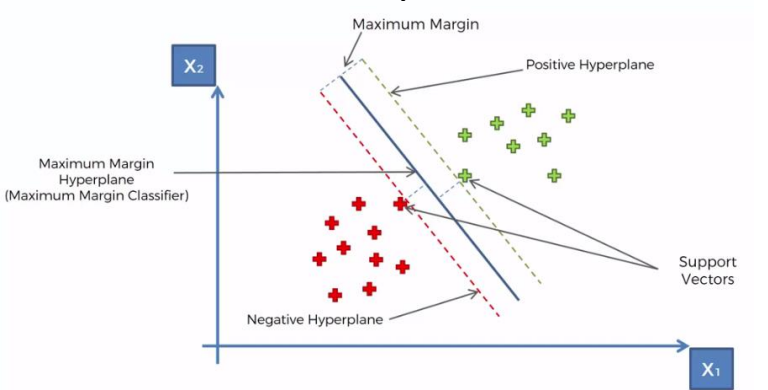
Ranking /Prioritisation

Cmpd	Docking Pose	Ligand Interaction
I		
II		
III		

Molecular Docking



Modello ML per la classificazione



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 multiparametrici, molti test in vivo sono poco affidabili perché il «sistema» animale è troppo differente da quello umano

Soluzione:

- ✓ Applicare multiclassificatori che considerino contemporaneamente 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

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}

^{*}Johns Hopkins University, Bloomberg School of Public Health, Center for Alternatives to Animal Testing (CAAT), Baltimore, Maryland 21205; [†]ToxTrack, Baltimore, Maryland 21209; [‡]UL Product Supply Chain Intelligence, Underwriters Laboratories (UL), Northbrook, Illinois 60062; and [§]University of Konstanz, CAAT-Europe, Konstanz 78464, Germany

nature

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

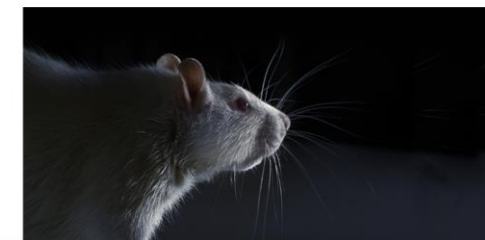
[nature](#) > [news](#) > [article](#)

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](#)

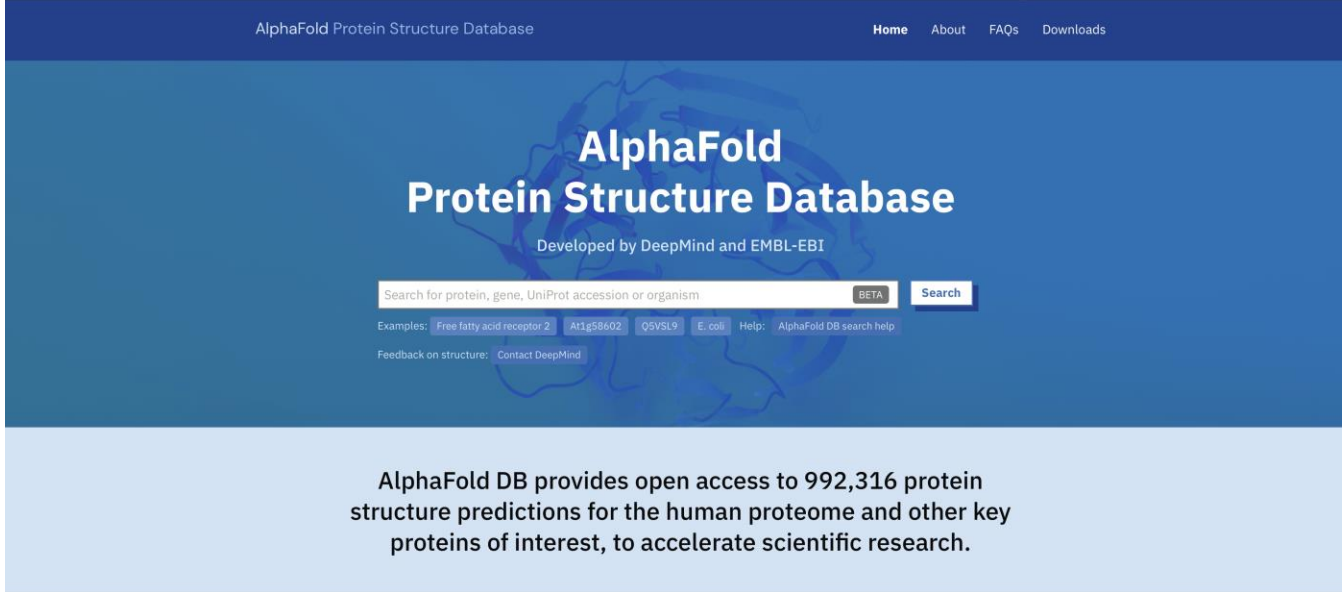


Problema:

- 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



AlphaFold Protein Structure Database

Home About FAQs Downloads

AlphaFold Protein Structure Database

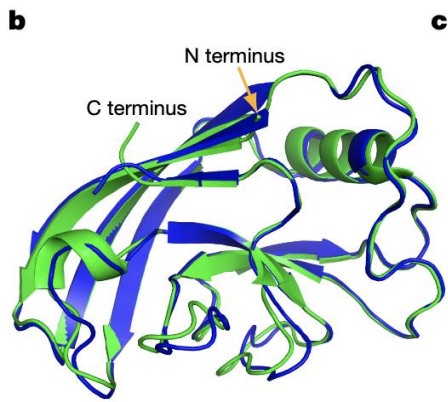
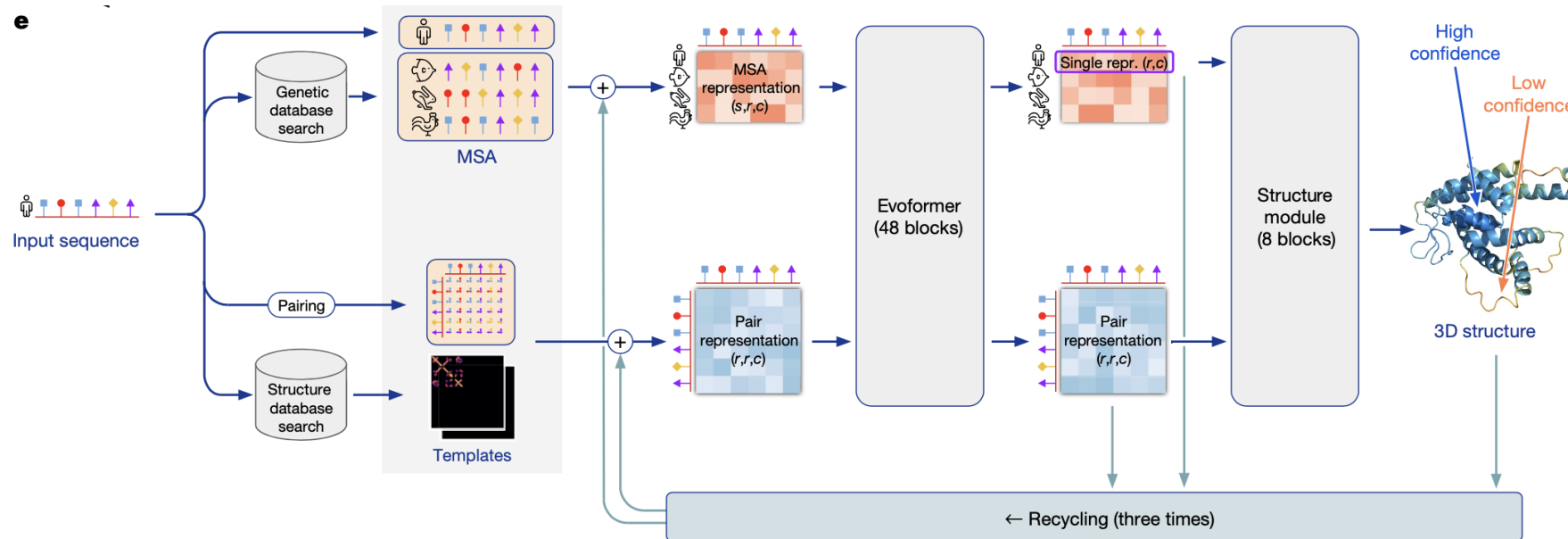
Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism BETA

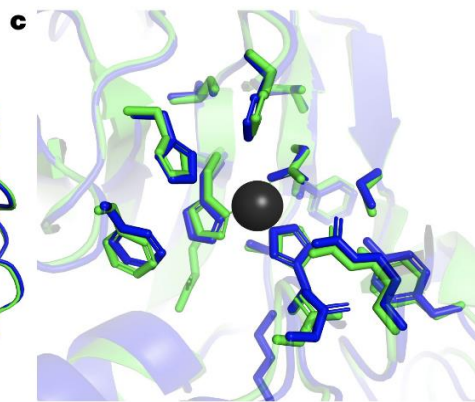
Examples: [Free fatty acid receptor 2](#) [A1g58602](#) [Q5VSL9](#) [E. coli](#) [Help: AlphaFold DB search help](#)

Feedback on structure: [Contact DeepMind](#)

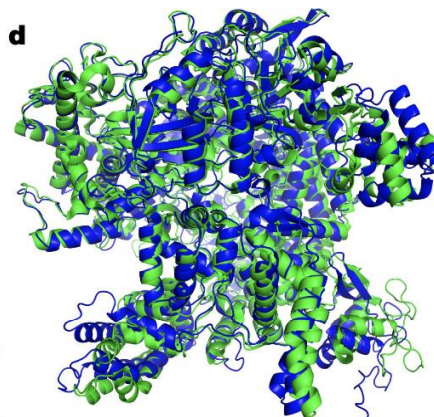
AlphaFold DB provides open access to 992,316 protein structure predictions for the human proteome and other key proteins of interest, to accelerate scientific research.



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

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.

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

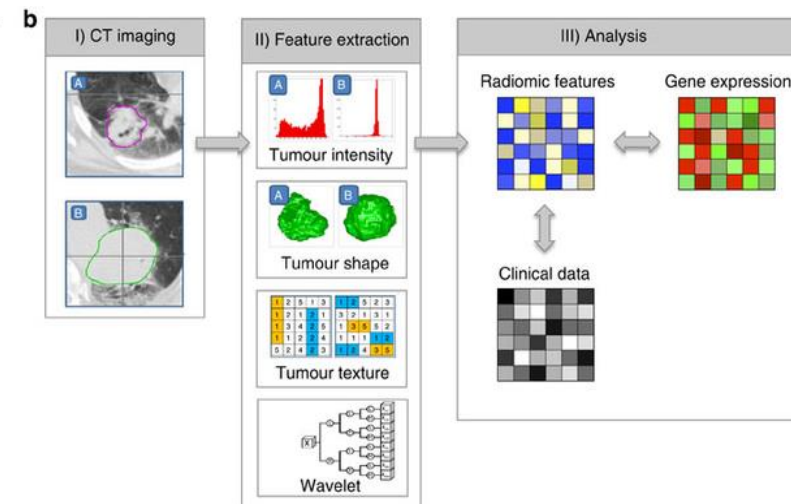
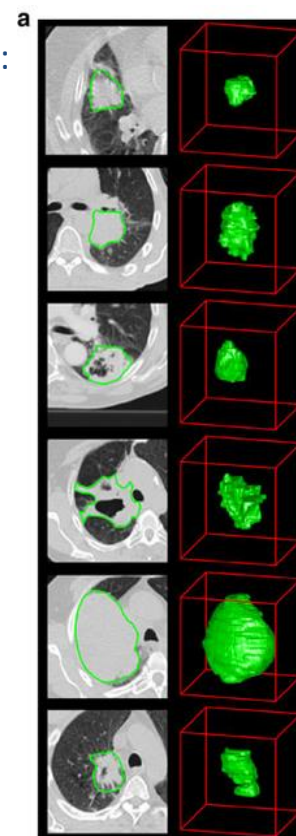
Identifica/crea aree (immagini 2D) o volumi di interesse (immagini 3D). Può essere eseguito manualmente, semiautomatico o completamente automatizzato utilizzando l'intelligenza artificiale (usato in big data)

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).



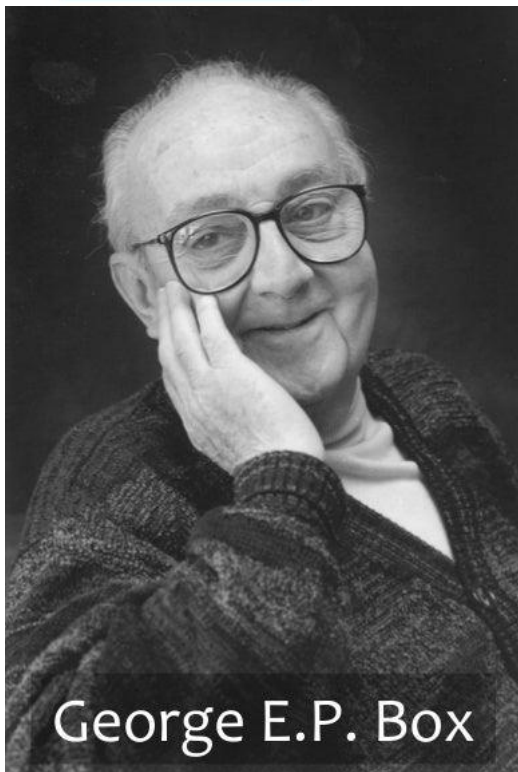
Case courtesy of Dr Mohammed Sultan, Radiopaedia.org, rID: 65398

Punti di forza:

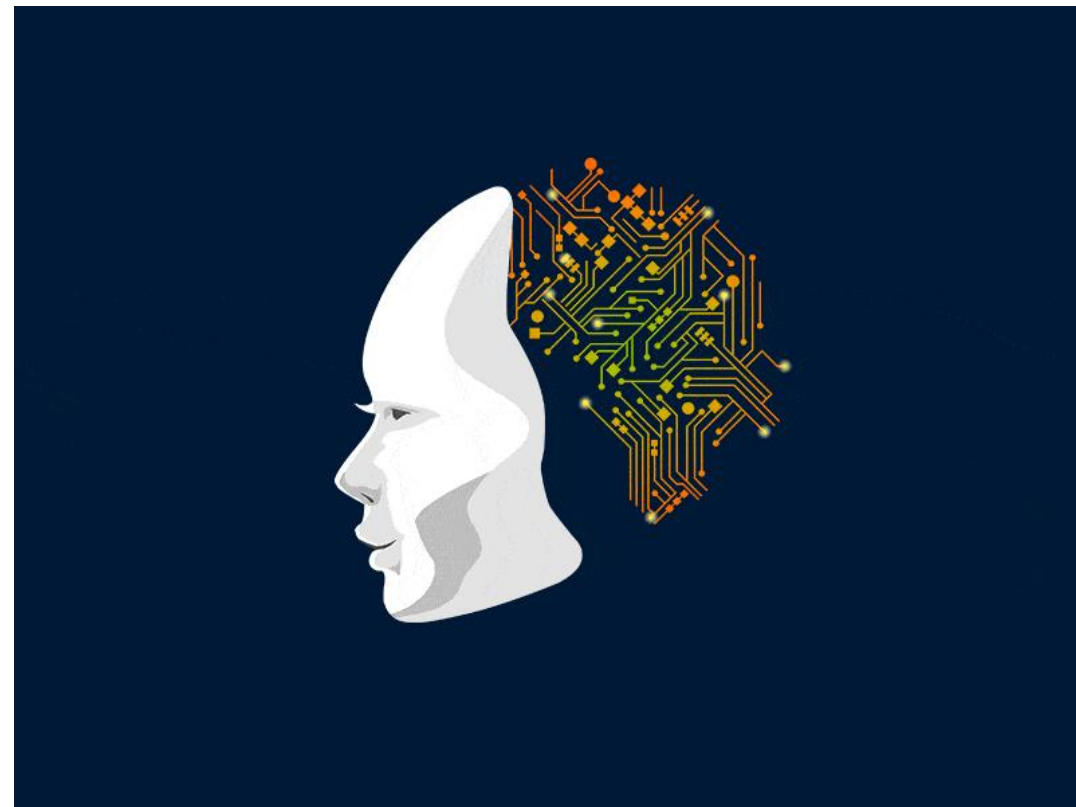
- Predire per molecole simili a quelle studiate la possibile tossicità e anticipare fallimenti nelle campagne di drug discovery
- Lavorare con dati umani opportunamente modellati e possibilità di trasferire i modelli da un comparto ad un altro
- Capacità di predire in largo anticipo situazioni che potrebbero 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



Essentially,
all models
are wrong,
but
some are useful



GRAZIE PER LA VOSTRA ATTENZIONE