Modelli di deep learning per il progetto *de-novo* di molecole

Nicola Ancona

nicola.ancona@cnr.it

Istituto di Sistemi e Tecnologie Industriali Intelligenti per il Manifatturiero Avanzato - CNR

Collaboratori

- Teresa Maria Creanza, CNR - Istituto di Sistemi e Tecnologie Industriali Intelligenti per il Manifatturiero Avanzato
- Giuseppe Felice Mangiatordi, Giuseppe Lamanna, Pietro Delre, Nicola Corriero, Michele Saviano, CNR - Istituto di Cristallografia
- Marialessandra Contino, Dipartimento di Farmacia - Scienze Farmaceutiche, Universita di Bari “Aldo Moro“

Sommario

- Contesto applicativo.
- Natura dei dati.
- Modello di linguaggio statistico.
- DeLA-Drug: Recurrent Neural Network con LSTM layers.
- Generazione di molecole: SFS e SWS.
- Valutazione *in silico* delle molecole generate.
- Conclusioni e sviluppi futuri.

Contesto applicativo

Lo sviluppo di nuovi farmaci è un processo complesso cha ha come obiettivo l'identificazione di nuove molecole con particolari proprietà chimiche per il trattamento di determinate patologie.

Tale processo ha costi estremamente elevati (miliardi di $), coinvolge un gran numero di risorse umane ed un lungo intervallo di tempo.

Lo sviluppo di nuovi farmaci parte dalla scoperta e ottimizzazione di sostanze in forma prototipale e prevede una lunga ricerca clinica con un elevato rischio di scoprire farmaci con scarsa efficienza.

Contesto applicativo

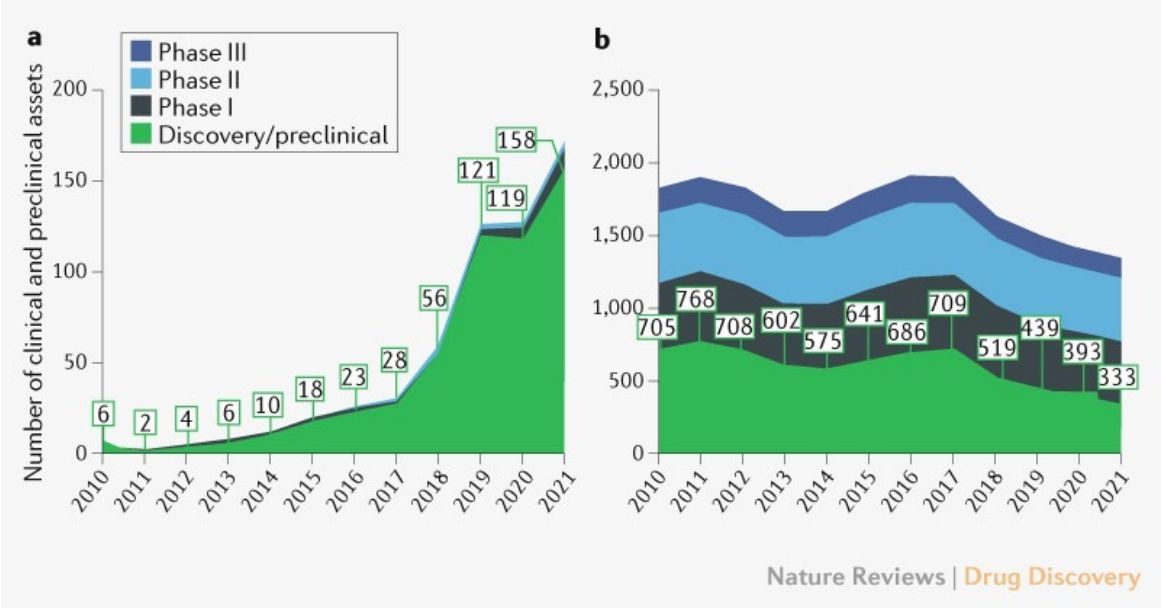
Il problema dell'ottimizzazione prevede la ricerca di sostanze che da un lato massimizzano alcune proprietà farmacocinetiche (solubilità e accessibilità sintetica), dall'altro siano bioattive verso determinati target.

Tale problema è estremamente sfidante in quanto la dimensione dello *spazio chimico* varia da 10^{23} a 10^{60} potenziali molecole, ma solo 10^{8} sostanze sono state finora sintetizzate.

Quindi sono necessarie tecniche computazionali (*virtual screening*) per esplorare lo spazio chimico *generando* nuove molecole e identificare sostanze che si legano ad un determinato target in modo più veloce ed efficace.

Contesto applicativo

Modelli di *deep learning* possono velocizzare la scoperta ed ottimizzazione di nuovi farmaci consentendo la generazione *data-driven* di sostanze con determinate proprietà chimico-fisiche.



AI in small-molecule drug discovery: a coming wave? *M. K. P. Jayatunga, W. Xie, L. Ruder, U. Schulze and C. Meier. Nature Reviews Drug Discovery Feb. 2022*

DeLA-Drug: a **Deep Learning Generative Model**

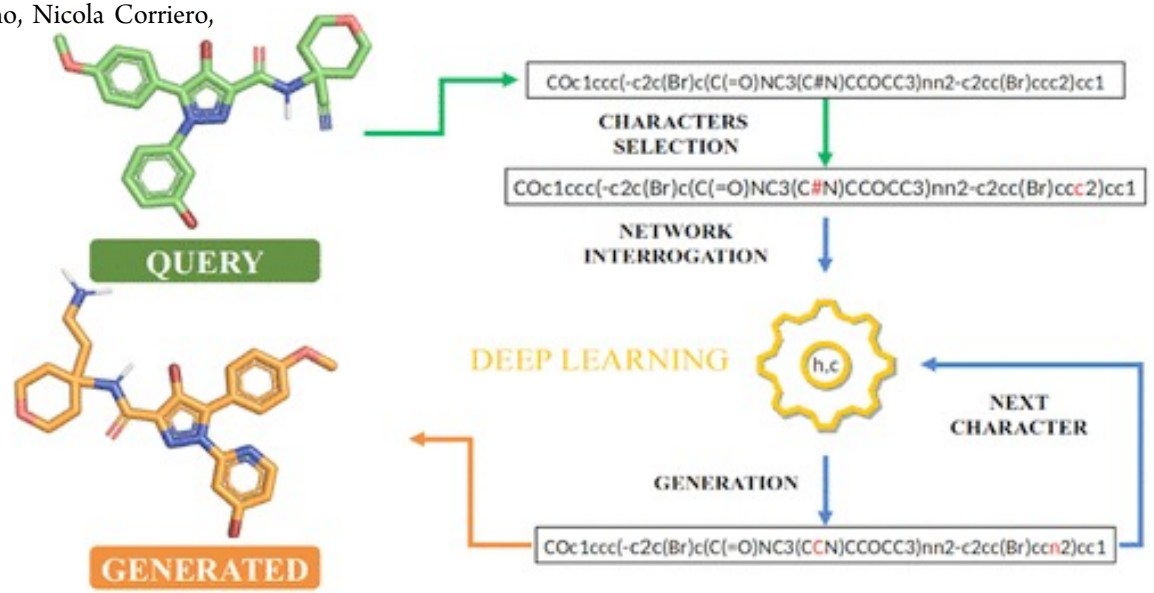
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

DeLA-Drug: A Deep Learning Algorithm for Automated Design of Druglike Analogues

Teresa Maria Creanza,[⊥] Giuseppe Lamanna,[⊥] Pietro Delre, Marialessandra Contino, Nicola Corriero, Michele Saviano, Giuseppe Felice Mangiatordi,* and Nicola Ancona

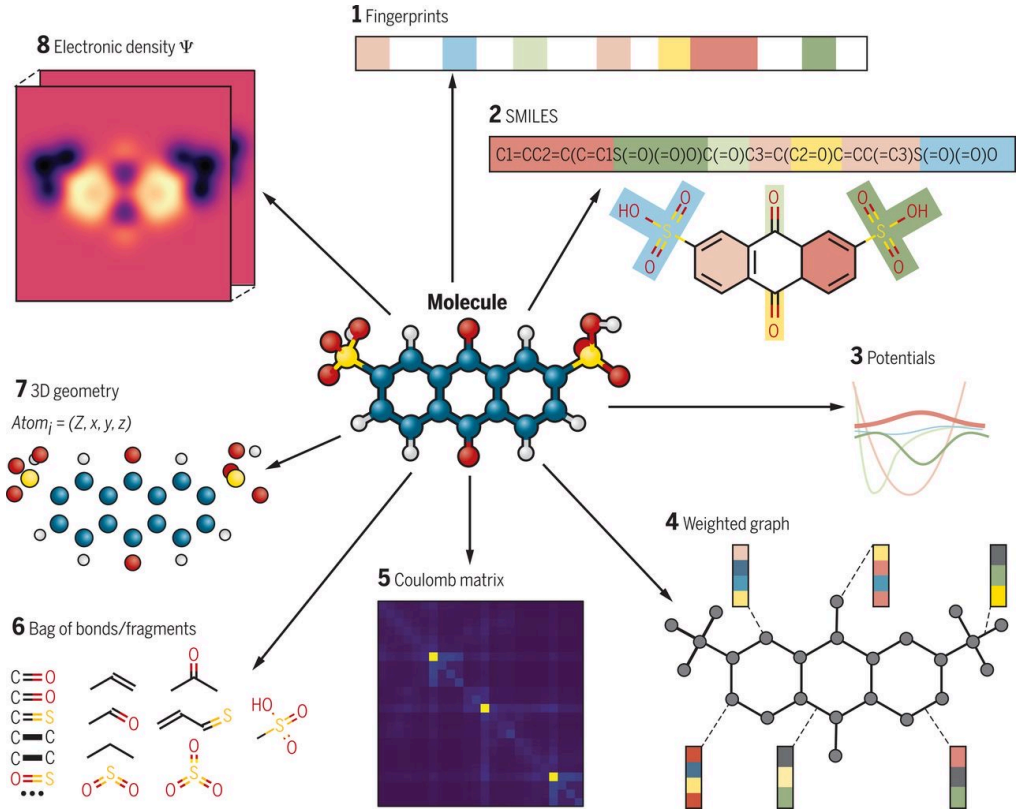
https://doi.org/10.1021/acs.jcim.2c00205



CNR - Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing
CNR – Institute of Crystallography
Department of Pharmacy - Pharmaceutical Sciences, University of Bari "Aldo Moro"

Natura dei dati

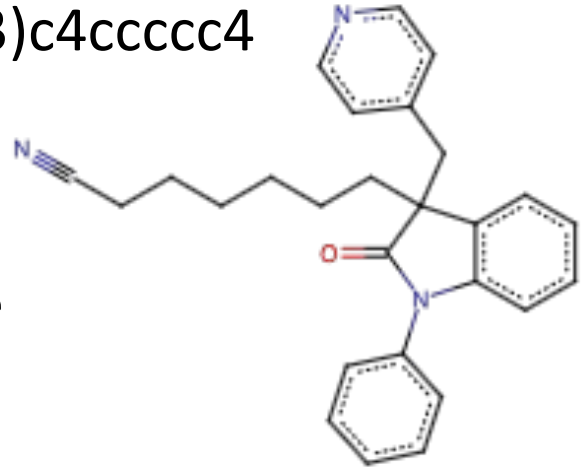
Differenti tipi di rappresentazioni molecolari



Natura dei dati

Un possibile modo per rappresentare molecole è attraverso stringhe di caratteri (canonical SMILES) appartenenti ad un determinato alfabeto.

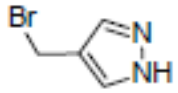
N#CCCCCCCC1(Cc2ccncc2)C(=O)N(c3c1cccc3)c4ccccc4



Repository: ChEMBL28 con 1.092.285 sostanze rappresentate da SMILES con 29-75 caratteri.
Alfabeto è composto da 37 caratteri:
alcuni (H,C, N, O, F, Br, I, Cl, P, S) indicano atomi,
altri (-, =, #) indicano legami.
Caratteri speciali: $ (BoS), ~ (EoS), € (padding).

Natura dei dati

Ogni carattere viene rappresentato da un *one-hot vector* costituito da un numero di componenti uguale al numero di caratteri dell'alfabeto.



SMILES BrCc1c[nH]nc1

One-hot encoding

Br	C	c	1	c	[	n	H	]	n	c	1
1	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	1	0	1	0	0	0	0	0	1	0
0	0	0	0	0	0	1	0	0	1	0	0
0	0	0	1	0	0	0	0	0	0	0	1
0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0

Modello di linguaggio statistico

Un SLM è una funzione densità di probabilità congiunta *f* su sequenze

$(x_1, x_2, \ldots, x_T)$ di caratteri in un alfabeto *A* di dimensione *S*.

Il nostro obiettivo è:

- Apprendere una stima di $f(x_1, x_2, \ldots, x_T)$ dai dati;
- Campionare *f* per generare nuove sequenze.

Ogni x_i e' una variabile aleatoria discreta che assume uno di *S* valori.

Se $T=3$ e $A=\{a,b\}$, dobbiamo stimare $p_{aaa}=P\{aaa\}, p_{aab}=P\{aab\}, \ldots, p_{bbb}=P\{bbb\}$.

Poiché $p_{aaa}+p_{aab}+\ldots+p_{bbb}=1$ allora il numero di parametri liberi è $FP=2^3-1$.

In generale $FP=S^T-1$.

Per $T=75$ $S=37$ si ha $FP=4x10^{117}$.

Modello di linguaggio statistico

Consideriamo sequenze di lunghezza T. La densità congiunta $f(x_1, x_2, \ldots, x_T)$ può essere espressa utilizzando la chain rule per densità condizionali:

$$f(x_1, x_2, \ldots, x_T) = f(x_T \mid x_{T-1}, x_{T-2}, \ldots, x_1)\, f(x_{T-1}, x_{T-2}, \ldots, x_1)$$

$$f(x_1, x_2, \ldots, x_T) = f(x_T \mid x_{T-1}, x_{T-2}, \ldots, x_1)\, f(x_{T-1} \mid x_{T-2}, x_{T-3}, \ldots, x_1)\, f(x_{T-2}, x_{T-3}, \ldots, x_1)$$

In conclusione:

$$f(x_1, x_2, \ldots, x_T) = f(x_T \mid x_{T-1}, x_{T-2}, \ldots, x_1)\, f(x_{T-1} \mid x_{T-2}, x_{T-3}, \ldots, x_1) \ldots f(x_2 \mid x_1)\, f(x_1)$$

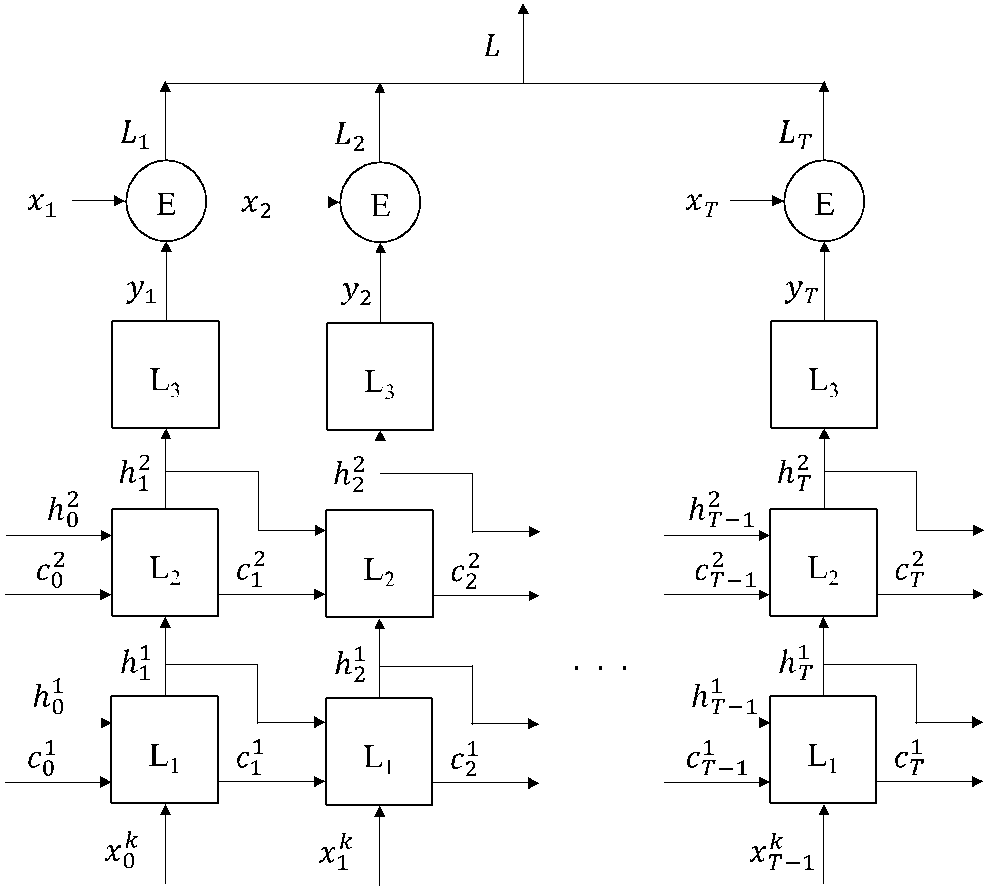
$$f(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} f(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_1)$$

Modello di linguaggio statistico

Il nostro SLM è rappresentato da una neural network costruita e addestrata come un classificatore probabilistico che impara a predire una distribuzione di probabilità $f(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_1)$ per tutti i simboli $x_t \in A$.

Abbiamo utilizzato una Recurrent Neural Network (RNN) con Long Short-Term Memory (LSTM) units per predire la distribuzione del carattere successivo della sequenza.

DeLA-Drug


L
L1
L2
LT
x1
E
x2
E
xT
E
y1
y2
yT
L3
L3
L3
h12
h22
hT2
h02
hT−12
c02
c12
L2
c22
L2
cT−12
cT2
L2
h11
h21
hT1
h01
hT−11
c01
c11
L1
c21
L1
cT−11
cT1
L1
x0k
x1k
xT−1k

LSTM unit

$i = \sigma(W_i[x_t, h_{t-1}] + b_i)$ **input gate**

$f = \sigma(W_f[x_t, h_{t-1}] + b_f)$ **forget gate**

$o = \sigma(W_o[x_t, h_{t-1}] + b_o)$ **output gate**

$g = tanh(W_g[x_t, h_{t-1}] + b_g)$ **modulation gate**

$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot tanh(c_t)$$

Output layer

$$z_t = W_z^t h_t + b_z^t$$

$$u_t = ReLU(z_t)$$

$$v_t = W_v^t u_t + b_v^t$$

$$y_t = s(v_t)$$

Dove s è la funzione di attivazione *softmax* .

Training

Per determinare i pesi della rete, consideriamo la likelihood $\mathcal{L}(X|Y)$ di osservare i dati $\boldsymbol{x}_t$ date le probabilità $\boldsymbol{y}_t$ assegnate dal modello:

$$\mathcal{L}(X|Y) = \prod_{t=1}^{T} \prod_{k=1}^{S} \left(\boldsymbol{y}_t^{(k)}\right)^{\boldsymbol{x}_t^{(k)}}$$

Dove $\boldsymbol{x}_t^{(k)}$ denota la k-th componente del vettore $\boldsymbol{x}_t$.

Cross-Entropy Loss

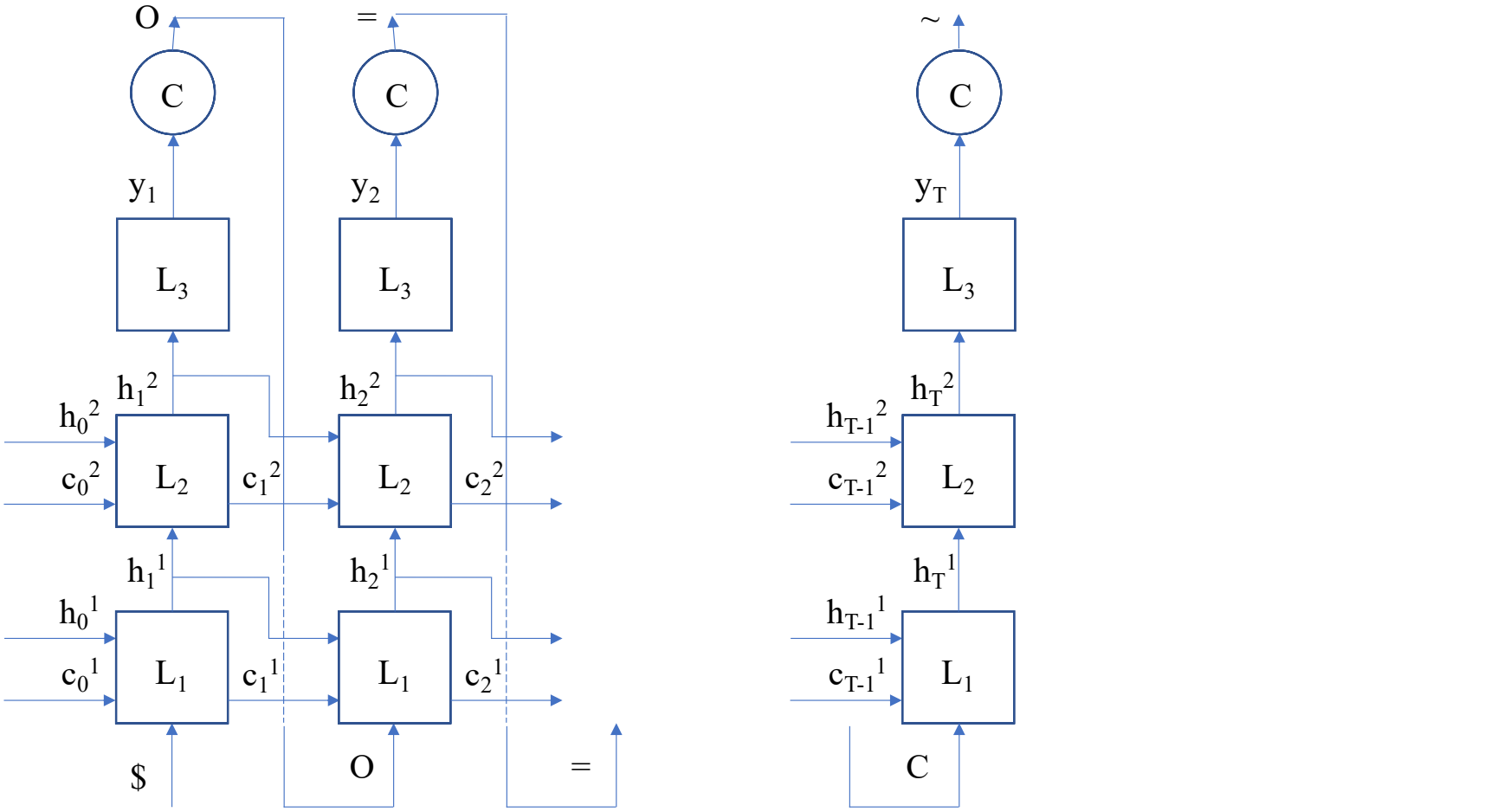
$$E = -\frac{1}{T}\log \mathcal{L}\,(X|Y)$$

$$E = \frac{1}{T}\sum_{t=1}^{T} L\,(\boldsymbol{x}_t, \boldsymbol{y}_t)$$

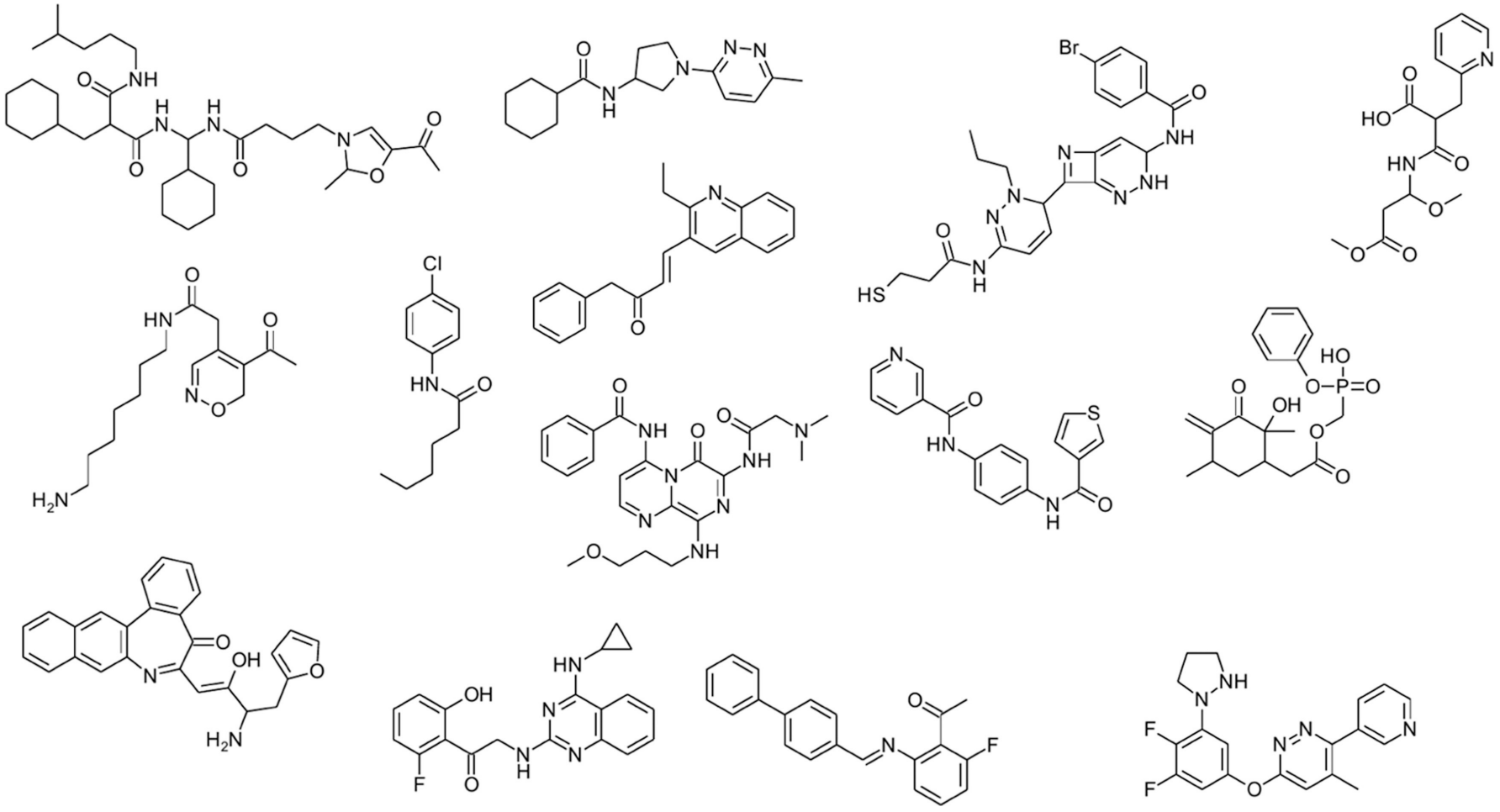
$$L(\boldsymbol{x}_t, \boldsymbol{y}_t) = -\sum_{k=1}^{S} \boldsymbol{x}_t^{(k)} \log\left(\boldsymbol{y}_t^{(k)}\right)$$

L è la *cross-entropy* tra $\boldsymbol{x}_t$ e $\boldsymbol{y}_t$.

Sampling From Scratch



Sampling From Scratch



SAMPLING WITH SUBSTITUTIONS

O = C (N C 1 C C C C 1) c 1 c c c n c 1 O C c 1 c c c c c 1

O = P

O = P (N C C

O = P (N C C 1

O = P (N C C 1 C C N

O = P (N C C 1 C C N C 1) c 1

O = P (N C C 1 C C N C 1) c 1 c c c c

O = P (N C C 1 C C N C 1) c 1 c c c c c 1 O

O = P (N C C 1 C C N C 1) c 1 c c c c c 1 O C c 1

O = P (N C C 1 C C N C 1) c 1 c c c c c 1 O C c 1 c c

O = P (N C C 1 C C N C 1) c 1 c c c c c 1 O C c 1 c c c c c

O = P (N C C 1 C C N C 1) c 1 c c c c c 1 O C c 1 c c c c c 1

Valutazione in silico (SFS)

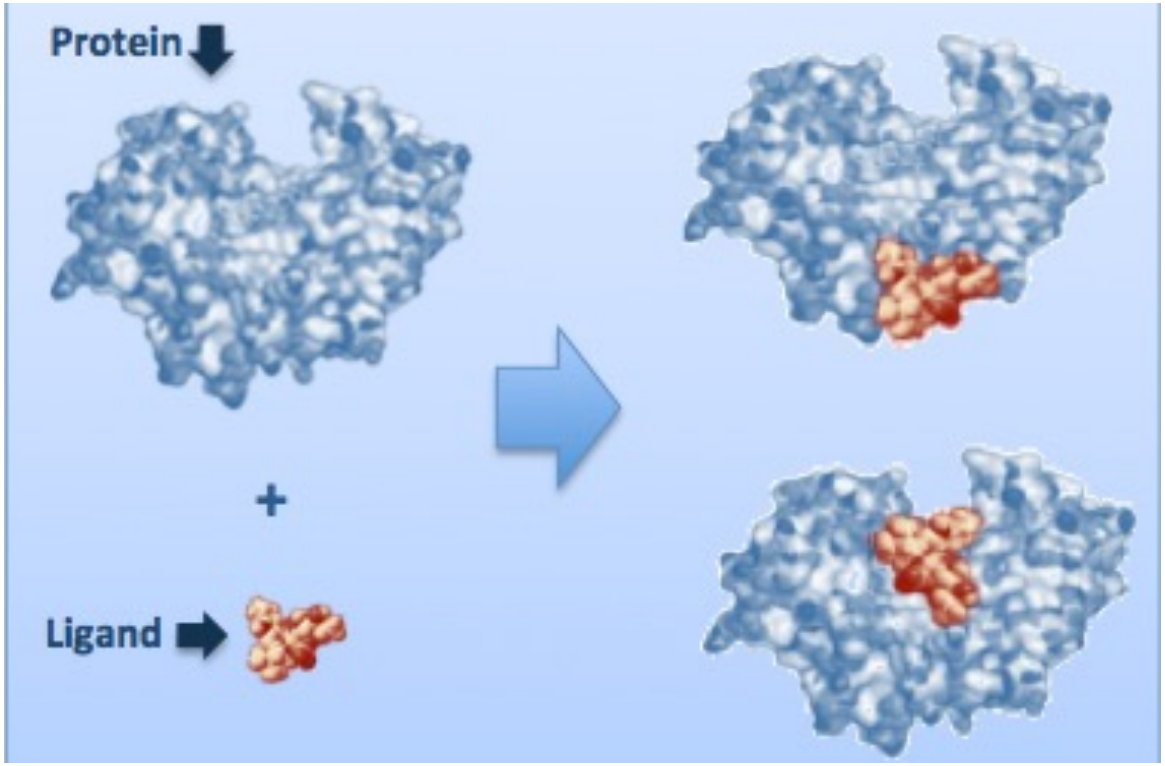
Set Name	Val (%)	Un (%)	ID (Td)	Nov (%)	QED	SA	nPAINS (%)
FSC1	85.14	85.05	0.87	99.14	0.57 ± 0.21	3.0±0.8	95.27
FSC2	83.80	87.93	0.87	99.15	0.58 ± 0.20	2.8 ± 0.7	95.32
FSC4	84.94	87.05	0.87	99.24	0.57 ± 0.21	2.9 ± 0.8	95.12
FSC6	79.43	88.51	0.87	99.29	0.58 ± 0.21	2.9 ± 0.8	94.67
FSC8	86.24	84.73	0.87	99.30	0.58 ±0.20	3.0 ± 0.7	95.30
FSC10	83.85	88.27	0.87	99.34	0.56± 0.21	2.9 ± 0.8	95.27
FSC12	80.90	87.84	0.87	99.21	0.58± 0.21	2.8 ± 0.7	95.78

Valutazione in silico (SWS)

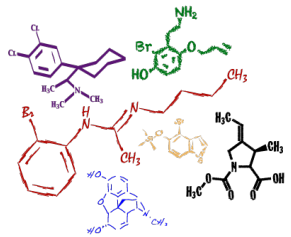
Set Name	n. valid molecules	Un (%)	ID	Nov (%)	QED	SA	nPAINS (%)	QS
CB2R-DB	**1,845**	**/**	**0.82**	**/**	**0.54 ± 0.19**	**2.90 ± 0.70**	**95.83**	**/**
S5C1	**108,740**	**60.60**	**0.84**	**99.13**	**0.53 ± 0.20**	**3.31± 0.84**	**95.83**	**0.88**
S5C8	**113,512**	**59.36**	**0.82**	**99.26**	**0.53 ± 0.20**	**3.32± 0.84**	**95.82**	**0.89**
S10C1	**34,426**	**89.35**	**0.87**	**98.83**	**0.54 ± 0.20**	**3.43 ± 0.97**	**96.06**	**0.61**
S10C8	**37,178**	**88.28**	**0.87**	**98.94**	**0.55 ± 0.20**	**3.48 ± 1.00**	**95.86**	**0.60**
S15C1	**11,824**	**96.63**	**0.88**	**98.29**	**0.54 ± 0.21**	**3.58 ± 1.00**	**96.04**	**0.40**
S15C8	**13,442**	**95.31**	**0.88**	**98.22**	**0.56 ± 0.21**	**3.67 ± 1.0**	**96.12**	**0.39**

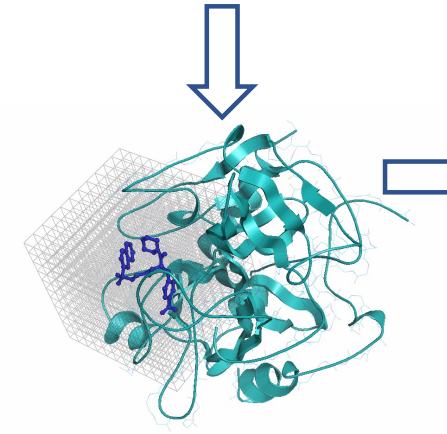
Molecular Docking

Tecnica in silico che cerca di predire la posizione e l'orientazione di un ligando (small molecule) quando si lega ad una proteina.

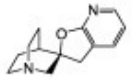
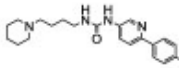
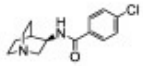
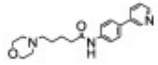


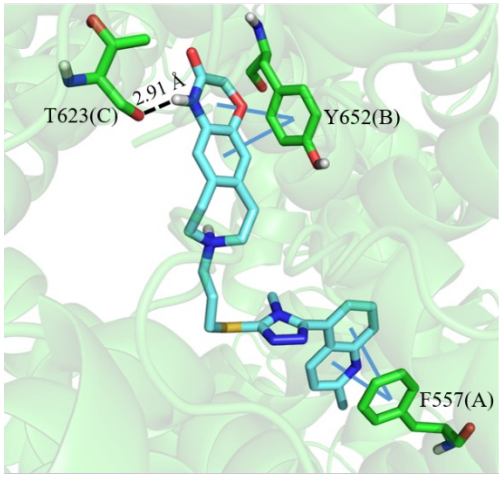
Molecular Docking





score + binding mode

1		-7.8 kcal/mol
2		-6.6 kcal/mol
3		-8.3 kcal/mol
...	...	
n		-5.8 kcal/mol



candidates selection

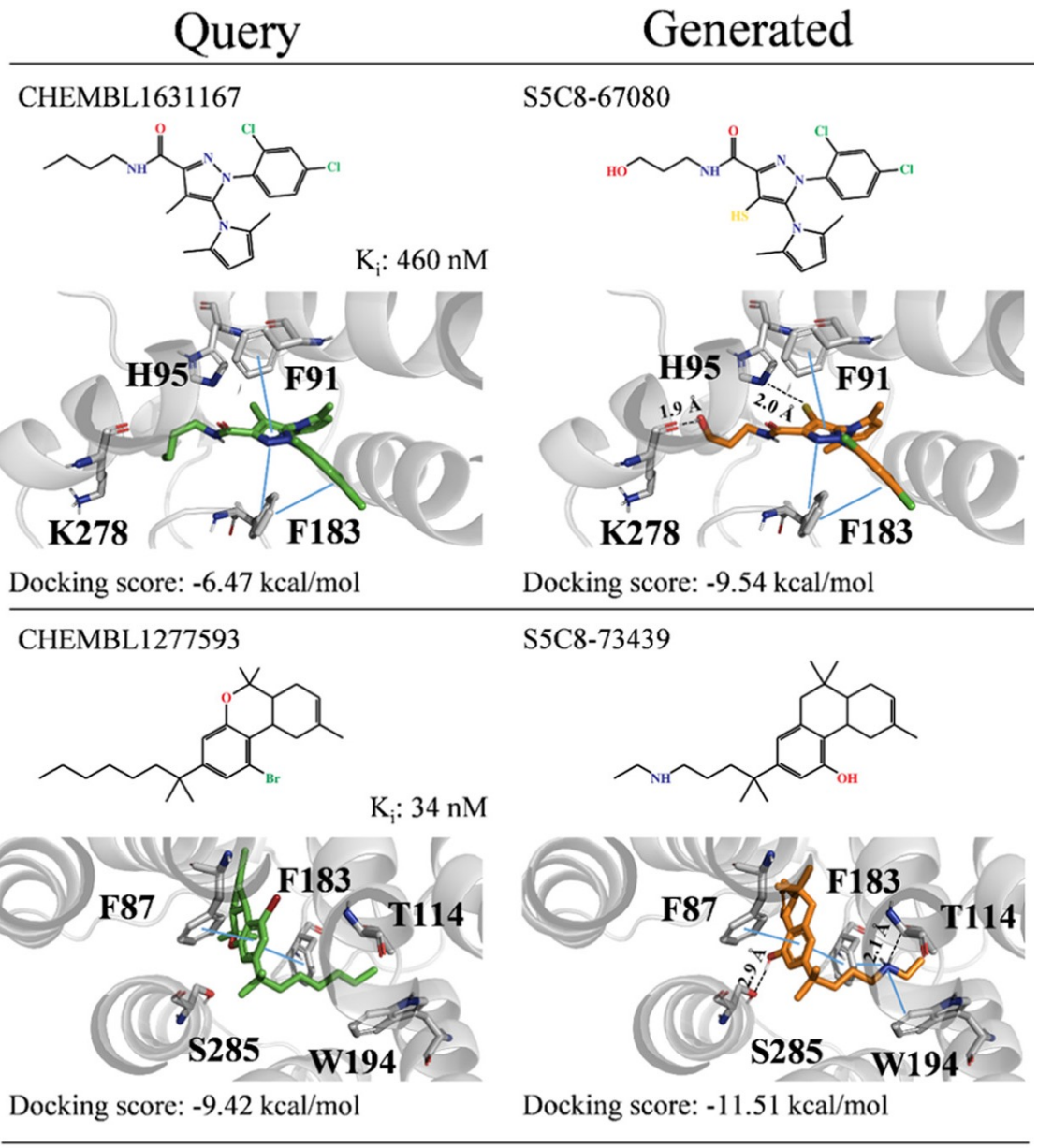
Validazione Statistica del Docking

- Consideriamo 3 dataset BD1, BD2, BD3 contenenti 50 molecole attive su CB2R e 833 decoy (5.66% attive/totale).
- Ordiniamo le molecole in ciascun dataset per lo score di docking e contiamo il numero di molecole attive nel primo decile.
- Abbiamo il 13.00% in BD1 (p-value=0.0073), il 15.89% in BD2 (p-value=1.6×10^{-4}), 22.80% in BD3 (p-value=4.9×10^{-9}).

Virtual Screening con Docking

Set pairs	Fisher's exact test P-values
CB2R-DB vs. randChEMBL	$3x10^{-28}$
randS5C8 vs. randChEMBL	$2.7x10^{-24}$
randS10C8 vs. randChEMBL	$1.3x10^{-14}$
randS15C8 vs. randChEMBL	$1.16x10^{-4}$

Molecole Generate



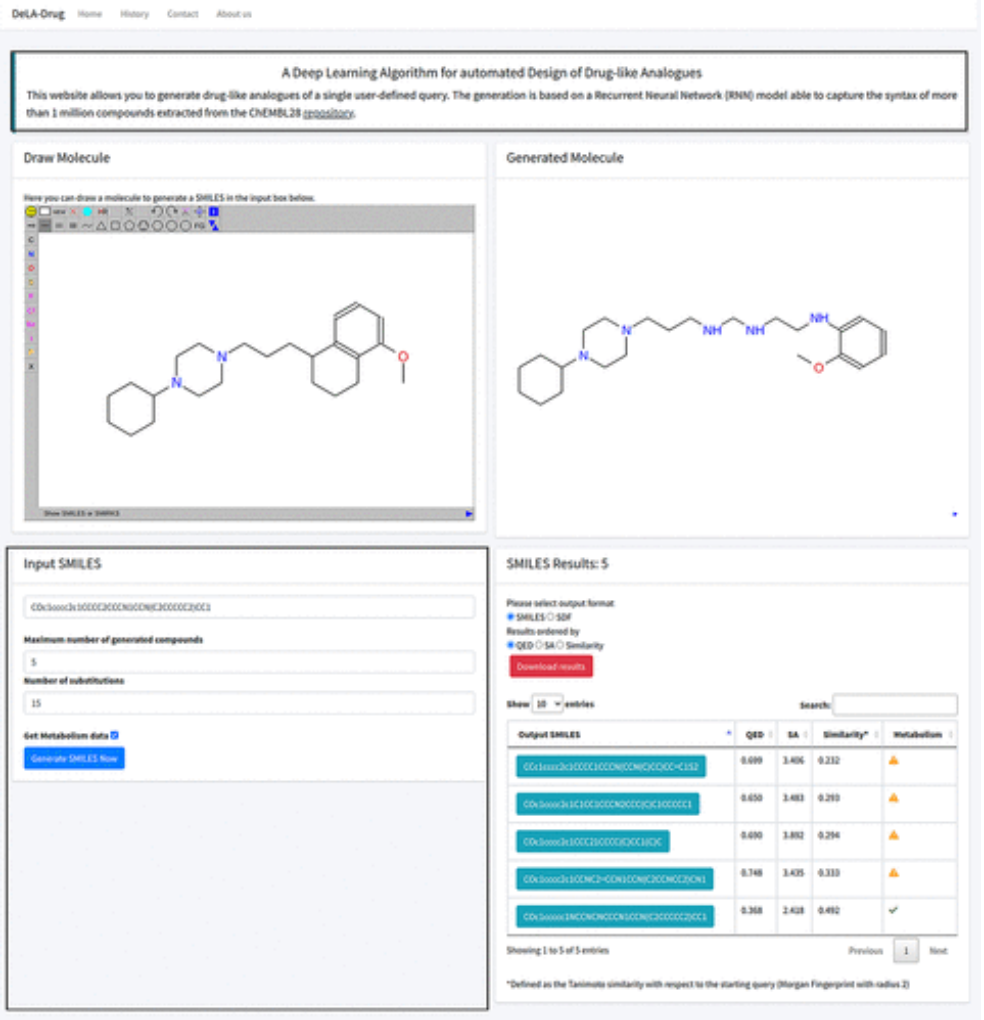
DeLA-Drug web platform

https://www.ba.ic.cnr.it/softwarec/deladrugportal/

Input

l'utente può:

- scrivere o disegnare una molecola query;
- scegliere il numero di caratteri da sostituire nella procedura SWB;
- vedere potenziali warning metabolici.



Output

l'utente può:

- ispezionare una molecola generata;
- scaricare i risultati in differenti formati ordinati per QED o SA.

Sviluppi Futuri

- Transformer e Autoencoder Variazionali

- Autoencoder Variazionali Condizionati

Tali modelli saranno addestrati condizionando su dati trascrittomici al fine di progettare molecole aventi un’alta probabilità di indurre un determinato profilo trascrittomico.