FREE CIRCLE

Linux Day 2023

Giornata nazionale a favore della diffusione del software libero e del sistema operativo GNU/Linux

LINUX DAY 2023

Applicazioni di Intelligenza artificiale alla scoperta di nuove terapie

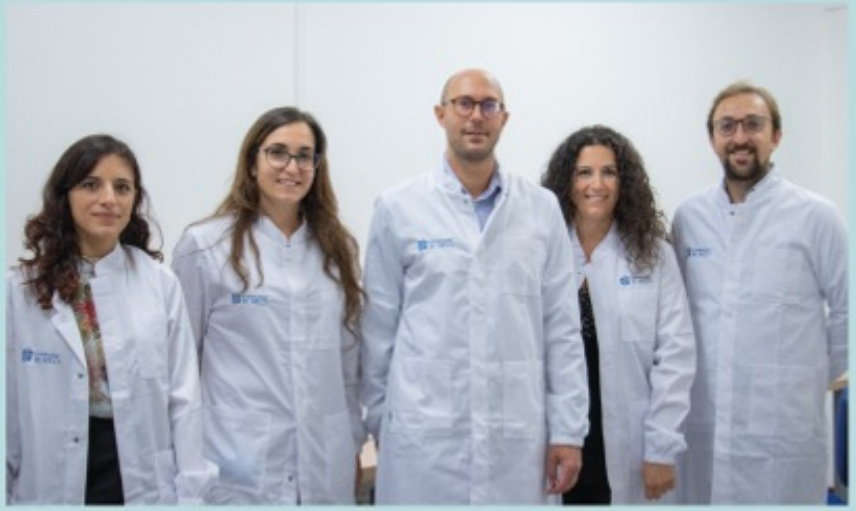
Dr. Ugo Perricone

Group Leader in Molecular Informatics @ FONDAZIONE Ri.MED

Chi Sono

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

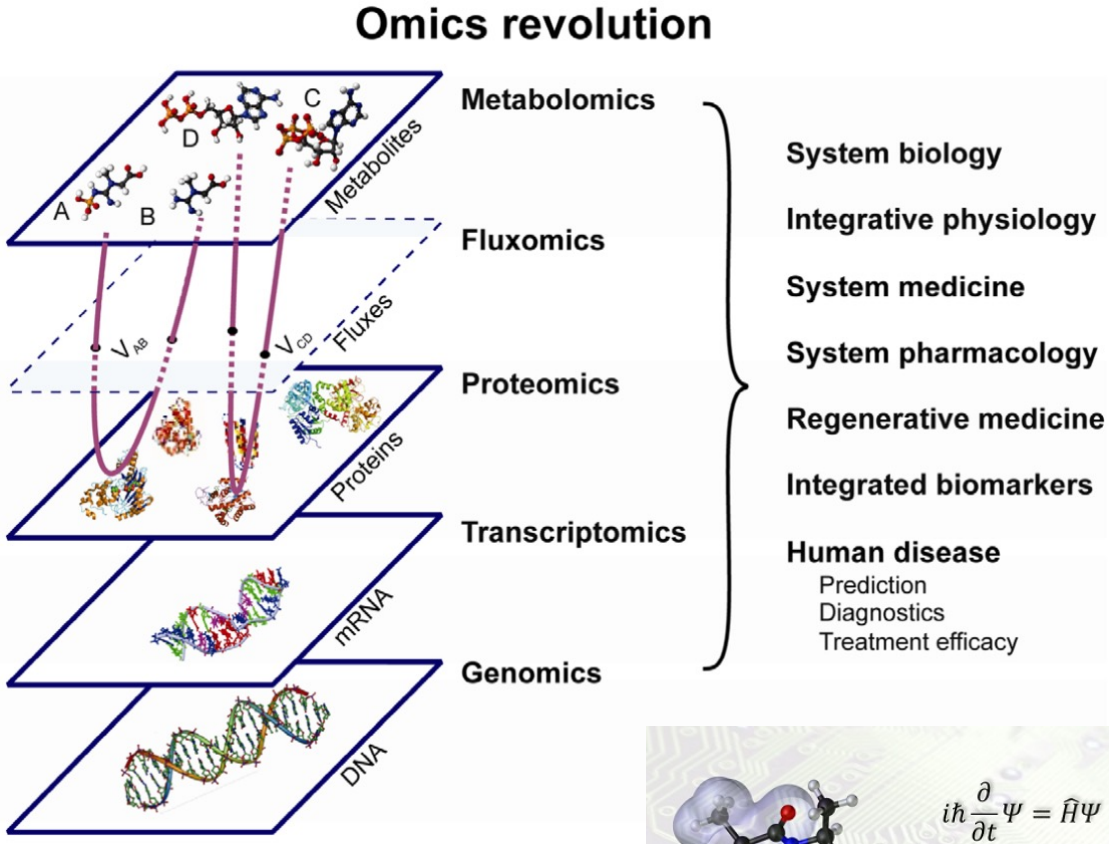
Molecular Informatics

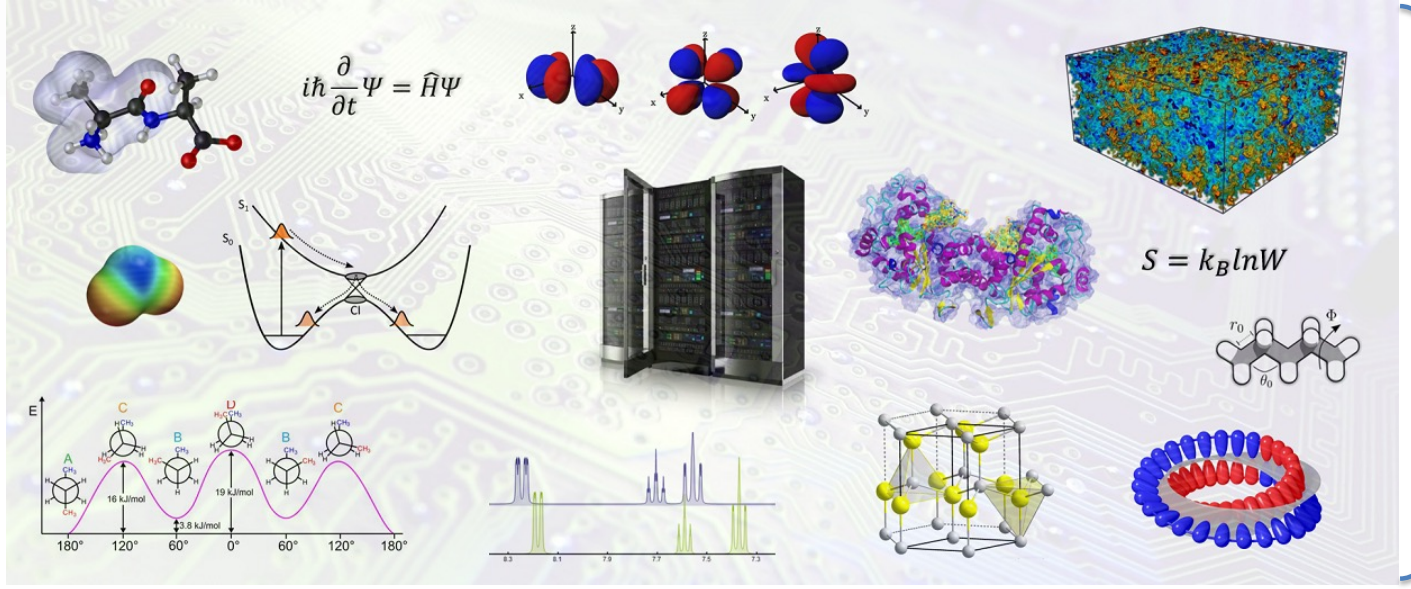


Ugo Perricone
Group Leader
in Molecular Informatics

4 FELLOWSHIPS
4 PhD STUDENTS

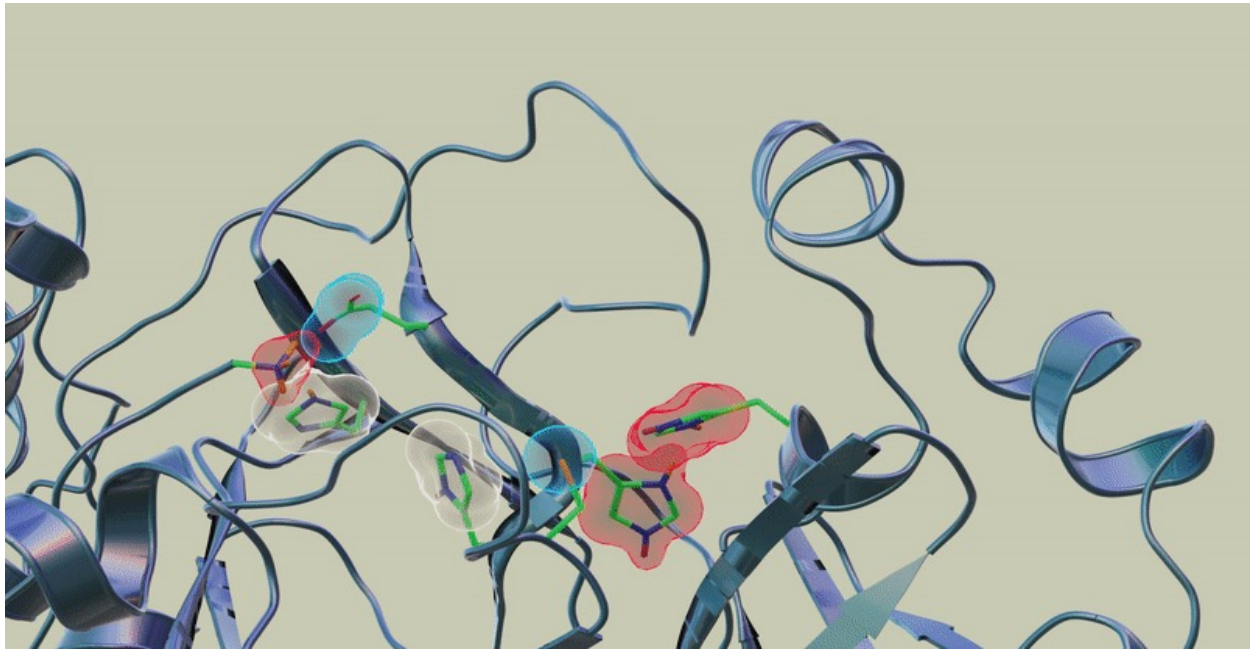
L'interdisciplinarietà di un problema scientifico oggi



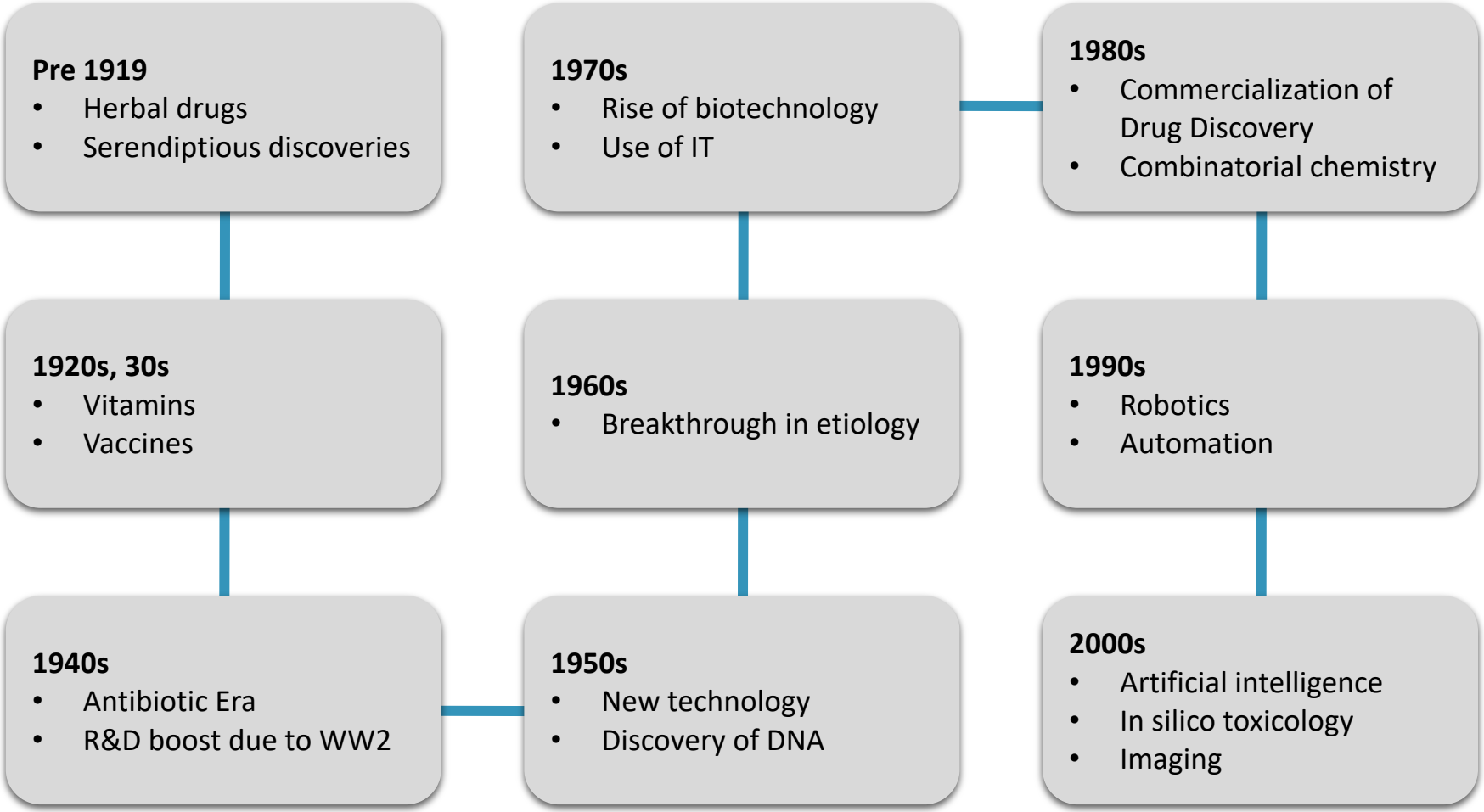


- ✓ Organic Chemistry
- ✓ Inorganic Chemistry
- ✓ Medicinal Chemistry
- ✓ Physical Chemistry
- ✓ Physics (QM & MM)
- ✓ Math
- ✓ Informatics
- ✓ Structural Biology

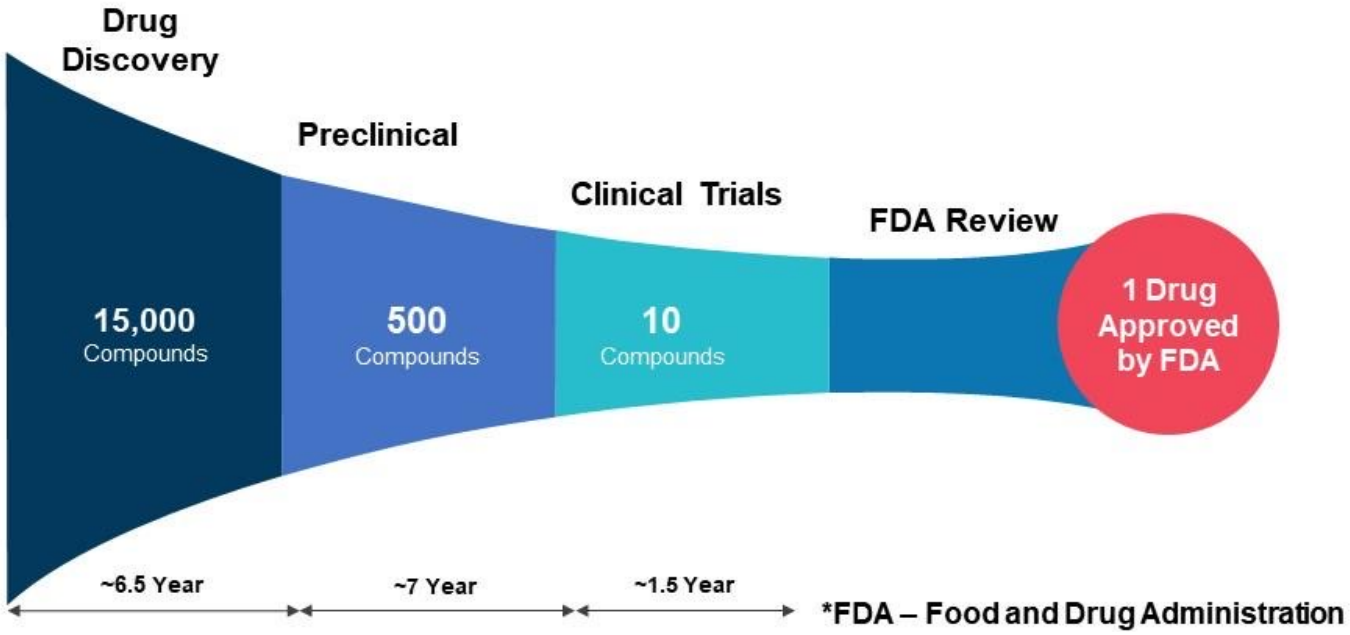
Un esempio di Big Data nelle scienze… La scoperta di un nuovo farmaco



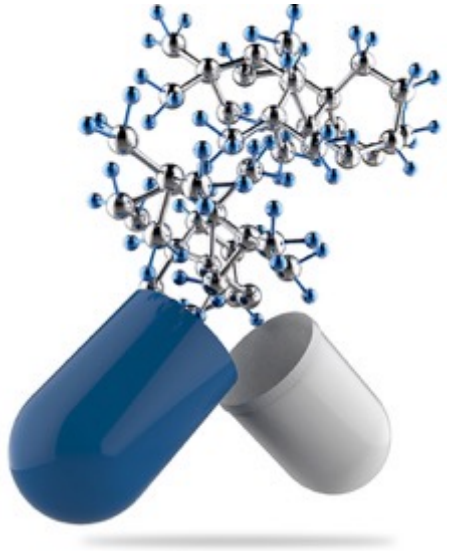
Processo di scoperta di nuovi farmaci



Drug Discovery & Development - Timeline

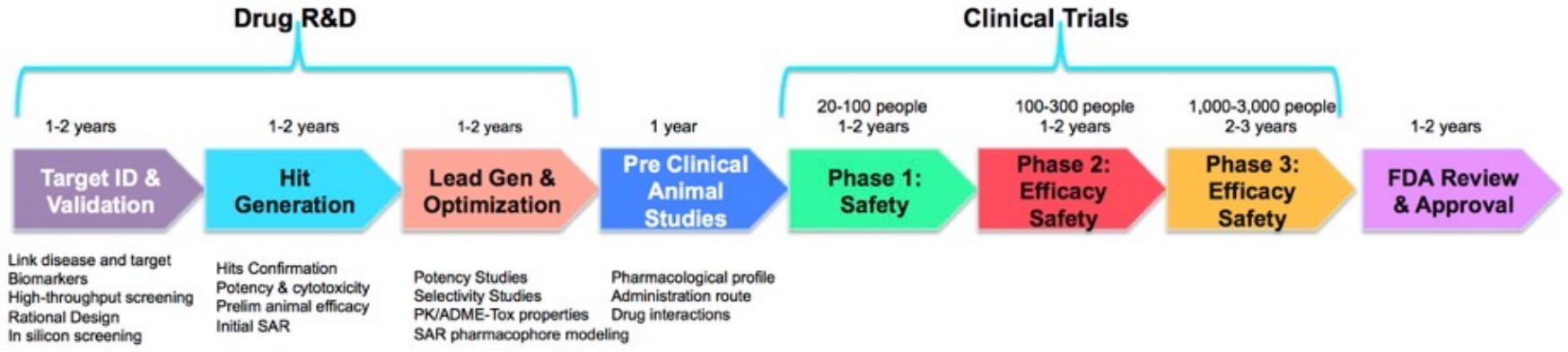


L’informatica Molecolare nella progettazione farmaceutica (Drug Design)

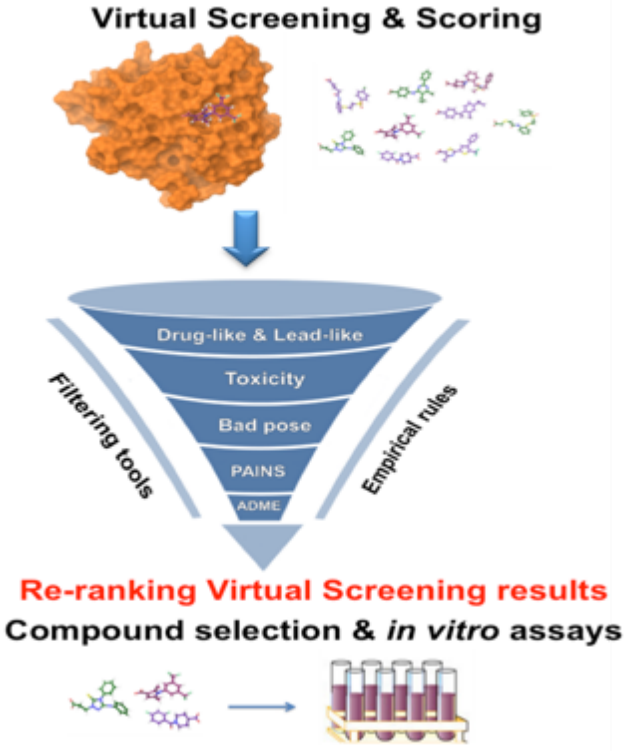


5 Ottobre 1981

“Next Industrial Revolution: Designing Drugs by Computer at Merck”



SCOPO PRINCIPALE: Trovare molecole sicure (non tossiche) e selettive per una patologia e per un target proteico (recettore) → SCREENING VIRTUALE



(1) Filtrare grandi librerie di molecole e trovare quelle che meglio colpiscono il bersaglio

(2) Approfondire il modo in cui il farmaco interagisce con il recettore

(3) Ottimizzare le molecole attive trovate e confermate dai saggi biologici per renderle ancora più sicure ed efficaci

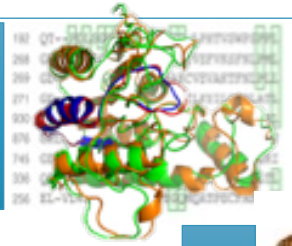
(4) Guidare nella progettazione razionale di nuove molecole

(5) Supportare nella terapia personalizzata

Molecular Modeling (Un esempio)

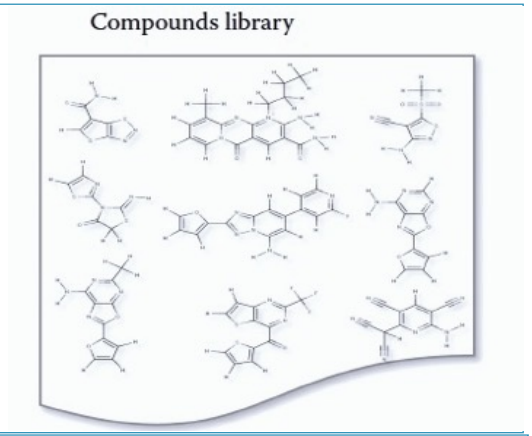
Input structure

HOMOLOGY MODELING



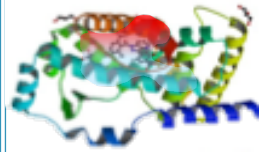
NMR

X-RAY

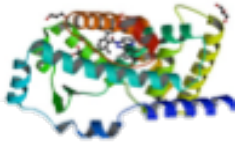


Preparation

- Hydrogen atoms assignement
- Histidine tautomers selection



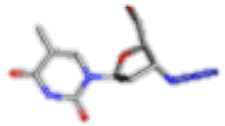
Binding site identification



Binding hypotesis generation

REOS
Rule of 5
PAINS

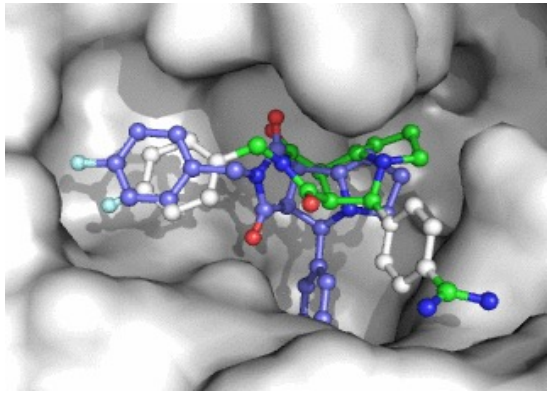
Filters application

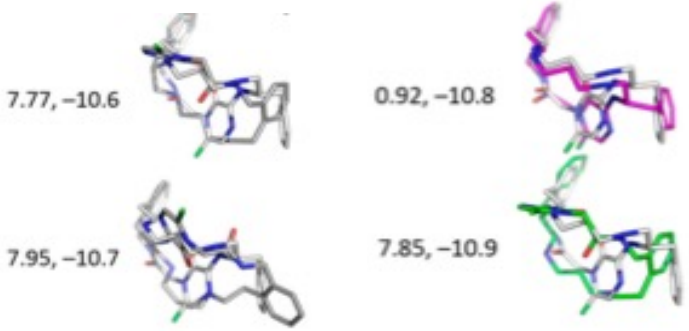


3D conformation

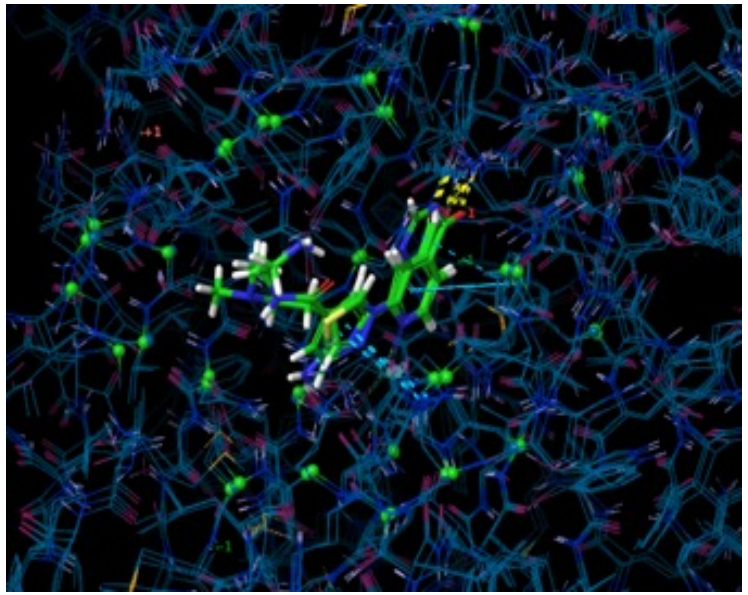
- Protonation states assignement
- Tautomers selection

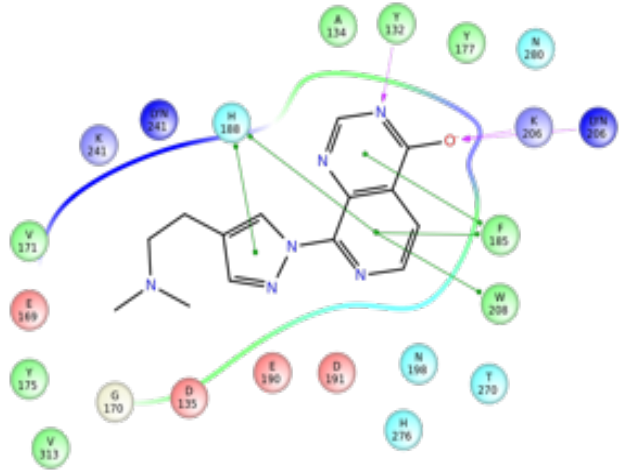
Docking & Scoring

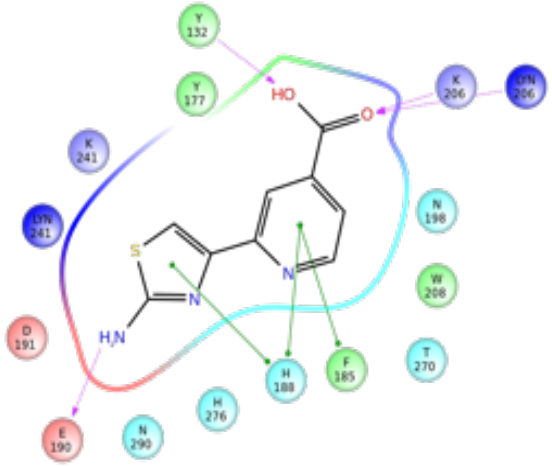




Structures Analysis



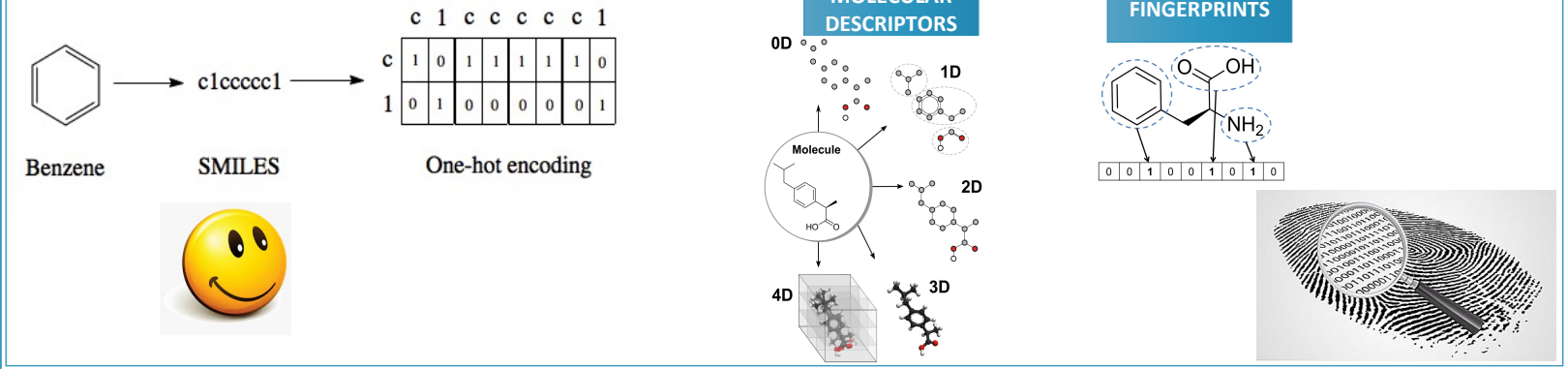




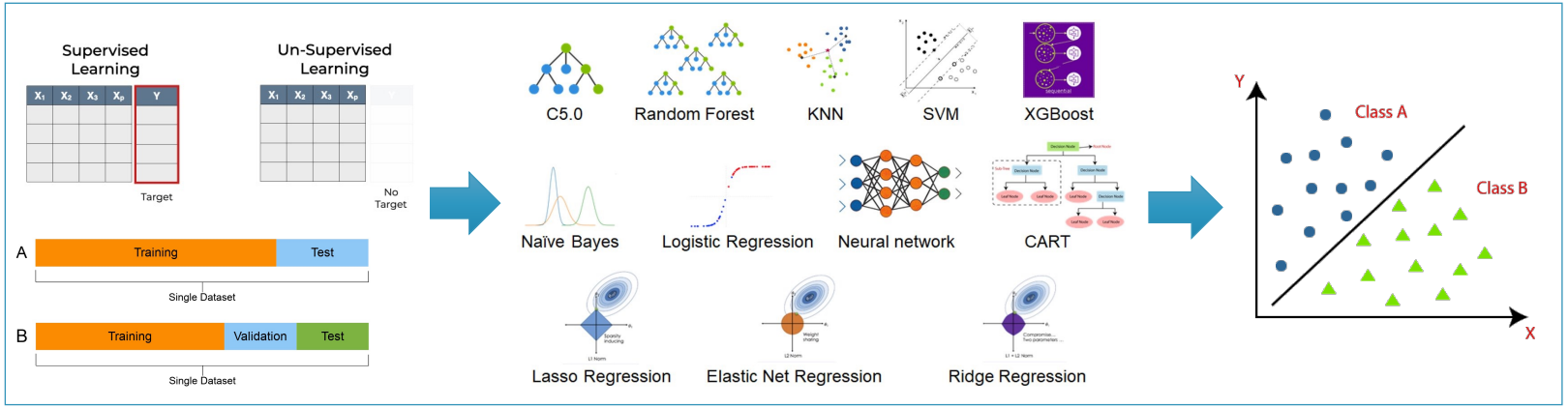
PDB Nes	Residues										
	LYS206:A	LYS241:A	TYR132:A	ASP135:A	PHE185:A	GLU190:A	CYS11:C	SER196:A	SER288:A	ASN198:A	HIS188:A
5F37	√	√	√	X	√	X	X	X	X	X	√
3PDQ	√	√	√	√	√	√	X	X	X	X	√
3U4S	√	X	√	X	X	X	√	√	√	√	X
4URA	X	√	√	X	√	X	X	X	X	X	X
4V2V	√	X	√	X	X	X	X	√	√	√	X
5A7N	√	X	√	X	√	√	X	X	X	X	√
5A7O	√	√	√	X	√	√	X	X	X	X	√
5A7P	√	X	√	X	√	√	X	X	X	X	X
5A7Q	√	X	√	X	√	√	X	X	X	√	√
5F2S	√	X	√	X	√	√	X	X	X	X	√
5F3C	√	X	√	√	√	X	X	X	X	X	√
5F3E	√	X	√	X	√	X	X	X	X	X	√
5F3I	√	X	√	X	√	X	X	X	X	X	√
5F32	√	X	√	X	√	√	X	X	X	X	√

Chemoinformatics (Un esempio)

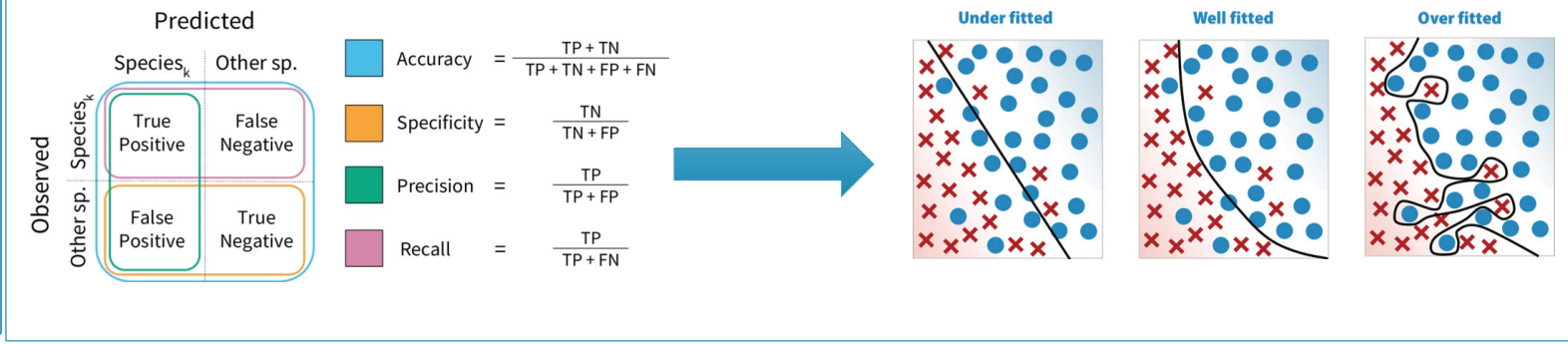
Input structure



Training



Prediction



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


www.nature.com/scientificreports

scientific reports

OPEN

KUALA: a machine learning-driven framework for kinase inhibitors repositioning

Giada De Simone, Davide Stefano Sardina, Maria Rita Gulotta & Ugo Perricone

The family of protein kinases comprises more than 500 genes involved in numerous functions. Hence, their physiological dysfunction has paved the way toward drug discovery for cancer, cardiovascular, and inflammatory diseases. As a matter of fact, Kinase binding sites high similarity has a double role. On the one hand it is a critical issue for selectivity, on the other hand, according to poly-pharmacology, a synergistic controlled effect on more than one target could be of great pharmacological interest. Another important aspect of binding similarity is the possibility of exploit it for repositioning of drugs on targets of the same family. In this study, we propose our approach called Kinase drUgs mAchine Learning frAmework (KUALA) to automatically identify kinase active ligands by using specific sets of molecular descriptors and provide a multi-target priority score and a repurposing threshold to suggest the best repurposable and non-repurposable molecules. The comprehensive list of all kinase-ligand pairs and their scores can be found at https://github.com/molinfrimed/multi-kinases.

Scientific Reports | (2022) 12:7877 | https://doi.org/10.1038/s41598-022-22326-8 nature portfolio


Mendolia et al. BMC Bioinformatics 2020, 21(Suppl 8):310
https://doi.org/10.1186/s12859-020-03645-9

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 challenge involving different aspects of Chemistry and Biology, which can be faced using Machine Learning (ML).

One key aspect for ML approaches gaining success in property prediction, is the possibility to access and mining large data sets that contain heterogeneous information. Until recent years, the best performing ML techniques were "shallow" ones [1] that is

BMC

International Journal of Molecular Sciences

Article

EMBER—Embedding Multi
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

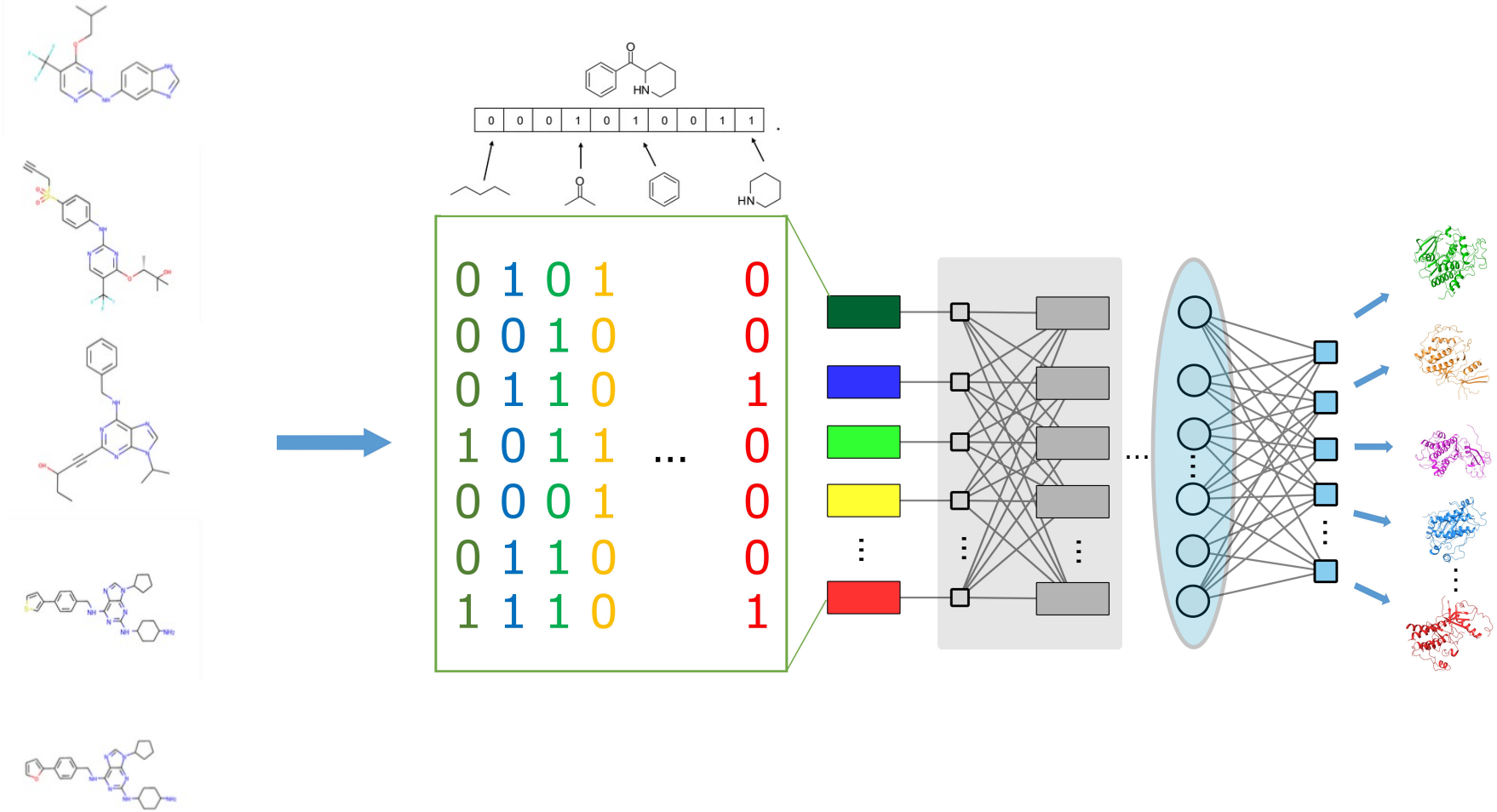
1. Introduction

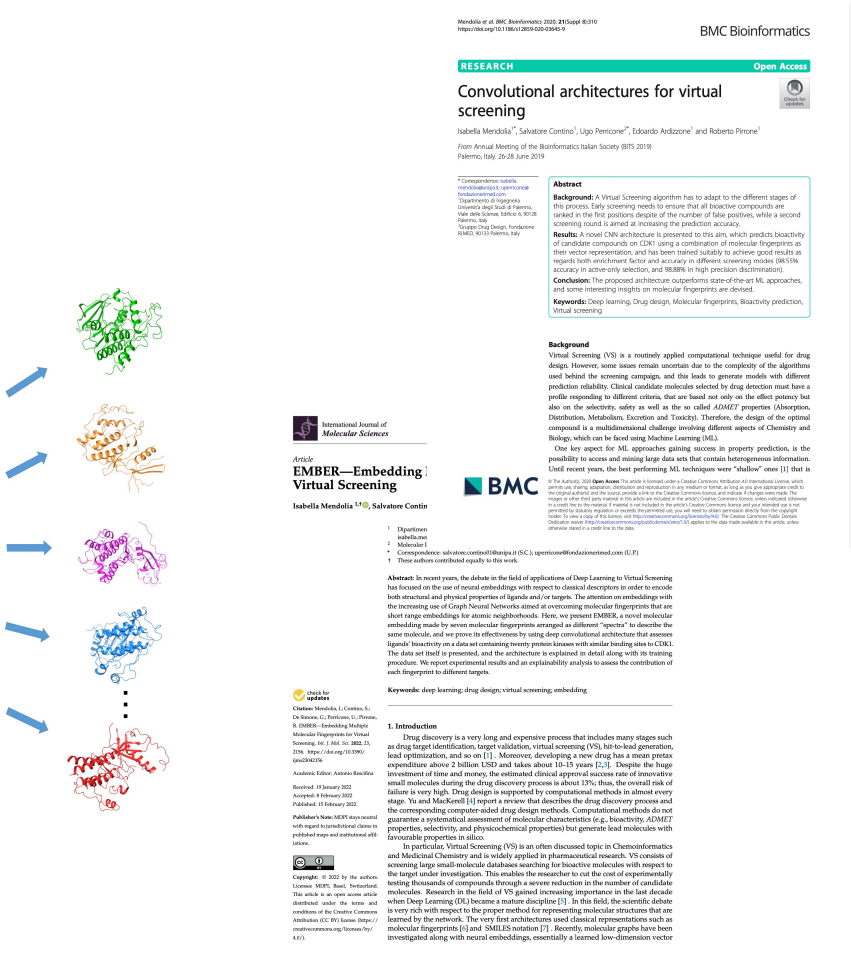
Drug discovery is a very long and expensive process that includes many stages such as drug target identification, target validation, virtual screening (VS), hit-to-lead generation, lead optimization, and so on [1]. Moreover, developing a new drug has a mean pretax expenditure above 2 billion USD and 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

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 challenge involving different aspects of Chemistry and Biology, which can be faced using Machine Learning (ML).

One key aspect for ML approaches gaining success in property prediction, is the possibility to access and mining large data sets that contain heterogeneous information. Until recent years, the best performing ML techniques were "shallow" ones [1] that is

International Journal of Molecular Sciences

Article

EMBER—Embedding Virtual Screening

Isabella Mendolia, Salvatore Contin

BMC

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.

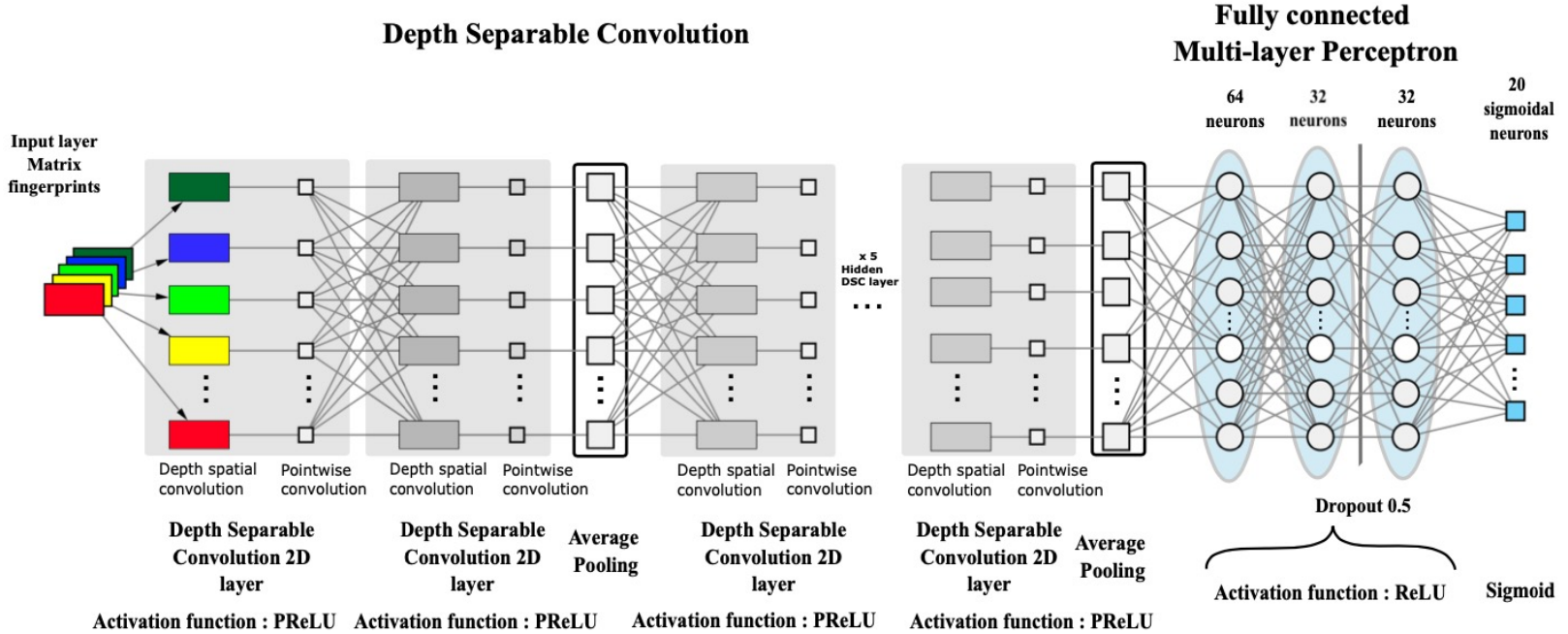
In particular, Virtual Screening (VS) is an often discussed topic in Chemoinformatics and Medicinal Chemistry and is widely applied in pharmaceutical research. VS consists of screening large small-molecule databases searching for bioactive molecules with respect to the target under investigation. This 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

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	2578	25,768
IRAK4	6EG9	OLI	2131	21,282
MAPK2K1	4AN9	ACP; 2P7	1254	12,508
PDK1	3NAX	MP7	1117	11,166

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

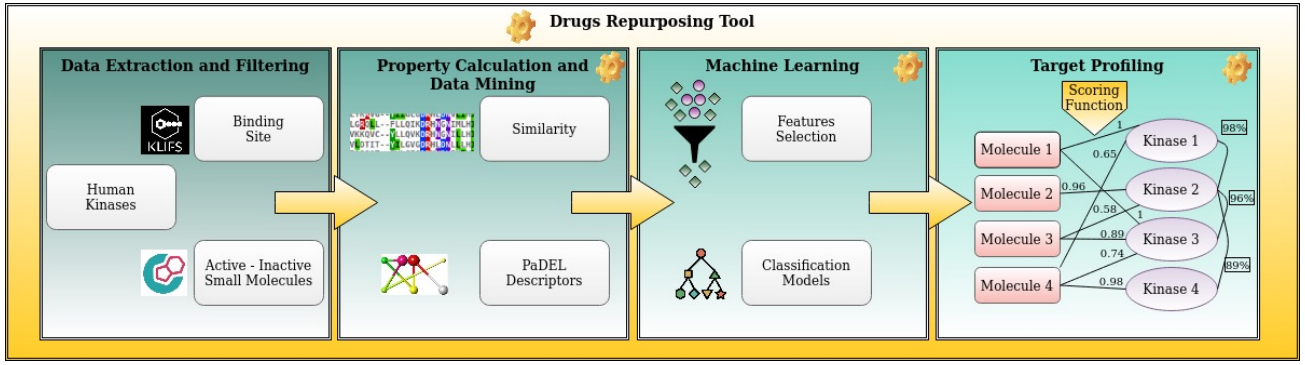


BMC Bioinformatics. 2020 Sep 16;21(Suppl 8):310.doi: 10.1186/s12859-020-03645-9 - Int. J. Mol. Sci. 2022, 23(4), 2156; https://doi.org/10.3390/ijms23042156

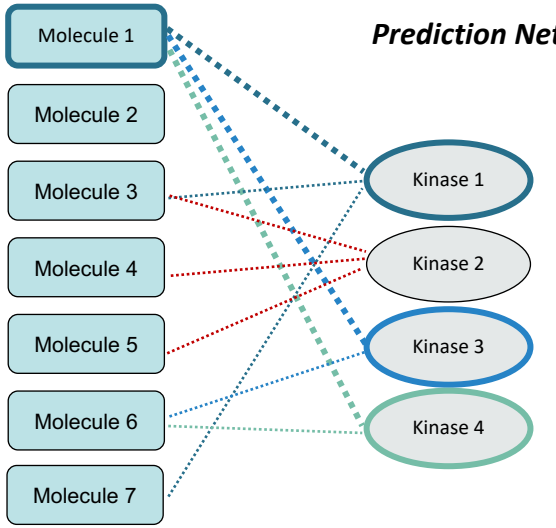
AI nella progettazione di nuovi antitumorali

Drug Repurposing: Kinase drUgs mAchine Learning frAmework (KUALA)- *Training*

Drug repurposing involves the investigation of existing drugs for new therapeutic purposes



Prediction Network




scientific reports

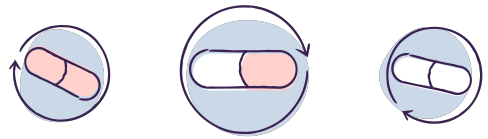
OPEN KUALA: a machine learning-driven framework for kinase inhibitors repositioning

Repurposing Threshold

Similarity Binding sites alignment

Targets
Number of known targets for the ligand

Performance
F1-measure of the predictive model



*De Simone, G., Sardina, D.S., Gulotta, M.R. et al. KUALA: a machine learning-driven framework for kinase inhibitors repositioning. Sci Rep **12**, 17877 (2022)*

AI nella progettazione di nuovi antitumorali

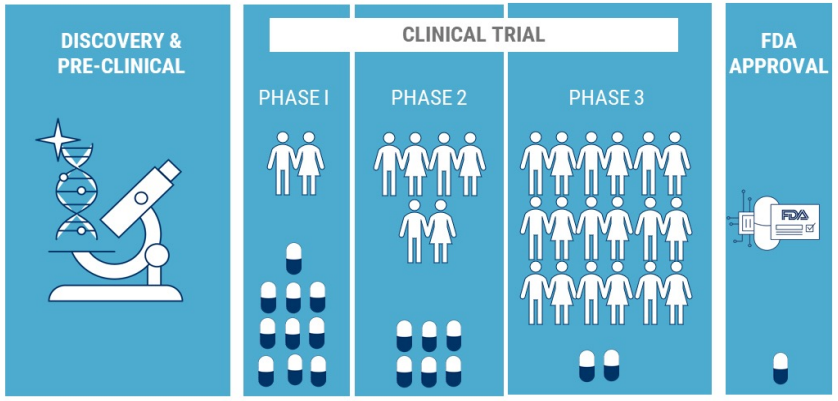

scientific reports

OPEN KUALA: a machine learning-driven framework for kinase inhibitors repositioning

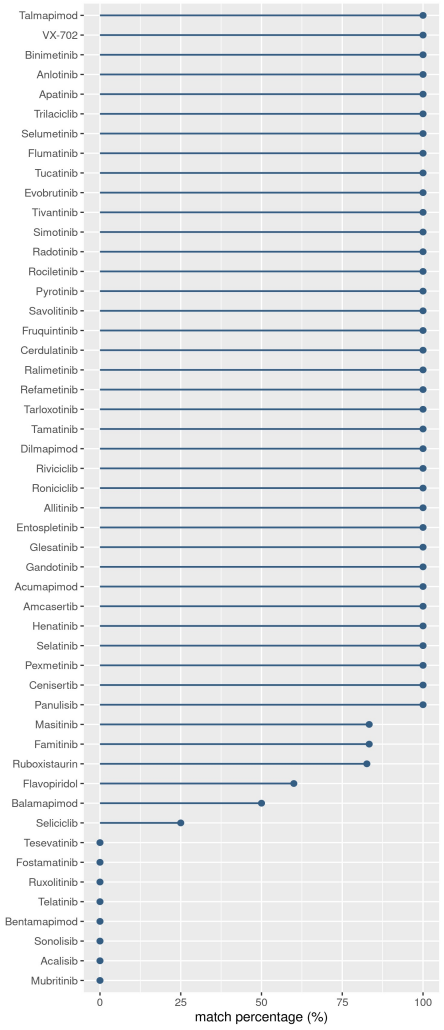
Drug Repurposing: Kinase drUgs mAchine Learning frAmework (KUALA)- *Validation*

MCE MedChem Express

PKIDB



Source: cbinsights.com

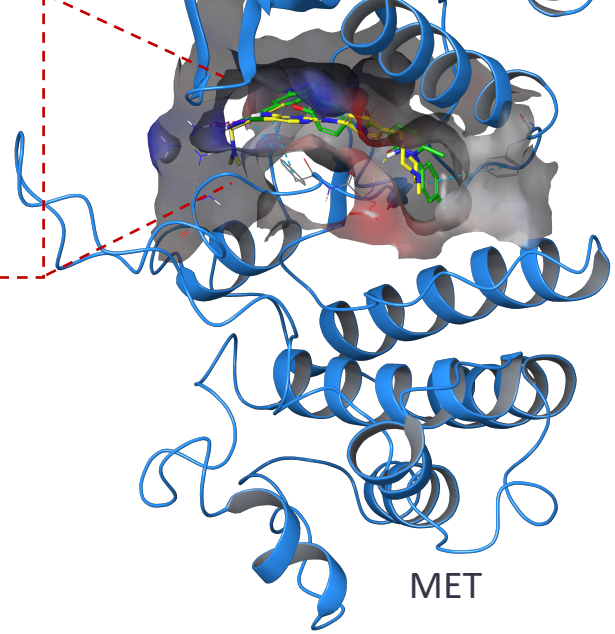


Model Validation with Kinase Inhibitors in Clinical Trials

Kinase models can correctly classify ***high percentages*** of active molecules

Example of repurposing

Dubermatinib



Predicted inhibitor

DUBERMATINIB (CHEMBL2022968)

- AXL inhibitor, highly effective in inducing apoptosis
- Phase I/II in Patients with Previously Treated CLL

De Simone, G., Sardina, D.S., Gulotta, M.R. et al. KUALA: a machine learning-driven framework for kinase inhibitors repositioning. Sci Rep ***12****, 17877 (2022)*

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: 15 July 2021
Published: 19 July 2021

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

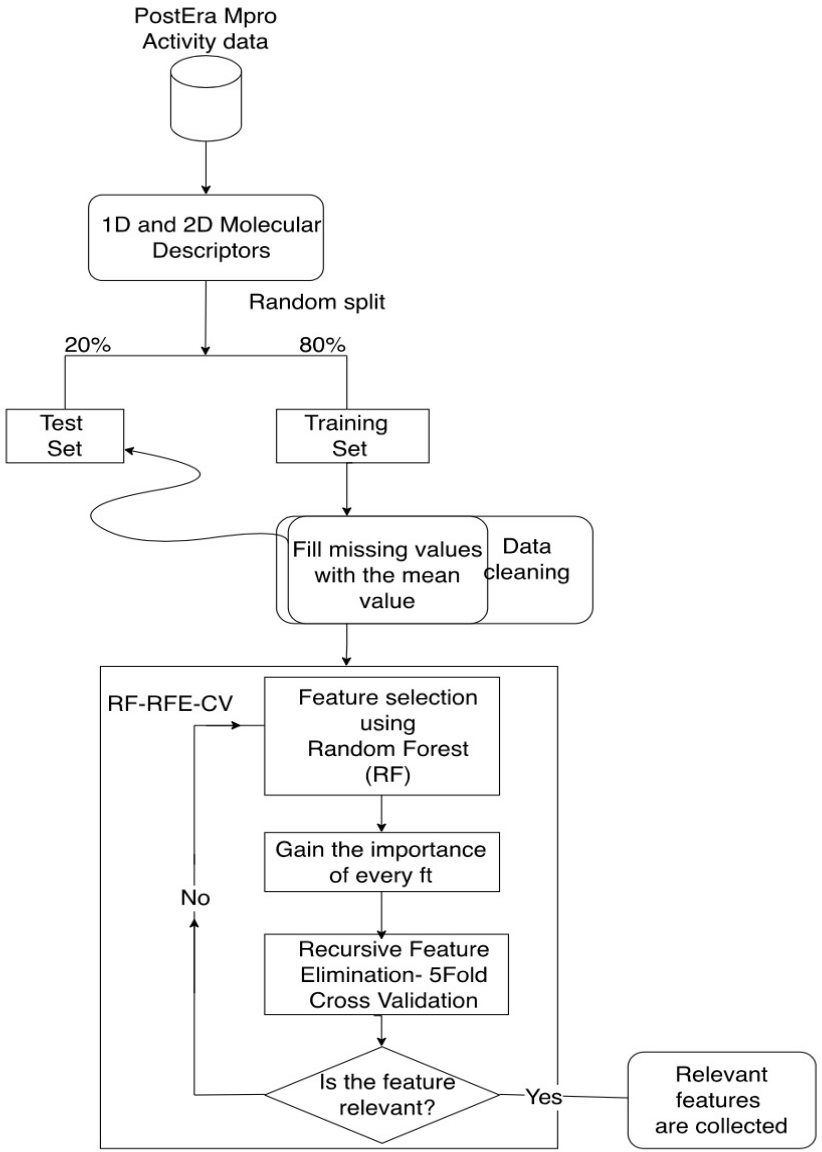
1. Introduction

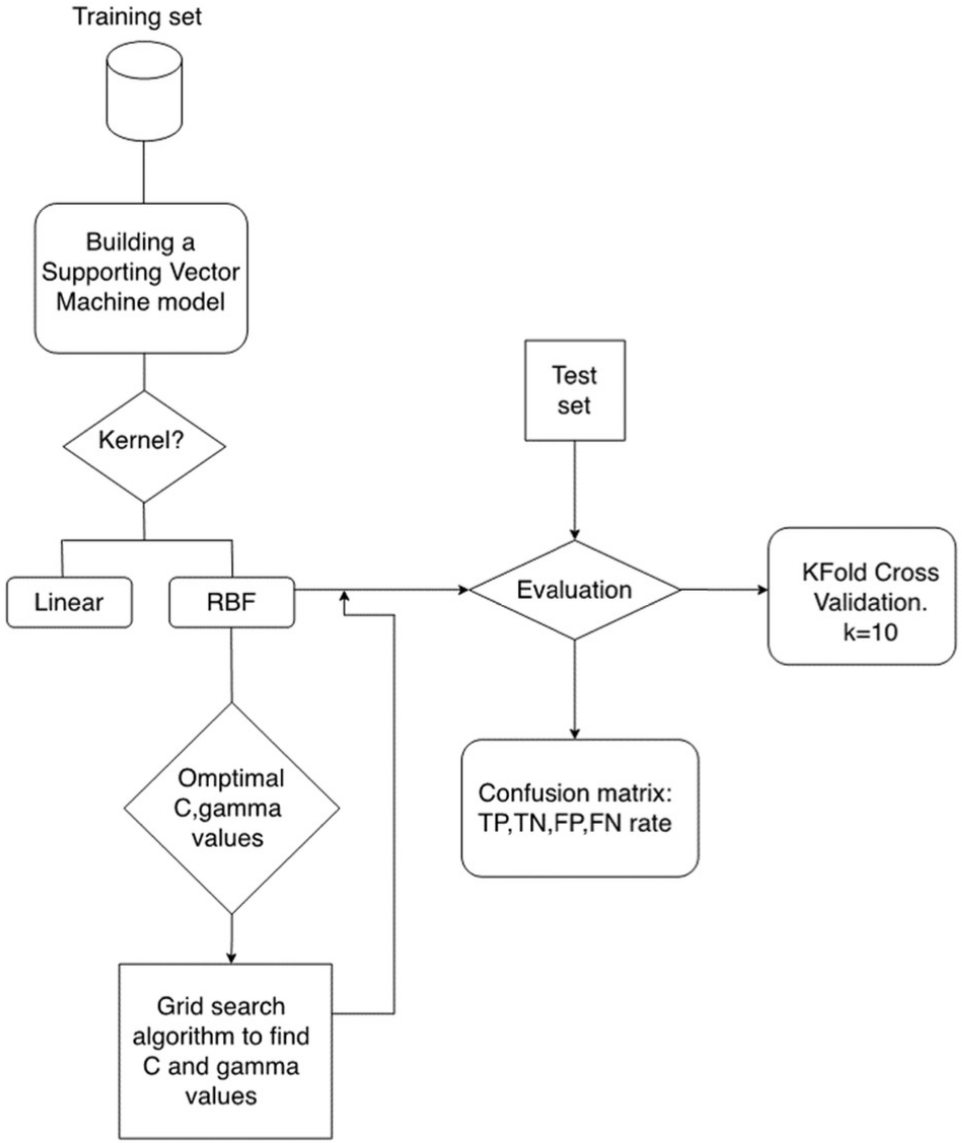
The COVID-19 pandemic, also known as Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is afflicting the health and routines of billions of people worldwide.

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 [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 RiMED, 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: 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.

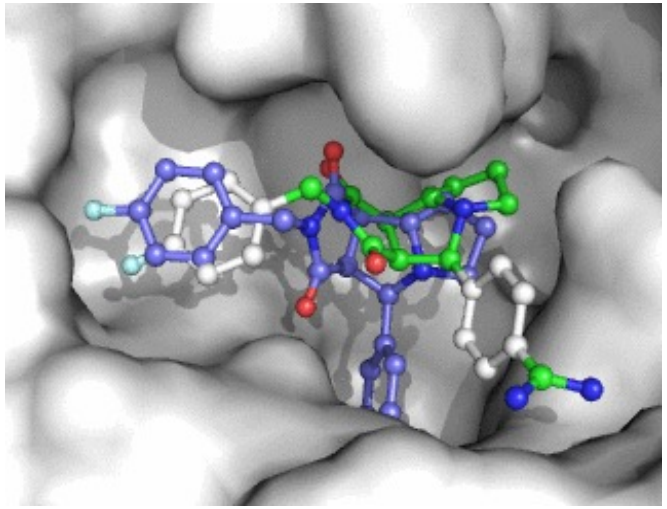
During the last few months, we are witnessing a race against time to vaccinate as many people as possible; however, the disparities in vaccine distribution between countries and the new emerging variants represent a further public health concern, making it hard to reach a full immunization [1,2].

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

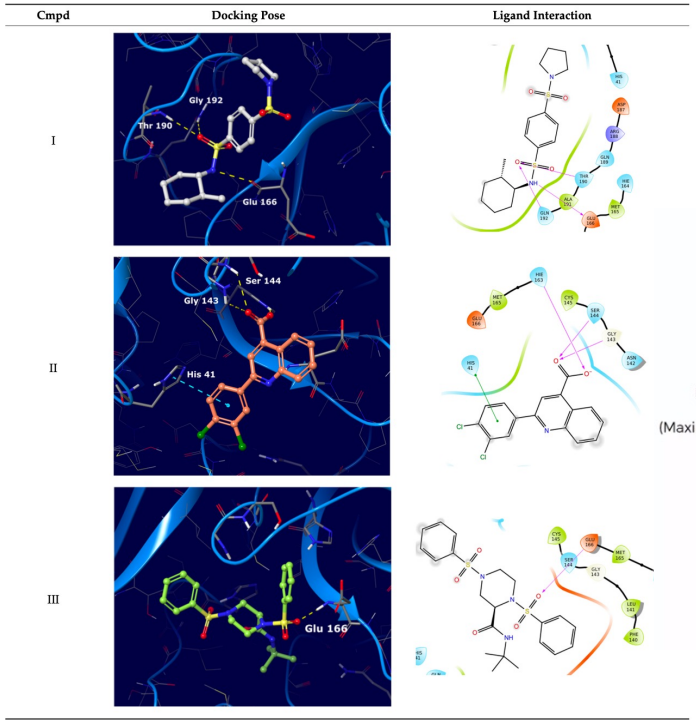
Int. J. Mol. Sci. 2021, 22, 7714. https://doi.org/10.3390/ijms22147714

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

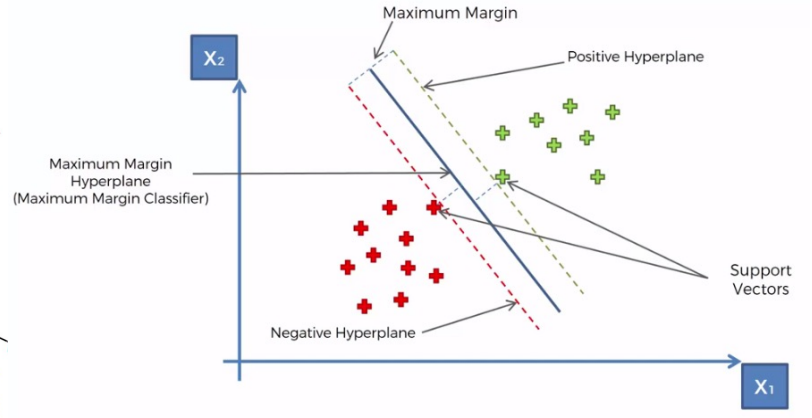
Molecular Docking

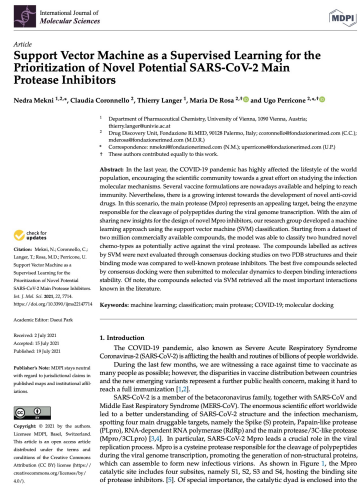


Ranking /Prioritisation



Modello ML per la classificazione



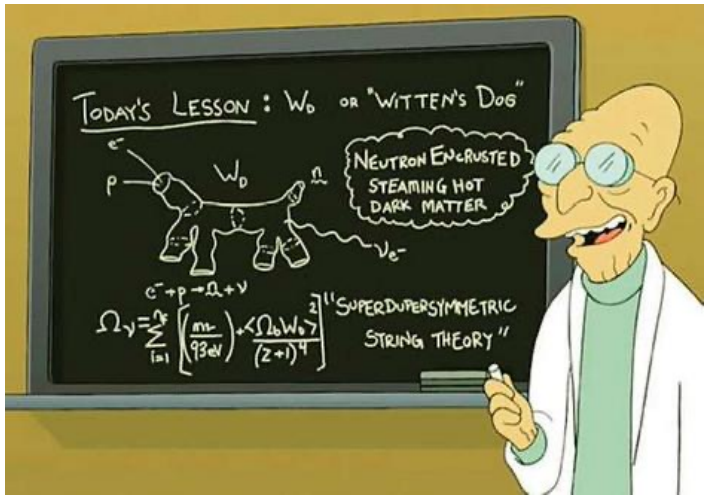


Support Vector Machine as a Supervised Learning for the Prioritization of Novel Potential SARS-CoV-2 Main Protease Inhibitors

Nedra Mekni [1,2,*], Claudia Coronnello [3], Thierry Langer [1], Maria De Rosa [4,†] and Ugo Perricone [3,*,†]

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

L’importanza di modelli computazionali (*In silico*) sempre più evoluti



ORIGINAL RESEARCH ARTICLE
Front. Physiol., 12 September 2017 | https://doi.org/10.3389/fphys.2017.00668

Human *In Silico* Drug Trials Demonstrate Higher Accuracy than Animal Models in Predicting Clinical Pro-Arrhythmic Cardiotoxicity

Elisa Passini[1*], Oliver J. Britton[1], Hua Rong Lu[2], Jutta Rohrbacher[2], An N. Hermans[2], David J. Gallacher[2], Robert J. H. Greig[3], Alfonso Bueno-Orovio[1] and Blanca Rodriguez[1]

[1]Computational Cardiovascular Science Group, Department of Computer Science, University of Oxford, Oxford, United Kingdom
[2]Global Safety, Pharmacology, Discovery Sciences, Janssen Research and Development, Janssen Pharmaceutica NV, Beerse, Belgium
[3]Oxford Computer Consultants Ltd., Oxford, United Kingdom

nature > news > article
a natureresearch journal
MENU nature International journal of science
Search E-alert Submit Login
News & Comment Research
News Opinion Research Analysis Careers Books & Culture

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 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:

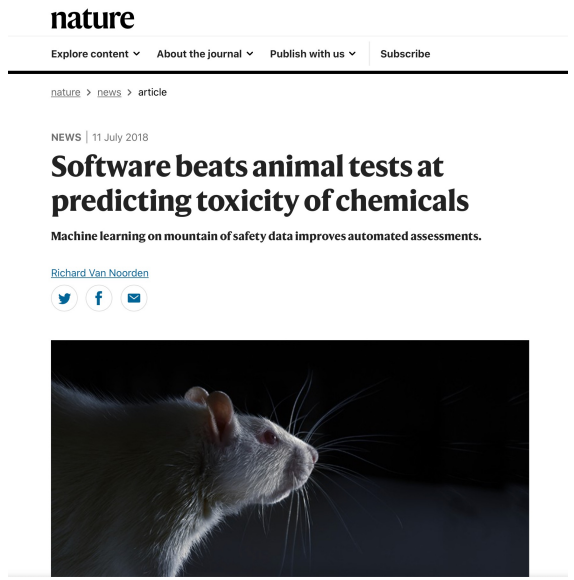
- ➢ Predire per molecole simili a quelle studiate la possibile tossicità e anticipare fallimenti nelle campagne di drug discovery
- ➢ Lavorare con dati umani opportunamente modellati e possibilità di trasferire i modelli da un comparto ad un altro


OXFORD SOT Society of Toxicology www.toxsci.oxfordjournals.org ToxSci 20 Years
TOXICOLOGICAL SCIENCES, 165(1), 2018, 198–212
doi: 10.1093/toxsci/kfy152
Advance Access Publication Date: July 11, 2018
Research Article

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*,§,¶

*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

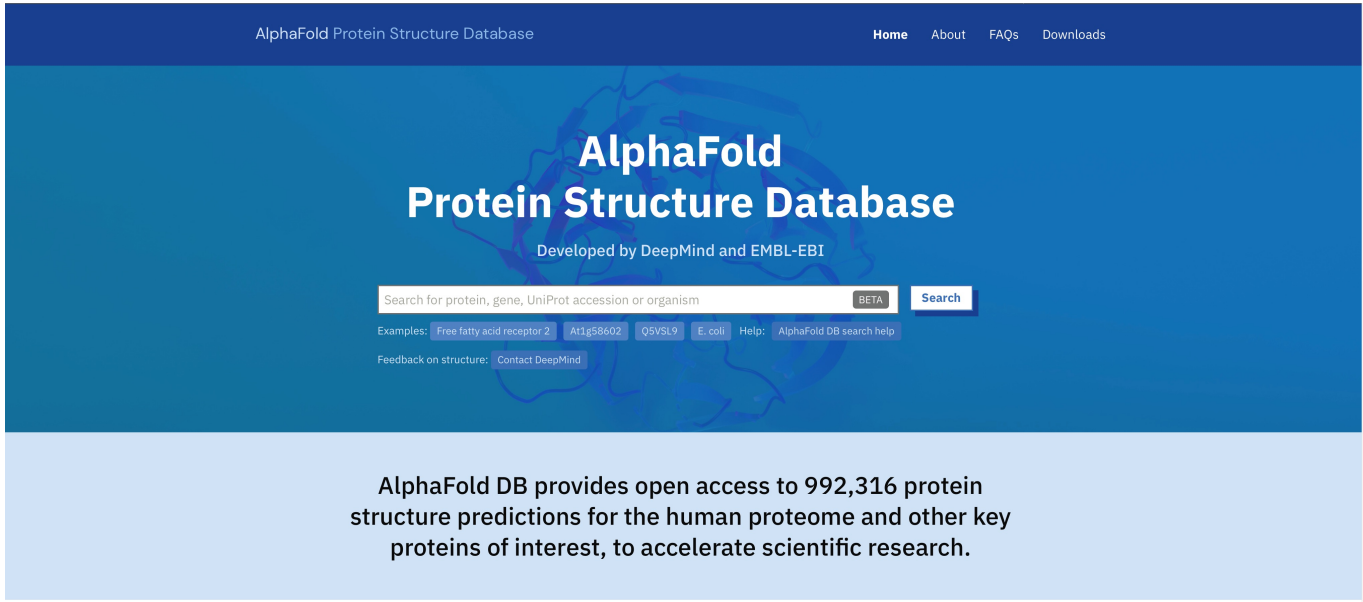
Machine learning on mountain of safety data improves automated assessments.

Richard Van Noorden

AI nella predizione di Strutture 3D…il caso Alphafold

Problema:

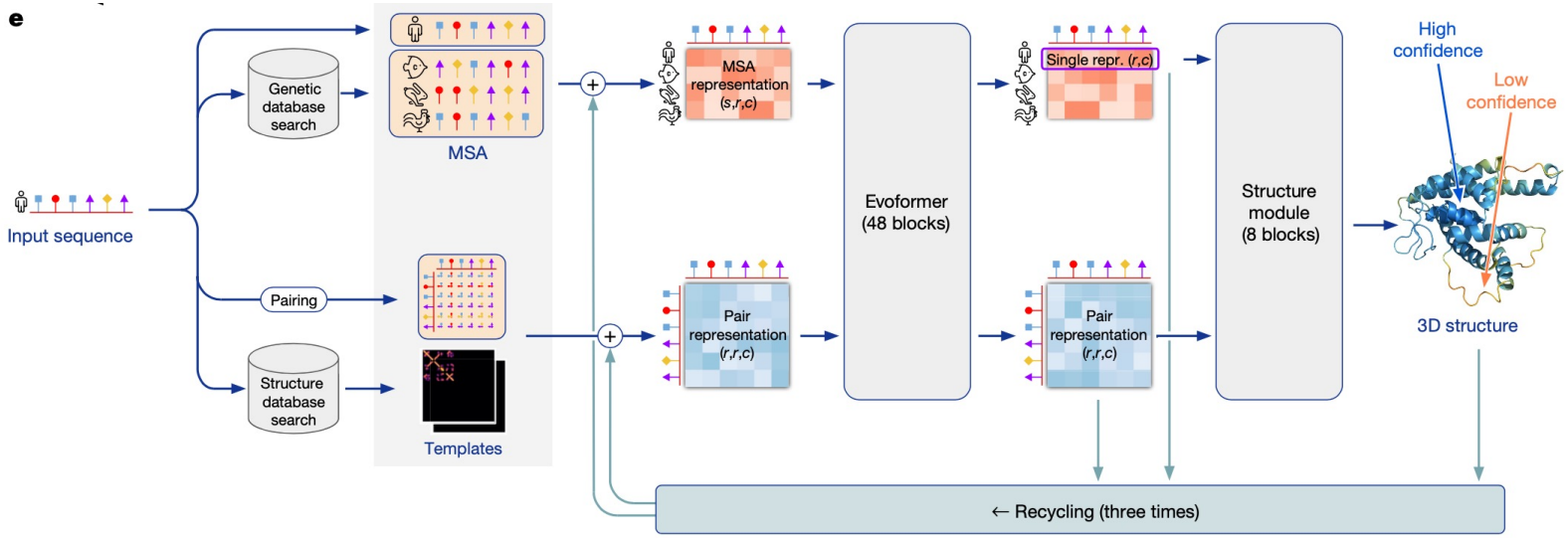
- Non tutte le strutture tridimensionali delle proteine sono disponibili (es. difficoltà legate ad isolamento e purificazione proteina)
- Gli studi strutturali stanno alla base della comprensione di processi biologici e patofisiologici

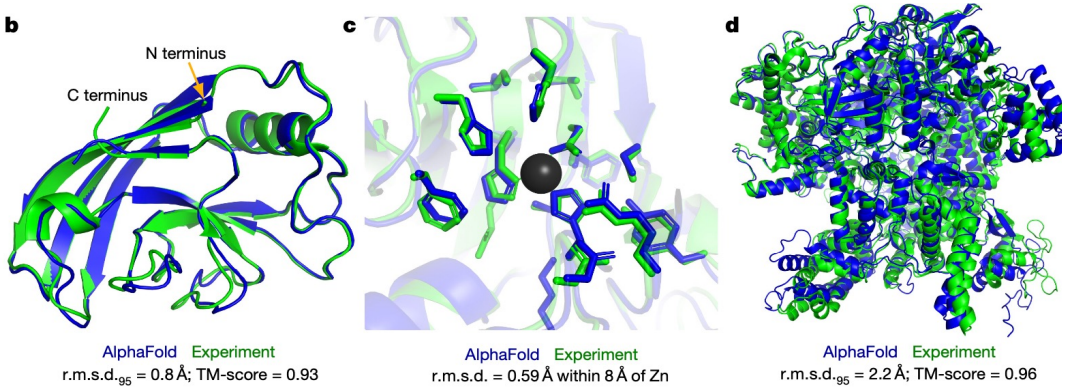


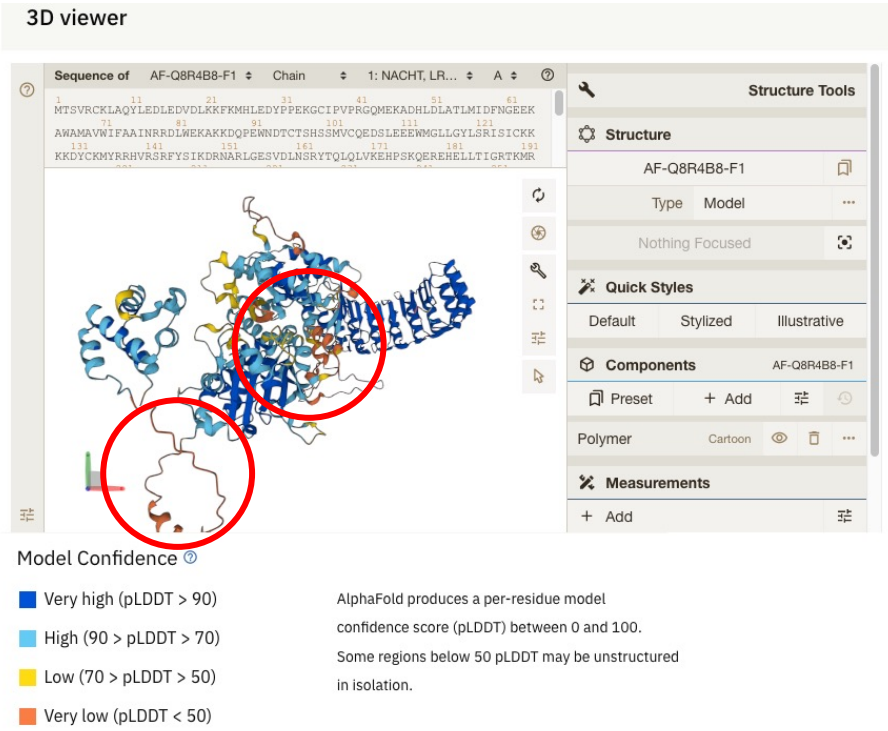
Soluzione:

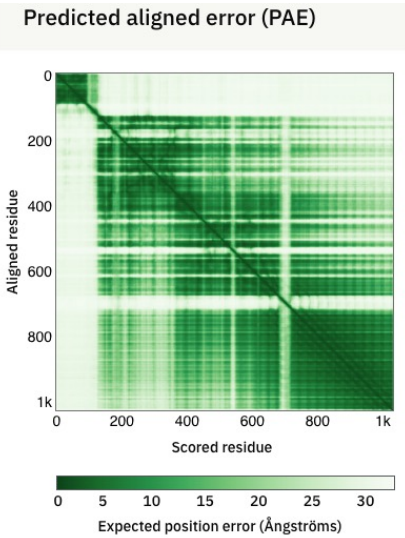
- ✓ Sfruttare tutti i dati disponibili in letteratura su sequenze e folding proteine (input : training/validation/test sets)
- ✓ Addestrare reti in grado di predire la struttura tridimensionale partendo dalla sequenza aminoacidica
- ✓ Predire strutture non disponibili

AI nella predizione di Strutture 3D…il caso Alphafold









Highly accurate protein structure prediction with AlphaFold https://doi.org/10.1038/s41586-021-03819-2

Un delle più grandi rivoluzioni nel drug discovery moderno?

TorchDrug

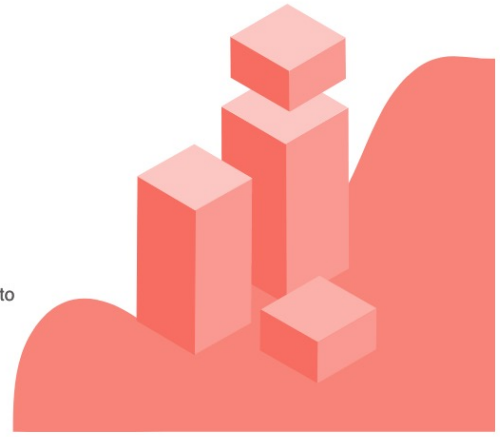
Get Started Tutorials Docs Contribute Publications About GitHub

A powerful and flexible machine learning platform for drug discovery

TorchDrug is a machine learning platform designed for drug discovery, covering techniques from graph machine learning (graph neural networks, geometric deep learning & knowledge graphs), deep generative models to reinforcement learning. It provides a comprehensive and flexible interface to support rapid prototyping of drug discovery models in PyTorch.

```
$ pip install torchdrug
```

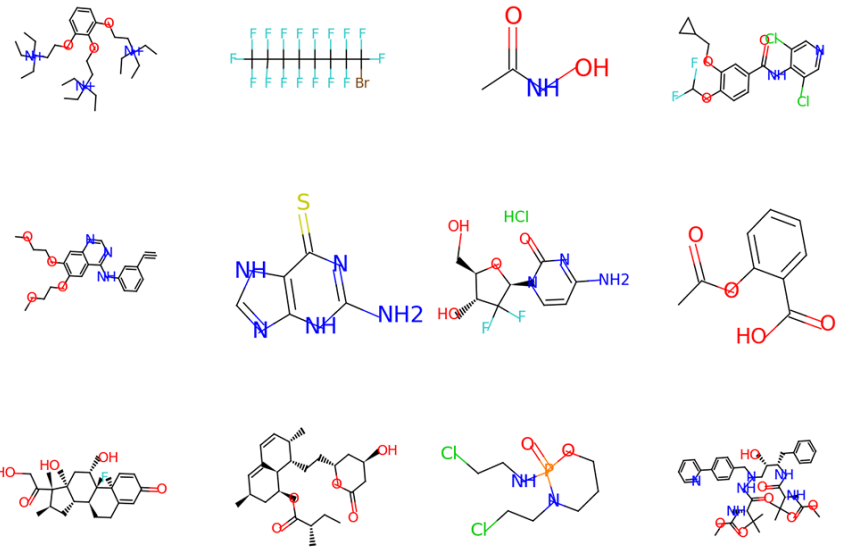
COPY



non-toxic
approved

toxic
not approved

toxic
approved



Key Features

Minimal Domain Knowledge

Build and train machine learning models for drug discovery with minimal domain knowledge.

Datasets and Building Blocks

Empower fast iteration of ideas by a large collection of common datasets and building blocks.

Comprehensive Benchmarks

Benchmarks provide a systematic comparison of deep learning architectures for drug discovery.

Scalable Training and Inference

Seamlessly scale models to multiple CPUs, multiple GPUs, or even distributed settings.

VIEW ALL FEATURES

Solutions to Drug Discovery Tasks

- Property Prediction
- Pretrained Molecular Representations
- De Novo Molecule Design & Optimization
- Reaction Prediction & Retrosynthesis
- Biomedical Knowledge Graph Reasoning
- Protein Representation Learning

Pro e contro di AI in drug discovery

- Taglio drastico dei tempi di ricerca e dei costi
- Maggiore precisione nella progettazione di terapie
- Anticipo nella predizione di effetti tossici per salute umana e ambiente
- Riduzione drastica di uso di modelli animali
- Impiego per terapie personalizzate

- Necessaria una profonda conoscenza di dominio medico-scientifico per applicazioni in questo campo…rischio alto di misinterpretazione dei risultati
- Necessità di grandi moli di dati per l'addestramento…non sempre disponibili
- Qualità dei dati presenti nei DB pubblcii non sempre adeguata

La più bella applicazione di AI in DRUG DISCOVERY

AI distinction

INS018_055 is Insilico's first wholly owned program in which AI was used to identify a novel target and generate novel small molecules through the company's Pharma.AI platform.

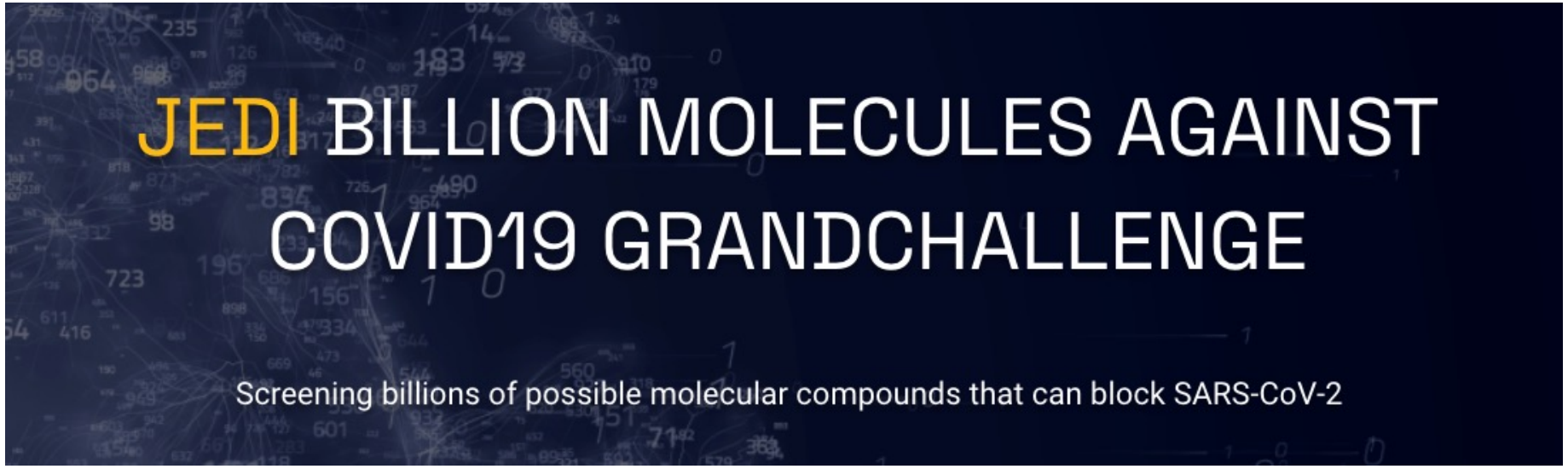
Insilico began development of INS018_055 in February 2021 using Pharma.AI. The platform incorporates a pair of specific-function platforms.

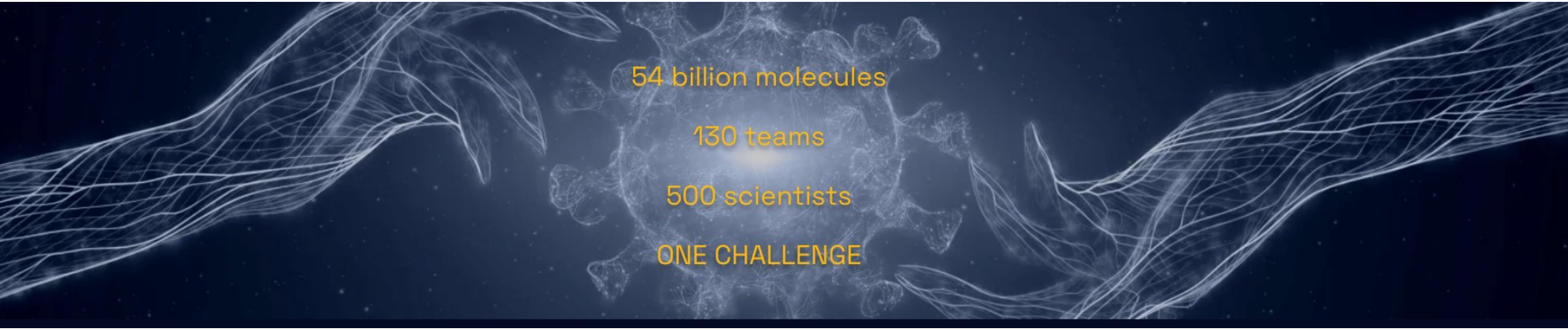
oronkov, PhD, Insilico Medicine founder and CEO; and Feng Ren, PhD, Co-CEO and CSO and Head of R&D, inside the company's AI-powered robotics lab. Based in Hong held Insilico supports every one of its development programs through an autonomous robotics laboratory in Suzhou, China. [Insilico Medicine]

Three clinical candidates

INS018_055 is one of three Insilico candidates to have advanced into clinical trials. The others are ISM3312, an oral small molecule 3CLPro inhibitor designed to treat COVID-19; and ISM3091, a small molecule ubiquitin specific protease 1 (USP1) inhibitor being developed to treat BRCA-mutant breast cancer. ISM3312 is now in a Phase I trial and actively dosing patients, Zhavoronkov said.

La rivoluzione del SARS-CoV-2 e l'importanza della condivisione





La rivoluzione del SARS-CoV-2 e l'importanza della condivisione

molecular informatics
models – molecules – systems

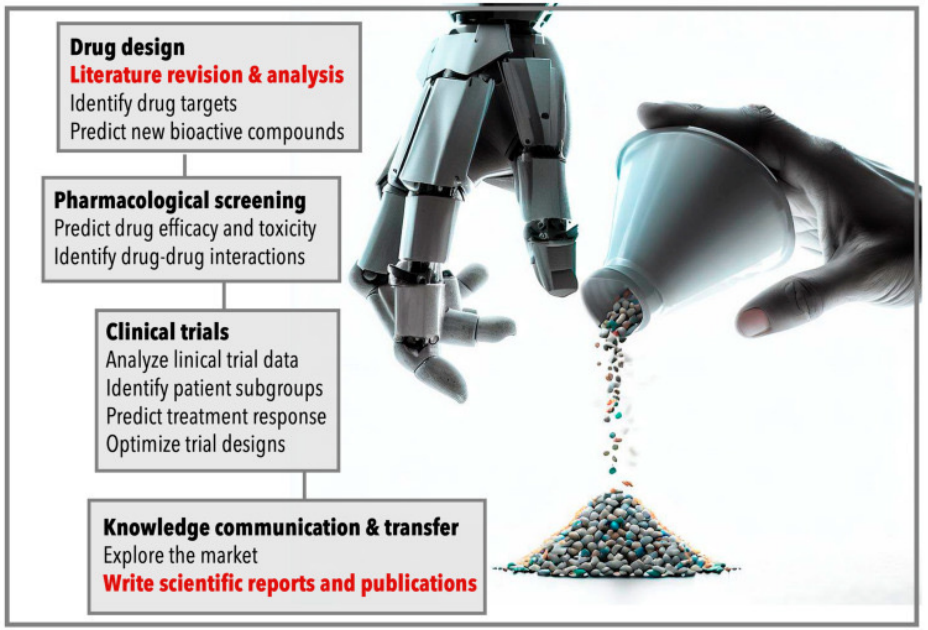
Research Article

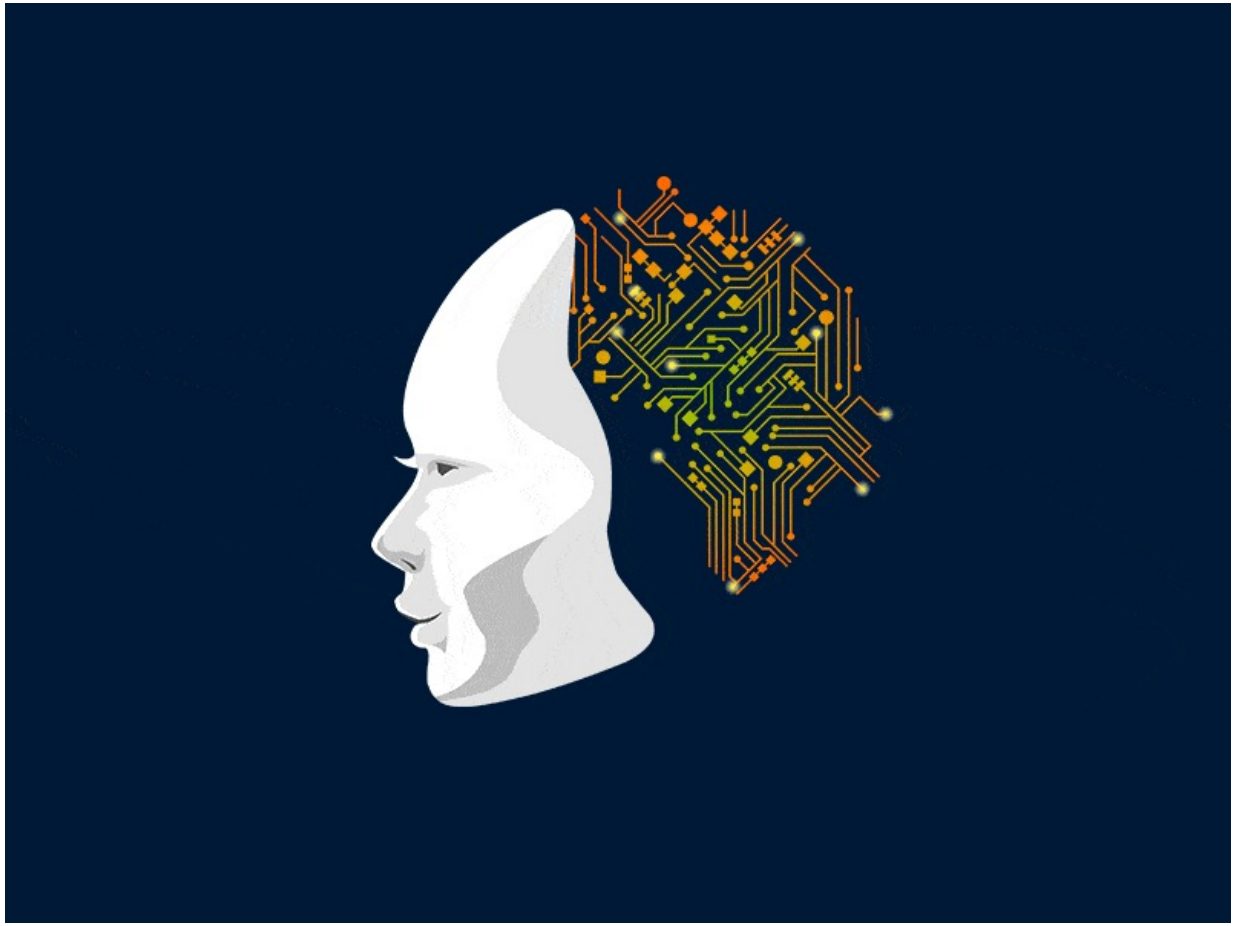
A community effort in SARS-CoV-2 drug discovery

Johannes Schimunek, Philipp Seidl, Katarina Elez, Tim Hempel, Tuan Le, Frank Noé, Simon Olsson, Lluís Raich, Robin Winter, Hatice Gokcan, Filipp Gusev, Evgeny M. Gutkin, Olexandr Isayev, Maria G. Kurnikova, Chamali H. Narangoda, Roman Zubatyuk, Ivan P. Bosko, Konstantin V. Furs, Anna D. Karpenko, Yury V. Kornoushenko, Mikita Shuldau, Artsemi Yushkevich, Mohammed Benabderrahmane, Patrick Bousquet-Melou, Ronan Bureau, Beatrice Charton, Bertrand Cirou, Gérard Gil, William J. Allen, Suman Sirimulla, Stanley Watowich, Nick Antonopoulos, Nikolaos Epitropakis, Agamemnon Krasoulis, Vassilis Pitsikalis, Stavros Theodorakis, Igor Kozlovskii, Anton Maliutin, Alexander Medvedev, Petr Popov, Mark Zaretckii, Hamid Eghbal-zadeh, Christina Halmich, Sepp Hochreiter, Andreas Mayr, Peter Ruch, Michael Widrich, Francois Berenger, Ashutosh Kumar, Yoshihiro Yamanishi, Kam Zhang, Emmanuel Bengio, Yoshua Bengio, Moksh Jain, Maksym Korablyov, Cheng-Hao Liu, Marcous Gilles, Enrico Glaab, Kelly Barnsley, Suhasini M. Iyengar, Mary Jo Ondrechen, V. Joachim Haupt, Florian Kaiser, Michael Schroeder, Luisa Pugliese, Simone Albani, Christina Athanasiou, Andrea Beccari, Paolo Carloni, Giulia D'Arrigo, Eleonora Gianquinto, Jonas Goßen, Anton Hanke, Benjamin P. Joseph, Daria B. Kokh, Sandra Kovachka, Candida Manelfi, Goutam Mukherjee, Abraham Muñiz-Chicharro, Francesco Musiani, Ariane Nunes-Alves, Giulia Paiardi, Giulia Rossetti, S. Kashif Sadiq, Francesca Spyrakis, Carmine Talarico, Alexandros Tsengenes, Rebecca Wade, Conner Copeland, Jeremiah Gaiser, Daniel R. Olson, Amitava Roy, Vishwesh Venkatraman, Travis J. Wheeler, Haribabu Arthanari, Klara Blaschitz, Marco Cespugli, Vedat Durmaz, Konstantin Fackeldey, Patrick D. Fischer, Christoph Gorgulla, Christian Gruber, Karl Gruber, Michael Hetmann, Jamie E. Kinney, Krishna M. Padmanabha Das, Shreya Pandita, Amit Singh, Georg Steinkellner, Guilhem Tesseyre, Gerhard Wagner, Zi-Fu Wang, Ryan J. Yust, Dmitry S. Druzhilovskiy, Dmitry Filimonov, Pavel V. Pogodin, Vladimir Poroikov, Anastassia V. Rudik, Leonid A. Stolbov, Alexander V. Veselovsky, Maria De Rosa, Giada De Simone, Maria R. Gulotta, Jessica Lombino, Nedra Mekni, Ugo Perricone, Arturo Casini, Amanda Embree, D. Benjamin Gordon, David Lei, Katelin Pratt, Christopher A. Voigt, Kuang-Yu Chen, Yves Jacob, Tim Krischuns, Pierre Lafaye, Agnès Zettor, M. Luis Rodríguez, Kris M. White, Daren Fearon, Frank von Delft, Martin A. Walsh, Dragos Horvath, Charles L. Brooks, Babak Falsafi, Bryan Ford, Adolfo García-Sastre, Sang Yup Lee, Nadia Naffakh, [illegible] Guenter Klambauer, Thomas M. Hermans ... See fewer authors

CONCLUSIONI

- AI è uno strumento sempre più potente…e come tale va gestito correttamente
- AI come strumento di collaborazioni multidisciplinari tra ricercatori in ambito medico scientifico ed esperti del settore informatico
- AI NON è una minaccia per il lavoro, NON rimpiazza gli umani, ma richiede competenze diverse dal passato
- Come qualsiasi novità AI non va temuta, ma va conosciuta e se gestita può portare ad un miglioramento globale!





GRAZIE PER LA VOSTRA ATTENZIONE

Per info o collaborazioni
uperricone@fondazionerimed.com