

# Aprendizado de Máquina Supervisionado I



Prof. Dr. Walter F. de Azevedo, Jr.

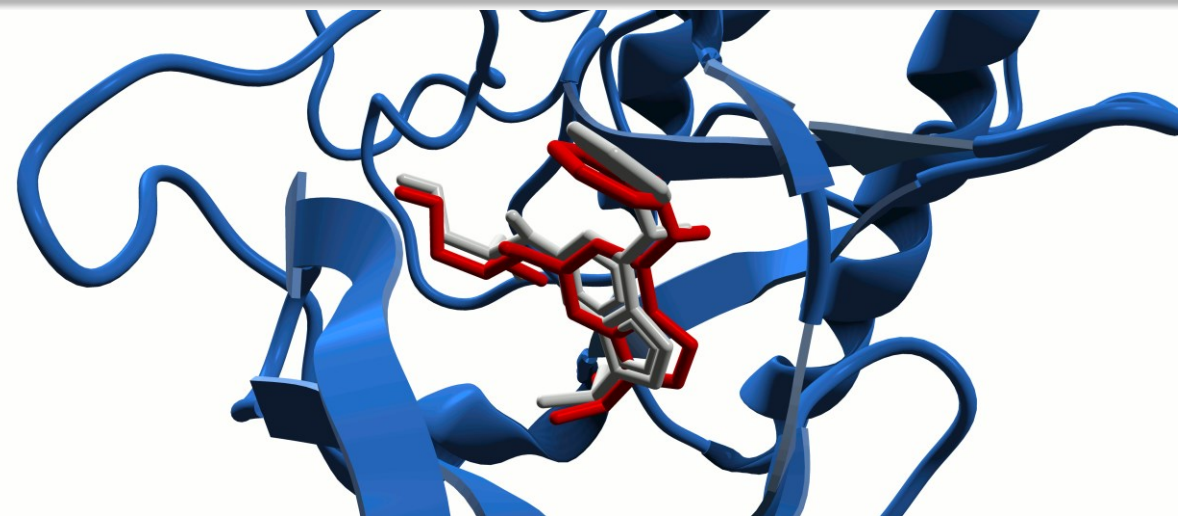
[walter@azevedolab.net](mailto:walter@azevedolab.net)

[Biography 01](#) ♥

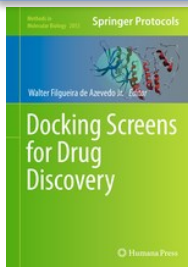
[Biography 02](#) ♥

[Biography 03](#) ♥

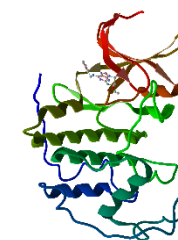
[Biography 04](#) ♥



Frontiers Section Editor (Bioinformatics and Biophysics) for the [Current Drug Targets](#) ISSN: 1873-5592  
 Section Editor (Bioinformatics in Drug Design and Discovery) for the [Current Medicinal Chemistry](#) ISSN: 1875-533X






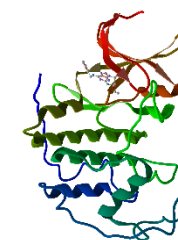
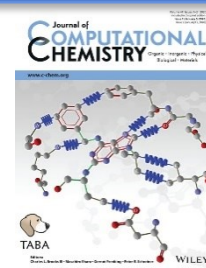
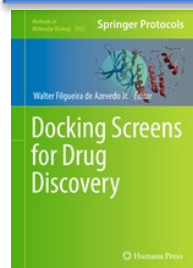
PROUD  
to be  
a Springer Author  
Read a free  
preview!





# Conteúdo

- [Resumo](#)
  - [Aprendizado de Máquina Supervisionado](#)
  - [Sistema Proteína-Fármaco](#)
  - [Quinase Dependente de Ciclina 2](#)
  - [Conjunto de Dados](#)
  - [Análise Estatística](#)
  - [Modelo de Regressão](#)
  - [Análise do Poder de Previsão do Modelo de Aprendizado de Máquina](#)
  - [Desafio 01](#)
  - [Referências](#)
- 
- 
- 



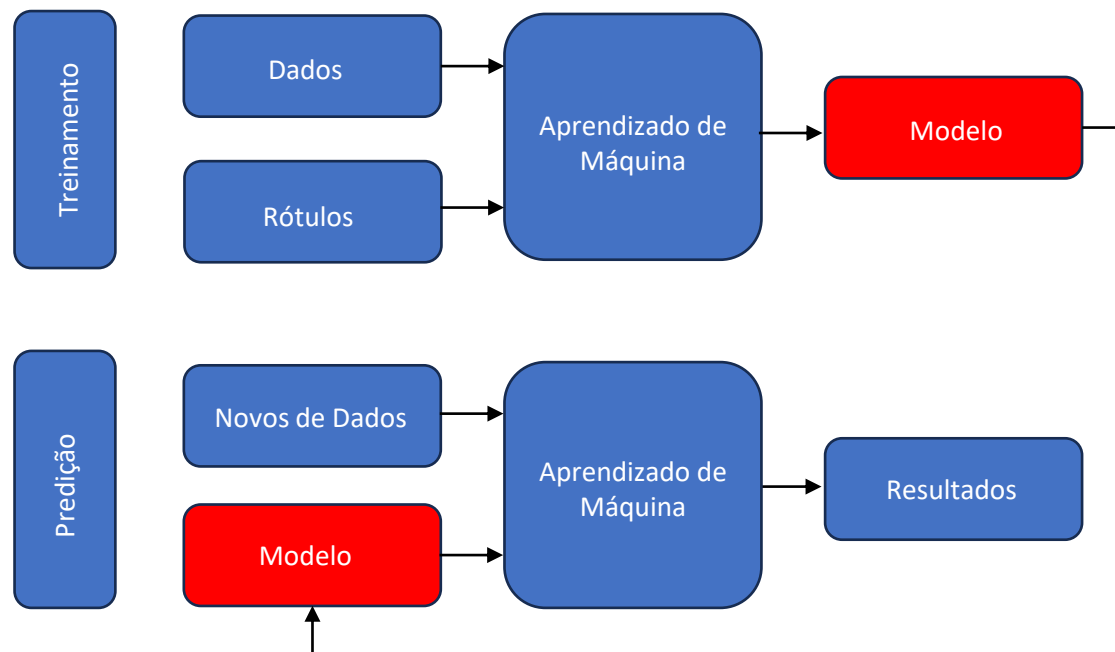
## Resumo

Nesta aula veremos como usar as características (*features*) da interação entre ligantes e uma proteína alvo para gerarmos um modelo de aprendizado de máquina supervisionado que prevê se o ligante inibe a proteína. Usando a analogia da chave e fechadura, o nosso modelo será treinado com diferentes chaves que se ligam na mesma fechadura. O modelo de aprendizado de máquina supervisionado irá prever se uma nova chave se encaixa ou não na fechadura. Na analogia o “encaixar da chave” representa a inibição da proteína. Essa abordagem tem grande potencial na descoberta de fármacos, pois uma vez treinado um modelo de aprendizado de máquina, podemos testar milhões de moléculas e focar os testes pré-clínicos e clínicos nas moléculas que tiveram melhores resultados previstos pelo modelo de aprendizado de máquina. O foco da aula é no estudo de inibidores da proteína quinase dependente de ciclina 2 (CDK2), onde iremos gerar modelos de regressão linear múltipla.

Palavras-chave: aprendizado de máquina, *machine learning*, modelo de aprendizado de máquina, biologia de sistemas, bioinformática, descoberta de fármacos, interação proteína-ligante, modelo chave-fechadura, quinase dependente de ciclina, CDK2, *Molegro Data Modeller*, regressão linear, regressão linear múltipla, predição, concentração inibitória a 50 %,  $IC_{50}$ .

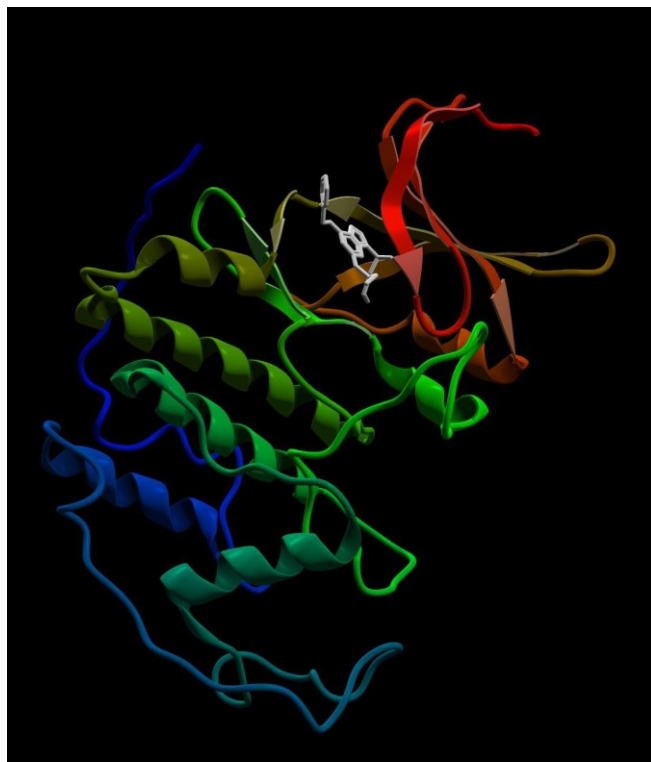
## Aprendizado de Máquina Supervisionado

O diagrama esquemático abaixo ilustra as principais etapas para a construção de um modelo a partir de abordagens de aprendizado de máquina supervisionado. Temos de entrada um conjunto de dados com os rótulos (valores experimentais). Esses dados são inseridos no algoritmo de aprendizado de máquina supervisionado que construirá um modelo. Este modelo é uma abstração matemática que é capaz de ler as variáveis independentes (*features*) e prever o valor da variável dependente (*target*). Usamos dados não empregados na construção do modelo (Novos Dados) para testar ou simplesmente realizar novas previsões ([de Azevedo, 2021](#)).



## Sistema Proteína-Fármaco

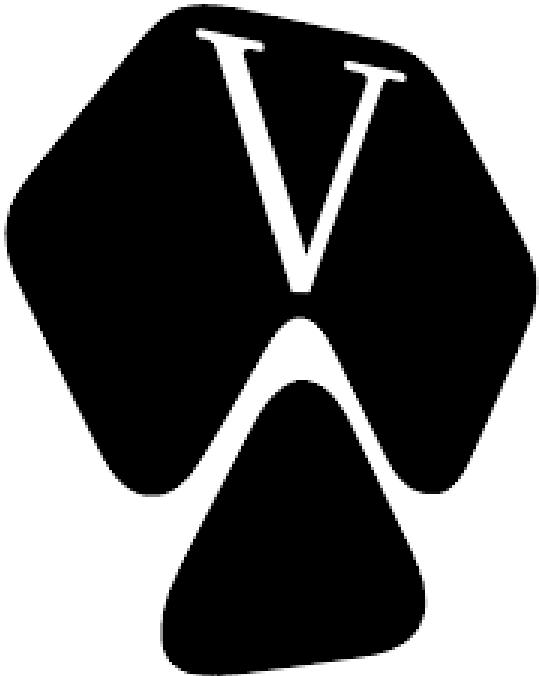
A forma mais simples de olhar a interação entre um fármaco e uma proteína é considerarmos o fármaco como uma chave e o bolsão de ligação da proteína como a fechadura. Chamamos de docagem molecular o processo de determinar computacionalmente a posição da chave na fechadura.



Analogia do sistema proteína-fármaco com o modelo chave-fechadura.

## Sistema Proteína-Fármaco

Todo programa de docagem molecular apresenta duas metodologias computacionais. Uma para calcular a energia de interação do fármaco com a proteína (energia de ligação) e outra para variar a posição do ligante na busca da posição de menor energia (algoritmo busca) ([Heberlé & de Azevedo, 2011](#); [de Azevedo & Dias, 2008](#)). Há diversas abordagens computacionais para determinarmos a energia da interação fármaco proteína, uma das mais usadas é por meio da função escore do programa AutoDock Vina ([Trott & Olson, 2010](#); [Quiroga & Villarreal, 2016](#); [Eberhardt et al., 2021](#)). A função escore do AutoDock Vina está calibrada para fornecer um valor previsto de afinidade proteína-fármaco em kcal/mol.



Logo do programa para simulação de docagem molecular AutoDock Vina ([Trott & Olson, 2010](#); [Eberhardt et al., 2021](#)).

## Sistema Proteína-Fármaco

A função escore do AutoDock Vina ([Trott & Olson, 2010](#); [Eberhardt et al., 2021](#)) é uma equação que tem a seguinte expressão:

$$c = \sum_{i < j} h_{t_i t_j}(d_{i,j})$$

Na somatória acima, temos todos os átomos  $i, j$  exceto aqueles envolvendo interação entre os átomos 1 e 4. O termo  $h_{t_i t_j}$  é a somatória ponderada para os pares de interações: *gauss1*, *gauss2*, *repulsion*, *hydrophobic*, *Hbond* e *torsional*. O *torsional* é baseado nos ângulos de torção do ligante. As outras variáveis independentes da equação acima têm as seguintes expressões.

$$gauss1(d_{i,j}) = e^{-(d_{i,j}/0.5\text{\AA})^2}$$

$$Hbond(d_{i,j}) = \begin{cases} 1, & \text{if } d_{i,j} < -0.7\text{\AA} \\ \text{interpolated if} & -0.7\text{\AA} < d_{i,j} < 0 \\ 0, & \text{if } d_{i,j} > 0 \end{cases}$$

$$gauss2(d_{i,j}) = e^{-((d_{i,j}-3\text{\AA})/2\text{\AA})^2}$$

$$repulsion(d_{i,j}) = \begin{cases} d_{i,j}^2, & \text{if } d_{i,j} < 0 \\ 0, & \text{if } d_{i,j} \geq 0 \end{cases}$$

$$hydrophobic(d_{i,j}) = \begin{cases} 1, & \text{if } d_{i,j} < 0.5\text{\AA} \\ \text{interpolated if} & 0.5\text{\AA} < d_{i,j} < 1.5\text{\AA} \\ 0, & \text{if } d_{i,j} > 1.5\text{\AA} \end{cases}$$

Nas equações  $d_{i,j} = r_{i,j} - R_{t_i} - R_{t_j}$  onde  $r_{i,j}$  é a distância entre dois átomos e  $R_t$  o raio de van der Waals do átomo.

## Quinase Dependente de Ciclina 2

Há centenas de estruturas cristalográficas da quinase dependente de ciclina 2 (*cyclin-dependent kinase 2*) (CDK2) ([de Azevedo, 2022](#)) depositadas no [Protein Data Bank](#). Neste estudo, usaremos as estruturas cristalográficas da CDK2 em complexo com inibidores com dados de  $IC_{50}$ . O parâmetro  $IC_{50}$  indica a concentração inibitória a 50 %, ou seja, é a concentração molar do inibidor necessária para eliminar 50 % da atividade enzimática da CDK2. Assim, quanto menor o  $IC_{50}$ , mais eficaz é o inibidor. Espera-se que do ponto de vista energético, quanto menor for a energia de ligação do inibidor, menor será o  $IC_{50}$ . Abaixo temos a estrutura cristalográfica da CDK2 em complexo com Roscovitine ([de Azevedo et al., 1997](#)) que tem um  $IC_{50} = 400 \text{ nM}$ .



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

Ligands 1 Unique

ID	Chains	Name / Formula / InChI Key	2D Diagram	3D Interactions
RRC <a href="#">Query on RRC</a>	B [auth A]	R-ROSCOVITINE C <sub>19</sub> H <sub>26</sub> N <sub>6</sub> O BTIHMVBBUGXLCJ-OAHLLOKOSA-N		<a href="#">Ligand Interaction</a>

Download Ideal Coordinates CCD File  
Download Instance Coordinates

Binding Affinity Annotations

ID	Source	Binding Affinity
RRC		Ki: 250 (nM) from 1 assay(s)
	BindingDB: 2A4L	Kd: min: 2900, max: 3400 (nM) from 2 assay(s)
		IC50: min: 0.1, max: 708 (nM) from 19 assay(s)
	Binding MOAD: 2A4L	IC50: 400 (nM) from 1 assay(s)
	PDBBind: 2A4L	IC50: 400 (nM) from 1 assay(s)



## Conjunto de Dados

Usaremos dados da CDK2 com  $IC_{50}$  (arquivo *CDK2\_IC50\_2022.csv*) preparado pelo SAnDReS ([Xavier et al., 2016](#)). Esse conjunto está numa planilha do Excel (formato CSV) e será analisado com o Molegro Data Modeller (MDM) ([Thomsen & Christensen, 2006](#); [Bitencourt-Ferreira & de Azevedo, 2019](#)). Abaixo temos as primeiras linhas do arquivo visto no programa MDM.

	PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128	
2	3WBL	PDY	A	302	2	1	2.3e-05	-4.63827	4.63827	4.027	6	-0.003	-0.0001	36.7885	
3	2VTH	LZ2	A	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267	
4	3IGG	EFQ	A	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014	
5	2VTA	LZ1	A	1301	2	1	0.000185	-3.73283	3.73283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467	
6	2VTP	LZ9	A	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188	
7	2VTO	LZ8	A	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992	
8	2VTN	LZ7	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516	
9	2VTM	LZM	A	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709	
10	3R8V	Z62	A	473	1.9	1	2.9e-06	-5.5376	5.5376	7.309	7	0.001	4.16667e-05	22.271	
11	2VTL	LZ5	A	1299	2	1	9.7e-05	-4.01323	4.01323	6.347	2	2.42861e-17	1.51788e-18	39.8371	
12	3R8U	Z31	A	465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014	
13	2VTI	LZ3	A	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345	
14	4LYN	1YG	A	301	2	1	6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956	
15	3UNJ	0BX	A	299	1.9	1	1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304	
16	3QTZ	X42	A	453	2	1	5e-08	-7.30103	7.30103	4.305	7	0	0	20.0535	
17	3QTX	X43	A	299	1.9	1	7e-08	-7.1549	7.1549	4.563	8	0.002	6.06061e-05	22.0461	
18	3QTW	X3A	A	451	1.8	1	6.5e-07	-6.18709	6.18709	7.581	5	0.002	8.33333e-05	27.2133	
19	4EZ3	0S0	A	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56	
20	3TIY	TIY	A	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956	
21	3QTU	X44	A	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117	
22	1PXL	CK4	A	500	2.5	0.59	9e-07	-6.04576	6.04576	8.826	4	0.002	8e-05	57.5554	

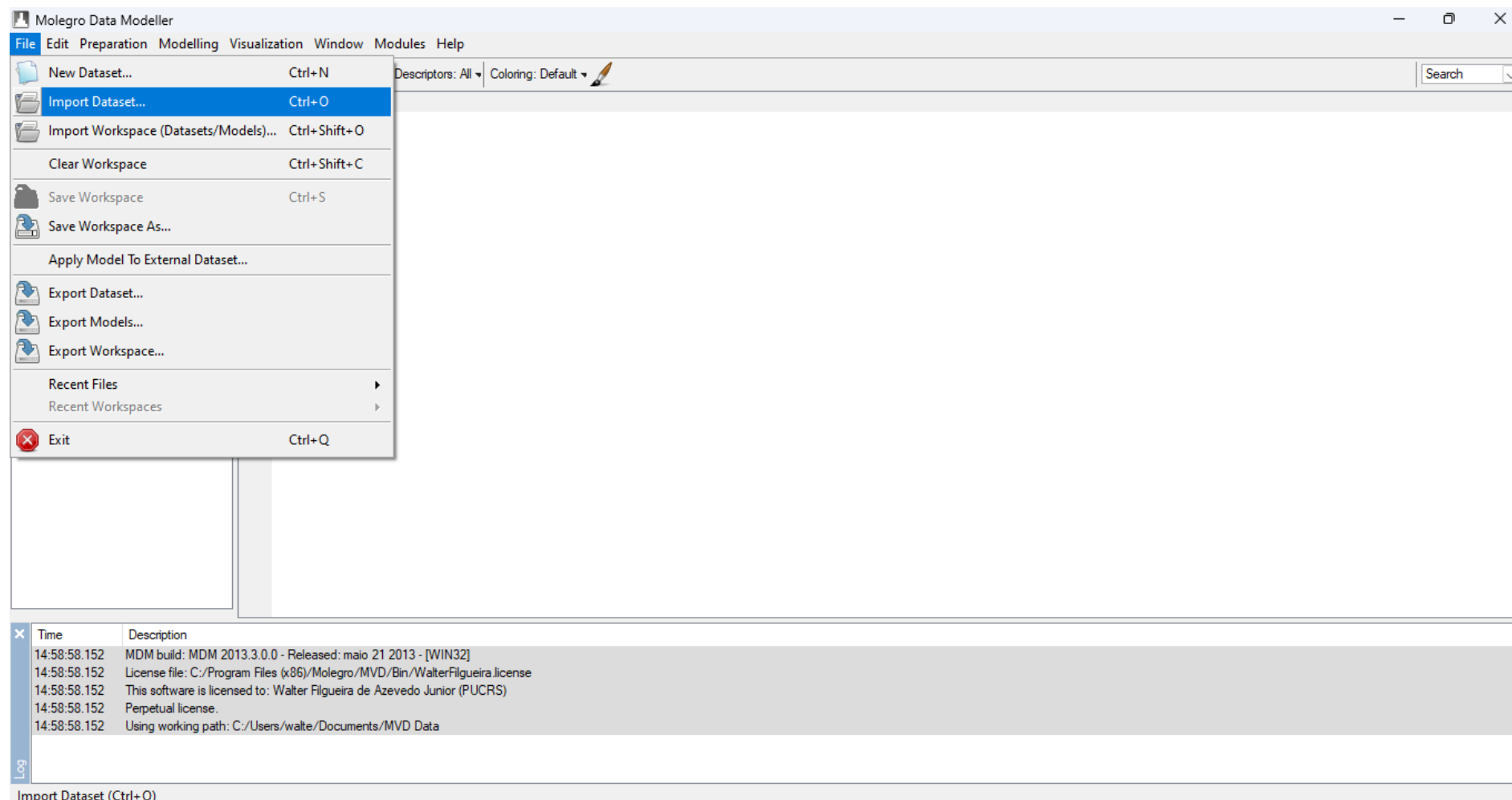
## Conjunto de Dados

Abaixo temos as colunas com os pares de interações da função escore do AutoDock Vina ([Trott & Olson, 2010](#); [Eberhardt et al., 2021](#)). Temos um total de 104 estruturas da CDK2 neste conjunto de dados. Na análise e modelagem dos dados usaremos o  $\log(IC_{50})$  como variável alvo (variável dependente). O motivo é que o  $IC_{50}$  tem uma faixa de variação grande e seu  $\log$  restringe a variação.

	r B-factor(A2)	B-factor ratio (Ligand/	C	N	O	P	S	F	Cl	Br	Affinity(kcal/mol)	Gauss 1	Gauss 2	Repulsion	Hydrophobic	Hydrogen	Torsional
1	3.9514	0.928173	18	4	3	0	0	0	0	0	-6.231	84.495	1376.13	5.707	33.609	2.533	0.845098
2	2.5843	1.62894	20	6	1	0	0	0	0	0	-7.306	78.552	1434.46	4.303	28.646	3.144	0.845098
3	3.8288	1.24529	10	1	3	0	1	0	0	0	-6.302	37.352	909.602	2.101	44.542	1.534	0.60206
4	3.2733	0.715329	16	3	3	0	0	0	0	0	-7.088	78.623	1183.91	3.513	36.453	2.993	0.845098
5	3.2011	0.87112	7	2	0	0	0	0	0	0	-5.133	32.84	586.044	1.601	30.253	1.706	0
6	3.5456	1.29183	17	4	2	0	0	3	0	0	-6.591	96.676	1451.69	7.34	34.674	2.873	0.69897
7	3.4637	1.33852	17	4	2	0	0	1	0	0	-7.034	80.177	1369.72	5.21	31.383	2.705	0.69897
8	3.6231	1.12502	12	4	2	0	0	1	0	0	-5.58	70.945	1090.2	5.076	20.681	2.705	0.60206
9	6.406	1.43852	7	4	0	0	0	0	0	0	-2.303	42.791	705.321	5.114	10.588	1.613	0.30103
10	4.4294	0.911644	13	4	3	0	1	0	0	0	-6.771	79.085	1186.29	3.279	16.911	3.718	0.90309
11	2.6262	1.22102	10	3	1	0	0	0	0	0	-4.706	56.781	792.541	4.097	19.025	2.828	0.477121
12	4.8021	0.979812	17	3	1	0	1	0	0	0	-6.803	82.841	1219.58	4.564	32.545	3.494	0.845098
13	7.5788	0.751829	14	4	3	0	1	0	0	0	-8.409	87.676	1208.97	3.969	51.119	3.251	0.69897
14	3.2829	1.33634	20	3	2	0	2	0	0	0	-7.232	89.489	1455.95	3.979	29.886	2.279	0.90309
15	7.9553	0.734401	17	4	2	0	0	0	0	0	-9.241	89.818	1342.89	2.274	57.043	2.603	0.845098
16	2.2616	0.62159	16	4	3	0	2	1	0	0	-8.633	101.423	1481.02	4.249	40.72	3.494	0.90309
17	2.0297	0.6883	16	5	5	0	2	0	0	0	-6.573	105.27	1578.48	7.348	26.939	3.547	0.954243
18	4.0268	0.799762	15	4	1	0	1	0	0	0	-7.416	76.285	1235.89	3.53	27.792	3.145	0.778151
19	0.0347	1.01749	11	4	4	0	1	0	0	0	-6.324	75.96	1112.94	3.825	22.646	2.397	0.778151
20	3.3832	1.35777	11	0	5	0	0	0	0	0	-5.962	58.098	1024.43	3.474	26.032	1.599	0.69897
21	3.5116	0.740392	16	5	5	0	3	0	0	0	-9.066	113.347	1627.35	4.992	39.47	4.909	1
22	3.5623	1.26323	16	4	0	0	1	3	0	0	-5.858	78.227	1295.88	5.81	30.455	1.972	0.69897

## Conjunto de Dados

Considerando que você tem o MDM instalado e aberto no seu computador, iremos carregar o arquivo *CDK2\_IC50\_2022.csv* no MDM. Para carregar um arquivo, clique em *File->Import Dataset...*



## Conjunto de Dados

Na nova janela, vá na pasta onde o arquivo *CDK2\_IC50\_2022.csv* se encontra. Selecione o arquivo e clique em Abrir.

The screenshot shows the Molegro Data Modeller application window. A 'Choose file' dialog box is open, displaying the file selection process. The dialog shows the current directory as 'Aula\_14 > Data'. A file named 'CDK2\_IC50\_2022' is selected, with a modification date of 28/10/2022 13:26 and a type of 'CSV File (Open V...'. The file name field contains 'CDK2\_IC50\_2022' and the file type dropdown is set to 'Text CSV (\*.csv;\*)'. The 'Abrir' (Open) button is highlighted.

Time	Description
14:58:58.152	MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
14:58:58.152	License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFigueira.license
14:58:58.152	This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
14:58:58.152	Perpetual license.
14:58:58.152	Using working path: C:/Users/walte/Documents/MVD Data

## Conjunto de Dados

Agora clique na opção *Import*.

The screenshot shows the Molegro Data Modeller interface with the 'Import Dataset from CSV' dialog box open. The dialog has two tabs: 'Import Settings' and 'Filtering'. The 'Import Settings' tab is active, showing a 'Dataset preview' table with 11 rows. The first row is selected. Below the table, there are options for 'Coloring' (textual columns are gray, numeric columns are white, and ignored columns are orange), 'Choose column format' (set to 'Text'), 'Choose CSV settings used during import' (Text encoding: UTF-8, Column separator type: Automatic detection, Use first row as header: checked, Create ID column: unchecked), and 'Workspace' (Replace or add to workspace: Add to current workspace). The 'Import' button is highlighted with a red dashed border.

	PDB	Ligand	Chain	Number	Resolution(A)	Ligar
1	3IG7	EFP	A	999	1.8	
2	3WBL	PDY	A	302	2	
3	2VTH	LZ2	A	1300	1.9	
4	3IGG	EFQ	A	999	1.8	
5	2VTA	LZ1	A	1301	2	
6	2VTP	LZ9	A	1299	2.1	
7	2VTO	LZ8	A	1299	2.1	
8	2VTN	LZ7	A	1299	2.2	
9	2VTM	LZM	A	1299	2.2	
10	3R8V	Z62	A	473	1.9	
11	2VTL	LZ5	A	1299	2	

## Conjunto de Dados

O MDM carrega a planilha de dados e a mostra parcialmente na tela, visto que temos um número grande de colunas (31) e linhas (104). Em aprendizado de máquina supervisionado é comum dividir o conjunto de dados em dois subconjuntos: o conjunto de treinamento e o conjunto de teste.

Molegro Data Modeller

File Edit Preparation Modelling Visualization Window Modules Help

Descriptors: All | Coloring: Default

Workspace Explorer

- Workspace: Unnamed
  - Datasets [1]
    - CDK2\_IC50\_2022

Properties

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128
2	3WBL	PDY	A	302	2	1	2.3e-05	-4.63827	4.63827	4.027	6	-0.003	-0.0001	36.7885
3	2VTH	LZ2	A	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267
4	3IGG	EFQ	A	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014
5	2VTA	LZ1	A	1301	2	1	0.000185	-3.73283	3.73283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467
6	2VTP	LZ9	A	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188
7	2VTO	LZ8	A	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992
8	2VTN	LZ7	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
9	2VTM	LZM	A	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709
10	3R8V	Z62	A	473	1.9	1	2.9e-06	-5.5376	5.5376	7.309	7	0.001	4.16667e-05	22.271
11	2VTL	LZ5	A	1299	2	1	9.7e-05	-4.01323	4.01323	6.347	2	2.42861e-17	1.51788e-18	39.8371
12	3R8U	Z31	A	465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014
13	2VTI	LZ3	A	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345
14	4LYN	1YG	A	301	2	1	6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956
15	3UNJ	0BX	A	299	1.9	1	1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304
16	3QTZ	X42	A	453	2	1	5e-08	-7.30103	7.30103	4.305	7	0	0	20.0535
17	3QTX	X43	A	299	1.9	1	7e-08	-7.1549	7.1549	4.563	8	0.002	6.06061e-05	22.0461
18	3QTW	X3A	A	451	1.8	1	6.5e-07	-6.18709	6.18709	7.581	5	0.002	8.33333e-05	27.2133
19	4EZ3	0S0	A	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56
20	3TIY	TIY	A	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956
21	3QTU	X44	A	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117
22	1PXL	CK4	A	500	2.5	0.59	9e-07	-6.04576	6.04576	8.826	4	0.002	8e-05	57.5554

Log

Time	Description
14:58:58.152	MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
14:58:58.152	License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFilgueira.license
14:58:58.152	This software is licensed to: Walter Filgueira de Azevedo Junior (PUCRS)
14:58:58.152	Perpetual license.
14:58:58.152	Using working path: C:/Users/walte/Documents/MVD Data

## Conjunto de Dados

Para dividir o conjunto de dados, clique em *Preparation->Create Subset using Random Selection->Write Subset IDs to 'Subset' Column...*

The screenshot shows the Molegro Data Modeller interface. The 'Preparation' menu is open, and the 'Create Subset using Random Selection' option is selected. A sub-menu is visible, showing the option 'Write Subset IDs to 'Subset' Column...' which is highlighted. The main window displays a table of data with columns: Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. The table contains 22 rows of data.

Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F			
A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128				
A	302	2	1	2.3e-05	-4.63827	4.63827	4.027	6	-0.003	-0.0001	36.7885				
A	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267				
A	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014				
A	1301	2	1	0.000185	-3.73283	3.73283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467				
A	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188				
A	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992				
A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516				
A	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709				
A	473	1.9	1	2.9e-06	-5.5376	5.5376	7.309	7	0.001	4.16667e-05	22.271				
A	1299	2	1	9.7e-05	-4.01323	4.01323	6.347	2	2.42861e-17	1.51788e-18	39.8371				
A	465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014				
A	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345				
					6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956			
					1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304			
					5e-08	-7.30103	7.30103	4.305	7	0	0	20.0535			
17	3QTX	X43	A	299	1.9	1	7e-08	-7.1549	7.1549	4.563	8	0.002	6.06061e-05	22.0461	
18	3QTW	X3A	A	451	1.8	1	6.5e-07	-6.18709	6.18709	7.581	5	0.002	8.33333e-05	27.2133	
19	4EZ3	0S0	A	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56	
20	3TIY	TIY	A	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956	
21	3QTU	X44	A	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117	
22	1PXL	CK4	A	500	2.5	0.59	9e-07	-6.04576	6.04576	8.826	4	0.002	8e-05	57.5554	

The Log window at the bottom shows the following entries:

- 14:58:58.152 MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
- 14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFigueira.license
- 14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
- 14:58:58.152 Perpetual license.
- 14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data

## Conjunto de Dados

Selecione 70 % dos dados para o conjunto de treinamento, como indicado abaixo. Clique em OK.

Molegro Data Modeller

File Edit Preparation Modelling Visualization Window Modules Help

Descriptors: All | Coloring: Default

Workspace Explorer

- Workspace: Unnamed
  - Datasets [1]
    - CDK2\_IC50\_2022

Properties

Property	Value

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128
2	3WBL	PDY	A	302	2	1	2.3e-05	-4.63827	4.63827	4.027	6	-0.003	-0.0001	36.7885
3	2VTH	LZ2	A	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267
4	3IGG	EFQ	A	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014
5	2VTA	LZ1	A	1301	2	1	0.000185	-3.73283	3.73283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467
6	2VTP	LZ9	A	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188
7	2VTO	LZ8	A	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992
8	2VTN	LZ7	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
9	2VTM	LZM	A	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709
10	3R8V	Z62	A	473				5.5376	7.309	7	0.001	4.16667e-05	22.271	
11	2VTL	LZ5	A	1299				4.01323	6.347	2	2.42861e-17	1.51788e-18	39.8371	
12	3R8U	Z31	A	465				5.30103	1.128	6	0.002	8e-05	24.3014	
13	2VTI	LZ3	A	1299				6.18046	1.509	4	0.001	3.84615e-05	20.7345	
14	4LYN	1YG	A	301				7.22185	4.71	7	-0.001	-3.57143e-05	37.7956	
15	3UNJ	0BX	A	299				4.95861	1.421	6	0.003	0.000115385	20.5304	
16	3QTZ	X42	A	453				7.30103	4.305	7	0	0	20.0535	
17	3QTX	X43	A	299				7.1549	4.563	8	0.002	6.06061e-05	22.0461	
18	3QTW	X3A	A	451	1.8	1	6.5e-07	-6.18709	6.18709	7.581	5	0.002	8.33333e-05	27.2133
19	4EZ3	0S0	A	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56
20	3TIY	TIY	A	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956
21	3QTU	X44	A	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117
22	1PXL	CK4	A	500	2.5	0.59	9e-07	-6.04576	6.04576	8.826	4	0.002	8e-05	57.5554

Create Subset using Random Selection

Choose number of records in subset:

Number of records: 72

Percentage: 70.00

OK Cancel

Log

Time	Description
14:58:58.152	MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
14:58:58.152	License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFigueira.license
14:58:58.152	This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
14:58:58.152	Perpetual license.
14:58:58.152	Using working path: C:/Users/walte/Documents/MVD Data



## Conjunto de Dados

Agora você clica com o botão direito do mouse no *Dataset CDK2\_IC50\_2022* que está no *Workspace Explorer*. Em seguida, você seleciona a opção *Split Dataset (Using 'Subset' Column)*.

The screenshot shows the Molegro Data Modeller interface. The main window displays a table of datasets with columns: PDB, Ligand, Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. A context menu is open over the 'CDK2\_IC50\_2022' dataset, with the option 'Split Dataset (Using 'Subset' Column)' selected. The 'Properties' panel on the left shows details for the selected dataset, including Name, Records, Columns, and Original Filename. The 'Log' panel at the bottom shows system messages.

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128
2	3WBL	PDY	A	302	2	1	2.3e-05	-4.63827	4.63827	4.027	6	-0.003	-0.0001	36.7885
3	2VTH	LZ2	A	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267
				999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014
				1301	2	1	0.000185	-3.73283	3.73283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467
				1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188
				1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992
				1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
				1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709
				473	1.9	1	2.9e-06	-5.5376	5.5376	7.309	7	0.001	4.16667e-05	22.271
				1299	2	1	9.7e-05	-4.01323	4.01323	6.347	2	2.42861e-17	1.51788e-18	39.8371
				465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014
13	2VTI	LZ3	A	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345
14	4LYN	1YG	A	301	2	1	6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956
15	3UNJ	0BX	A	299	1.9	1	1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304
16	3QTZ	X42	A	453	2	1	5e-08	-7.30103	7.30103	4.305	7	0	0	20.0535
17	3QTX	X43	A	299	1.9	1	7e-08	-7.1549	7.1549	4.563	8	0.002	6.06061e-05	22.0461
18	3QTW	X3A	A	451	1.8	1	6.5e-07	-6.18709	6.18709	7.581	5	0.002	8.33333e-05	27.2133
19	4EZ3	0S0	A	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56
20	3TIY	TIY	A	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956
21	3QTU	X44	A	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117
22	1PXL	CK4	A	500	2.5	0.59	9e-07	-6.04576	6.04576	8.826	4	0.002	8e-05	57.5554

**Dataset Context Menu:**

- Export Dataset...
- Rename Dataset...
- Clone Dataset (Creates Copy)
- Split Dataset (Using 'Subset' Column)**
- Extract One Subset (Using 'Subset' Column)
- Revert to Original Sorting Order
- Delete Dataset From Workspace

**Properties Panel:**

Property	Value
Name	CDK2_IC50_2...
Records	104
Columns (total)	32
Numerical Descri...	29
Predictions / Clas...	0
Original Filename	C:/Users/walte/D...

**Log Panel:**

Time	Description
14:58:58.152	MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
14:58:58.152	License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFilgueira.license
14:58:58.152	This software is licensed to: Walter Filgueira de Azevedo Junior (PUCRS)
14:58:58.152	Perpetual license.
14:58:58.152	Using working path: C:/Users/walte/Documents/MVD Data
15:36:26.743	Subset created with Subset Id: 1 (records: 72)

## Conjunto de Dados

O MDM cria dois novos conjuntos de dados e mostra no *Workspace Explorer*. Foram criados os conjuntos: *CDK2\_IC50\_2022\_0* e *CDK2\_IC50\_2022\_1*. O primeiro tem 30 % dos dados e será usado como conjunto de teste e o segundo tem 70 % dos dados e será usado como conjunto de treinamento.

The screenshot displays the Molegro Data Modeller interface. The main window shows a table of protein-ligand complexes with the following columns: PDB, Ligand, Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. The table lists 22 entries, with the first entry (PDB: 3IG7, Ligand: EFP) highlighted. The left sidebar shows the Workspace Explorer with a tree view containing 'Workspace: Unnamed' and 'Datasets [3]' (CDK2\_IC50\_2022, CDK2\_IC50\_2022\_0, CDK2\_IC50\_2022\_1). The bottom window shows a log of system events.

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128
2	3WBL	PDY	A	302	2	1	2.3e-05	-4.63827	4.63827	4.027	6	-0.003	-0.0001	36.7885
3	2VTH	LZ2	A	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267
4	3IGG	EFQ	A	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014
5	2VTA	LZ1	A	1301	2	1	0.000185	-3.73283	3.73283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467
6	2VTP	LZ9	A	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188
7	2VTO	LZ8	A	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992
8	2VTN	LZ7	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
9	2VTM	LZM	A	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709
10	3R8V	Z62	A	473	1.9	1	2.9e-06	-5.5376	5.5376	7.309	7	0.001	4.16667e-05	22.271
11	2VTL	LZ5	A	1299	2	1	9.7e-05	-4.01323	4.01323	6.347	2	2.42861e-17	1.51788e-18	39.8371
12	3R8U	Z31	A	465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014
13	2VTI	LZ3	A	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345
14	4LYN	1YG	A	301	2	1	6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956
15	3UNJ	OBX	A	299	1.9	1	1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304
16	3QTZ	X42	A	453	2	1	5e-08	-7.30103	7.30103	4.305	7	0	0	20.0535
17	3QTX	X43	A	299	1.9	1	7e-08	-7.1549	7.1549	4.563	8	0.002	6.06061e-05	22.0461
18	3QTW	X3A	A	451	1.8	1	6.5e-07	-6.18709	6.18709	7.581	5	0.002	8.33333e-05	27.2133
19	4EZ3	OS0	A	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56
20	3TIY	TIY	A	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956
21	3QTU	X44	A	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117
22	1PXL	CK4	A	500	2.5	0.59	9e-07	-6.04576	6.04576	8.826	4	0.002	8e-05	57.5554

Log window content:

```

x Time Description
14:58:58.152 MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFigueira.license
14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
14:58:58.152 Perpetual license.
14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
15:36:26.743 Subset created with Subset Id: 1 (records: 72)
15:41:23.678 Creating new dataset 'CDK2_IC50_2022_0' from subset: 0
15:41:23.678 Creating new dataset 'CDK2_IC50_2022_1' from subset: 1
  
```

## Análise Estatística

Podemos aferir a métricas (e.g., raiz quadrada do erro médio, *RMSE*) da função do AutoDock Vina (*Affinity*) para o  $\log(IC_{50})$  para os conjuntos de treinamento e de teste. Para isso, selecione o conjunto no *Workspace Explorer* clicando no conjunto escolhido. Selecionamos o conjunto *CDK2\_IC50\_2022\_1* (conjunto de treinamento).

Molegro Data Modeller

File Edit Preparation Modelling Visualization Window Modules Help

Workspace Explorer

Items

- Workspace: Unnamed
  - Datasets [3]
    - CDK2\_IC50\_2022
    - CDK2\_IC50\_2022\_0
    - CDK2\_IC50\_2022\_1

Properties

Property	Value
Name	CDK2_IC50_2022...
Records	72
Columns (total)	32
Numerical Descri...	29
Predictions / Clas...	0
Original Filename	C:/Users/walte/D...

	PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128	
2	2VTH	LZ2	A	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267	
3	3IGG	EFQ	A	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014	
4	2VTP	LZ9	A	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188	
5	2VTO	LZ8	A	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992	
6	2VTN	LZ7	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516	
7	2VTM	LZM	A	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709	
8	3R8U	Z31	A	465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014	
9	2VTI	LZ3	A	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345	
10	4LYN	1YG	A	301	2	1	6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956	
11	3UNJ	0BX	A	299	1.9	1	1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304	
12	4EZ3	0S0	A	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56	
13	3TIY	TIY	A	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956	
14	3QTU	X44	A	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117	
15	3QTR	X36	A	497	1.8	1	9.3e-07	-6.03152	6.03152	5.165	5	-1.56125e-17	-6.50521e-19	22.3224	
16	3QU0	X40	A	454	1.9	1	2e-08	-7.69897	7.69897	0.999	7	0.003	0.0001	20.7888	
17	3TI1	B49	A	299	1.9	1	0.00013	-3.88606	3.88606	0.611	7	-3.46945e-18	-1.0842e-19	34.0538	
18	1PXK	CK3	A	500	2.8	0.69	2.2e-06	-5.65758	5.65758	8.622	3	1.11022e-16	5.84328e-18	68.0518	
19	2W1H	LOF	A	1299	2.1	1	5.2e-08	-7.284	7.284	5.296	3	-0.001	-3.84615e-05	27.5857	
20	5D1J	56H	A	4000	1.8	1	4.8e-08	-7.31876	7.31876	4.845	6	-0.001	-3.7037e-05	35.5056	
21	3EZV	EZV	A	300	1.9	1	1.04e-06	-5.98297	5.98297	4.31	4	-0.001	-2.77778e-05	50.6353	
22	2W05	FRT	A	1299	1.9	1	1e-09	-9	9	1.117	9	0.001	3.125e-05	33.7253	

Log

Time	Description
14:58:58.152	MUM build: MUM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
14:58:58.152	License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFigueira.license
14:58:58.152	This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
14:58:58.152	Perpetual license.
14:58:58.152	Using working path: C:/Users/walte/Documents/MVD Data
15:36:26.743	Subset created with Subset Id: 1 (records: 72)
15:41:23.678	Creating new dataset 'CDK2_IC50_2022_0' from subset: 0
15:41:23.678	Creating new dataset 'CDK2_IC50_2022_1' from subset: 1

# Análise Estatística

Clicamos em *Modelling* -> *Bivariate Statistics...*

The screenshot shows the Molegro Data Modeller interface. The 'Modelling' menu is open, highlighting 'Bivariate Statistics...'. The main window displays a table with the following columns: Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. The table contains 22 rows of data. A context menu is also visible over the table, listing various statistical and machine learning methods such as Multiple Linear Regression, Support Vector Regression, KNN Classification, etc.

	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F		
	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128			
	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267			
	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014			
	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188			
	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992			
	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516			
	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709			
	465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014			
	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345			
	301	2	1	6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956			
	299	1.9	1	1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304			
	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56			
	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956			
	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117			
	497	1.8	1	9.3e-07	-6.03152	6.03152	5.165	5	-1.56125e-17	-6.50521e-19	22.3224			
	454	1.9	1	2e-08	-7.69897	7.69897	0.999	7	0.003	0.0001	20.7888			
	299	1.9	1	0.00013	-3.88606	3.88606	0.611	7	-3.46945e-18	-1.0842e-19	34.0538			
18	1PXK	CK3	A	500	2.8	0.69	2.2e-06	-5.65758	5.65758	8.622	3	1.11022e-16	5.84328e-18	68.0518
19	2W1H	LOF	A	1299	2.1	1	5.2e-08	-7.284	7.284	5.296	3	-0.001	-3.84615e-05	27.5857
20	5D1J	56H	A	4000	1.8	1	4.8e-08	-7.31876	7.31876	4.845	6	-0.001	-3.7037e-05	35.5056
21	3EZV	EZV	A	300	1.9	1	1.04e-06	-5.98297	5.98297	4.31	4	-0.001	-2.77778e-05	50.6353
22	2W05	FRT	A	1299	1.9	1	1e-09	-9	9	1.117	9	0.001	3.125e-05	33.7253

Log

Time	Description
14:58:58.152	MUM build: MUM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
14:58:58.152	License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFigueira.license
14:58:58.152	This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
14:58:58.152	Perpetual license.
14:58:58.152	Using working path: C:/Users/walte/Documents/MVD Data
15:36:26.743	Subset created with Subset Id: 1 (records: 72)
15:41:23.678	Creating new dataset 'CDK2_IC50_2022_0' from subset: 0
15:41:23.678	Creating new dataset 'CDK2_IC50_2022_1' from subset: 1

## Análise Estatística

Na nova janela, selecionamos duas colunas (variáveis) para a análise estatística.

**Bivariate Statistics**

First column ('Reference'): PDB      Second column ('Prediction'): Ligand

Name	Value
Entries	72
Equal Entries	0
Accuracy (%)	0
Macro-Averaged F (%)	0

Class	TPR(%)	TNR(%)	FPR(%)	FNR(%)	Precision(%)	Recall(%)	F-Measure (%)
SUM							
SUM							

Close      Copy All to Clipboard

PDB	Ligand	Chain
1 3IG7	EFP	A
2 2VTH	LZ2	A
3 3IGG	EFQ	A
4 2VTP	LZ9	A
5 2VTO	LZ8	A
6 2VTN	LZ7	A
7 2VTM	LZM	A
8 3R8U	Z31	A
9 2VTI	LZ3	A
10 4LYN	1YG	A
11 3UNJ	0BX	A
12 4EZ3	0S0	A
13 3TIY	TIY	A
14 3QTU	X44	A
15 3QTR	X36	A
16 3QU0	X40	A
17 3TI1	B49	A
18 1PXK	CK3	A
19 2W1H	LOF	A
20 5D1J	56H	A
21 3EZV	EZV	A
22 2W05	FRT	A

MSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
9.044	6	0.001	3.7037e-05	31.5128	
2.916	3	0	0	42.1267	
0.715	6	-0.001	-4.16667e-05	23.8014	
5.251	4	-0.003	-0.000103448	45.9188	
7.495	4	-0.001	-3.7037e-05	38.0992	
4.009	3	-0.001	-4.54545e-05	43.4516	
2.494	1	5.55112e-17	5.04647e-18	52.3709	
1.128	6	0.002	8e-05	24.3014	
1.509	4	0.001	3.84615e-05	20.7345	
4.71	7	-0.001	-3.57143e-05	37.7956	
1.421	6	0.003	0.000115385	20.5304	
2.539	5	0.001	4.16667e-05	30.56	
3.422	4	-0.002	-0.0001	39.8956	
0.962	9	0.001	2.77778e-05	24.8117	
5.165	5	-1.56125e-17	-6.50521e-19	22.3224	
0.999	7	0.003	0.0001	20.7888	
0.611	7	-3.46945e-18	-1.0842e-19	34.0538	
8.622	3	1.11022e-16	5.84328e-18	68.0518	
5.296	3	-0.001	-3.84615e-05	27.5857	
4.845	6	-0.001	-3.7037e-05	35.5056	
4.31	4	-0.001	-2.77778e-05	50.6353	
1.117	9	0.001	3.125e-05	33.7253	

Time      Description

14:58:58.152      MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]

14:58:58.152      License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFigueira

14:58:58.152      This software is licensed to: Walter Figueira de Azevedo Junior (PUC

14:58:58.152      Perpetual license.

14:58:58.152      Using working path: C:/Users/walte/Documents/MVD Data

15:36:26.743      Subset created with Subset Id: 1 (records: 72)

15:41:23.678      Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0

15:41:23.678      Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1

## Análise Estatística

Selecionamos  $\log(IC_{50})$  para a primeira coluna e  $Affinity(kcal/mol)$  para a segunda. O MDM gera um gráfico de dispersão e realiza uma análise estatística, como mostrado abaixo. Vemos um  $RMSE = 2,22765$  (o MDM chama de  $RMSD$ , mas é o  $RMSE$  previamente discutido). Podemos usar os pares de interações da função do AutoDock Vina para gerar um modelo específico para CDK2. Clique em *Close* para retornar.

The screenshot shows the Molegro Data Modeller interface. A 'Bivariate Statistics' window is open, displaying a scatter plot of  $\log(IC_{50})$  vs  $Affinity(kcal/mol)$ . The plot shows a positive correlation with a red regression line. Below the plot, statistical values are listed:

Name	Value
Pearson Correlation (r)	0.378
Pearson Correlation Squared (r <sup>2</sup> )	0.142542
Spearman Rank Correlation (ρ)	0.301876
Mean Squared Deviation (MSD)	4.96241
Root Mean Squared Deviation (RMSD)	2.22765
Least Square Fit	y = -4.53184 + 0.49309x
Cross Validated Squared CC (q <sup>2</sup> ) [1]	-1.64866

Annotations with arrows point to these values:

- Coeficiente de correlação de Spearman (ρ) → Spearman Rank Correlation (ρ)
- Coeficiente de correlação de Pearson ao quadrado (r<sup>2</sup>) → Pearson Correlation Squared (r<sup>2</sup>)
- RMSE (root mean squared error) (raiz quadrada do erro médio) → Root Mean Squared Deviation (RMSD)
- Coeficiente de determinação (R<sup>2</sup>) → Cross Validated Squared CC (q<sup>2</sup>) [1]

The background shows a table of protein-ligand complexes:

PDB	Ligand	Chain
1	3IG7	EFP A
2	2VTH	LZ2 A
3	3IGG	EFQ A
4	2VTP	LZ9 A
5	2VTO	LZ8 A
6	2VTN	LZ7 A
7	2VTM	LZM A
8	3R8U	Z31 A
9	2VTI	LZ3 A
10	4LYN	1YG A
11	3UNJ	0BX A
12	4EZ3	0S0 A
13	3TIY	TIY A
14	3QTU	X44 A
15	3QTR	X36 A
16	3QII	X40 A
18	1PFX	CK3 A
19	2W1H	L0F A
20	5D1J	56H A
21	3EZV	EZV A
22	2W05	FRT A

Observação: Provavelmente seus resultados são diferentes dos mostrados aqui, pois o MDM fez uma divisão distinta entre conjunto de treinamento e conjunto de teste. Mesmo usando a relação entre teste e treinamento de 30 % e 70 %, o MDM seleciona linhas distintas para cada conjunto. Uma forma de uniformizar os resultados é gerar a divisão (conjunto de treinamento e conjunto de teste) fora do MDM.

## Análise Estatística

Repetindo o processo para o conjunto *CDK2\_IC50\_2022\_0* (conjunto de teste), temos os resultados abaixo, com um *RMSE* = 2,02138. Clique em *Close* para retornar.

Molegro Data Modeller

File Edit Preparation Modelling Visualization Window Modules Help

Workspace Explorer

Items

- Workspace: Unnamed
  - Datasets [3]
    - CDK2\_IC50\_2022
    - CDK2\_IC50\_2022\_0**
    - CDK2\_IC50\_2022\_1

Properties

Property	Value
Name	CDK2_IC50_2022...
Records	32
Columns (total)	32
Numerical Descri...	29
Predictions / Clas...	0
Original Filename	C:/Users/walte/D...

PDB	Ligand	Chain
1	3WBL	PDY A
2	2VTA	LZ1 A
3	3R8V	Z62 A
4	2VTL	LZ5 A
5	3QTZ	X42 A
6	3QTX	X43 A
7	3QTW	X3A A
8	1PXL	CK4 A
9	3QTS	X46 A
10	3QTQ	X35 A
11	2W17	I19 A
12	1Y91	CT9 A
13	1G5S	I17 A
14	2W06	FRV A
15	3EZR	EZR A
16	3LFN	A27 A
17	3RAL	04Z A
18	2C6L	DT4 A
19	2C6M	DT5 A
20	1P2A	5BN A
21	2BTS	U32 A
22	2R3M	SCX A

Bivariate Statistics

First column ('Reference'): log(IC50)

Second column ('Prediction'): Affinity(kcal/mol)

Name	Value
Pearson Correlation (r)	0.518
Pearson Correlation Squared (r <sup>2</sup> )	0.268509
Spearman Rank Correlation (ρ)	0.382423
Mean Squared Deviation (MSD)	4.08596
Root Mean Squared Deviation (RMSD)	2.02138
Least Square Fit	y = -3.54942 + 0.676638 * x
Cross Validated Squared CC (q <sup>2</sup> ) [1]	-1.87171

[1] Second column must be a cross-validated prediction.

Close Copy All to Clipboard

MSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
4.027	6	-0.003	-0.0001	36.7885	
4.164	0	-2.77556e-17	-2.77556e-18	40.2467	
7.309	7	0.001	4.16667e-05	22.271	
6.347	2	2.42861e-17	1.51788e-18	39.8371	
4.305	7	0	0	20.0535	
4.563	8	0.002	6.06061e-05	22.0461	
7.581	5	0.002	8.33333e-05	27.2133	
8.826	4	0.002	8e-05	57.5554	
1.023	6	-0.001	-3.84615e-05	25.6057	
4.233	6	0.002	9.52381e-05	24.7356	
0.494	7	0.001	2.94118e-05	45.6897	
1.791	8	0.001	2.7027e-05	41.3221	
1.054	7	-0.002	-5.88235e-05	35.8557	
0.562	6	6.93889e-18	2.39272e-19	35.2556	
5.467	5	-0.004	-0.000105263	46.0829	
0.62	6	-9.19403e-17	-2.96582e-18	41.8704	
0.94	8	0	0	18.9422	
2.075	7	0.004	0.000117647	45.0029	
0.577	6	0.003	0.0001	34.117	
4.797	5	0.001	3.57143e-05	38.1561	
0.453	5	3.81639e-17	1.73472e-18	29.6805	
1.018	6	5.55112e-17	1.79068e-18	30.1736	

Time Description

- 14:58:58.152 MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
- 14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFigueira
- 14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUC
- 14:58:58.152 Perpetual license.
- 14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
- 15:36:26.743 Subset created with Subset Id: 1 (records: 72)
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1

## Modelo de Regressão

Para criarmos um modelo de regressão, usaremos o conjunto de treinamento. Depois de selecionado o conjunto de treinamento (*CDK2\_IC50\_2022\_1*), clicamos em *Modelling->Multiple Linear Regression...*

The screenshot shows the Molegro Data Modeller interface. The 'Modelling' menu is open, highlighting 'Multiple Linear Regression...'. The main window displays a table with columns: Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. Below the table, a log window shows the following entries:

```

x Time Description
14:58:58.152 MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFigueira.license
14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
14:58:58.152 Perpetual license.
14:58:58.152 Using working path: C:/Users/walfe/Documents/MVD Data
15:36:26.743 Subset created with Subset Id: 1 (records: 72)
15:41:23.678 Creating new dataset 'CDK2_IC50_2022_0' from subset: 0
15:41:23.678 Creating new dataset 'CDK2_IC50_2022_1' from subset: 1
  
```

Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F			
999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128				
1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267				
999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014				
1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188				
1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992				
1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516				
1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709				
465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014				
1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345				
301	2	1	6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956				
299	1.9	1	1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304				
301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56				
311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956				
451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117				
497	1.8	1	9.3e-07	-6.03152	6.03152	5.165	5	-1.56125e-17	-6.50521e-19	22.3224				
454	1.9	1	2e-08	-7.69897	7.69897	0.999	7	0.003	0.0001	20.7888				
299	1.9	1	0.00013	-3.88606	3.88606	0.611	7	-3.46945e-18	-1.0842e-19	34.0538				
18	1PXK	CK3	A	500	2.8	0.69	2.2e-06	-5.65758	5.65758	8.622	3	1.11022e-16	5.84328e-18	68.0518
19	2W1H	LOF	A	1299	2.1	1	5.2e-08	-7.284	7.284	5.296	3	-0.001	-3.84615e-05	27.5857
20	5D1J	56H	A	4000	1.8	1	4.8e-08	-7.31876	7.31876	4.845	6	-0.001	-3.7037e-05	35.5056
21	3EZV	EZV	A	300	1.9	1	1.04e-06	-5.98297	5.98297	4.31	4	-0.001	-2.77778e-05	50.6353
22	2W05	FRT	A	1299	1.9	1	1e-09	-9	9	1.117	9	0.001	3.125e-05	33.7253



## Modelo de Regressão

Agora selecionamos a variável dependente (*target variable*) que será o  $\log(IC_{50})$ . Depois clicamos em *Next*.

Molegro Data Modeller

File Edit Preparation Modelling Visualization Window Modules Help

Workspace Explorer

Items

- Workspace: Unnamed
  - Datasets [3]
    - CDK2\_IC50\_2022
    - CDK2\_IC50\_2022\_0
    - CDK2\_IC50\_2022\_1

Properties

Property	Value
Name	CDK2_IC50_2022...
Records	72
Columns (total)	32
Numerical Descri...	29
Predictions / Clas...	0
Original Filename	C:/Users/walte/D...

Regression Wizard (using Multiple Linear Regression)

Select Dataset and Target Variable

Dataset

Select dataset used for building model: CDK2\_IC50\_2022\_1

Select subset: All

Select target variable:

Target variable (dependent variable):

- Number
- Resolution(A)
- Ligand Occupation Factor
- IC50(M)
- log(IC50)
- pIC50
- RMSD(A)
- Torsions
- Q
- Average Q
- Ligand B-factor(A2)
- Receptor B-factor(A2)
- B-factor ratio (Ligand/Receptor)
- C
- N
- O

Disabled items indicate either constant value columns or invalid columns.

< Back Next > Cancel

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A							044	6	0.001	3.7037e-05	31.5128
2	2VTH	LZ2	A							316	3	0	0	42.1267
3	3IGG	EFQ	A							715	6	-0.001	-4.16667e-05	23.8014
4	2VTP	LZ9	A							251	4	-0.003	-0.000103448	45.9188
5	2VTO	LZ8	A							495	4	-0.001	-3.7037e-05	38.0992
6	2VTN	LZ7	A							309	3	-0.001	-4.54545e-05	43.4516
7	2VTM	LZM	A							494	1	5.55112e-17	5.04647e-18	52.3709
8	3R8U	Z31	A							128	6	0.002	8e-05	24.3014
9	2VTI	LZ3	A							509	4	0.001	3.84615e-05	20.7345
10	4LYN	1YG	A							71	7	-0.001	-3.57143e-05	37.7956
11	3UNJ	0BX	A							421	6	0.003	0.000115385	20.5304
12	4EZ3	0S0	A							539	5	0.001	4.16667e-05	30.56
13	3TIY	TIY	A							422	4	-0.002	-0.0001	39.8956
14	3QTU	X44	A							362	9	0.001	2.77778e-05	24.8117
15	3QTR	X36	A							165	5	-1.56125e-17	-6.50521e-19	22.3224
16	3QU0	X40	A							999	7	0.003	0.0001	20.7888
17	3TI1	B49	A							511	7	-3.46945e-18	-1.0842e-19	34.0538
18	1PXK	CK3	A							522	3	1.11022e-16	5.84328e-18	68.0518
19	2W1H	LOF	A							296	3	-0.001	-3.84615e-05	27.5857
20	5D1J	56H	A							345	6	-0.001	-3.7037e-05	35.5056
21	3EZV	EZV	A							31	4	-0.001	-2.77778e-05	50.6353
22	2W05	FRT	A							117	9	0.001	3.125e-05	33.7253

Time Description

- 14:58:58.152 MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WINS...
- 14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFi...
- 14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
- 14:58:58.152 Perpetual license.
- 14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
- 15:36:26.743 Subset created with Subset Id: 1 (records: 72)
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1

## Modelo de Regressão

Nesta etapa selecionamos as variáveis independentes (*features*). Mas antes é mais seguro limpar as seleções, clique em *Clear*.

The screenshot shows the Molegro Data Modeller interface. A 'Regression Wizard (using Multiple Linear Regression)' dialog box is open, prompting the user to 'Select Descriptors Manually or by Feature Selection'. The dialog includes a dropdown menu for 'Descriptor selection' set to 'Manual selection from list below' and a list of available descriptors. The background data table is as follows:

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A							044	6	0.001	3.7037e-05	31.5128
2	2VTH	LZ2	A							316	3	0	0	42.1267
3	3IGG	EFQ	A							715	6	-0.001	-4.16667e-05	23.8014
4	2VTP	LZ9	A							251	4	-0.003	-0.000103448	45.9188
5	2VTO	LZ8	A							495	4	-0.001	-3.7037e-05	38.0992
6	2VTN	LZ7	A							009	3	-0.001	-4.54545e-05	43.4516
7	2VTM	LZM	A							494	1	5.55112e-17	5.04647e-18	52.3709
8	3R8U	Z31	A							128	6	0.002	8e-05	24.3014
9	2VTI	LZ3	A							509	4	0.001	3.84615e-05	20.7345
10	4LYN	1YG	A							71	7	-0.001	-3.57143e-05	37.7956
11	3UNJ	0BX	A							421	6	0.003	0.000115385	20.5304
12	4EZ3	0S0	A							539	5	0.001	4.16667e-05	30.56
13	3TIY	TIY	A							422	4	-0.002	-0.0001	39.8956
14	3QTU	X44	A							362	9	0.001	2.77778e-05	24.8117
15	3QTR	X36	A							165	5	-1.56125e-17	-6.50521e-19	22.3224
16	3QU0	X40	A							999	7	0.003	0.0001	20.7888
17	3TI1	B49	A							511	7	-3.46945e-18	-1.0842e-19	34.0538
18	1PXK	CK3	A							522	3	1.11022e-16	5.84328e-18	68.0518
19	2W1H	LOF	A							296	3	-0.001	-3.84615e-05	27.5857
20	5D1J	56H	A							345	6	-0.001	-3.7037e-05	35.5056
21	3EZV	EZV	A							31	4	-0.001	-2.77778e-05	50.6353
22	2W05	FRT	A							117	9	0.001	3.125e-05	33.7253

The regression wizard dialog box contains the following text and controls:

Regression Wizard (using Multiple Linear Regression)

Select Descriptors Manually or by Feature Selection

Descriptors (independent variables)

Descriptor selection: Manual selection from list below

Number of descriptors selected: 24

Descriptors

- Number
- Resolution(A)
- Ligand Occupation Factor
- IC50(M)
- pIC50
- RMSD(A)
- Torsions
- Q
- Average Q
- Ligand B-factor(A2)
- Receptor B-factor(A2)
- B-factor ratio (Ligand/Receptor)
- C
- N
- O
- P [Constant]

Disabled items indicate either constant value columns or invalid columns.

Buttons: Select All, Invert Selection, Clear, < Back, Next >, Cancel

Log window (bottom left):

- 14:58:58.152 MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
- 14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterF...
- 14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
- 14:58:58.152 Perpetual license.
- 14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
- 15:36:26.743 Subset created with Subset Id: 1 (records: 72)
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1

## Modelo de Regressão

Agora selecionamos as variáveis da função escore do AutoDock Vina, como mostrado abaixo. Temos seis variáveis selecionadas. Iremos gerar um modelo de regressão de seis dimensões (6D). Clicamos *Next*.

The screenshot shows the Molegro Data Modeller interface. A 'Regression Wizard (using Multiple Linear Regression)' dialog box is open, allowing for manual selection of descriptors. The background table lists various ligands and their associated properties.

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A							044	6	0.001	3.7037e-05	31.5128
2	2VTH	LZ2	A							316	3	0	0	42.1267
3	3IGG	EFQ	A							715	6	-0.001	-4.16667e-05	23.8014
4	2VTP	LZ9	A							251	4	-0.003	-0.000103448	45.9188
5	2VTO	LZ8	A							495	4	-0.001	-3.7037e-05	38.0992
6	2VTN	LZ7	A							309	3	-0.001	-4.54545e-05	43.4516
7	2VTM	LZM	A							494	1	5.55112e-17	5.04647e-18	52.3709
8	3R8U	Z31	A							128	6	0.002	8e-05	24.3014
9	2VTI	LZ3	A							509	4	0.001	3.84615e-05	20.7345
10	4LYN	1YG	A							71	7	-0.001	-3.57143e-05	37.7956
11	3UNJ	0BX	A							421	6	0.003	0.000115385	20.5304
12	4EZ3	0S0	A							539	5	0.001	4.16667e-05	30.56
13	3TIY	TIY	A							422	4	-0.002	-0.0001	39.8956
14	3QTU	X44	A							362	9	0.001	2.77778e-05	24.8117
15	3QTR	X36	A							165	5	-1.56125e-17	-6.50521e-19	22.3224
16	3QU0	X40	A							999	7	0.003	0.0001	20.7888
17	3TI1	B49	A							511	7	-3.46945e-18	-1.0842e-19	34.0538
18	1PXK	CK3	A							522	3	1.11022e-16	5.84328e-18	68.0518
19	2W1H	L0F	A							296	3	-0.001	-3.84615e-05	27.5857
20	5D1J	56H	A							345	6	-0.001	-3.7037e-05	35.5056
21	3EZV	EZV	A							31	4	-0.001	-2.77778e-05	50.6353
22	2W05	FRT	A							117	9	0.001	3.125e-05	33.7253

The regression wizard dialog box shows the following options:

- Descriptor selection: Manual selection from list below
- Number of descriptors selected: 6
- Selected Descriptors: Gauss 1, Gauss 2, Repulsion, Hydrophobic, Hydrogen, Torsional

Buttons: Select All, Invert Selection, Clear, < Back, Next >, Cancel

Log messages at the bottom:

- 14:58:58.152 MDM build: MDM 2013.3.0.0 - Released: maio 21 2013 - [WINS]
- 14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterF
- 14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (FUCRS)
- 14:58:58.152 Perpetual license.
- 14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
- 15:36:26.743 Subset created with Subset Id: 1 (records: 72)
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1

## Modelo de Regressão

Para gerar o modelo de regressão, clicamos *Next*.

The screenshot shows the Molegro Data Modeller interface. A 'Regression Wizard (using Multiple Linear Regression)' dialog box is open, allowing the user to customize the training algorithm. The dialog box includes the following fields:

- Training algorithm:** Multiple Linear Regression (selected in a dropdown menu)
- Shuffle dataset before model training:**
- Random seed used in model training:** 334222673 (with a 'New Seed' button)
- Parameter settings:** No parameters available for Multiple Linear Regression algorithm.

The background shows a table with columns: PDB, Ligand, Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. The table contains 22 rows of data.

At the bottom of the dialog box, there are three buttons: '< Back', 'Next >', and 'Cancel'.

## Modelo de Regressão

Não altere as opções e clique em *Start*.

The screenshot shows the Molegro Data Modeller interface. A 'Regression Wizard (using Multiple Linear Regression)' dialog box is open, displaying experimental setup options. The background shows a table with columns: PDB, Ligand, Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F.

**Regression Wizard (using Multiple Linear Regression) - Experimental Setup**

Experimental settings

Create new model and prediction:

- Using 'CDK2\_IC50\_2022\_1' (subset: All) as training set

Validate model building parameters (creates a prediction but no model):

- Using Leave one out
  - Include subset with index '0'
- Using N-fold cross validation. N: 
  - Overwrite 'Subset' column with fold subsets
- Percentage split: Training set percentage: 
  - Overwrite 'Subset' column with train/test subsets
- Perform feature selection to identify relevant descriptors:
  - Feature selection method:
  - Descriptor relevance:
  - Model selection criterion:

Buttons: < Back, Start, Cancel

**Table Data (Visible Rows):**

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A						044	6	0.001	3.7037e-05	31.5128	
2	2VTH	LZ2	A						316	3	0	0	42.1267	
3	3IGG	EFQ	A						715	6	-0.001	-4.16667e-05	23.8014	
4	2VTP	LZ9	A						251	4	-0.003	-0.000103448	45.9188	
5	2VTO	LZ8	A						495	4	-0.001	-3.7037e-05	38.0992	
6	2VTN	LZ7	A						309	3	-0.001	-4.54545e-05	43.4516	
7	2VTM	LZM	A						494	1	5.55112e-17	5.04647e-18	52.3709	
8	3R8U	Z31	A						128	6	0.002	8e-05	24.3014	
9	2VTI	LZ3	A						509	4	0.001	3.84615e-05	20.7345	
10	4LYN	1YG	A						71	7	-0.001	-3.57143e-05	37.7956	
11	3UNJ	0BX	A						421	6	0.003	0.000115385	20.5304	
12	4EZ3	0S0	A						539	5	0.001	4.16667e-05	30.56	
13	3TIY	TIY	A						422	4	-0.002	-0.0001	39.8956	
14	3QTU	X44	A						362	9	0.001	2.77778e-05	24.8117	
15	3QTR	X36	A						165	5	-1.56125e-17	-6.50521e-19	22.3224	
16	3QU0	X40	A						999	7	0.003	0.0001	20.7888	
17	3TI1	B49	A						511	7	-3.46945e-18	-1.0842e-19	34.0538	
18	1PXK	CK3	A						522	3	1.11022e-16	5.84328e-18	68.0518	
19	2W1H	L0F	A						296	3	-0.001	-3.84615e-05	27.5857	
20	5D1J	56H	A						345	6	-0.001	-3.7037e-05	35.5056	
21	3EZV	EZV	A						31	4	-0.001	-2.77778e-05	50.6353	
22	2W05	FRT	A						117	9	0.001	3.125e-05	33.7253	

**Properties Panel:**

Property	Value
Name	CDK2_IC50_2022_...
Records	72
Columns (total)	32
Numerical Descri...	29
Predictions / Clas...	0
Original Filename	C:/Users/walte/D...

**Log Panel:**

Time	Description
14:58:58.152	MUM build: MUM 2013.3.0.0 - Released: maio 21 2013 - [WIN32]
14:58:58.152	License file: C:/Program Files (x86)/Molegro/MVD/Bin/WalterFile
14:58:58.152	This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
14:58:58.152	Perpetual license.
14:58:58.152	Using working path: C:/Users/walte/Documents/MVD Data
15:36:26.743	Subset created with Subset Id: 1 (records: 72)
15:41:23.678	Creating new dataset 'CDK2_IC50_2022_0' from subset: 0
15:41:23.678	Creating new dataset 'CDK2_IC50_2022_1' from subset: 1

## Modelo de Regressão

O MDM gerou um modelo de regressão chamado MLR(6D) (*Multiple Linear Regression with 6 Dimensions*), clique OK.

The screenshot displays the Molegro Data Modeller interface. The main window shows a table of protein-ligand complexes with columns for PDB, Ligand, Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. A dialog box titled 'Prediction Results...' is overlaid on the table, indicating that a prediction named 'MLR-Train (6D)' has been added to the 'CDK2\_IC50\_2022\_1' dataset. The Pearson's correlation coefficient (r<sup>2</sup>) is 0.217894. The dialog also provides instructions to inspect the regression model by selecting 'Show Details...' from the model context menu. The dialog has 'OK' and 'Show More Statistics...' buttons.

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128
2	2VTH	LZ2	A	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267
3	3IGG	EFQ	A	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014
4	2VTP	LZ9	A	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188
5	2VTO	LZ8	A	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992
6	2VTN	LZ7	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
7	2VTM	LZM	A	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709
8	3R8U	Z31	A	465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014
9	2VTI	LZ3	A	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345
10	4LYN	1YG	A	1299	2.2	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992
11	3UNJ	0BX	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
12	4EZ3	0S0	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
13	3TIY	TIY	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
14	3QTU	X44	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
15	3QTR	X36	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
16	3QU0	X40	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
17	3TI1	B49	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
18	1PXK	CK3	A	500	2.8	0.69	2.2e-06	-5.65758	5.65758	8.622	3	1.11022e-16	5.84328e-18	68.0518
19	2W1H	LOF	A	1299	2.1	1	5.2e-08	-7.284	7.284	5.296	3	-0.001	-3.84615e-05	27.5857
20	5D1J	56H	A	4000	1.8	1	4.8e-08	-7.31876	7.31876	4.845	6	-0.001	-3.7037e-05	35.5056
21	3EZV	EZV	A	300	1.9	1	1.04e-06	-5.98297	5.98297	4.31	4	-0.001	-2.77778e-05	50.6353
22	2W05	FRT	A	1299	1.9	1	1e-09	-9	9	1.117	9	0.001	3.125e-05	33.7253

Time | Description

- 14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira.license
- 14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
- 14:58:58.152 Perpetual license.
- 14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
- 15:36:26.743 Subset created with Subset Id: 1 (records: 72)
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1
- 16:20:08.629 Using random seed: 334222673 for model training

## Modelo de Regressão

O modelo gerado está disponível no *Workspace Explorer* e chama-se *MLR(6D)*.

The screenshot displays the Molegro Data Modeller software interface. The main window shows a table of protein-ligand complexes with the following columns: PDB, Ligand, Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. The table contains 22 rows of data. On the left side, the Workspace Explorer panel shows a tree view with 'Workspace: Unnamed', 'Datasets [3]', and 'Models [1]'. The 'Models [1]' folder is expanded, showing 'MLR (6D)'. An arrow points from the text above to this model. Below the table, the Properties panel shows details for the selected model, including Name, Records, Columns, Numerical Descriptions, Predictions, and Original Filename. At the bottom, a Log panel shows the software's activity, including license information and dataset creation steps.

	PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128	
2	2VTH	LZ2	A	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267	
3	3IGG	EFQ	A	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014	
4	2VTP	LZ9	A	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188	
5	2VTO	LZ8	A	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992	
6	2VTN	LZ7	A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516	
7	2VTM	LZM	A	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709	
8	3R8U	Z31	A	465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014	
9	2VTI	LZ3	A	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345	
10	4LYN	1YG	A	301	2	1	6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956	
11	3UNJ	0BX	A	299	1.9	1	1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304	
12	4EZ3	0S0	A	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56	
13	3TIY	TIY	A	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956	
14	3QTU	X44	A	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117	
15	3QTR	X36	A	497	1.8	1	9.3e-07	-6.03152	6.03152	5.165	5	-1.56125e-17	-6.50521e-19	22.3224	
16	3QU0	X40	A	454	1.9	1	2e-08	-7.69897	7.69897	0.999	7	0.003	0.0001	20.7888	
17	3TI1	B49	A	299	1.9	1	0.00013	-3.88606	3.88606	0.611	7	-3.46945e-18	-1.0842e-19	34.0538	
18	1PXK	CK3	A	500	2.8	0.69	2.2e-06	-5.65758	5.65758	8.622	3	1.11022e-16	5.84328e-18	68.0518	
19	2W1H	LOF	A	1299	2.1	1	5.2e-08	-7.284	7.284	5.296	3	-0.001	-3.84615e-05	27.5857	
20	5D1J	56H	A	4000	1.8	1	4.8e-08	-7.31876	7.31876	4.845	6	-0.001	-3.7037e-05	35.5056	
21	3EZV	EZV	A	300	1.9	1	1.04e-06	-5.98297	5.98297	4.31	4	-0.001	-2.77778e-05	50.6353	
22	2W05	FRT	A	1299	1.9	1	1e-09	-9	9	1.117	9	0.001	3.125e-05	33.7253	

## Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Podemos fazer uma análise estatística do poder de previsão do modelo gerado. Clique em *Modelling->Bivariate Statistics...*

The screenshot shows the Molegro Data Modeller interface. The 'Modelling' menu is open, highlighting 'Bivariate Statistics...'. The main window displays a table with the following columns: Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. The table contains 22 rows of data. Below the table, there is a small table with columns for ID, Name, and Value, showing records for CDK2. At the bottom, a log window displays system messages.

	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F		
	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128			
	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267			
	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014			
	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188			
	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992			
	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516			
	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709			
	465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014			
	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345			
	301	2	1	6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956			
	299	1.9	1	1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304			
	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56			
	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956			
	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117			
	497	1.8	1	9.3e-07	-6.03152	6.03152	5.165	5	-1.56125e-17	-6.50521e-19	22.3224			
	454	1.9	1	2e-08	-7.69897	7.69897	0.999	7	0.003	0.0001	20.7888			
	299	1.9	1	0.00013	-3.88606	3.88606	0.611	7	-3.46945e-18	-1.0842e-19	34.0538			
18	1PXK	CK3	A	500	2.8	0.69	2.2e-06	-5.65758	5.65758	8.622	3	1.11022e-16	5.84328e-18	68.0518
19	2W1H	LOF	A	1299	2.1	1	5.2e-08	-7.284	7.284	5.296	3	-0.001	-3.84615e-05	27.5857
20	5D1J	56H	A	4000	1.8	1	4.8e-08	-7.31876	7.31876	4.845	6	-0.001	-3.7037e-05	35.5056
21	3EZV	EZV	A	300	1.9	1	1.04e-06	-5.98297	5.98297	4.31	4	-0.001	-2.77778e-05	50.6353
22	2W05	FRT	A	1299	1.9	1	1e-09	-9	9	1.117	9	0.001	3.125e-05	33.7253

Property	Value
Name	CDK2
Records	72
Columns (total)	32
Numerical Descri...	29
Predictions / Clas...	0
Original Filename	C:/Us...

```

Log
x Time Description
14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira.license
14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
14:58:58.152 Perpetual license.
14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
15:36:26.743 Subset created with Subset Id: 1 (records: 72)
15:41:23.678 Creating new dataset 'CDK2_IC50_2022_0' from subset: 0
15:41:23.678 Creating new dataset 'CDK2_IC50_2022_1' from subset: 1
16:20:08.629 Using random seed: 334222673 for model training
  
```



## Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Na nova janela, selecione para primeira coluna o  $\log(IC_{50})$  e como segunda coluna o *MLR-Train (6D)*.

**Bivariate Statistics**

First column ('Reference'):  $\log(IC_{50})$

Second column ('Prediction'): Affinity(kcal/mol)

Legend:

- Br
- Affinity(kcal/mol)
- Gauss 1
- Gauss 2
- Repulsion
- Hydrophobic
- Hydrogen
- Torsional
- Subset
- MLR-Train (6D)**

Statistics:

Name	Value
Pearson Correlation (r)	0.378
Pearson Correlation Squared (r <sup>2</sup> )	0.142542
Spearman Rank Correlation (ρ)	0.301876
Mean Squared Deviation (MSD)	4.96241
Root Mean Squared Deviation (RMSD)	2.22765
Least Square Fit	$y = -4.53184 + 0.493097 \cdot x$
Cross Validated Squared CC (q <sup>2</sup> ) [1]	-1.64866

[1] Second column must be a cross-validated prediction.

Buttons: Close, Copy All to Clipboard

MSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
9.044	6	0.001	3.7037e-05	31.5128	
2.916	3	0	0	42.1267	
0.715	6	-0.001	-4.16667e-05	23.8014	
5.251	4	-0.003	-0.000103448	45.9188	
7.495	4	-0.001	-3.7037e-05	38.0992	
4.009	3	-0.001	-4.54545e-05	43.4516	
2.494	1	5.55112e-17	5.04647e-18	52.3709	
1.128	6	0.002	8e-05	24.3014	
1.509	4	0.001	3.84615e-05	20.7345	
4.71	7	-0.001	-3.57143e-05	37.7956	
1.421	6	0.003	0.000115385	20.5304	
2.539	5	0.001	4.16667e-05	30.56	
3.422	4	-0.002	-0.0001	39.8956	
0.962	9	0.001	2.77778e-05	24.8117	
5.165	5	-1.56125e-17	-6.50521e-19	22.3224	
0.999	7	0.003	0.0001	20.7888	
0.611	7	-3.46945e-18	-1.0842e-19	34.0538	
8.622	3	1.11022e-16	5.84328e-18	68.0518	
5.296	3	-0.001	-3.84615e-05	27.5857	
4.845	6	-0.001	-3.7037e-05	35.5056	
4.31	4	-0.001	-2.77778e-05	50.6353	
1.117	9	0.001	3.125e-05	33.7253	

## Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Os resultados estão indicados abaixo, como era de se esperar, houve uma melhora no *RMSE*, comparado com a função *Affinity*. Agora temos um  $RMSE = 1,2105$  (contra  $2,22765$  para o *Affinity*). Mas a avaliação mais importante é para o conjunto de teste. Clique *Close*.

Molegro Data Modeller

File Edit Preparation Modelling Visualization Window Modules Help

Workspace Explorer

Items

- Workspace: Unnamed
  - Datasets [3]
    - CDK2\_IC50\_2022
    - CDK2\_IC50\_2022\_0
    - CDK2\_IC50\_2022\_1
  - Models [1]
    - MLR (6D)

Properties

Property	Value
Name	CDK2_IC50_2022...
Records	72
Columns (total)	32
Numerical Descri...	29
Predictions / Clas...	0
Original Filename	C:/Users/walte/D...

PDB	Ligand	Chain
1	3IG7	EFP A
2	2VTH	LZ2 A
3	3IGG	EFQ A
4	2VTP	LZ9 A
5	2VTO	LZ8 A
6	2VTN	LZ7 A
7	2VTM	LZM A
8	3R8U	Z31 A
9	2VTI	LZ3 A
10	4LYN	1YG A
11	3UNJ	0BX A
12	4EZ3	0S0 A
13	3TIY	TIY A
14	3QTU	X44 A
15	3QTR	X36 A
16	3QU0	X40 A
17	3TI1	B49 A
18	1PXK	CK3 A
19	2W1H	LOF A
20	5D1J	56H A
21	3EZV	EZV A
22	2W05	FRT A

Bivariate Statistics

First column ('Reference'): log(IC50)

Second column ('Prediction'): MLR-Train (6D)

Name	Value
Pearson Correlation (r)	0.467
Pearson Correlation Squared (r <sup>2</sup> )	0.217894
Spearman Rank Correlation (ρ)	0.42317
Mean Squared Deviation (MSD)	1.46532
Root Mean Squared Deviation (RMSD)	1.2105
Least Square Fit	y = -4.9566 + 0.217894 * x
Cross Validated Squared CC (q <sup>2</sup> ) [1]	0.217894

[1] Second column must be a cross-validated prediction.

Close Copy All to Clipboard

MSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
9.044	6	0.001	3.7037e-05	31.5128	
2.916	3	0	0	42.1267	
0.715	6	-0.001	-4.16667e-05	23.8014	
5.251	4	-0.003	-0.000103448	45.9188	
7.495	4	-0.001	-3.7037e-05	38.0992	
4.009	3	-0.001	-4.54545e-05	43.4516	
2.494	1	5.55112e-17	5.04647e-18	52.3709	
1.128	6	0.002	8e-05	24.3014	
1.509	4	0.001	3.84615e-05	20.7345	
4.71	7	-0.001	-3.57143e-05	37.7956	
1.421	6	0.003	0.000115385	20.5304	
2.539	5	0.001	4.16667e-05	30.56	
3.422	4	-0.002	-0.0001	39.8956	
0.962	9	0.001	2.77778e-05	24.8117	
5.165	5	-1.56125e-17	-6.50521e-19	22.3224	
0.999	7	0.003	0.0001	20.7888	
0.611	7	-3.46945e-18	-1.0842e-19	34.0538	
8.622	3	1.11022e-16	5.84328e-18	68.0518	
5.296	3	-0.001	-3.84615e-05	27.5857	
4.845	6	-0.001	-3.7037e-05	35.5056	
4.31	4	-0.001	-2.77778e-05	50.6353	
1.117	9	0.001	3.125e-05	33.7253	

Time Description

- 14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira
- 14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUC)
- 14:58:58.152 Perpetual license.
- 14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
- 15:36:26.743 Subset created with Subset Id: 1 (records: 72)
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1
- 16:20:08.629 Using random seed: 334222673 for model training

## Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Para aplicarmos o modelo de regressão gerado aos dados do conjunto de teste (*CDK2\_IC50\_2022\_0*), vamos ao *Workspace Explorer* e clicamos com o botão direito do mouse sobre o modelo *MLR (6D)* e escolhemos a opção *Make Prediction...*

The screenshot displays the Molegro Data Modeller interface. The main window shows a table of datasets with columns: PDB, Ligand, Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. A context menu is open over the 'MLR (6D)' model, with 'Make Prediction...' selected. The 'Log' window at the bottom shows the software's activity.

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F	
1	3WBL	PDY	A	302	2	2.3e-05	-4.63827	4.63827	4.027	6	-0.003	-0.0001	36.7885		
2	2VTA	LZ1	A	1301	2	0.000185	-3.73283	3.73283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467		
3	3R8V	Z62	A	473	1.9	2.9e-06	-5.5376	5.5376	7.309	7	0.001	4.16667e-05	22.271		
4	2VTL	LZ5	A	1299	2	9.7e-05	-4.01323	4.01323	6.347	2	2.42861e-17	1.51788e-18	39.8371		
5	3QTZ	X42	A	453	2	5e-08	-7.30103	7.30103	4.305	7	0	0	20.0535		
6	3QTX	X43	A	299	1.9	7e-08	-7.1549	7.1549	4.563	8	0.002	6.06061e-05	22.0461		
			A	451	1.8	6.5e-07	-6.18709	6.18709	7.581	5	0.002	8.33333e-05	27.2133		
			A	500	2.5	9e-07	-6.04576	6.04576	8.826	4	0.002	8e-05	57.5554		
			A	299	1.9	3.1e-06	-5.50864	5.50864	1.023	6	-0.001	-3.84615e-05	25.6057		
			A	476	1.8	3.1e-06	-5.50864	5.50864	4.233	6	0.002	9.52381e-05	24.7356		
			A	1300	2.1	2e-09	-8.69897	8.69897	0.494	7	0.001	2.94118e-05	45.6897		
			A	401	2.1	5.9e-08	-7.22915	7.22915	1.791	8	0.001	2.7027e-05	41.3221		
			A	400	2.6	4.8e-08	-7.31876	7.31876	1.054	7	-0.002	-5.88235e-05	35.8557		
			A	1300	2	8.3e-08	-7.08092	7.08092	0.562	6	6.93889e-18	2.39272e-19	35.2556		
			A	300	1.9	1e-05	-5	5	5.467	5	-0.004	-0.000105263	46.0829		
16	3LFN	A27	A	299	2.2	3.1e-06	-5.50031	5.50031	0.62	6	-9.19403e-17	-2.96582e-18	41.8704		
17	3RAL	04Z	A	499	1.7	1e-07	-7	7	0.94	8	0	0	18.9422		
18	2C6L	DT4	A	1299	2.3	0.9375	2.7e-07	-6.56864	6.56864	2.075	7	0.004	0.000117647	45.0029	
19	2C6M	DT5	A	1297	1.9	3.5e-07	-6.45593	6.45593	0.577	6	0.003	0.0001	34.117		
20	1P2A	5BN	A	301	2.5	1.2e-08	-7.92082	7.92082	4.797	5	0.001	3.57143e-05	38.1561		
21	2BTS	U32	A	1299	1.9	2e-08	-7.69897	7.69897	0.453	5	3.81639e-17	1.73472e-18	29.6805		
22	2R3M	SCX	A	501	1.7	1e-08	-8	8	1.018	6	5.55112e-17	1.79068e-18	30.1736		

**Log**

- 14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira.license
- 14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
- 14:58:58.152 Perpetual license.
- 14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
- 15:36:26.743 Subset created with Subset Id: 1 (records: 72)
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1
- 16:20:08.629 Using random seed: 334222673 for model training

## Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Agora selecionamos o conjunto de teste (*CDK2\_IC50\_2022\_0*) e clicamos OK.

The screenshot displays the Molegro Data Modeller software interface. The main window shows a table of datasets with columns: PDB, Ligand, Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. A dialog box titled 'Make Model Prediction' is open, showing the following fields:

- Select dataset: CDK2\_IC50\_2022\_0
- Select subset: All
- Name of prediction: MLR (6D)
- Create class probabilities for each class (as numerical columns)
- Create column containing highest probability (as numerical column)

The dialog box has 'OK' and 'Cancel' buttons. The background table shows the following data (rows 1-22):

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3WBL	PDY	A	302	2	1	2.3e-05	-4.63827	4.63827	4.027	6	-0.003	-0.0001	36.7885
2	2VTA	LZ1	A	1301	2	1	0.000185	-3.73283	3.73283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467
3	3R8V	Z62	A	473	1.9	1	2.9e-06	-5.5376	5.5376	7.309	7	0.001	4.16667e-05	22.271
4	2VTL	LZ5	A	1299	2	1	9.7e-05	-4.01323	4.01323	6.347	2	2.42861e-17	1.51788e-18	39.8371
5	3Q TZ	X42	A	453	2	1	5e-08	-7.30103	7.30103	4.305	7	0	0	20.0535
6	3Q TX	X43	A	299	1.9	1	7e-08	-7.1549	7.1549	4.563	8	0.002	6.06061e-05	22.0461
7	3Q TW	X3A	A	451	1.8	1	6.5e-07	-6.18709	6.18709	7.581	5	0.002	8.33333e-05	27.2133
8	1PXL	CK4	A	500	2.5	0.59	9e-07	-6.04576	6.04576	8.826	4	0.002	8e-05	57.5554
9	3Q TS	X46	A	299				5.50864	1.023	6	-0.001	-3.84615e-05	25.6057	
10	3Q TQ	X35	A	476				5.50864	4.233	6	0.002	9.52381e-05	24.7356	
11	2W17	I19	A	1300				8.69897	0.494	7	0.001	2.94118e-05	45.6897	
12	1Y91	CT9	A	401				7.22915	1.791	8	0.001	2.7027e-05	41.3221	
13	1G5S	I17	A	400				7.31876	1.054	7	-0.002	-5.88235e-05	35.8557	
14	2W06	FRV	A	1300				7.08092	0.562	6	6.93889e-18	2.39272e-19	35.2556	
15	3EZR	EZR	A	300				5	5.467	5	-0.004	-0.000105263	46.0829	
16	3LFN	A27	A	299				5.50031	0.62	6	-9.19403e-17	-2.96582e-18	41.8704	
17	3RAL	04Z	A	499				7	0.94	8	0	0	18.9422	
18	2C6L	DT4	A	1299				6.56864	2.075	7	0.004	0.000117647	45.0029	
19	2C6M	DT5	A	1297	1.9	1	3.5e-07	-6.45593	6.45593	0.577	6	0.003	0.0001	34.117
20	1P2A	5BN	A	301	2.5	1	1.2e-08	-7.92082	7.92082	4.797	5	0.001	3.57143e-05	38.1561
21	2BTS	U32	A	1299	1.9	1	2e-08	-7.69897	7.69897	0.453	5	3.81639e-17	1.73472e-18	29.6805
22	2R3M	SCX	A	501	1.7	1	1e-08	-8	8	1.018	6	5.55112e-17	1.79068e-18	30.1736

The bottom status bar shows a log of actions:

- 14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira.license
- 14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
- 14:58:58.152 Perpetual license.
- 14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
- 15:36:26.743 Subset created with Subset Id: 1 (records: 72)
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1
- 16:20:08.629 Using random seed: 334222673 for model training

# Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Clique em *Show More Statistics...*

Molegro Data Modeller

File Edit Preparation Modelling Visualization Window Modules Help

Descriptors: All | Coloring: Default

Workspace Explorer

- Workspace: Unnamed
  - Datasets [3]
    - CDK2\_IC50\_2022
    - CDK2\_IC50\_2022\_0
    - CDK2\_IC50\_2022\_1
  - Models [1]
    - MLR (6D)

Properties

Property	Value
Model Name	MLR (6D)
Type	MLR (regression)
Target variable	log(IC50)
Descriptors	[6]

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F	
1	3WBL	PDY	A	302	2	2.3e-05	-4.63827	4.63827	4.027	6	-0.003	-0.0001	36.7885		
2	2VTA	LZ1	A	1301	2	0.000185	-3.73283	3.73283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467		
3	3R8V	Z62	A	473	1.9	2.9e-06	-5.5376	5.5376	7.309	7	0.001	4.16667e-05	22.271		
4	2VTL	LZ5	A	1299	2	9.7e-05	-4.01323	4.01323	6.347	2	2.42861e-17	1.51788e-18	39.8371		
5	3QTZ	X42	A	453	2	5e-08	-7.30103	7.30103	4.305	7	0	0	20.0535		
6	3QTX	X43	A	299	1.9	7e-08	-7.1549	7.1549	4.563	8	0.002	6.06061e-05	22.0461		
7	3QTW	X3A	A	451	1.8	6.5e-07	-6.18709	6.18709	7.581	5	0.002	8.33333e-05	27.2133		
8	1PXL	CK4	A	500	2.5	9e-07	-6.04576	6.04576	8.826	4	0.002	8e-05	57.5554		
9	3QTS	X46	A	299	1.9	3.1e-06	-5.50864	5.50864	1.023	6	-0.001	-3.84615e-05	25.6057		
10	3QTQ	X35	A						4.233	6	0.002	9.52381e-05	24.7356		
11	2W17	I19	A						0.494	7	0.001	2.94118e-05	45.6897		
12	1Y91	CT9	A						1.791	8	0.001	2.7027e-05	41.3221		
13	1G5S	I17	A						1.054	7	-0.002	-5.88235e-05	35.8557		
14	2W06	FRV	A						0.562	6	6.93889e-18	2.39272e-19	35.2556		
15	3EZR	EZR	A						5.467	5	-0.004	-0.000105263	46.0829		
16	3LFN	A27	A						0.62	6	-9.19403e-17	-2.96582e-18	41.8704		
17	3RAL	04Z	A	499	1.7	1e-07	-7	7	0.94	8	0	0	18.9422		
18	2C6L	DT4	A	1299	2.3	0.9375	2.7e-07	-6.56864	6.56864	2.075	7	0.004	0.000117647	45.0029	
19	2C6M	DT5	A	1297	1.9	1	3.5e-07	-6.45593	6.45593	0.577	6	0.003	0.0001	34.117	
20	1P2A	5BN	A	301	2.5	1	1.2e-08	-7.92082	7.92082	4.797	5	0.001	3.57143e-05	38.1561	
21	2BTS	U32	A	1299	1.9	1	2e-08	-7.69897	7.69897	0.453	5	3.81639e-17	1.73472e-18	29.6805	
22	2R3M	SCX	A	501	1.7	1	1e-08	-8	8	1.018	6	5.55112e-17	1.79068e-18	30.1736	

Prediction Results...

The Prediction (named: MLR (6D)) has been added to the CDK2\_IC50\_2022\_0 dataset.  
Pearson's correlation coefficient (r<sup>2</sup>): 0.309926

OK Show More Statistics...

Time Description

- 14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira.license
- 14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
- 14:58:58.152 Perpetual license.
- 14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
- 15:36:26.743 Subset created with Subset Id: 1 (records: 72)
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1
- 16:20:08.629 Using random seed: 334222673 for model training

## Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Agora temos um  $RMSE = 1,00179$  para o conjunto de teste, contra um  $RMSE = 2,02138$  para o *Affinity*. Assim vemos uma melhora no poder de previsão. Destaco, os pontos experimentais do conjunto teste não foram usados para treinar o modelo, por isso o conjunto de teste é considerado significativo.

Molegro Data Modeller

File Edit Preparation Modelling Visualization Window Modules Help

Coloring: Default

Descriptors: All

Workspace Explorer

Items

- Workspace: Unnamed
  - Datasets [3]
    - CDK2\_IC50\_2022
    - CDK2\_IC50\_2022\_0
    - CDK2\_IC50\_2022\_1
  - Models [1]
    - MLR (6D)

Properties

Property	Value
Name	MLR (6D)
Type	MLR (regression)
Target variable	log(IC50)
Descriptors	[6]

PDB	Ligand	Chain
1	3WBL	PDY
2	2VTA	LZ1
3	3R8V	Z62
4	2VTL	LZ5
5	3QTZ	X42
6	3QTX	X43
7	3QTW	X3A
8	1PXL	CK4
9	3QTS	X46
10	3QTQ	X35
11	2W17	I19
12	1Y91	CT9
13	1G5S	I17
14	2W06	FRV
15	3EZR	EZR
16	3LFN	A27
17	3RAL	04Z
18	2C6L	DT4
19	2C6M	DT5
20	1P2A	5BN
21	2BTS	U32
22	2R3M	SCX

Bivariate Statistics

First column ('Reference'): log(IC50)

Second column ('Prediction'): MLR (6D)

Name	Value
Pearson Correlation (r)	0.557
Pearson Correlation Squared (r <sup>2</sup> )	0.309926
Spearman Rank Correlation (ρ)	0.489644
Mean Squared Deviation (MSD)	1.00358
Root Mean Squared Deviation (RMSD)	1.00179
Least Square Fit	y = -3.99439 + 0.364915 * x
Cross Validated Squared CC (q <sup>2</sup> ) [1]	0.294662

MSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
4.027	6	-0.003	-0.0001	36.7885	
4.164	0	-2.77556e-17	-2.77556e-18	40.2467	
7.309	7	0.001	4.16667e-05	22.271	
6.347	2	2.42861e-17	1.51788e-18	39.8371	
4.305	7	0	0	20.0535	
4.563	8	0.002	6.06061e-05	22.0461	
7.581	5	0.002	8.33333e-05	27.2133	
8.826	4	0.002	8e-05	57.5554	
1.023	6	-0.001	-3.84615e-05	25.6057	
4.233	6	0.002	9.52381e-05	24.7356	
0.494	7	0.001	2.94118e-05	45.6897	
1.791	8	0.001	2.7027e-05	41.3221	
1.054	7	-0.002	-5.88235e-05	35.8557	
0.562	6	6.93889e-18	2.39272e-19	35.2556	
5.467	5	-0.004	-0.000105263	46.0829	
0.62	6	-9.19403e-17	-2.96582e-18	41.8704	
0.94	8	0	0	18.9422	
2.075	7	0.004	0.000117647	45.0029	
0.577	6	0.003	0.0001	34.117	
4.797	5	0.001	3.57143e-05	38.1561	
0.453	5	3.81639e-17	1.73472e-18	29.6805	
1.018	6	5.55112e-17	1.79068e-18	30.1736	

[1] Second column must be a cross-validated prediction.

Close Copy All to Clipboard

Time Description

14:58:58.152	License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira
14:58:58.152	This software is licensed to: Walter Figueira de Azevedo Junior (PUC
14:58:58.152	Perpetual license.
14:58:58.152	Using working path: C:/Users/walte/Documents/MVD Data
15:36:26.743	Subset created with Subset Id: 1 (records: 72)
15:41:23.678	Creating new dataset 'CDK2_IC50_2022_0' from subset: 0
15:41:23.678	Creating new dataset 'CDK2_IC50_2022_1' from subset: 1
16:20:08.629	Using random seed: 334222673 for model training

## Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Alguns comentários sobre a análise estatística realizada pelo MDM. Nem todas métricas discutidas estão disponíveis no MDM, programas mais completos como o [SAnDReS](#) (Xavier et al., 2016) têm uma análise estatística (Walsh et al., 2021) e modelagens mais completas. Clique *Close*.

Molegro Data Modeller

File Edit Preparation Modelling Visualization Window Modules Help

Coloring: Default

Descriptors: All

Workspace Explorer

Items

- Workspace: Unnamed
  - Datasets [3]
    - CDK2\_IC50\_2022
    - CDK2\_IC50\_2022\_0
    - CDK2\_IC50\_2022\_1
  - Models [1]
    - MLR (6D)

Properties

Property	Value
Name	MLR (6D)
Type	MLR (regression)
Target variable	log(IC50)
Descriptors	[6]

PDB	Ligand	Chain
1	3WBL	PDY
2	2VTA	LZ1
3	3R8V	Z62
4	2VTL	LZ5
5	3QTZ	X42
6	3QTX	X43
7	3QTW	X3A
8	1PXL	CK4
9	3QTS	X46
10	3QTQ	X35
11	2W17	I19
12	1Y91	CT9
13	1G5S	I17
14	2W06	FRV
15	3EZR	EZR
16	3LFN	A27
17	3RAL	04Z
18	2C6L	DT4
19	2C6M	DT5
20	1P2A	5BN
21	2BTS	U32
22	2R3M	SCX

Bivariate Statistics

First column ('Reference'): log(IC50)

Second column ('Prediction'): MLR (6D)

Name	Value
Pearson Correlation (r)	0.557
Pearson Correlation Squared (r <sup>2</sup> )	0.309926
Spearman Rank Correlation (ρ)	0.489644
Mean Squared Deviation (MSD)	1.00358
Root Mean Squared Deviation (RMSD)	1.00179
Least Square Fit	y = -3.99439 + 0.364915 * x
Cross Validated Squared CC (q <sup>2</sup> ) [1]	0.294662

[1] Second column must be a cross-validated prediction.

Close Copy All to Clipboard

MSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
4.027	6	-0.003	-0.0001	36.7885	
4.164	0	-2.77556e-17	-2.77556e-18	40.2467	
7.309	7	0.001	4.16667e-05	22.271	
6.347	2	2.42861e-17	1.51788e-18	39.8371	
4.305	7	0	0	20.0535	
4.563	8	0.002	6.06061e-05	22.0461	
7.581	5	0.002	8.33333e-05	27.2133	
8.826	4	0.002	8e-05	57.5554	
1.023	6	-0.001	-3.84615e-05	25.6057	
4.233	6	0.002	9.52381e-05	24.7356	
0.494	7	0.001	2.94118e-05	45.6897	
1.791	8	0.001	2.7027e-05	41.3221	
1.054	7	-0.002	-5.88235e-05	35.8557	
0.562	6	6.93889e-18	2.39272e-19	35.2556	
5.467	5	-0.004	-0.000105263	46.0829	
0.62	6	-9.19403e-17	-2.96582e-18	41.8704	
0.94	8	0	0	18.9422	
2.075	7	0.004	0.000117647	45.0029	
0.577	6	0.003	0.0001	34.117	
4.797	5	0.001	3.57143e-05	38.1561	
0.453	5	3.81639e-17	1.73472e-18	29.6805	
1.018	6	5.55112e-17	1.79068e-18	30.1736	

Time Description

- 14:58:58.152 License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira
- 14:58:58.152 This software is licensed to: Walter Figueira de Azevedo Junior (PUC
- 14:58:58.152 Perpetual license.
- 14:58:58.152 Using working path: C:/Users/walte/Documents/MVD Data
- 15:36:26.743 Subset created with Subset Id: 1 (records: 72)
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0
- 15:41:23.678 Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1
- 16:20:08.629 Using random seed: 334222673 for model training

## Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Para salvar o modelo gerado marque o MLR(6D) no *Workspace explorer*, com o botão direito do mouse ative o menu e selecione *Model-> Export Model...*

The screenshot shows the Molegro Data Modeller interface. The main window displays a table of models with columns: PDB, Ligand, Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. A context menu is open over the 'MLR (6D)' model, with 'Export Model...' selected.

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EPF	A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128
2	3WBL	PDY	A	302	2	1	2.3e-05	-4.63827	4.63827	4.027	6	-0.003	-0.0001	36.7885
3	2VTH	LZ2	A	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267
4	3IGG	EFQ	A	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014
5	2VTA	LZ1	A	1301	2	1	0.000185	-3.73283	3.73283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467
6	2VTP	LZ9	A	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188
			A	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992
			A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516
			A	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709
			A	473	1.9	1	2.9e-06	-5.5376	5.5376	7.309	7	0.001	4.16667e-05	22.271
			A	1299	2	1	9.7e-05	-4.01323	4.01323	6.347	2	2.42861e-17	1.51788e-18	39.8371
			A	465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014
			A	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345
			A	301	2	1	6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956
			A	299	1.9	1	1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304
16	3QTZ	X42	A	453	2	1	5e-08	-7.30103	7.30103	4.305	7	0	0	20.0535
17	3QTX	X43	A	299	1.9	1	7e-08	-7.1549	7.1549	4.563	8	0.002	6.06061e-05	22.0461
18	3QTW	X3A	A	451	1.8	1	6.5e-07	-6.18709	6.18709	7.581	5	0.002	8.33333e-05	27.2133
19	4EZ3	OS0	A	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56
20	3TIY	TIY	A	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956
21	3QTU	X44	A	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117
22	1PXL	CK4	A	500	2.5	0.59	9e-07	-6.04576	6.04576	8.826	4	0.002	8e-05	57.5554

The context menu for the 'MLR (6D)' model includes the following options:

- Show Details...
- Make Prediction...
- Apply Model to External Dataset...
- Export Model...**
- Rename Model...
- Delete Model From Workspace

The Log window at the bottom shows the following entries:

```

19:14:44.566 License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira.license
19:14:44.566 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
19:14:44.566 Perpetual license.
19:14:44.566 Using working path: C:/Users/walte/Documents/MVD Data
19:15:06.310 Subset created with Subset Id: 1 (records: 72)
19:15:12.588 Creating new dataset 'CDK2_IC50_2022_0' from subset: 0
19:15:12.603 Creating new dataset 'CDK2_IC50_2022_1' from subset: 1
19:19:55.794 Using random seed: 4056628687 for model training
  
```



# Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

No novo menu clique em *Export...*

The screenshot displays the Molegro Data Modeller interface. The main window shows a table of model results with columns for PDB, Ligand, Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F. An 'Export Models...' dialog box is open, showing a tree view of models with 'MLR (6D)' selected. The bottom status bar shows a log of recent actions.

PDB	Ligand	Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F
1	3IG7	EFP	A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128
2	3WBL	PDY	A	302					3827	4.027	6	-0.003	-0.0001	36.7885
3	2VTH	LZ2	A	1300					2082	2.916	3	0	0	42.1267
4	3IGG	EFQ	A	999					7718	0.715	6	-0.001	-4.16667e-05	23.8014
5	2VTA	LZ1	A	1301					3283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467
6	2VTP	LZ9	A	1299					2288	5.251	4	-0.003	-0.000103448	45.9188
7	2VTO	LZ8	A	1299					5387	7.495	4	-0.001	-3.7037e-05	38.0992
8	2VTN	LZ7	A	1299					7058	4.009	3	-0.001	-4.54545e-05	43.4516
9	2VTM	LZM	A	1299					3	2.494	1	5.55112e-17	5.04647e-18	52.3709
10	3R8V	Z62	A	473					3376	7.309	7	0.001	4.16667e-05	22.271
11	2VTL	LZ5	A	1299					1323	6.347	2	2.42861e-17	1.51788e-18	39.8371
12	3R8U	Z31	A	465					0103	1.128	6	0.002	8e-05	24.3014
13	2VTI	LZ3	A	1299					8046	1.509	4	0.001	3.84615e-05	20.7345
14	4LYN	1YG	A	301					2185	4.71	7	-0.001	-3.57143e-05	37.7956
15	3UNJ	0BX	A	299					5861	1.421	6	0.003	0.000115385	20.5304
16	3QTZ	X42	A	453					0103	4.305	7	0	0	20.0535
17	3QTX	X43	A	299					1549	4.563	8	0.002	6.06061e-05	22.0461
18	3QTW	X3A	A	451					8709	7.581	5	0.002	8.33333e-05	27.2133
19	4EZ3	0S0	A	301					4679	2.539	5	0.001	4.16667e-05	30.56
20	3TIY	TIY	A	311					6955	3.422	4	-0.002	-0.0001	39.8956
21	3QTU	X44	A	451					1549	0.962	9	0.001	2.77778e-05	24.8117
22	1PXL	CK4	A	500					4576	8.826	4	0.002	8e-05	57.5554

Log:

- 19:14:44.566 License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira.license
- 19:14:44.566 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
- 19:14:44.566 Perpetual license.
- 19:14:44.566 Using working path: C:/Users/walte/Documents/MVD Data
- 19:15:06.310 Subset created with Subset Id: 1 (records: 72)
- 19:15:12.588 Creating new dataset 'CDK2\_IC50\_2022\_0' from subset: 0
- 19:15:12.603 Creating new dataset 'CDK2\_IC50\_2022\_1' from subset: 1
- 19:19:55.794 Using random seed: 4056628687 for model training

## Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Escolha a pasta onde será salvo o modelo e coloque um nome nele. Aqui eu escolhi o nome: *MLR\_6D\_Modelo\_01.mdm*. Clique no botão Salvar. Vocês agora têm um modelo de aprendizado de máquina salvo no seu computador.

The screenshot displays the Molegro Data Modeller interface. The main window shows a table with columns: PDB, Ligand, Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, and Ligand B-factor(A2). The table lists 22 entries, with the first entry being PDB 3IG7, Ligand EFP, Chain A, Number 999, Resolution 1.8, and Average Q 3.7037e-05.

An "Export Models..." dialog box is open, showing a "Choose Filename" window. The file name is set to "MLR\_6D\_Modelo\_01" and the type is "Molegro Data Modeling format (\*.mdm)". The dialog is positioned over a file explorer view showing a folder structure with "SML\_01" and "Data" folders.

On the left side of the interface, there is a "Workspace Explorer" showing a tree view with "Datasets [3]" and "Models [1]". The "Properties" panel shows the model name as "MLR (6D)", type as "MLR (regression)", and target variable as "log(IC50)".

At the bottom left, a log window shows the following entries:

Time	Description
19:14:44.566	License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira.license
19:14:44.566	This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
19:14:44.566	Perpetual license.
19:14:44.566	Using working path: C:/Users/walte/Documents/MVD Data
19:15:06.310	Subset created with Subset Id: 1 (records: 72)
19:15:12.588	Creating new dataset 'CDK2_IC50_2022_0' from subset: 0
19:15:12.603	Creating new dataset 'CDK2_IC50_2022_1' from subset: 1
19:19:55.794	Using random seed: 4056628687 for model training

Observação: O modelo de aprendizado de máquina é um arquivo que você pode enviar por e-mail, disponibilizar num site ou colocar como material suplementar de um artigo científico. O modelo pode ser lido pelo MDM e aplicado a outro conjunto de dados. A única condição é que o novo conjunto de dados tenha os mesmos *features* (características) usados na construção do modelo. Do ponto de vista prático isto significa ter uma planilha com as colunas: Gauss 1, Gauss 2, Repulsion, Hydrophobic, Hydrogen e Torsional

## Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Para encerrar clique *File->Exit*. Alternativamente, você pode pressionar as teclas <Ctrl> Q para finalizar.

The screenshot displays the Molegro Data Modeller application window. The 'File' menu is open, showing options like 'New Dataset...', 'Import Dataset...', 'Export Dataset...', and 'Exit'. The 'Exit' option is highlighted with a blue background and the keyboard shortcut 'Ctrl+Q'. The main window contains a table of dataset records with the following columns: Chain, Number, Resolution(A), Ligand Occupation Fa, IC50(M), log(IC50), pIC50, RMSD(A), Torsions, Q, Average Q, Ligand B-factor(A2), and F.

Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)	F			
A	999	1.8	1	6.3e-08	-7.20066	7.20066	9.044	6	0.001	3.7037e-05	31.5128				
A	302	2	1	2.3e-05	-4.63827	4.63827	4.027	6	-0.003	-0.0001	36.7885				
A	1300	1.9	1	0.00012	-3.92082	3.92082	2.916	3	0	0	42.1267				
A	999	1.8	1	6.65e-08	-7.17718	7.17718	0.715	6	-0.001	-4.16667e-05	23.8014				
A	1301	2	1	0.000185	-3.73283	3.73283	4.164	0	-2.77556e-17	-2.77556e-18	40.2467				
A	1299	2.1	1	3e-09	-8.52288	8.52288	5.251	4	-0.003	-0.000103448	45.9188				
A	1299	2.1	1	1.4e-07	-6.85387	6.85387	7.495	4	-0.001	-3.7037e-05	38.0992				
A	1299	2.2	1	8.5e-07	-6.07058	6.07058	4.009	3	-0.001	-4.54545e-05	43.4516				
A	1299	2.2	1	0.001	-3	3	2.494	1	5.55112e-17	5.04647e-18	52.3709				
A	473	1.9	1	2.9e-06	-5.5376	5.5376	7.309	7	0.001	4.16667e-05	22.271				
A	1299	2	1	9.7e-05	-4.01323	4.01323	6.347	2	2.42861e-17	1.51788e-18	39.8371				
A	465	2	1	5e-06	-5.30103	5.30103	1.128	6	0.002	8e-05	24.3014				
A	1299	2	1	6.6e-07	-6.18046	6.18046	1.509	4	0.001	3.84615e-05	20.7345				
A	301	2	1	6e-08	-7.22185	7.22185	4.71	7	-0.001	-3.57143e-05	37.7956				
A	299	1.9	1	1.1e-05	-4.95861	4.95861	1.421	6	0.003	0.000115385	20.5304				
A	453	2	1	5e-08	-7.30103	7.30103	4.305	7	0	0	20.0535				
17	3QTX	X43	A	299	1.9	1	7e-08	-7.1549	7.1549	4.563	8	0.002	6.06061e-05	22.0461	
18	3QTW	X3A	A	451	1.8	1	6.5e-07	-6.18709	6.18709	7.581	5	0.002	8.33333e-05	27.2133	
19	4EZ3	0S0	A	301	2	1	4.5e-05	-4.34679	4.34679	2.539	5	0.001	4.16667e-05	30.56	
20	3TIY	TIY	A	311	1.8	1	1.7e-05	-4.76955	4.76955	3.422	4	-0.002	-0.0001	39.8956	
21	3QTU	X44	A	451	1.8	1	7e-08	-7.1549	7.1549	0.962	9	0.001	2.77778e-05	24.8117	
22	1PXL	CK4	A	500	2.5	0.59	9e-07	-6.04576	6.04576	8.826	4	0.002	8e-05	57.5554	

The bottom status bar shows a log of recent actions:

```

x Time Description
19:14:44.566 License file: C:/Program Files (x86)/Molegro/MVD/bin/WalterFigueira.license
19:14:44.566 This software is licensed to: Walter Figueira de Azevedo Junior (PUCRS)
19:14:44.566 Perpetual license.
19:14:44.566 Using working path: C:/Users/walte/Documents/MVD Data
19:15:06.310 Subset created with Subset Id: 1 (records: 72)
19:15:12.588 Creating new dataset 'CDK2_IC50_2022_0' from subset: 0
19:15:12.603 Creating new dataset 'CDK2_IC50_2022_1' from subset: 1
19:19:55.794 Using random seed: 4056628687 for model training
  
```

## Desafio 01






Agora use o conjunto de dados visto nesta aula (arquivo *CDK2\_IC50\_2022.csv*) e tente gerar novos modelos de regressão linear múltipla. Escolha outro conjunto de variáveis independentes (*features*) para a construção dos novos modelos. Lembrem-se de treinar o modelo com os dados do conjunto de treinamento. Antes de gerar novos modelos de regressão, você pode verificar a correlação de variáveis isoladas com o  $\log(IC_{50})$ . Por exemplo, a carga Q pode ter correlação com o  $\log(IC_{50})$ . **Tente gerar um modelo de regressão linear com poder de previsão melhor do que o visto hoje.** Construa vários modelos e anote as variáveis independentes usadas e as métricas para cada conjunto de *features*. **As métricas são avaliadas para o conjunto de teste.**

Prepare uma tabela de resultados com as seguintes informações. O Modelo 01 é o que acabamos de construir.

Modelo	Features	$r^2$	$\rho$	RMSE	$R^2$
01	Gauss 1, Gauss 2, Repulsion, Hydrophobic, Hydrogen, Torsional	0,309926	0,489644	1,00179	0,294662
02					
03					
..					

Mande por e-mail ([walter@azevedolab.net](mailto:walter@azevedolab.net)) o melhor modelo gerado até o dia 27/03/2024.

## Referências

-  Bitencourt-Ferreira G, de Azevedo WF Jr. Molegro Virtual Docker for Docking. *Methods Mol Biol.* 2019;2053:149-167.
- Bitencourt-Ferreira G, de Azevedo WF Jr. Machine Learning to Predict Binding Affinity. *Methods Mol Biol.* 2019;2053:251-273.
- De Azevedo WF, Leclerc S, Meijer L, Havlicek L, Strnad M, Kim SH. Inhibition of cyclin-dependent kinases by purine analogues: crystal structure of human cdk2 complexed with roscovitine. *Eur J Biochem.* 1997; 243(1-2):518-526.
-  De Azevedo WF Jr, Dias R. Computational methods for calculation of ligand-binding affinity. *Curr Drug Targets.* 2008;9(12):1031-1039.
- De Azevedo WF. Application of Machine Learning Techniques for Drug Discovery. *Curr Med Chem.* 2021;28(38):7805-7807.
-  De Azevedo WF. Protein-Ligand Interactions: High-Resolution Structures of CDK2. *Curr Drug Targets.* 2022;23(5):438-440.
- Heberlé G, de Azevedo WF Jr. Bio-inspired algorithms applied to molecular docking simulations. *Curr Med Chem.* 2011;18(9):1339-1352.
- Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model.* 2021; 61(8):3891-3898.
-  Quiroga R, Villarreal MA. Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PLoS One.* 2016 May 12;11(5):e0155183.
- Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem.* 2010; 31(2):455-461.
-  Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G; ELIXIR Machine Learning Focus Group, Harrow J, Psomopoulos FE, Tosatto SCE. DOME: recommendations for supervised machine learning validation in biology. *Nat Methods.* 2021;18(10):1122-1127.
- Thomsen R, Christensen MH. MolDock: a new technique for high-accuracy molecular docking. *J Med Chem.* 2006; 49(11): 3315–3321.
- Xavier MM, Heck GS, Avila MB, Levin NMB, Pintro VO, Carvalho NL, Azevedo WF. SANdReS a Computational Tool for Statistical Analysis of Docking Results and Development of Scoring Functions. *Comb Chem High Throughput Screen.* 2016;19(10):801-812.



Que a luz da ciência acabe com  
as trevas do negacionismo.