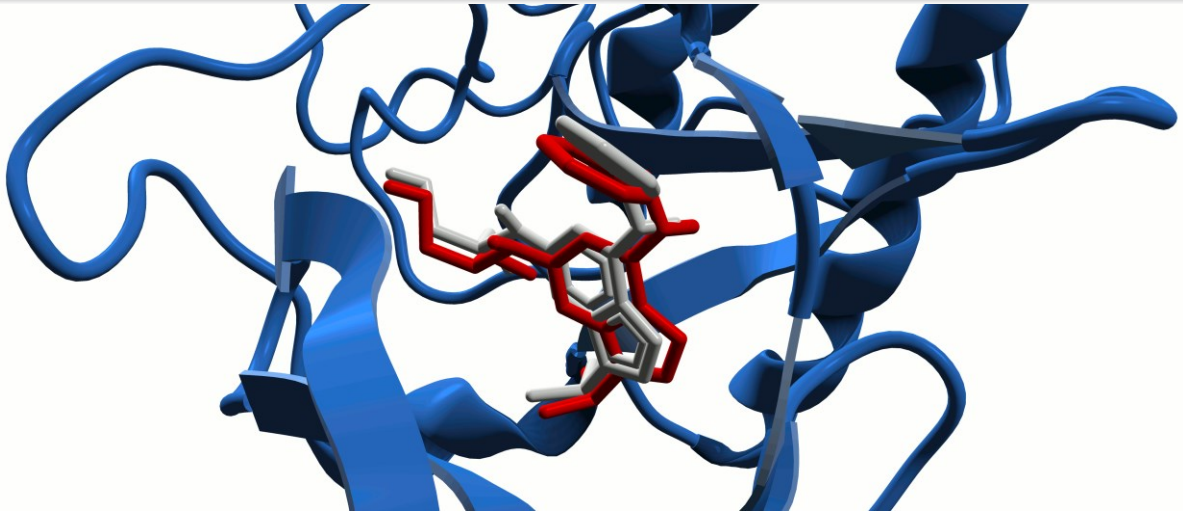
Aprendizado de Máquina Supervisionado I

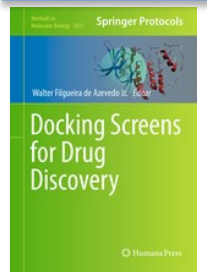
Prof. Dr. Walter F. de Azevedo, Jr.

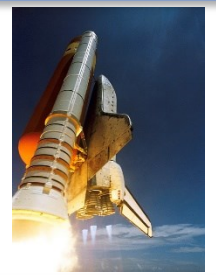
walter@azevedolab.net

Biography 01♥
Biography 02♥
Biography 03♥
Biography 04♥

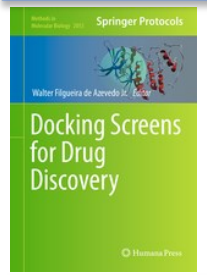


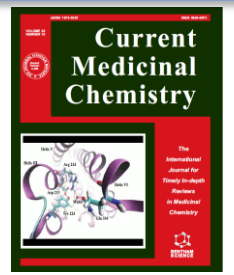
Frontiers Section Editor (Bioinformatics and Biophysics) for the Current Drug Targets ISSN: 1873-5592
Section Editor (Bioinformatics in Drug Design and Discovery) for the Current Medicinal Chemistry ISSN: 1875-533X

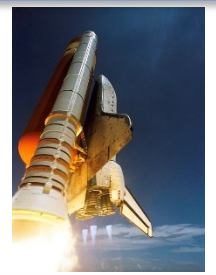




Conteúdo







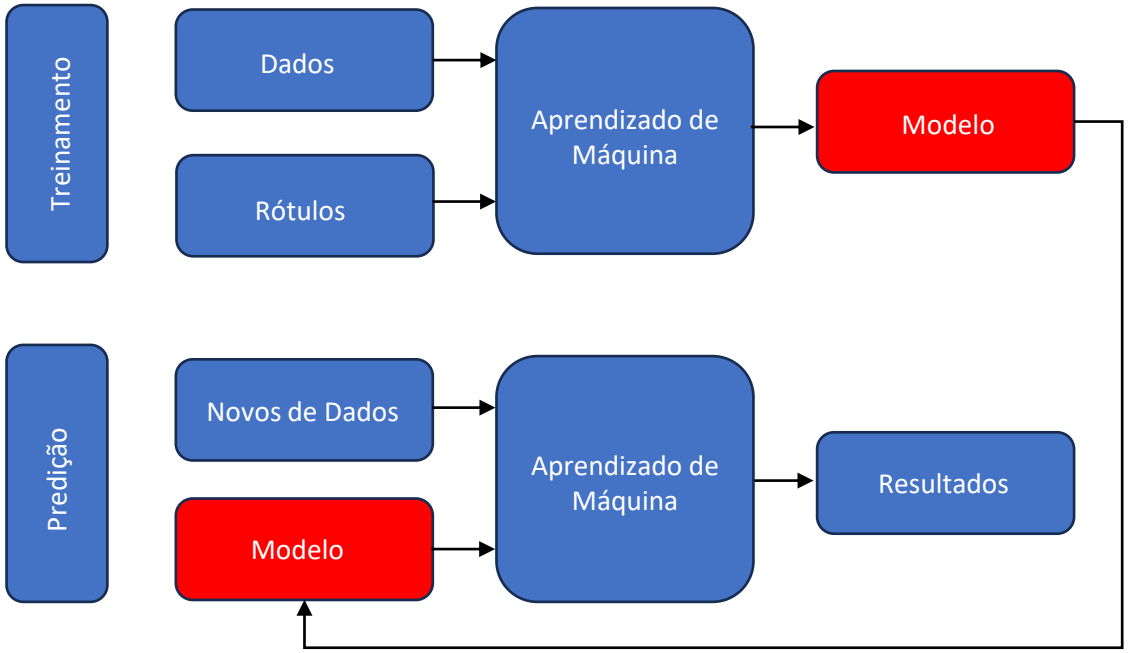
Resumo

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}.

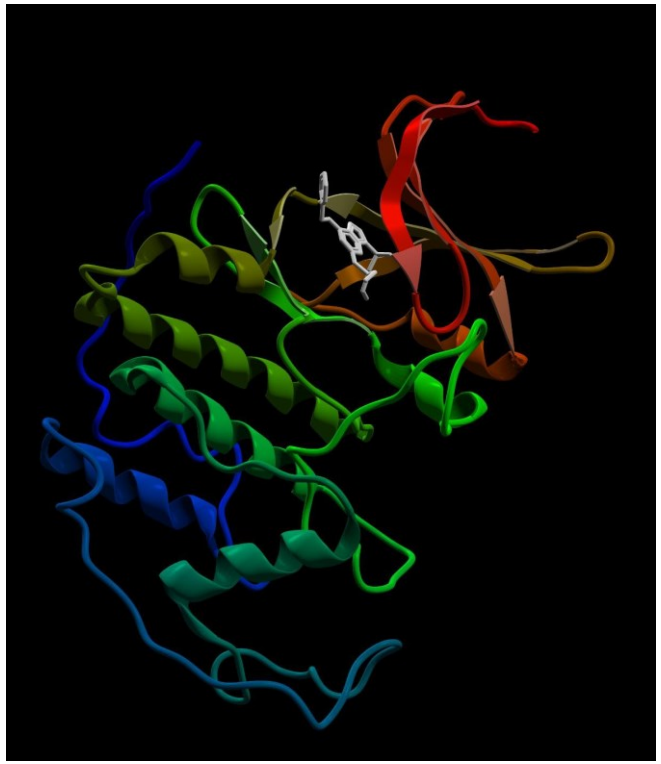
Aprendizado de Máquina Supervisionado

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 predições (de Azevedo, 2021).



Sistema Proteína-Fármaco

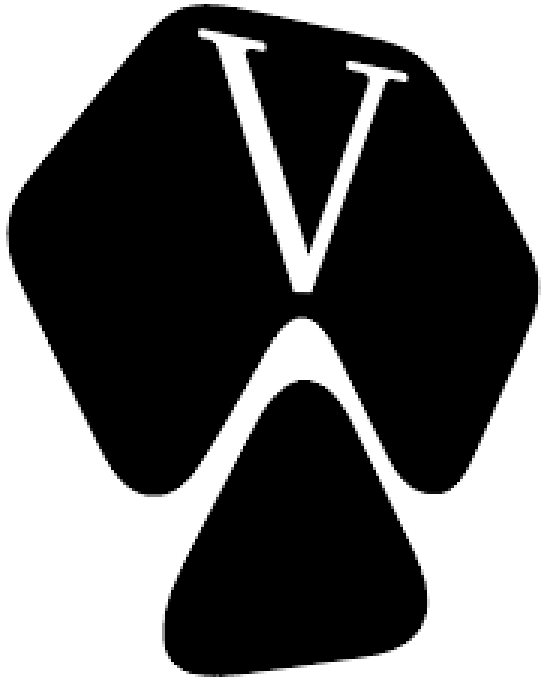
A forma mais simples de olhar a interação entre um fármaco e uma proteína é considerarmos o fármaco como uma chave e o bolsão de ligação da proteína como a fechadura. Chamamos de docagem molecular o processo de determinar computacionalmente a posição da chave na fechadura.



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.

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

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

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

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

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

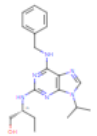
Nas equações $d_{i,j} = r_{i,j} - R_{t_i} - R_{t_j}$ onde $r_{i,j}$ é a distância entre dois átomos e R_t o raio de van der Waals do átomo.

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 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 Query on RRC Download Ideal Coordinates CCD File Download Instance Coordinates	B [auth A]	**R-ROSCOVITINE** $C_{19}H_{26}N_6O$ BTIHMVBBUGXLCJ-OAHLLOKOSA-N		Ligand Interaction

Binding Affinity Annotations

ID	Source	Binding Affinity
RRC	BindingDB: 2A4L	**Ki:** 250 (nM) from 1 assay(s)
		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	

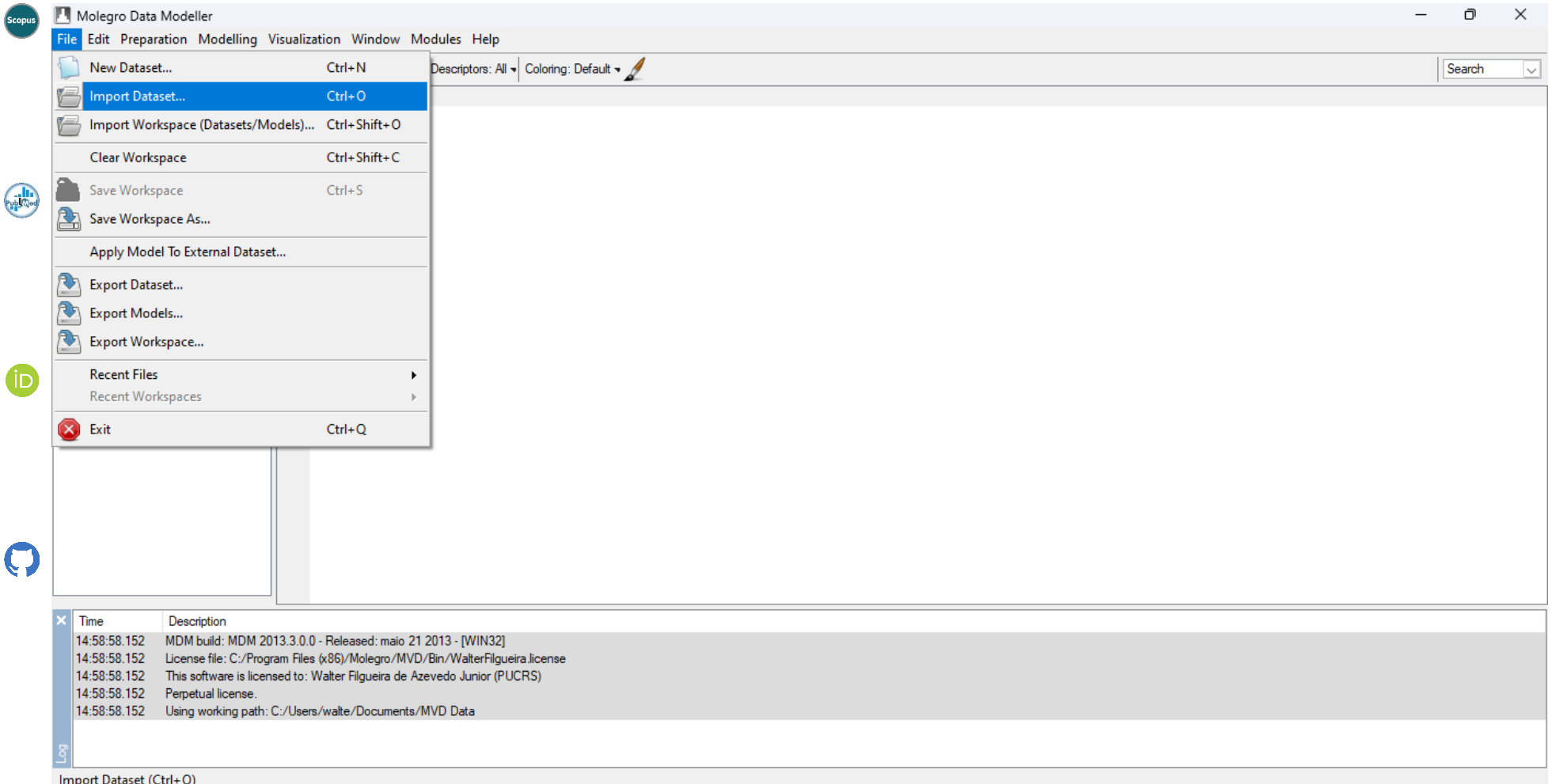
Conjunto de Dados

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.

	ır 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	5.2011	0.87112	7	2	0	0	0	0	0	0	-5.133	32.84	586.044	1.601	30.253	1.706	0
6	5.5456	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	9.3832	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	5.5623	1.26323	16	4	0	0	1	3	0	0	-5.858	78.227	1295.88	5.81	30.455	1.972	0.69897

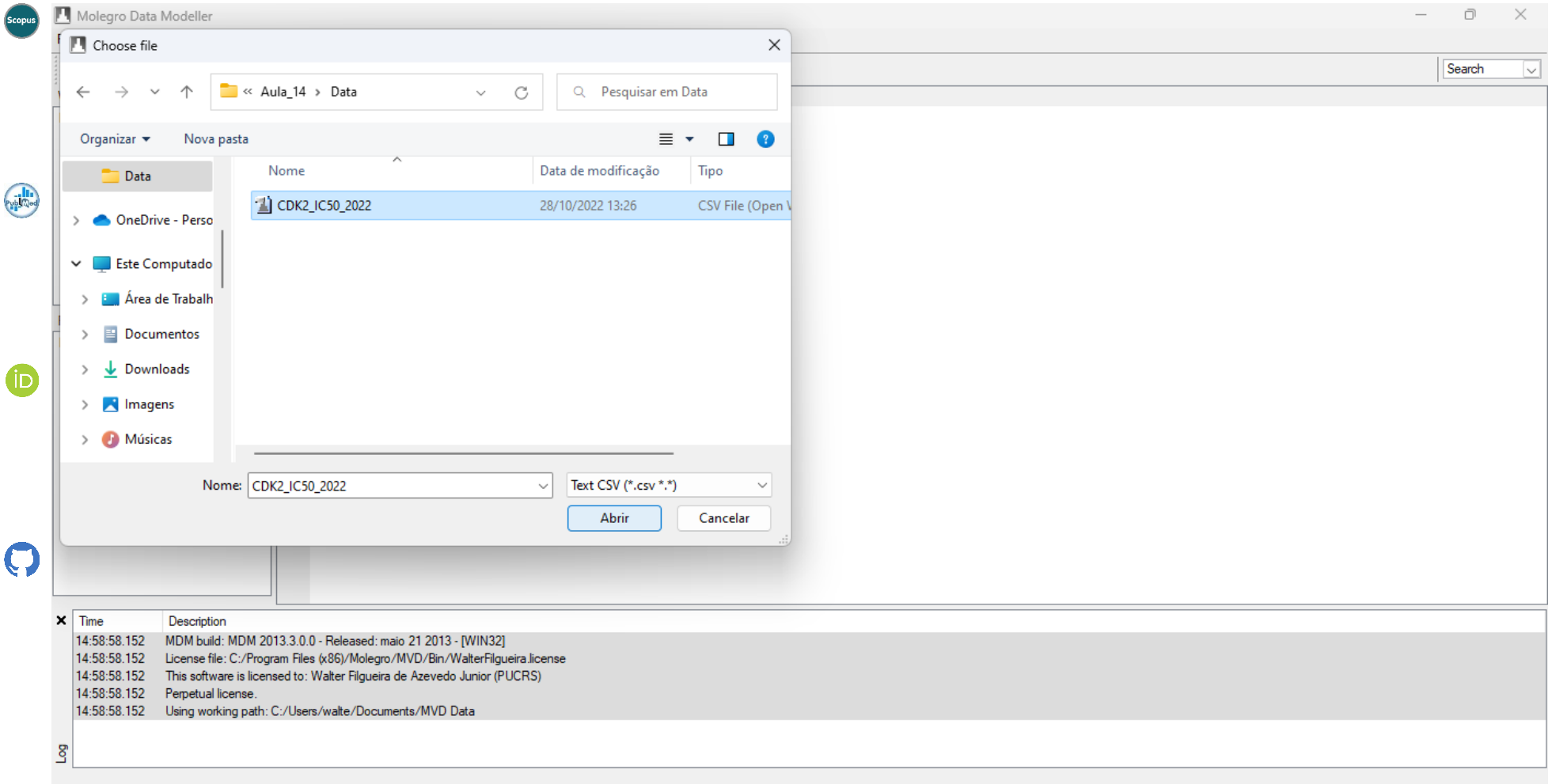
Conjunto de Dados

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...*



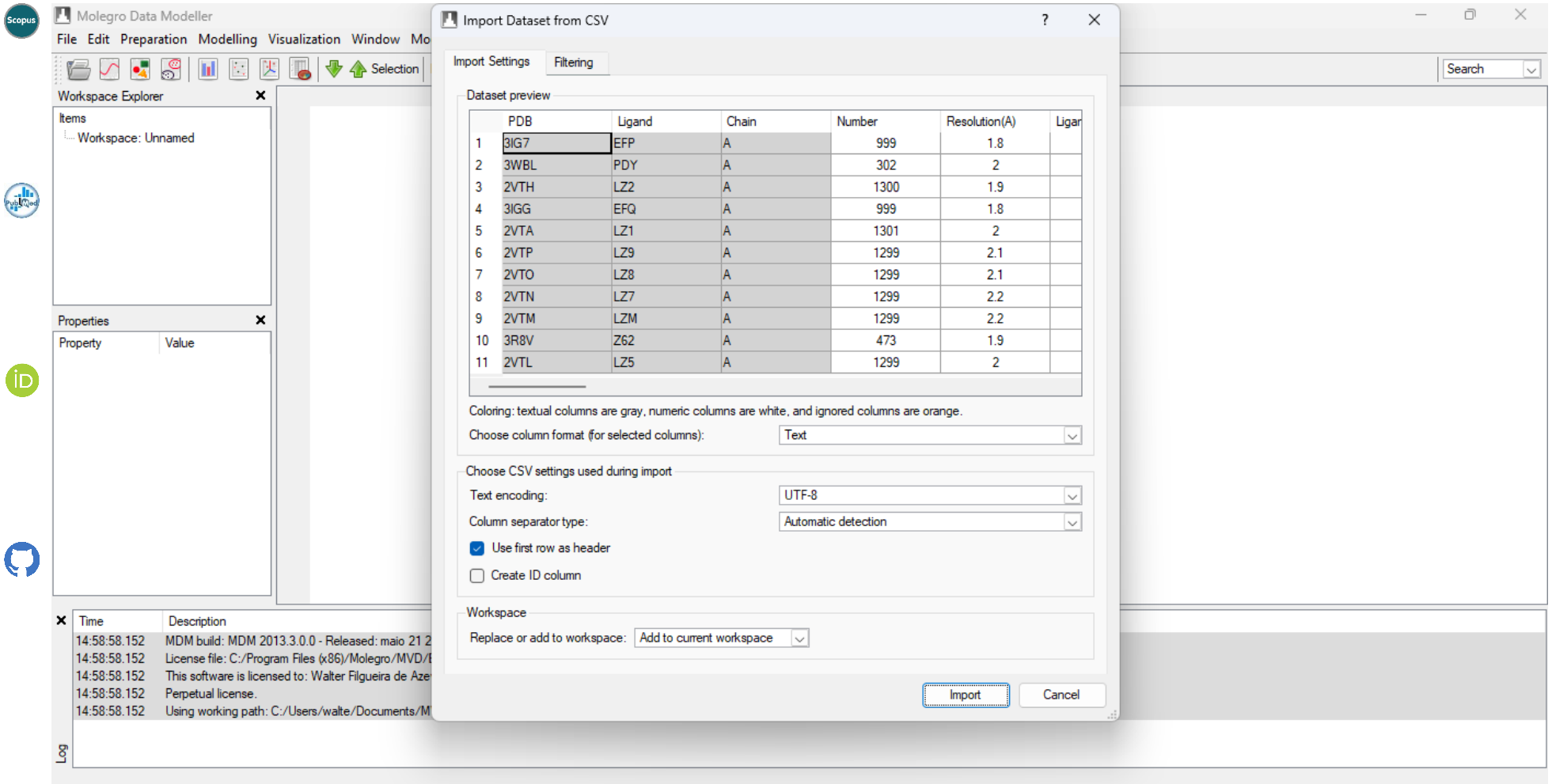
Conjunto de Dados

- Na nova janela, vá na pasta onde o arquivo *CDK2_IC50_2022.csv* se encontra. Selecione o arquivo e clique em Abrir.



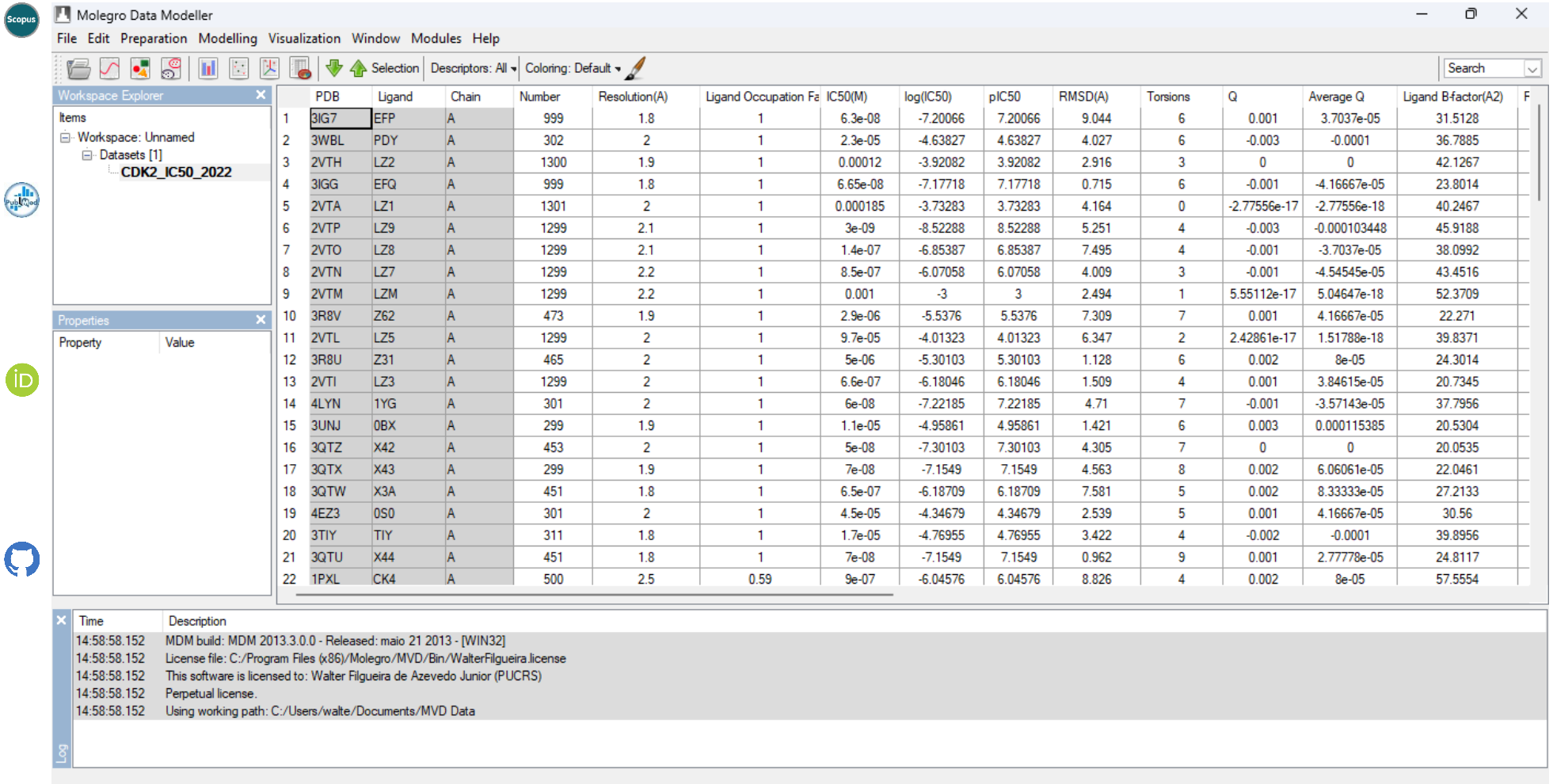
Conjunto de Dados

- Agora clique na opção *Import*.



Conjunto de Dados

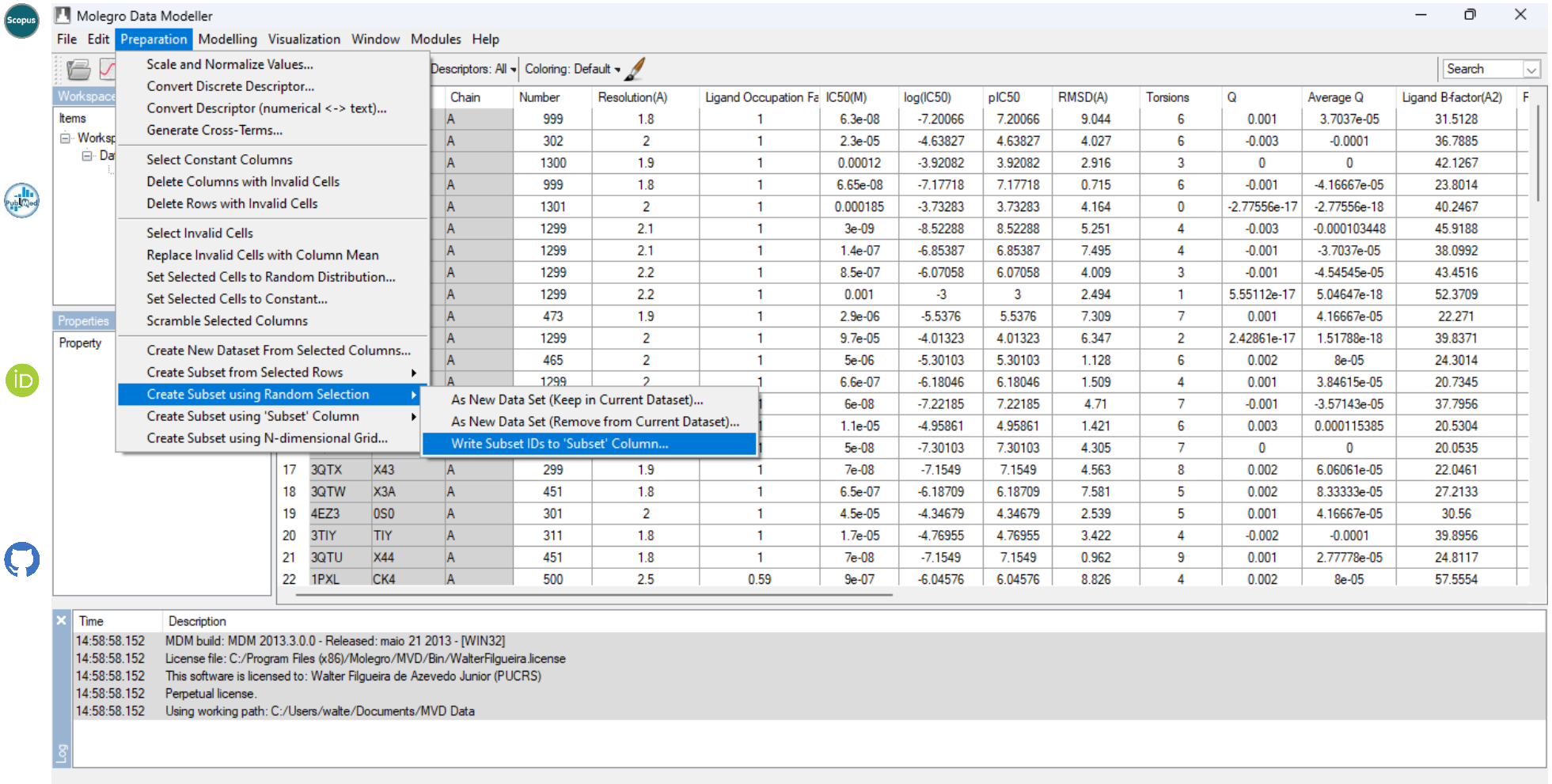
- O MDM carrega a planilha de dados e a mostra parcialmente na tela, visto que temos um número grande de colunas (31) e linhas (104). Em aprendizado de máquina supervionado é comum dividir o conjunto de dados em dois subconjuntos: o conjunto de treinamento e o conjunto de teste.



	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

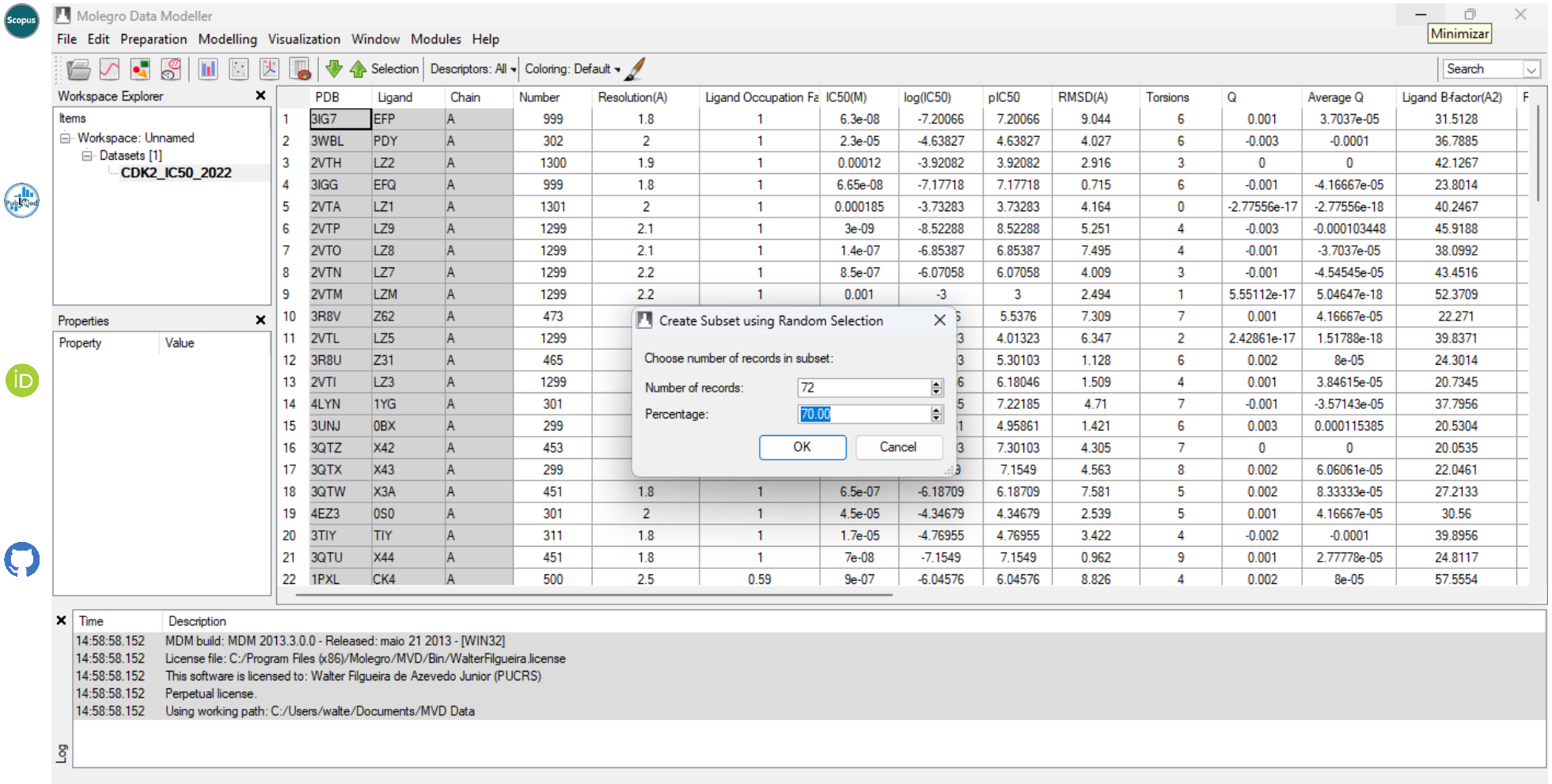
- Para dividir o conjunto de dados, clique em *Preparation->Create Subset using Random Selection->Write Subset IDs to 'Subset' Column...*



			Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)
			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

Conjunto de Dados

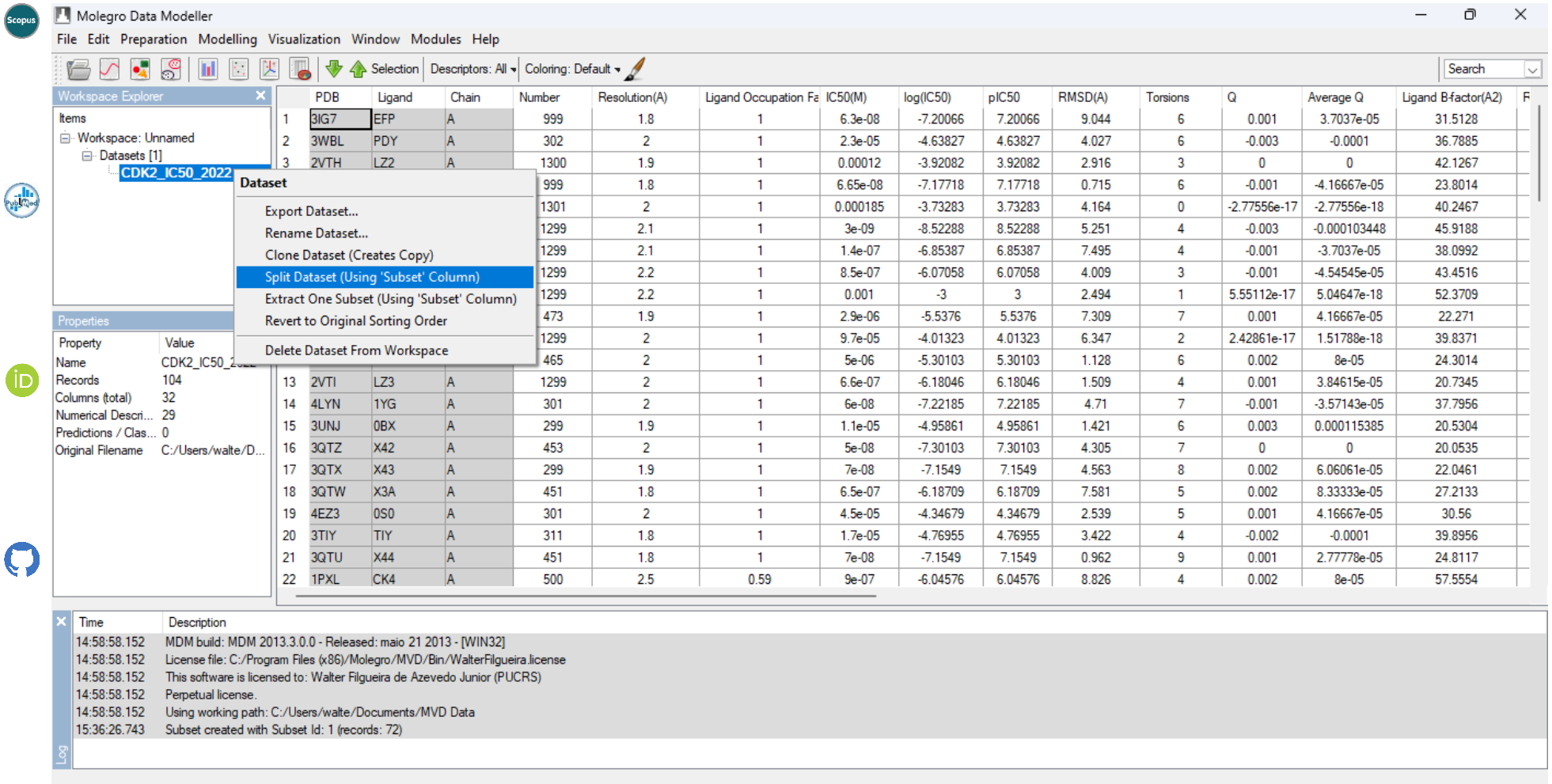
- Selecione 70 % dos dados para o conjunto de treinamento, como indicado abaixo. Clique em OK.



	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	

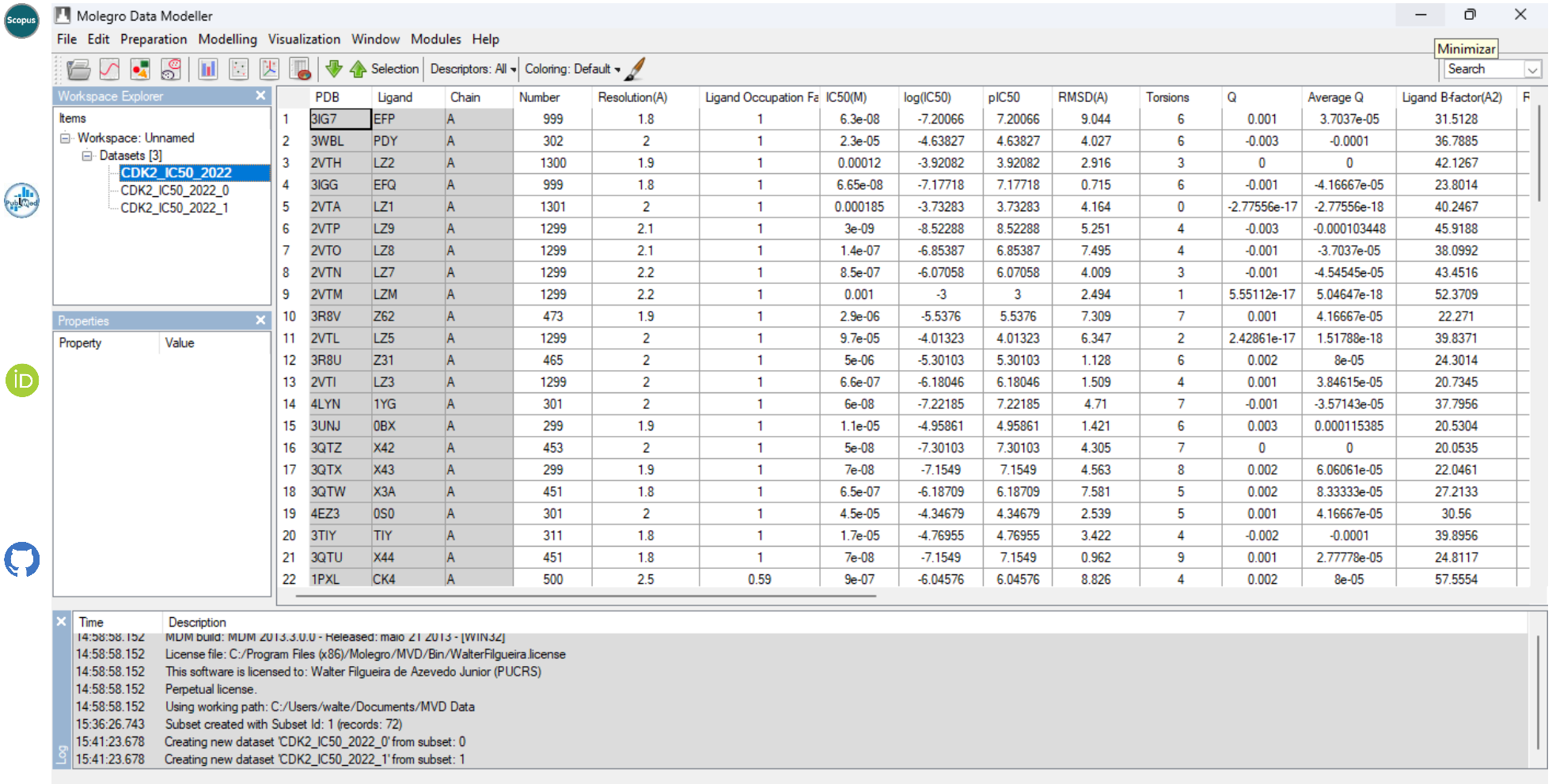
Conjunto de Dados

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)*.



Conjunto de Dados

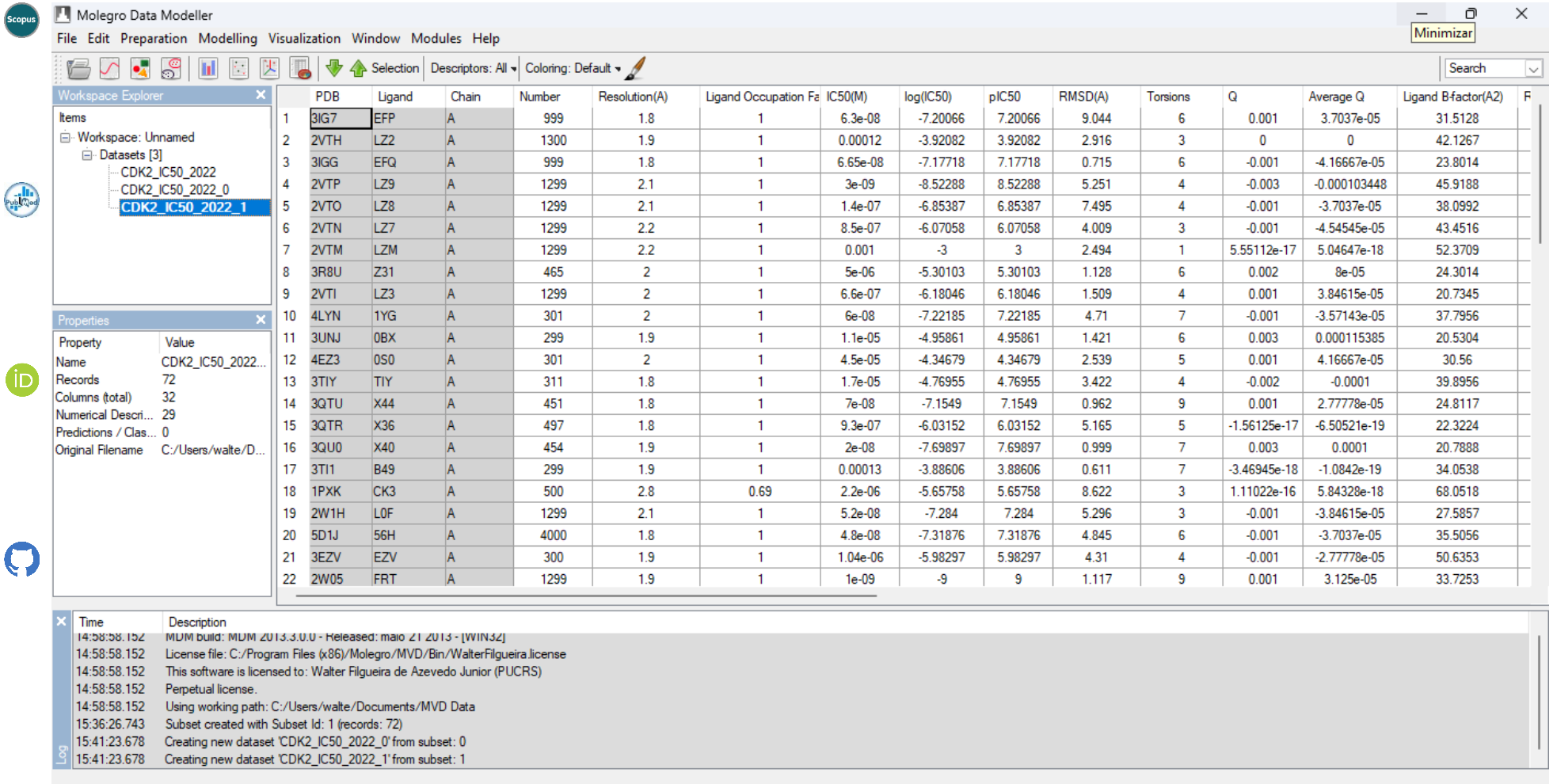
O MDM cria dois novos conjuntos de dados e mostra no *Workspace Explorer*. Foram criados os conjuntos: *CDK2_IC50_2022_0* e *CDK2_IC50_2022*_1. O primeiro tem 30 % dos dados e será usado como conjunto de teste e o segundo tem 70 % dos dados e será usado como conjunto de treinamento.



	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	

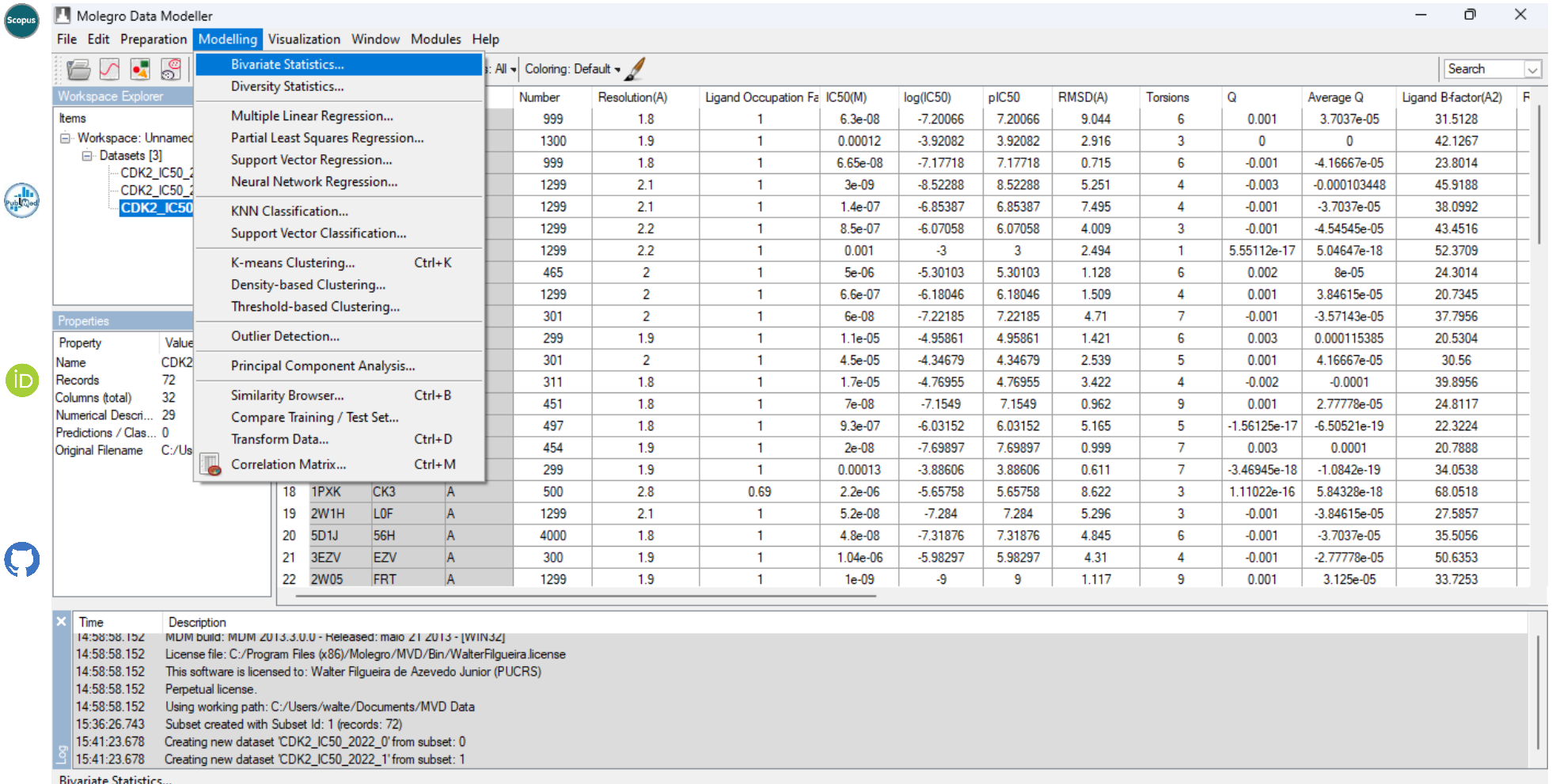
Análise Estatística

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).



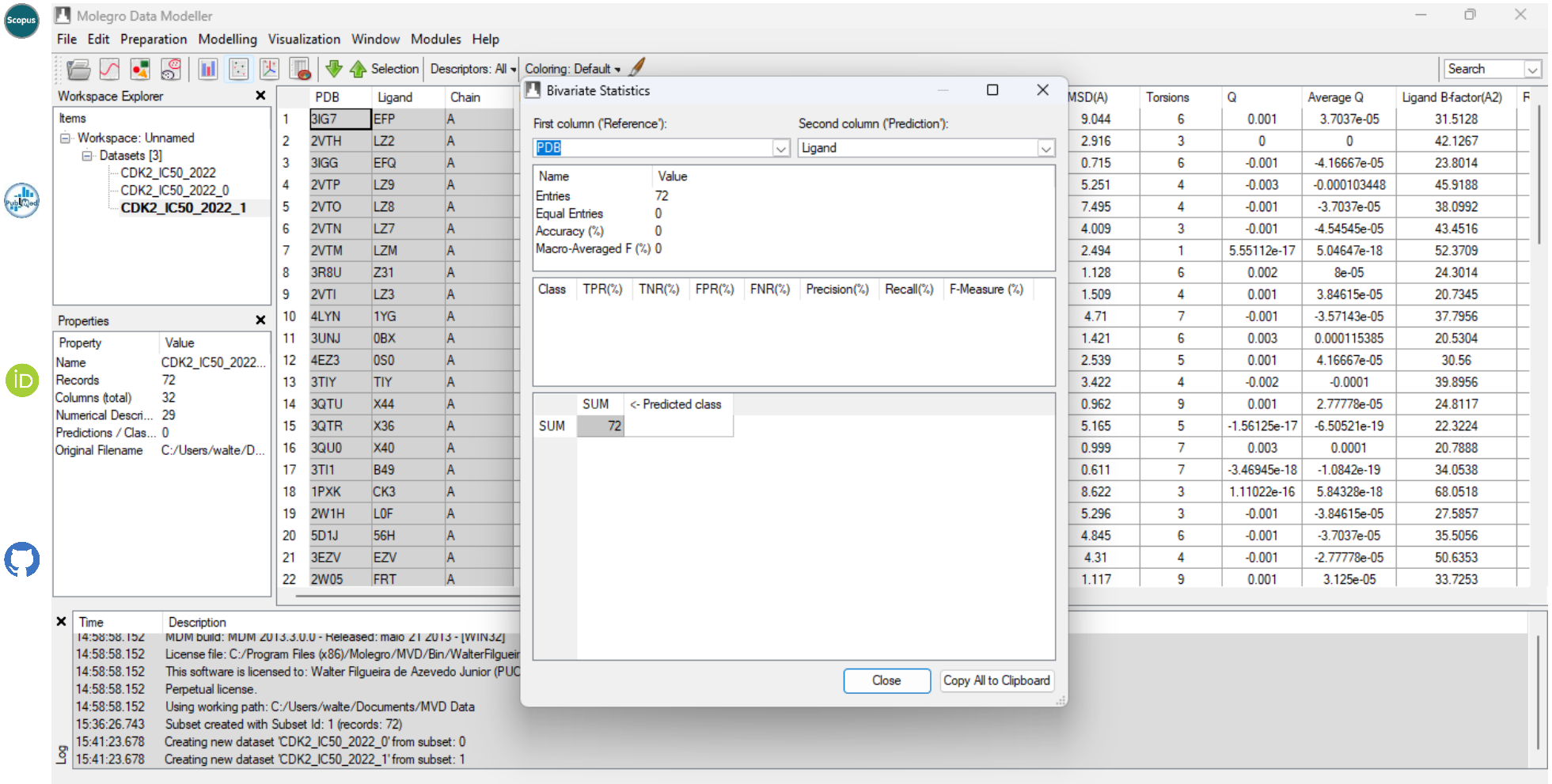
Análise Estatística

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



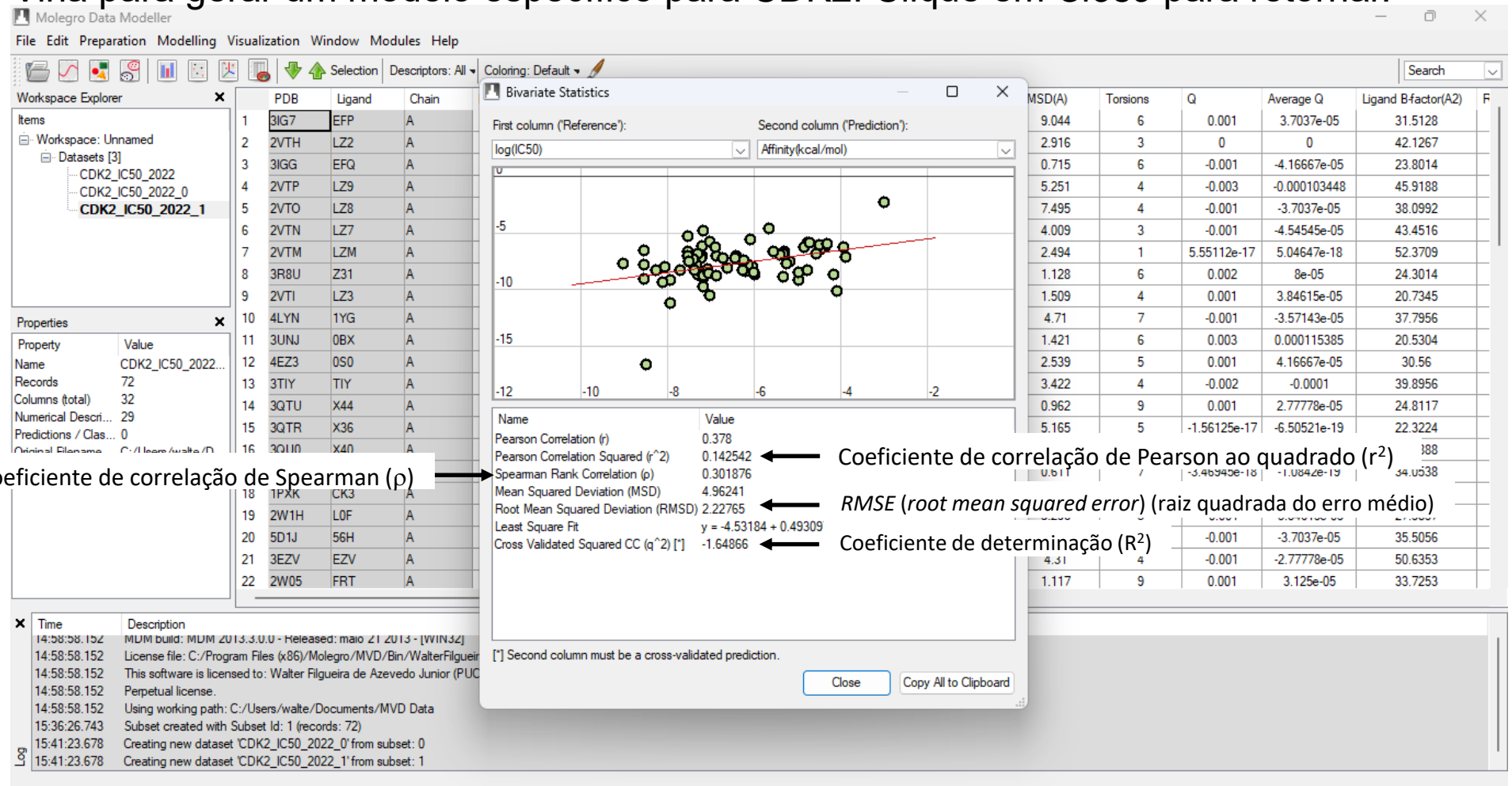
Análise Estatística

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



Análise Estatística

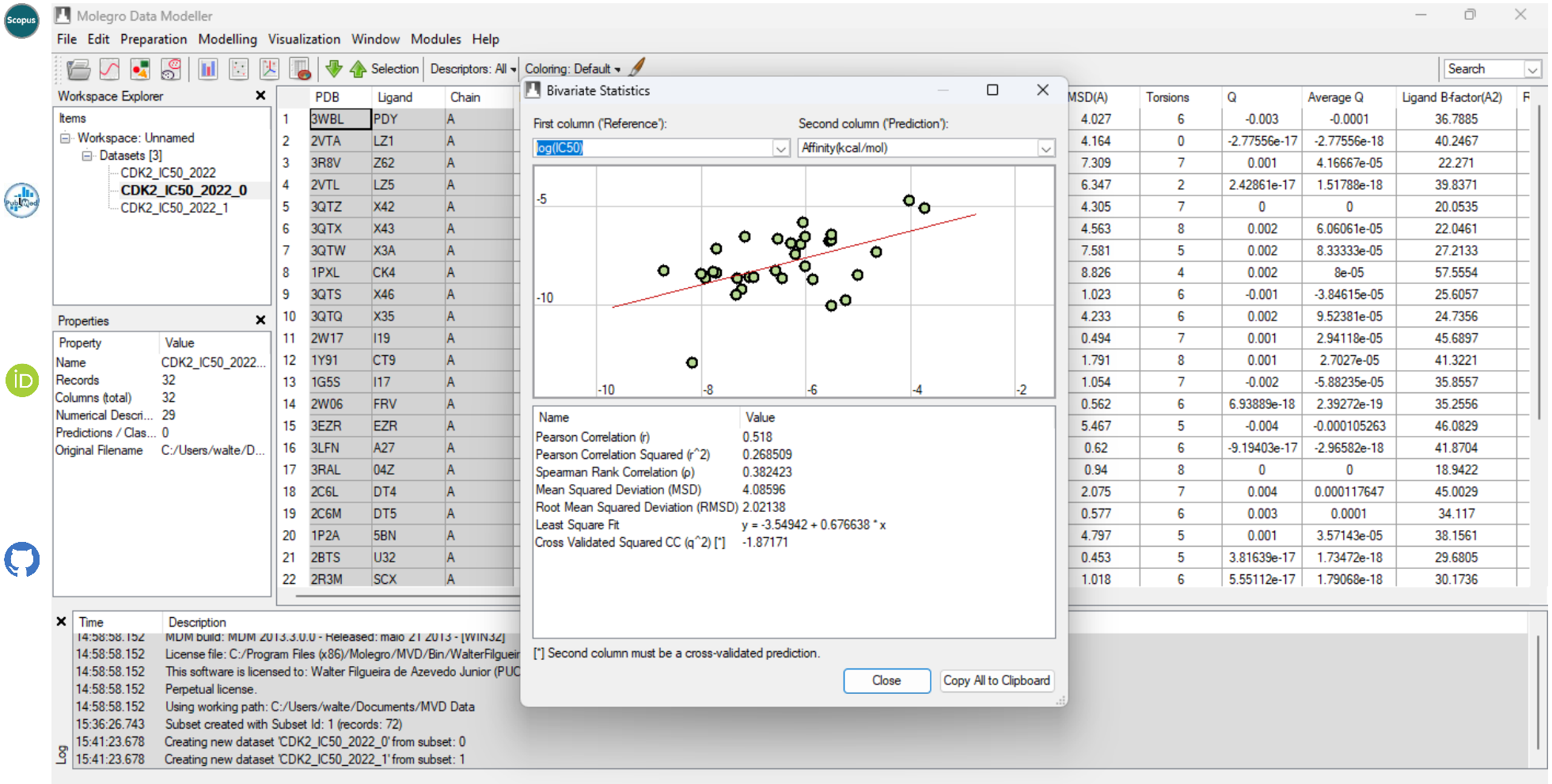
Selecionamos *log(IC_{50})* para a primeira coluna e *Affinity(kcal/mol)* para a segunda. O MDM gera um gráfico de dispersão e realiza uma análise estatística, como mostrado abaixo. Vemos um *RMSE = 2,22765* (o MDM chama de *RMSD*, mas é o *RMSE* previamente discutido). Podemos usar os pares de interações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



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.

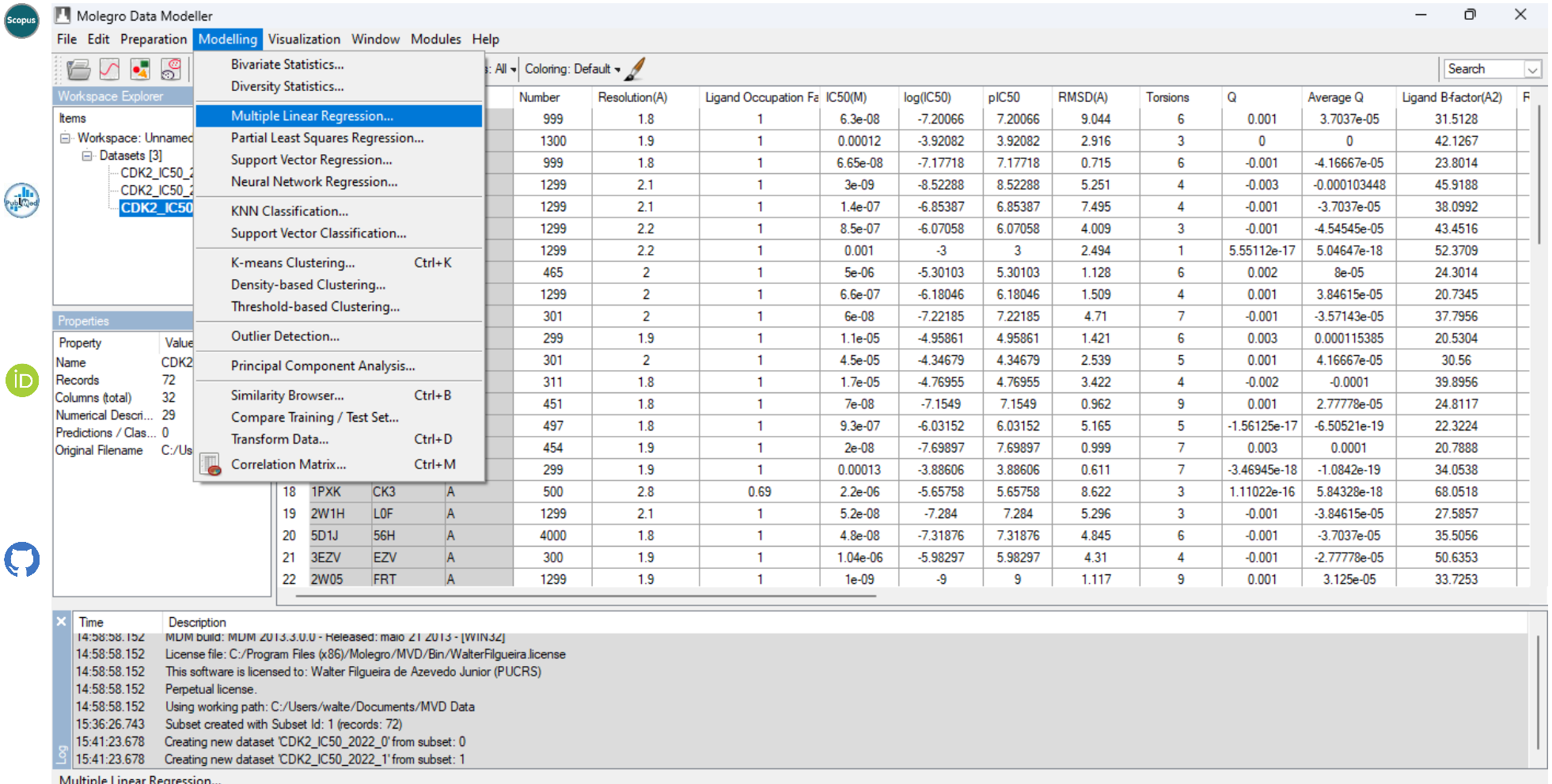
Análise Estatística

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.



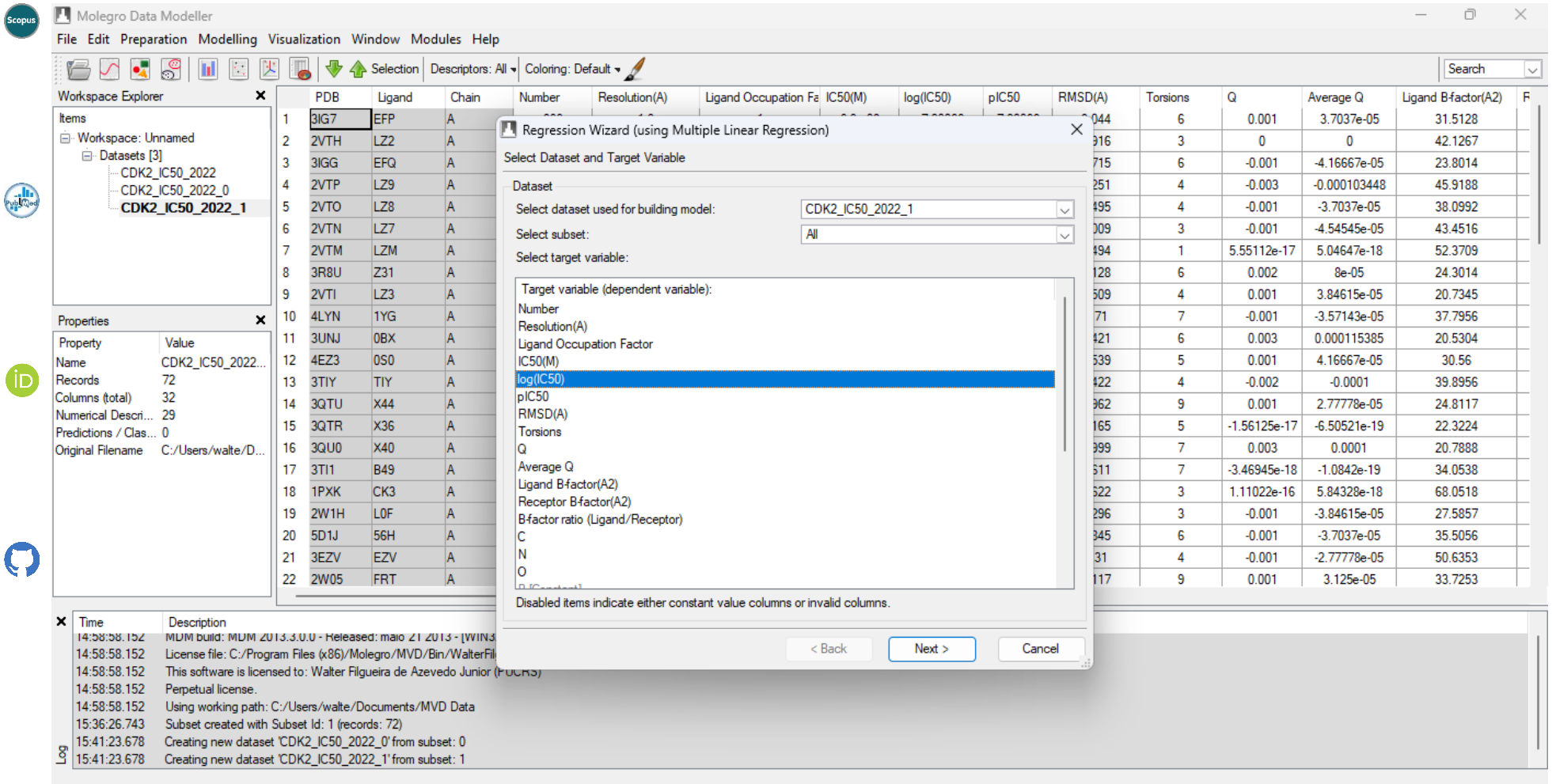
Modelo de Regressão

- 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...*



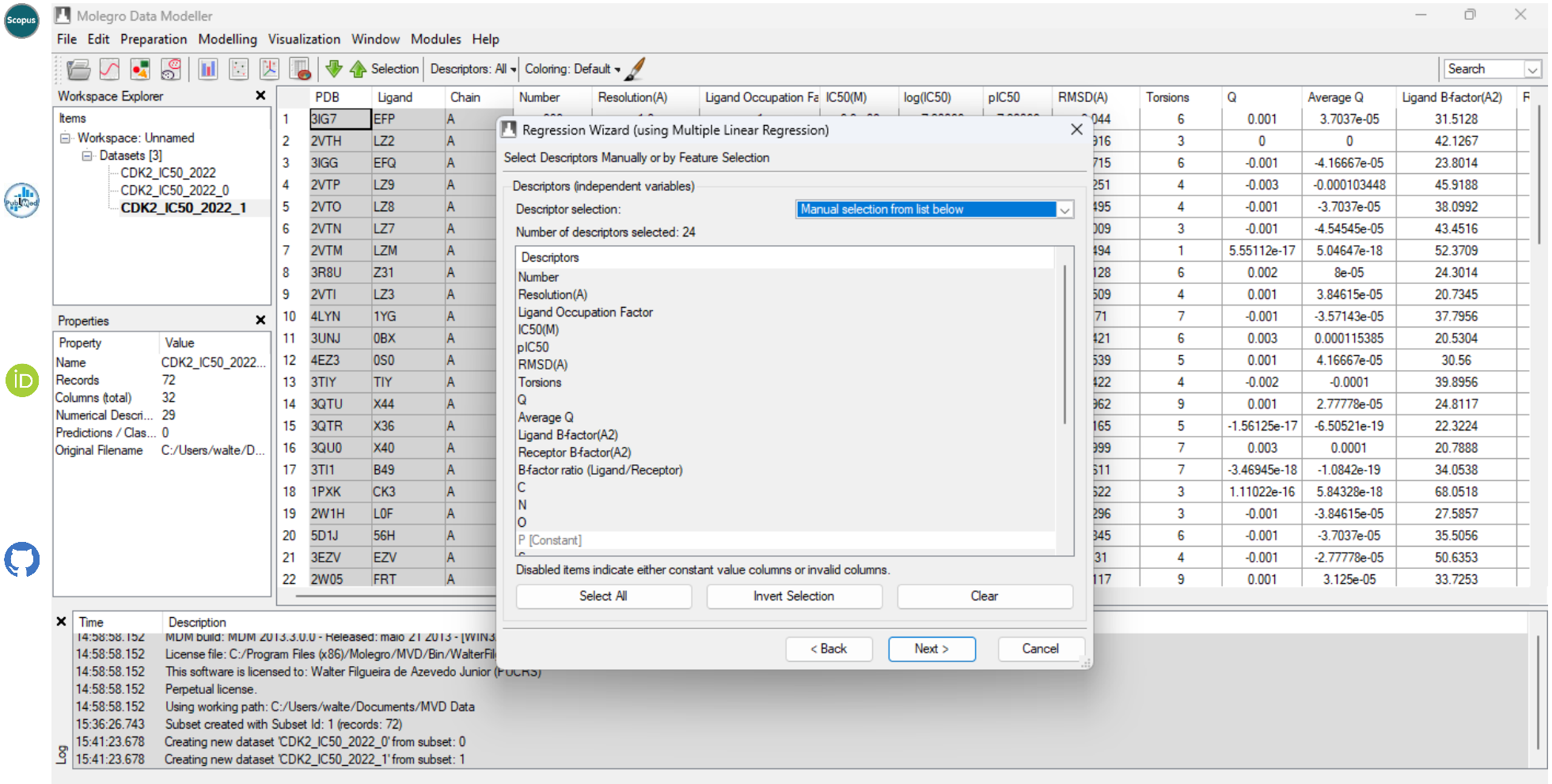
Modelo de Regressão

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



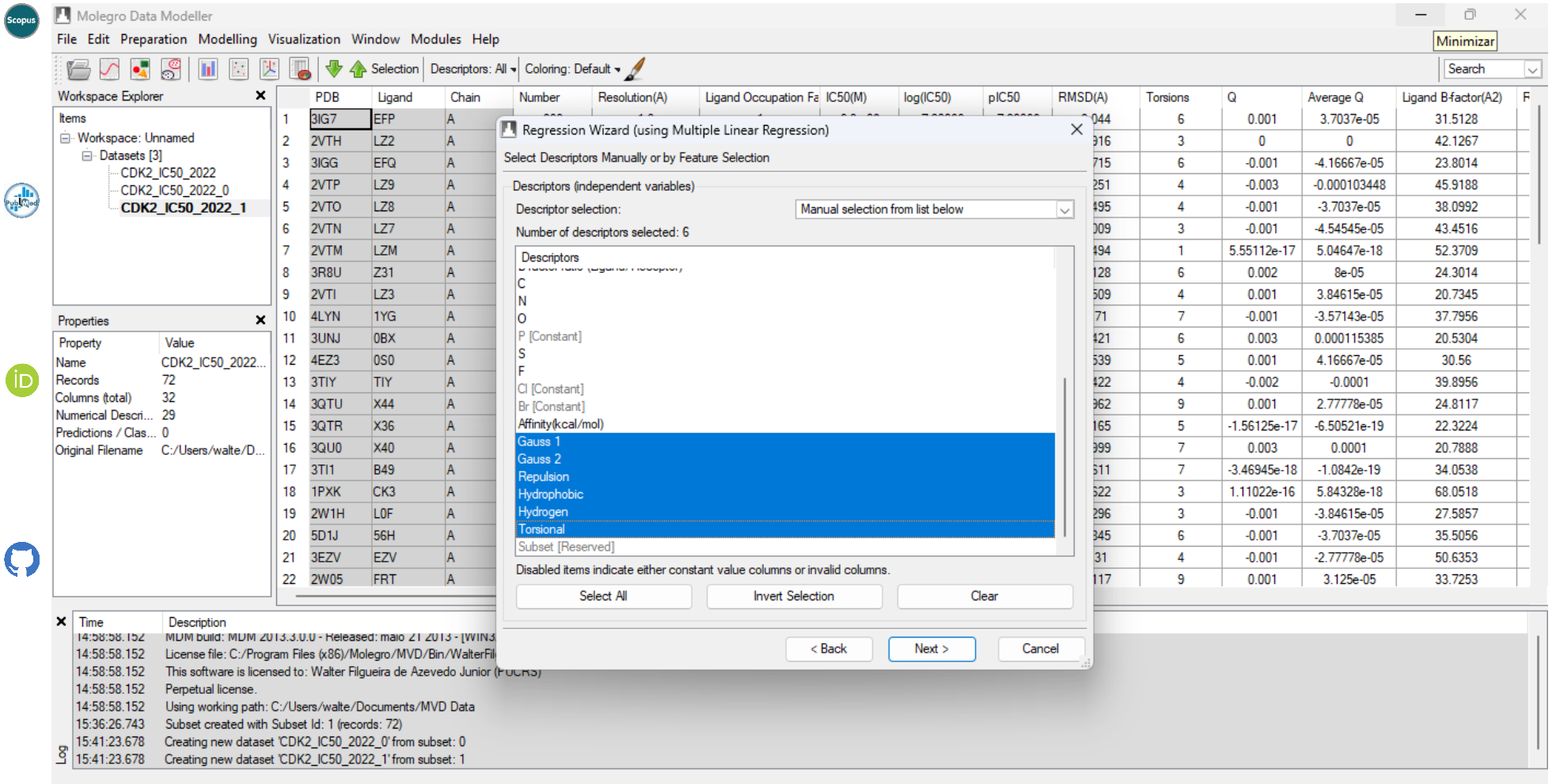
Modelo de Regressão

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



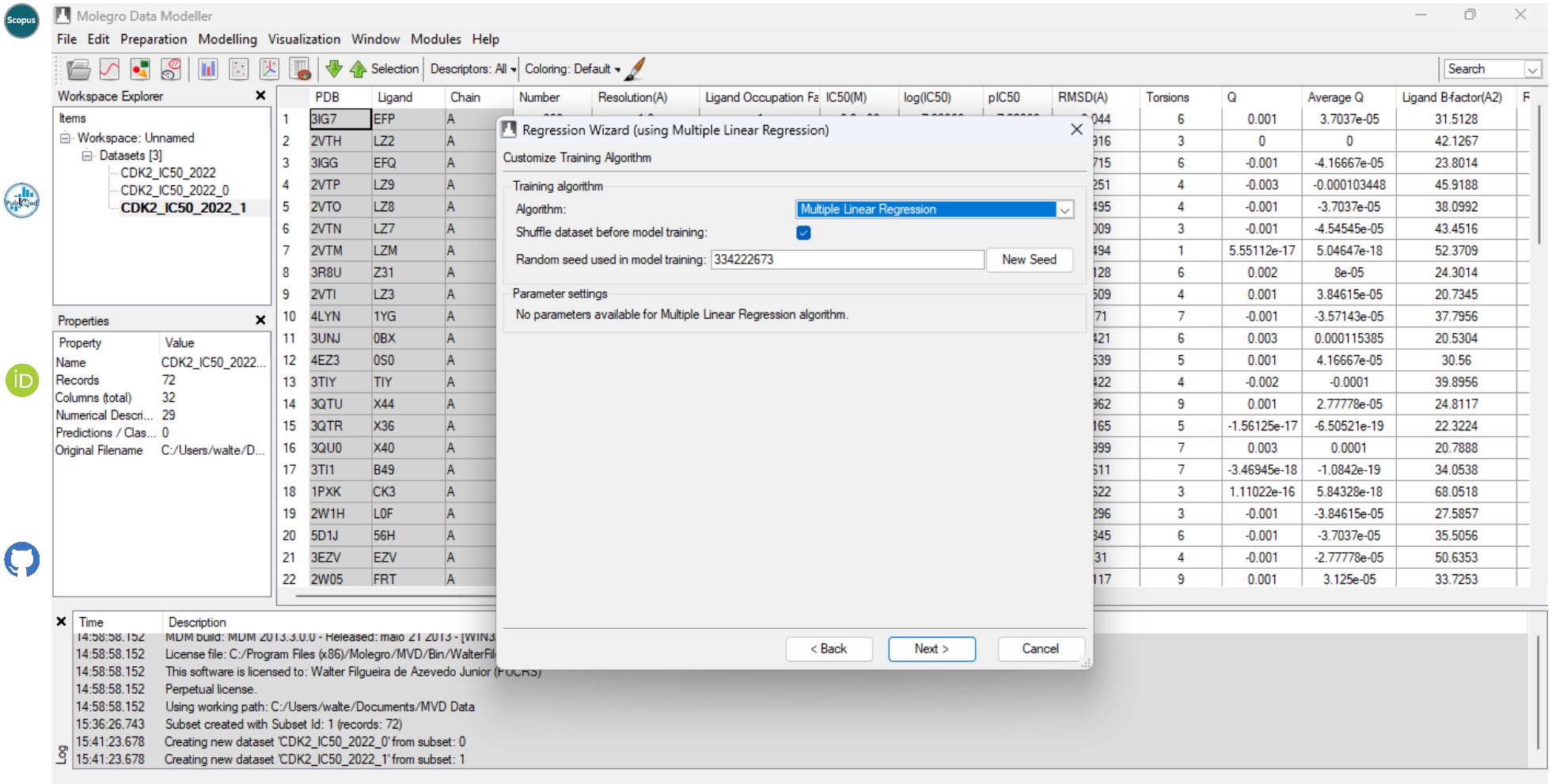
Modelo de Regressão

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*.



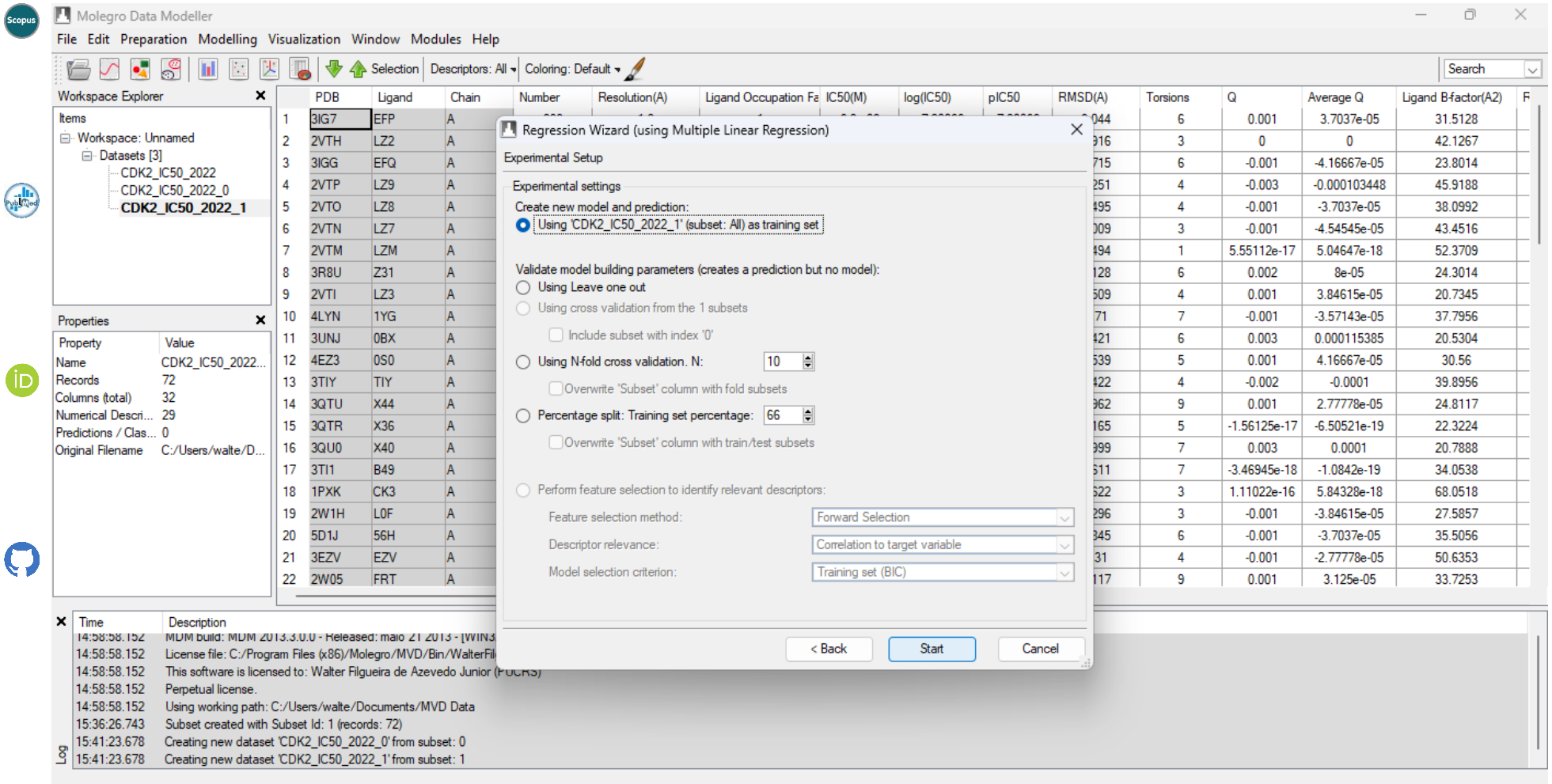
Modelo de Regressão

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



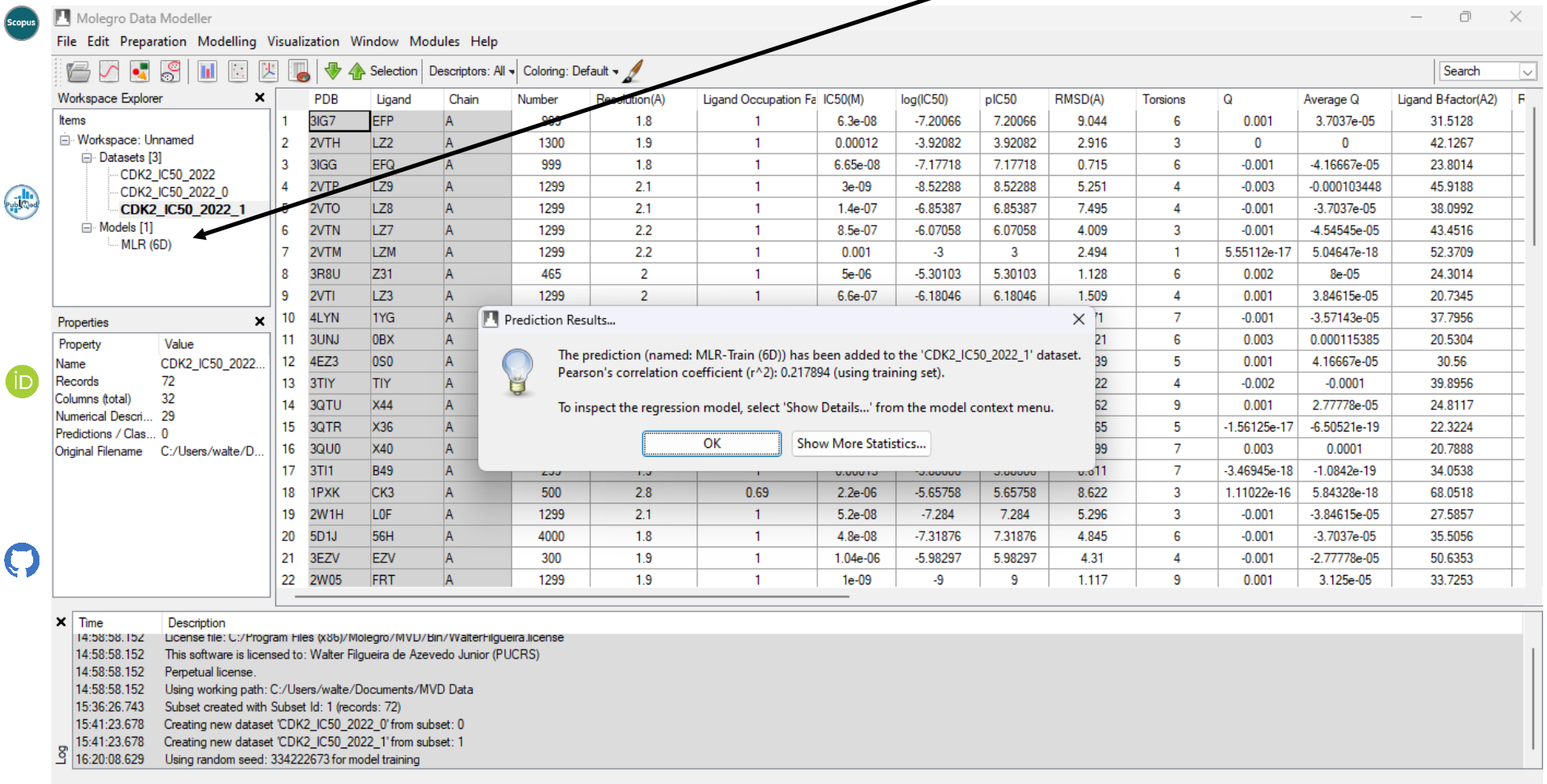
Modelo de Regressão

- Não altere as opções e clique em *Start*.



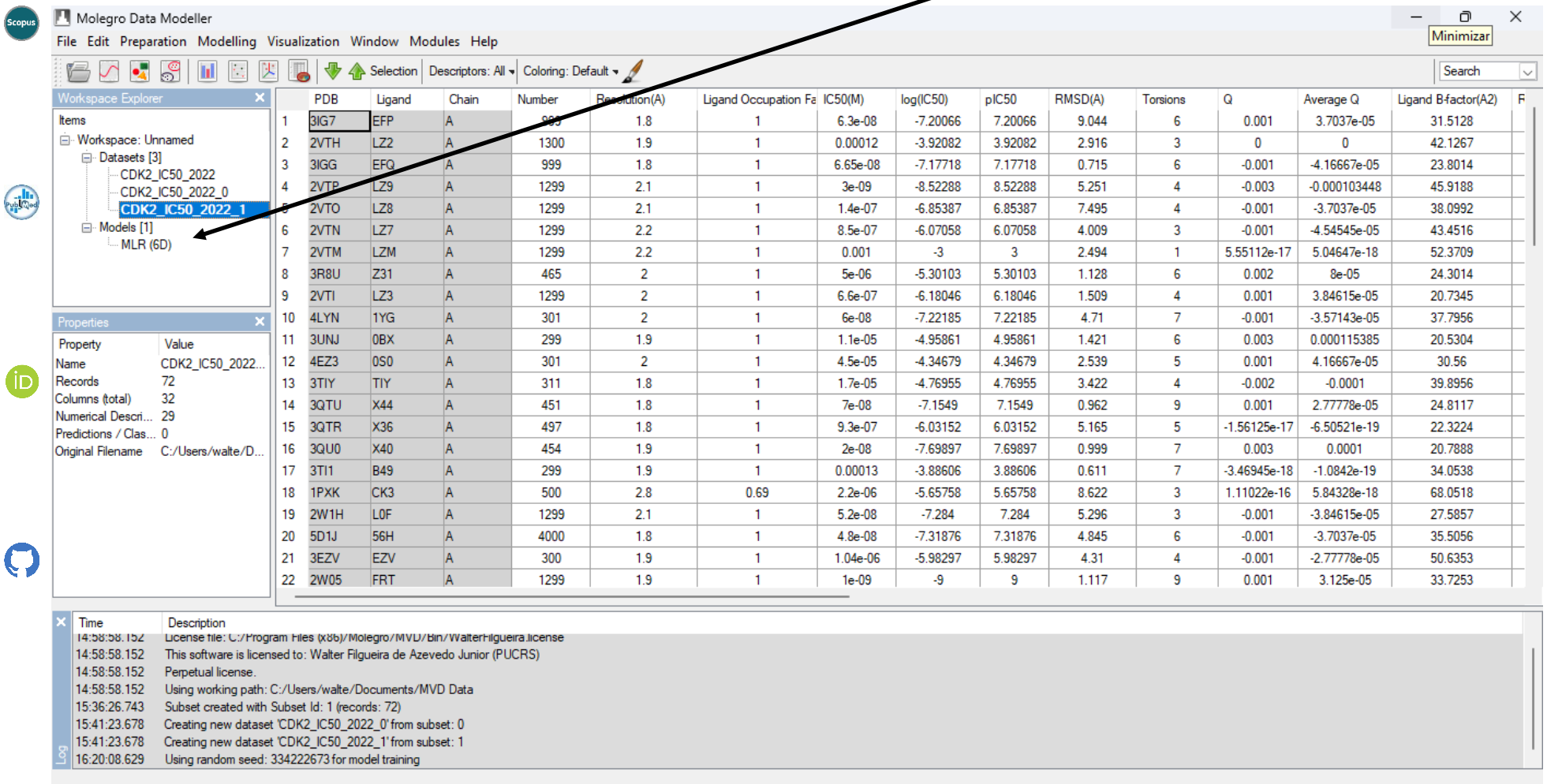
Modelo de Regressão

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



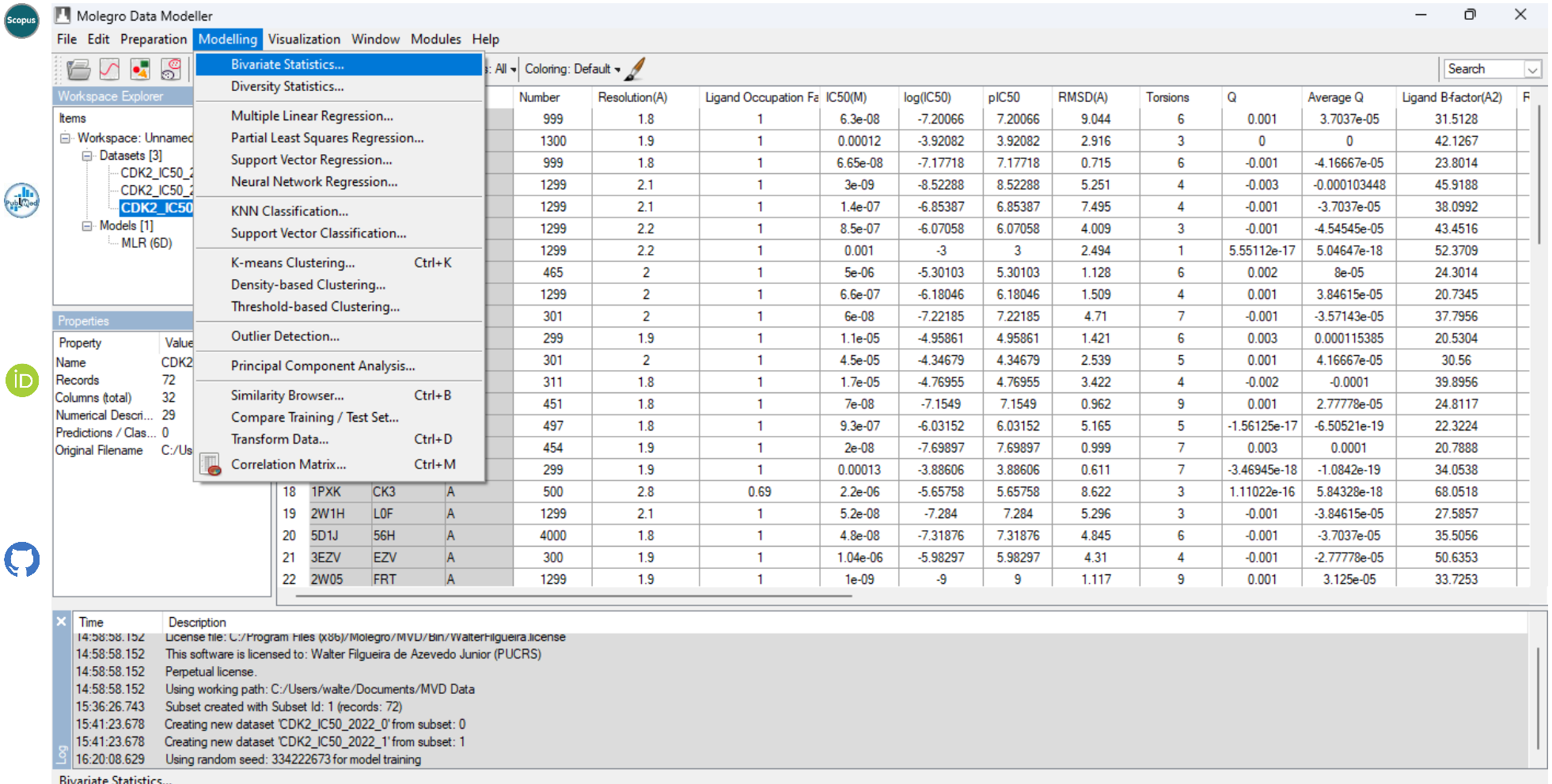
Modelo de Regressão

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



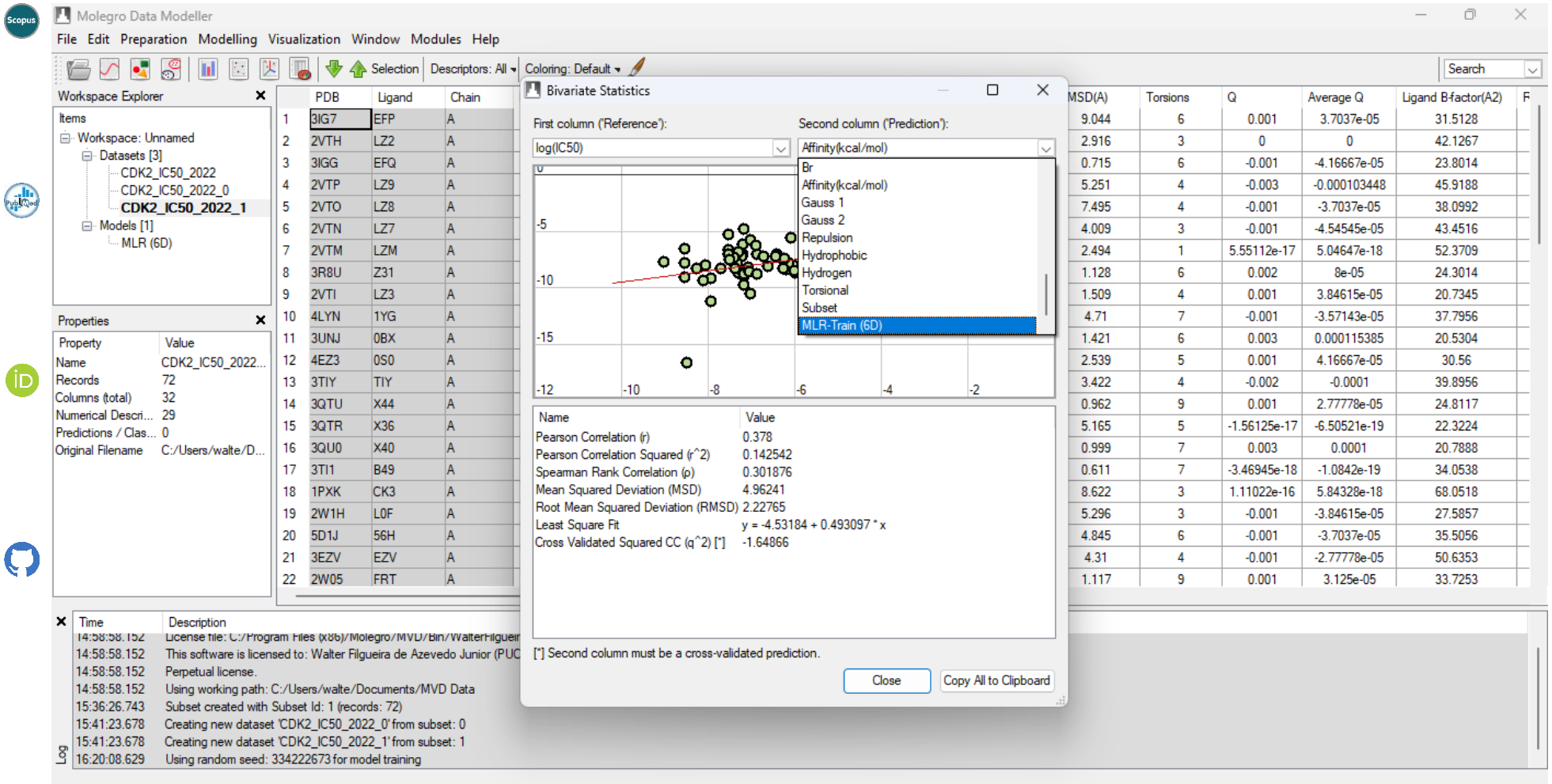
Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

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



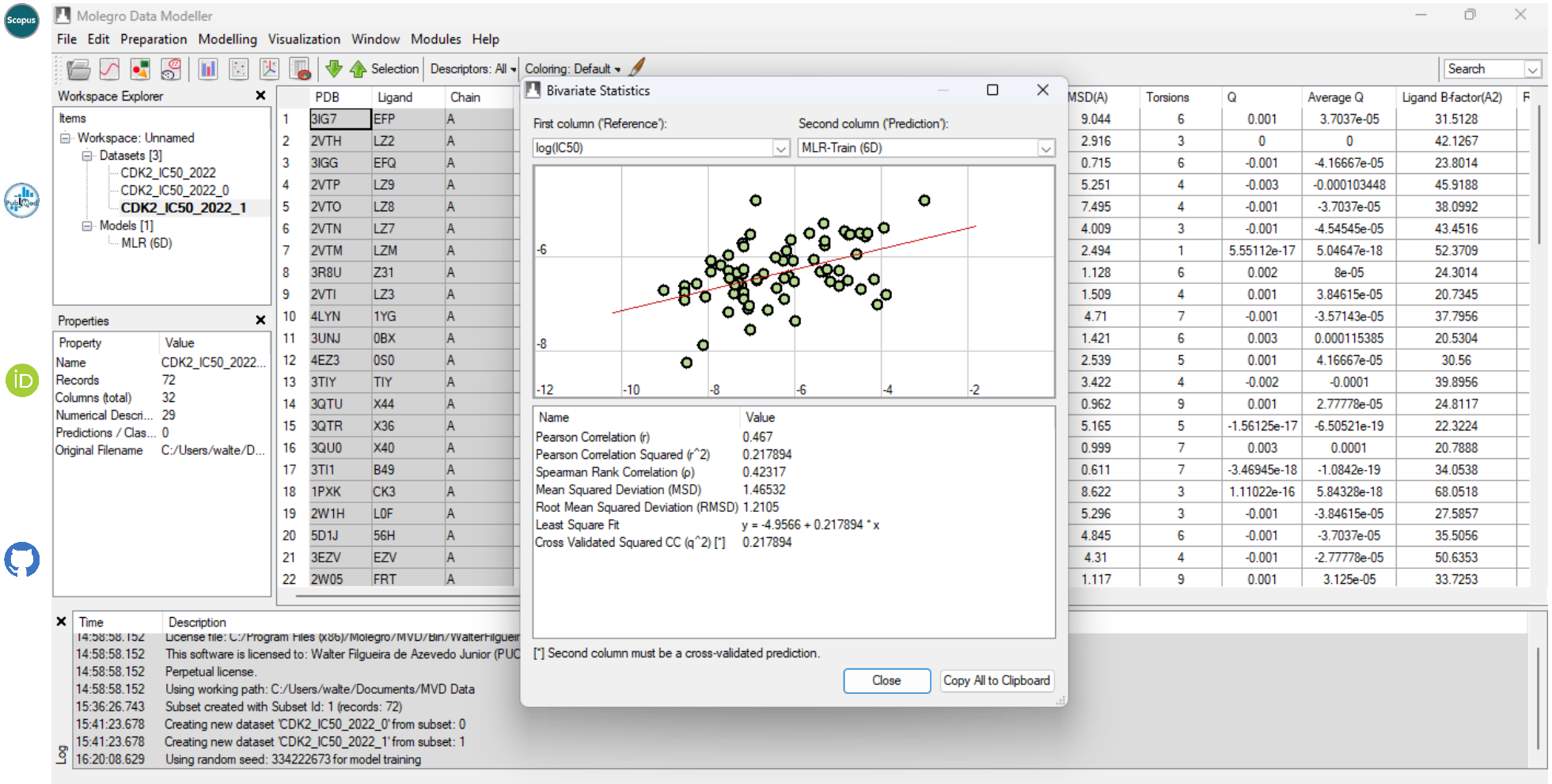
Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

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



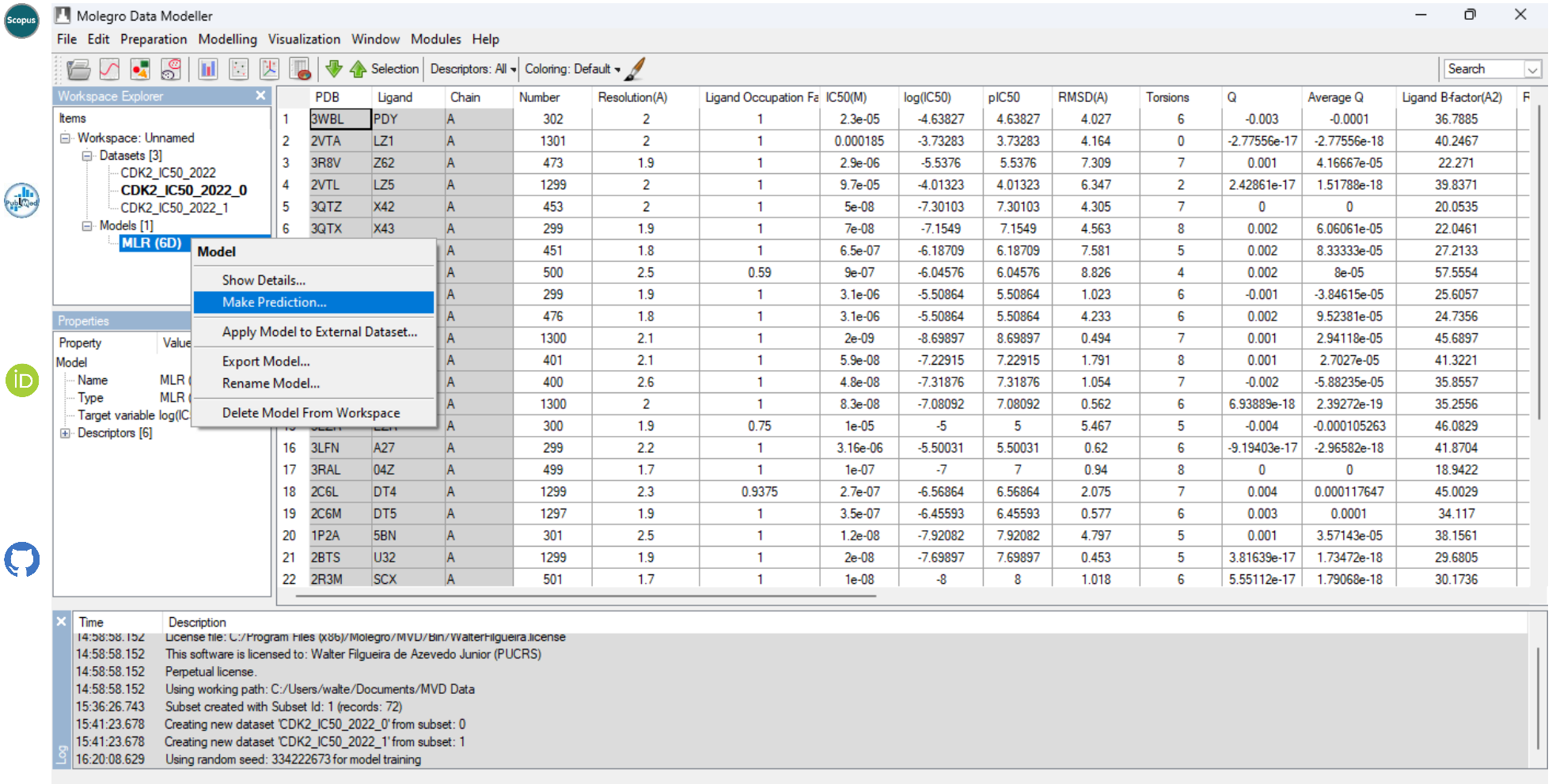
Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

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*.



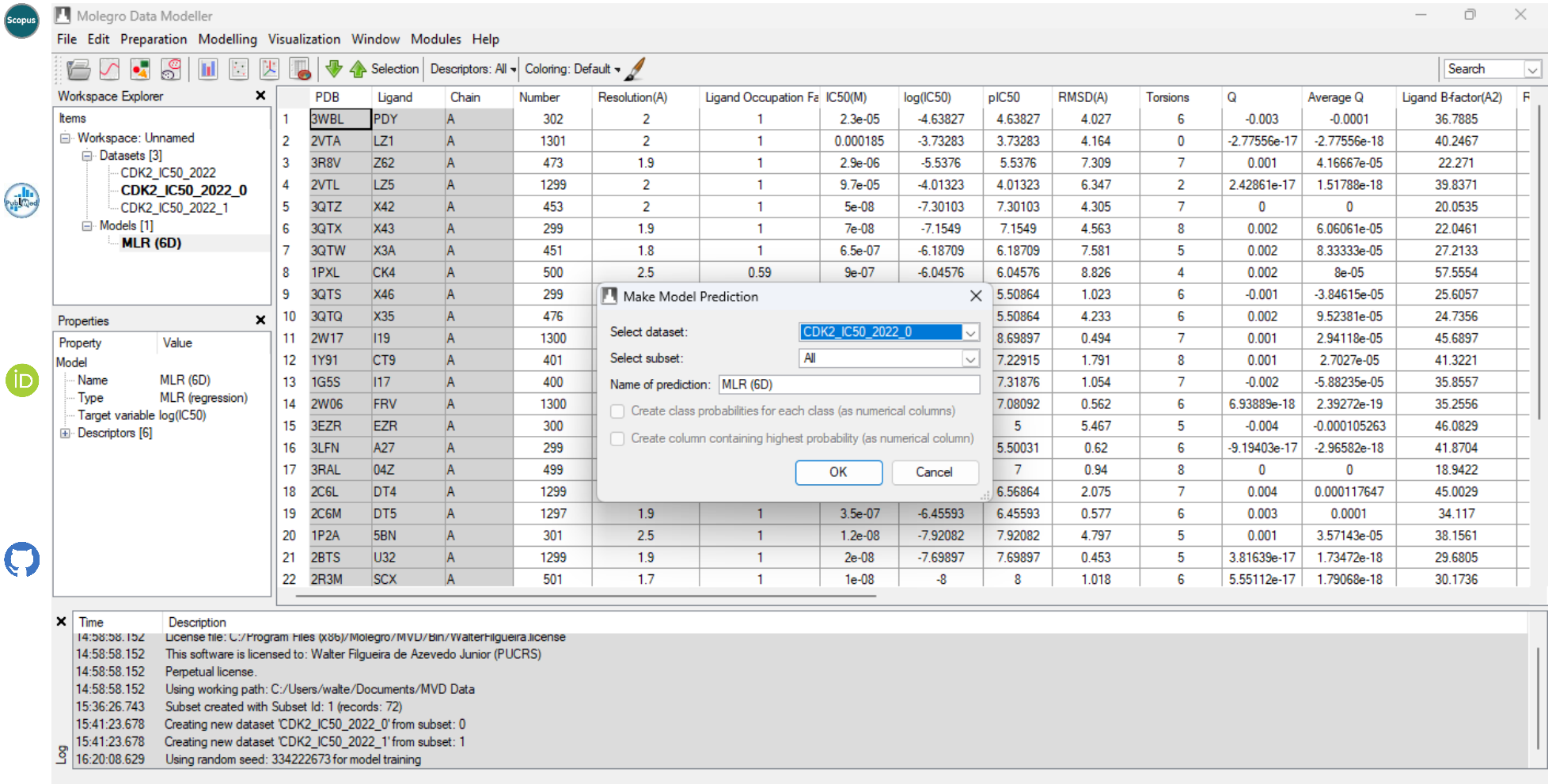
Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

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...*



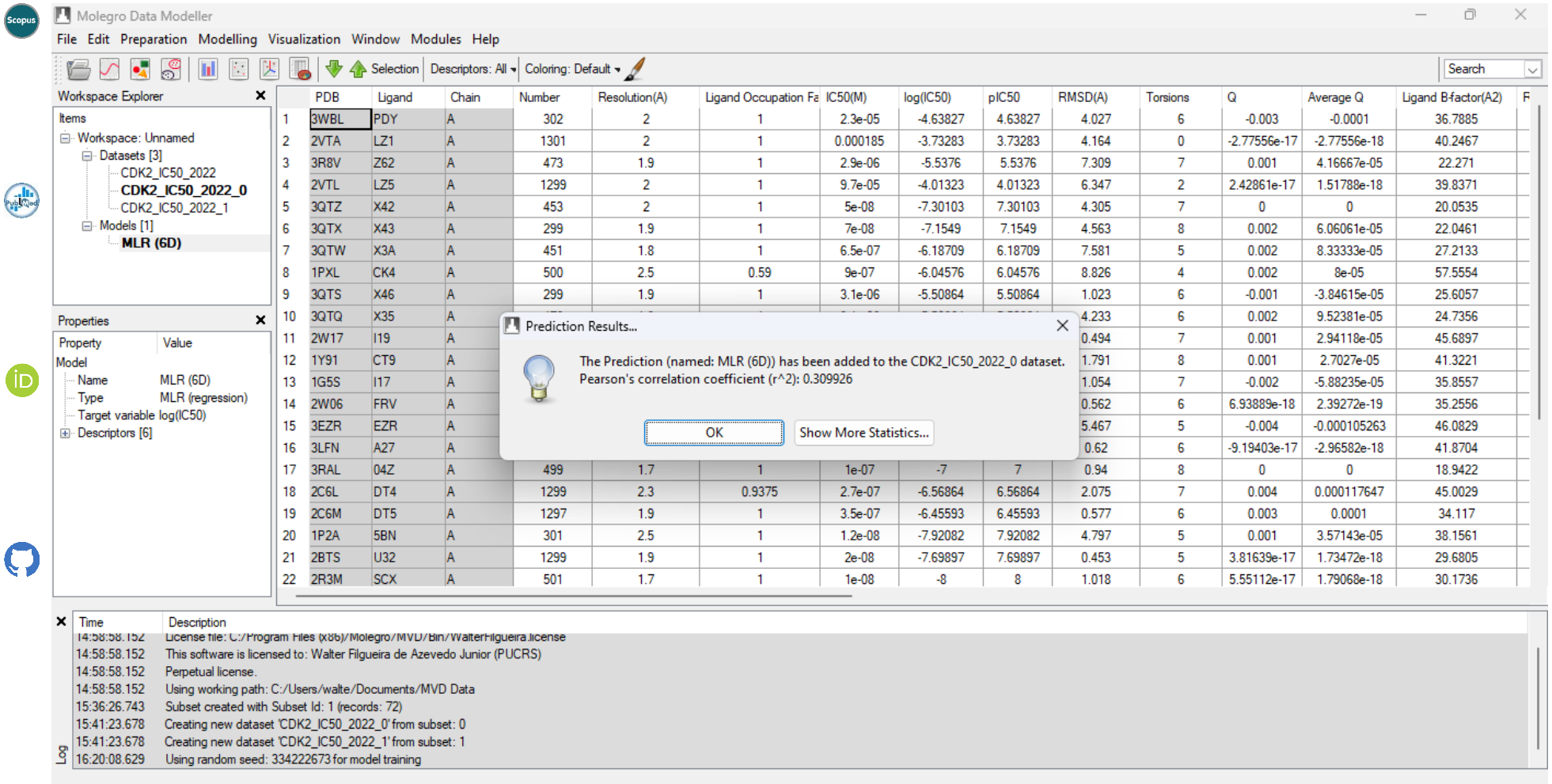
Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

Agora selecionamos o conjunto de teste (*CDK2_IC50_2022_0*) e clicamos OK.



Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

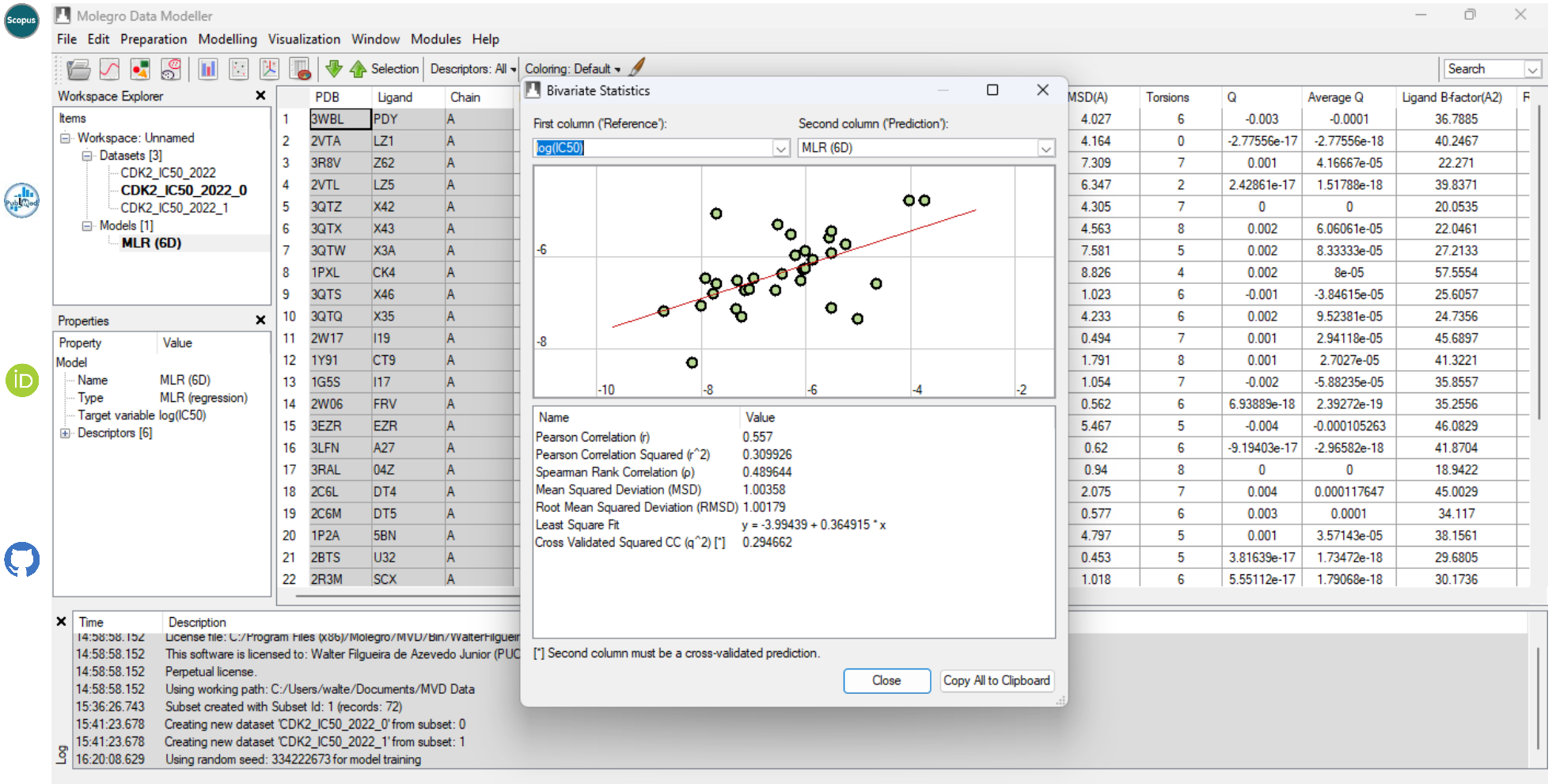
- Clique em *Show More Statistics...*



	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	3QTZ	X42	A	453	2	1	5e-08	-7.30103	7.30103	4.305	7	0	0	20.0535	
6	3QTX	X43	A	299	1.9	1	7e-08	-7.1549	7.1549	4.563	8	0.002	6.06061e-05	22.0461	
7	3QTW	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	3QTS	X46	A	299	1.9	1	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	1	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	

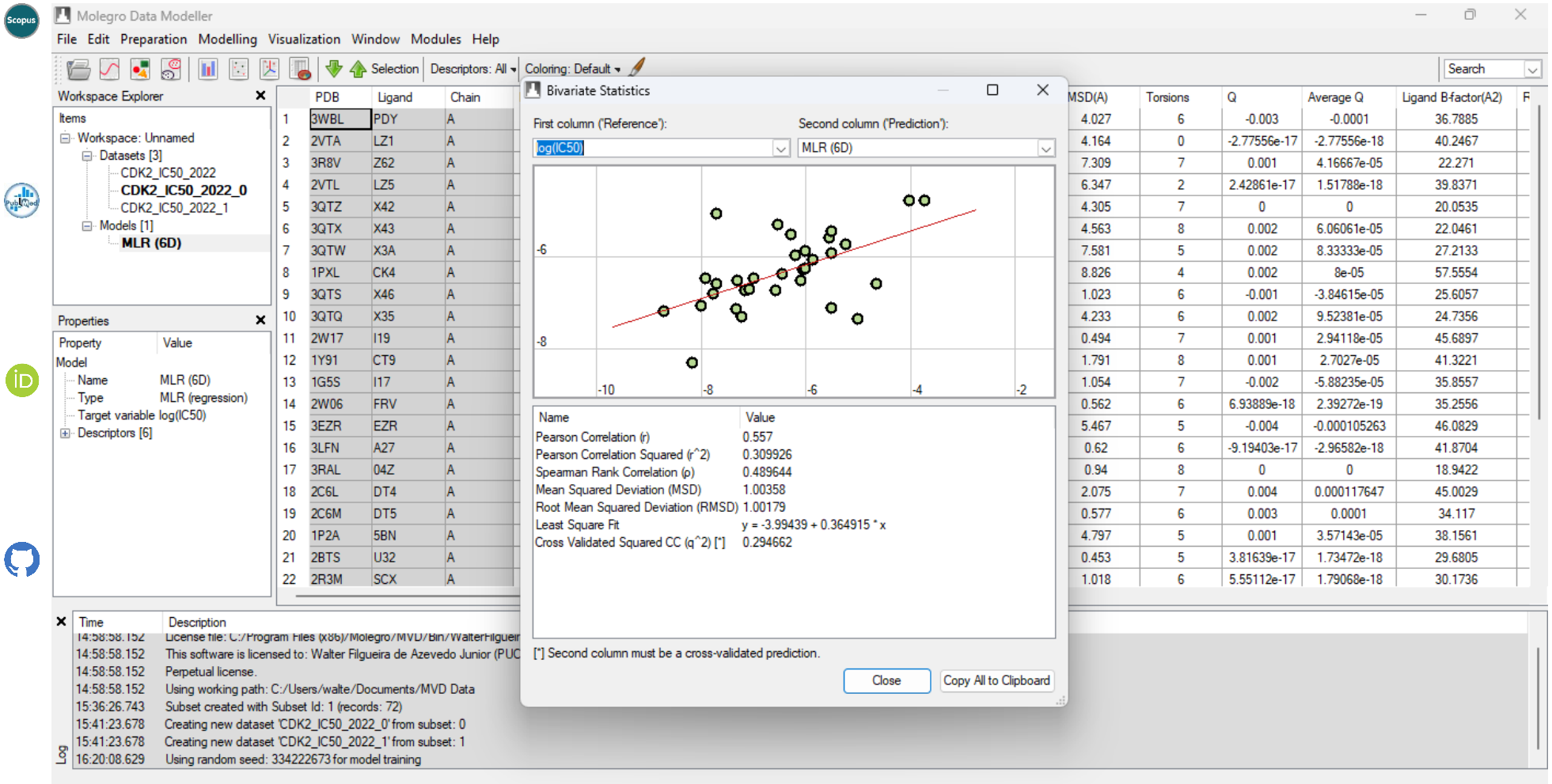
Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

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.



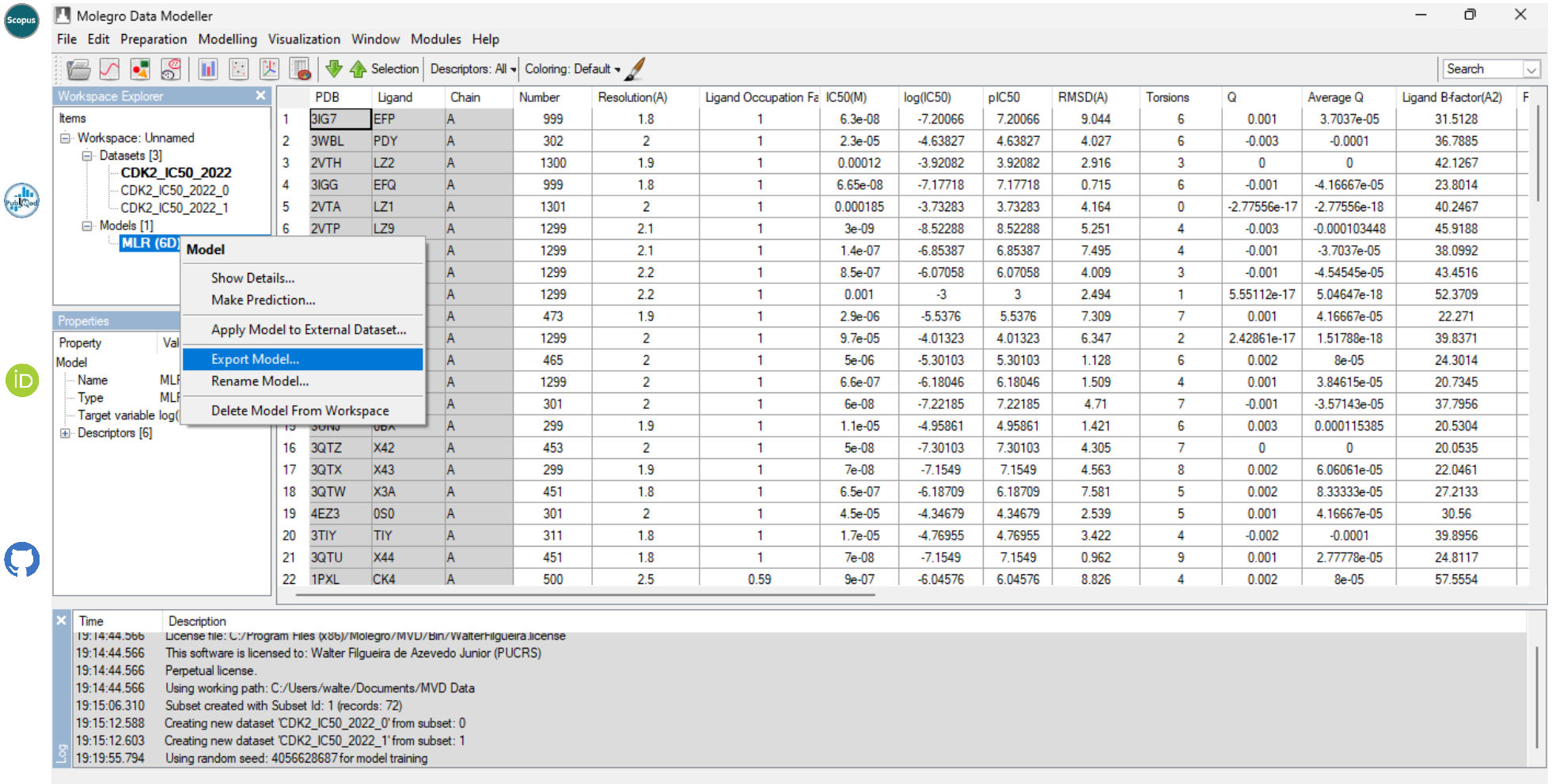
Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

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*.



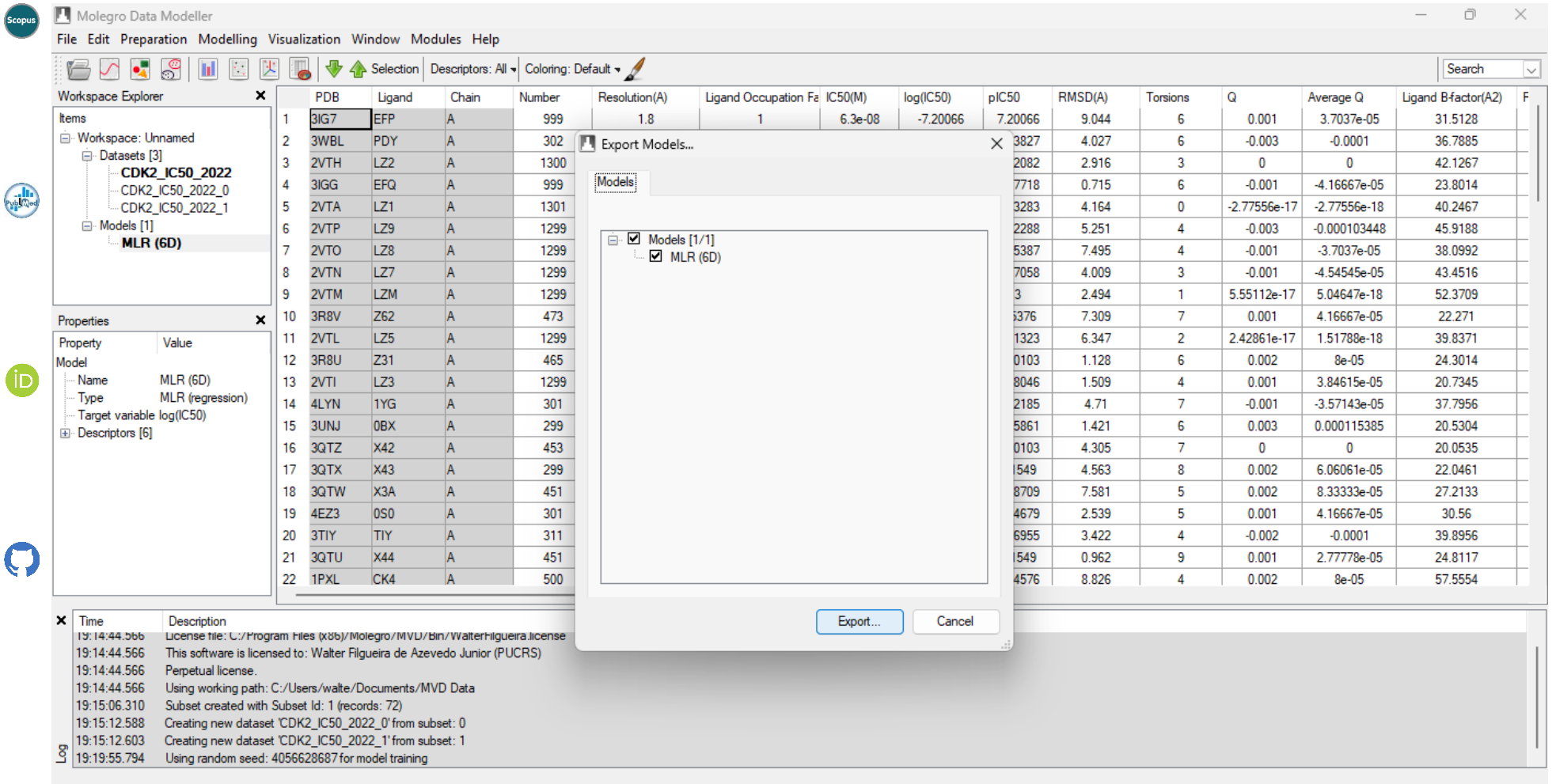
Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

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...*



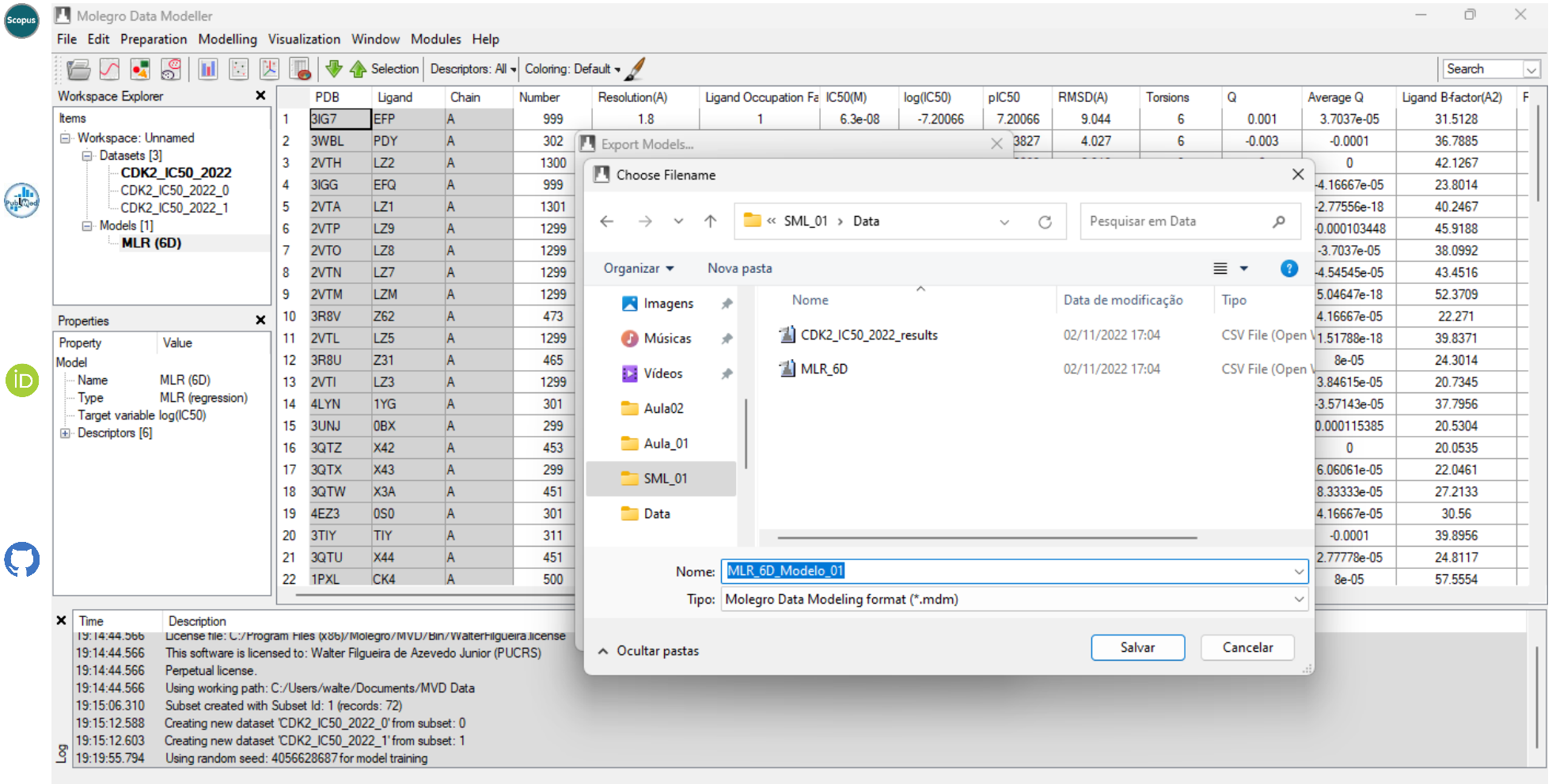
Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

- No novo menu clique em *Export...*



Análise do Poder de Previsão do Modelo de Aprendizado de Máquina

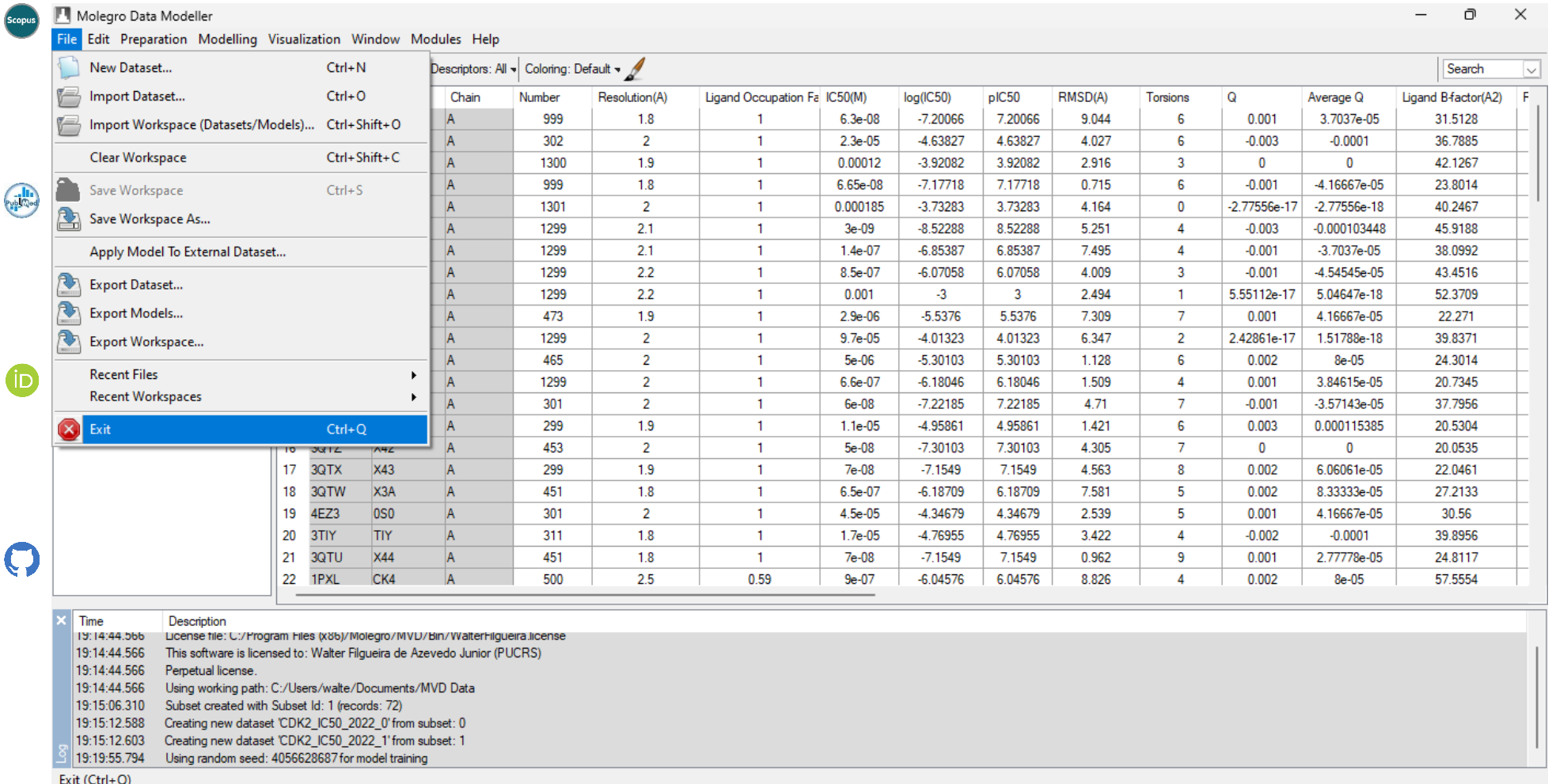
Escolha a pasta onde será salvo o modelo e coloque um nome nele. Aqui eu escolhi o nome: *MLR_6D_Modelo_01.mdm*. Clique no botão Salvar. Vocês agora têm um modelo de aprendizado de máquina salvo no seu computador.



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 <*Crtl*> *Q* para finalizar.



			Chain	Number	Resolution(A)	Ligand Occupation Fa	IC50(M)	log(IC50)	pIC50	RMSD(A)	Torsions	Q	Average Q	Ligand B-factor(A2)
			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

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	ρ	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) 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.