

# Aprendizado de Máquina Supervisionado IV



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

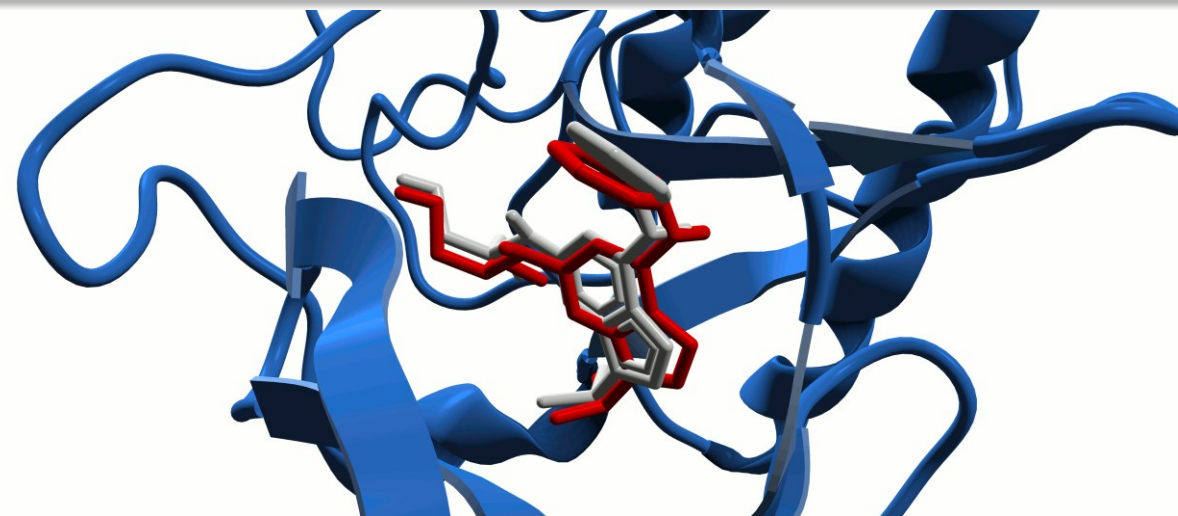
[walter@azevedolab.net](mailto:walter@azevedolab.net)

[Biography 01](#) ♥

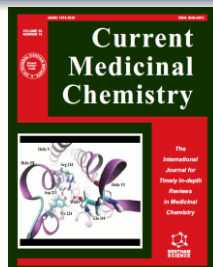
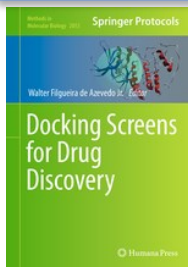
[Biography 02](#) ♥

[Biography 03](#) ♥

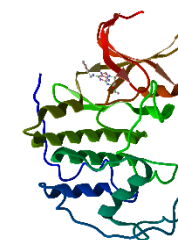
[Biography 04](#) ♥



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

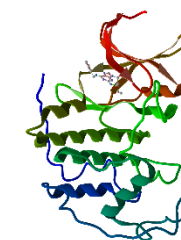
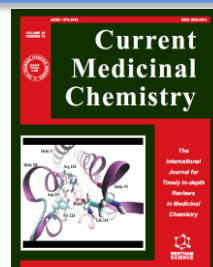
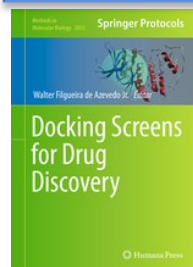


PROUD to be a Springer Author  
Read a free preview!



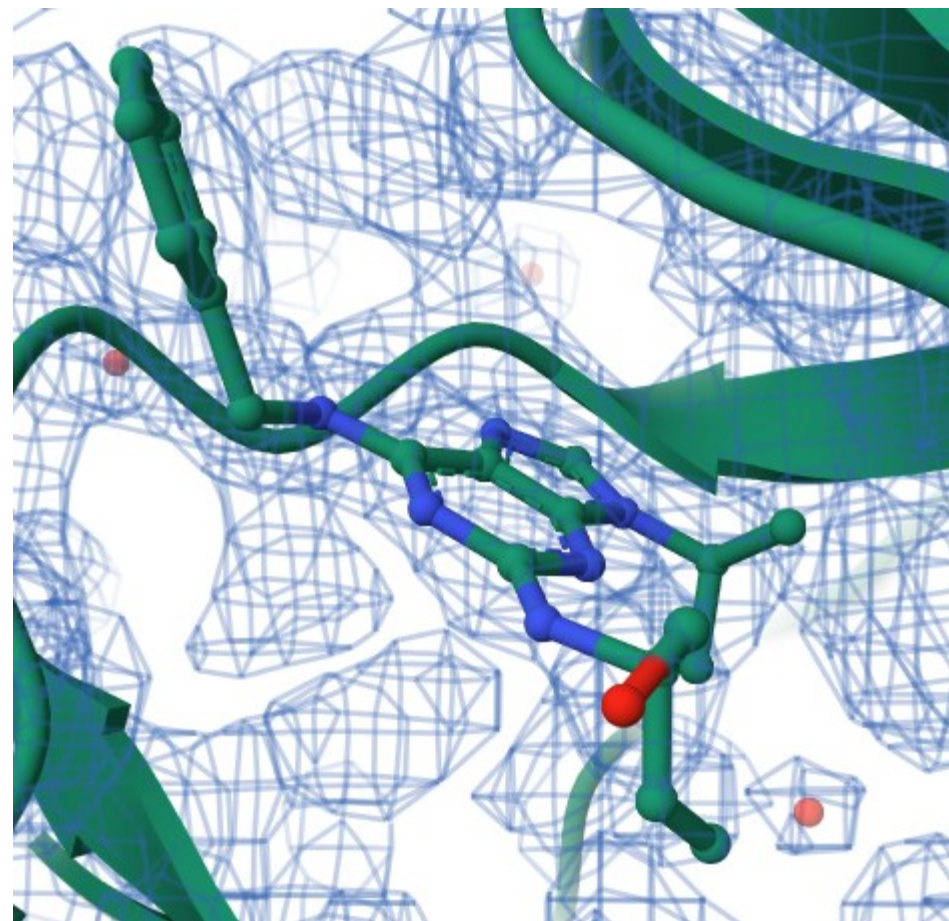
# Conteúdo

- [Resumo](#)
- [Sistema Proteína-Ligante](#)
- [Estruturas no PDB](#)
- [Espaço Químico](#)
- [Espaço de Proteínas](#)
- [Espaço de Funções Escores](#)
- [Taba](#)
- [SAnDReS](#)
- [Rede Internacional de Pesquisadores](#)
- [Exemplos de Projetos Propostos](#)



## Resumo

Nesta aula mostraremos como métodos de aprendizado de máquina supervisionados podem contribuir para o desenvolvimento de pesquisas com foco na descoberta de fármacos. A partir das informações sobre a estrutura tridimensional de proteínas relacionadas a alguma patologia, podemos usar as informações experimentais para treinarmos modelos de aprendizado de máquinas. Uma vez treinados e validados, esses modelos são usados para busca por moléculas que podem apresentar ação farmacológica. A partir da analogia chave-fechadura, podemos pensar que os métodos de aprendizado de máquina nos auxiliam na busca de novas chaves (fármacos em potencial) que se encaixam numa dada fechadura (parte da proteína que interage com o fármaco). Iremos apresentar o conceito de espaço de funções escores e como usá-lo para construir modelos de aprendizado de máquina direcionados para proteínas de interesse.

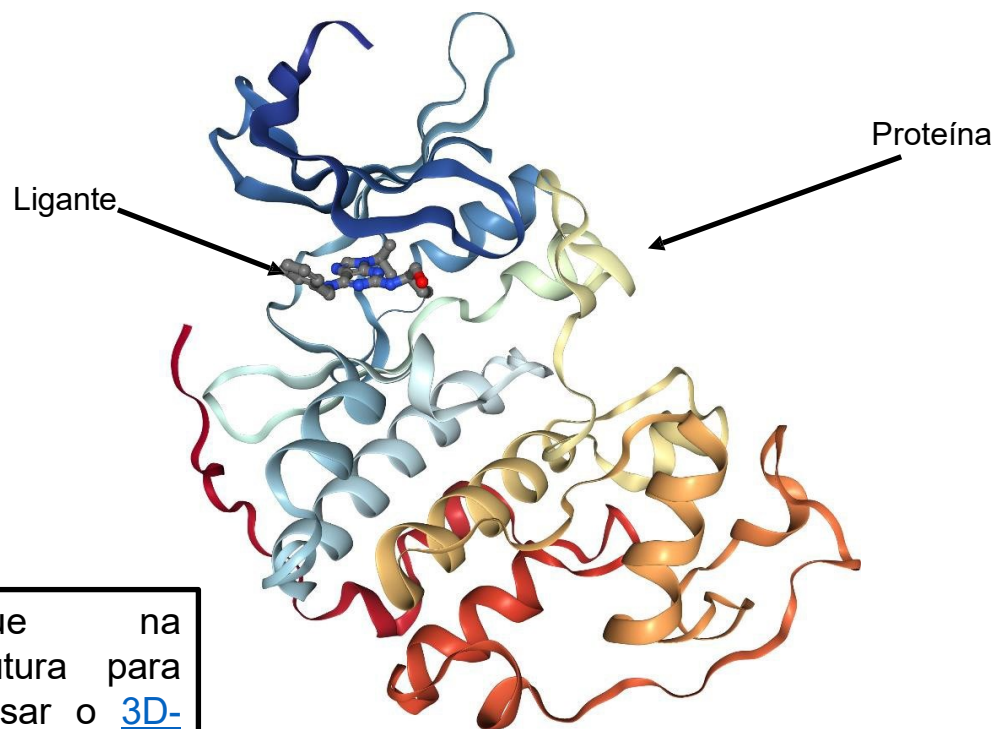


Mapa de densidade eletrônica da estrutura da CDK2 em complexo com Roscovitina. A região mostrada é do sítio de ligação (fechadura).



## Sistema Proteína-Ligante

Abaixo temos a estrutura do complexo da CDK2 com o fármaco Roscovitina. A molécula de Roscovitina é um inibidor competitivo, que impede que a molécula de ATP se ligue à proteína CDK2. A estrutura abaixo foi obtida a partir da técnica chamada cristalografia por difração de raios X. A principal informação obtida a partir da difração de raios X é a estrutura da molécula. Quando dizemos estrutura, nos referimos às coordenadas atômicas de todos os átomos da proteína e do ligante.

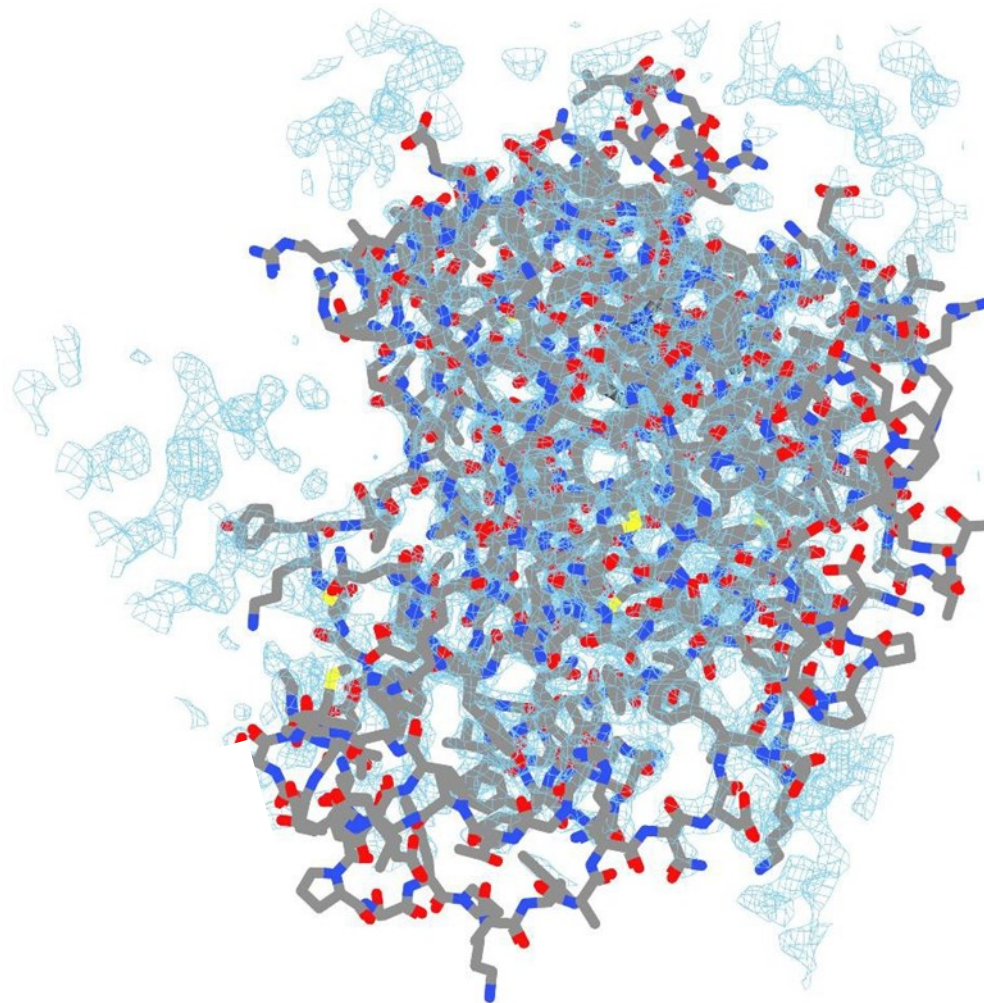


Clique na  
estrutura para  
acessar o [3D-  
View](#)



## Sistema Proteína-Ligante

As coordenadas atômicas são obtidas a partir do mapa de densidade eletrônica, indicado com gradeado azul abaixo. Esse gradeado é uma informação experimental derivada dos dados de cristalografia por difração de raios X.

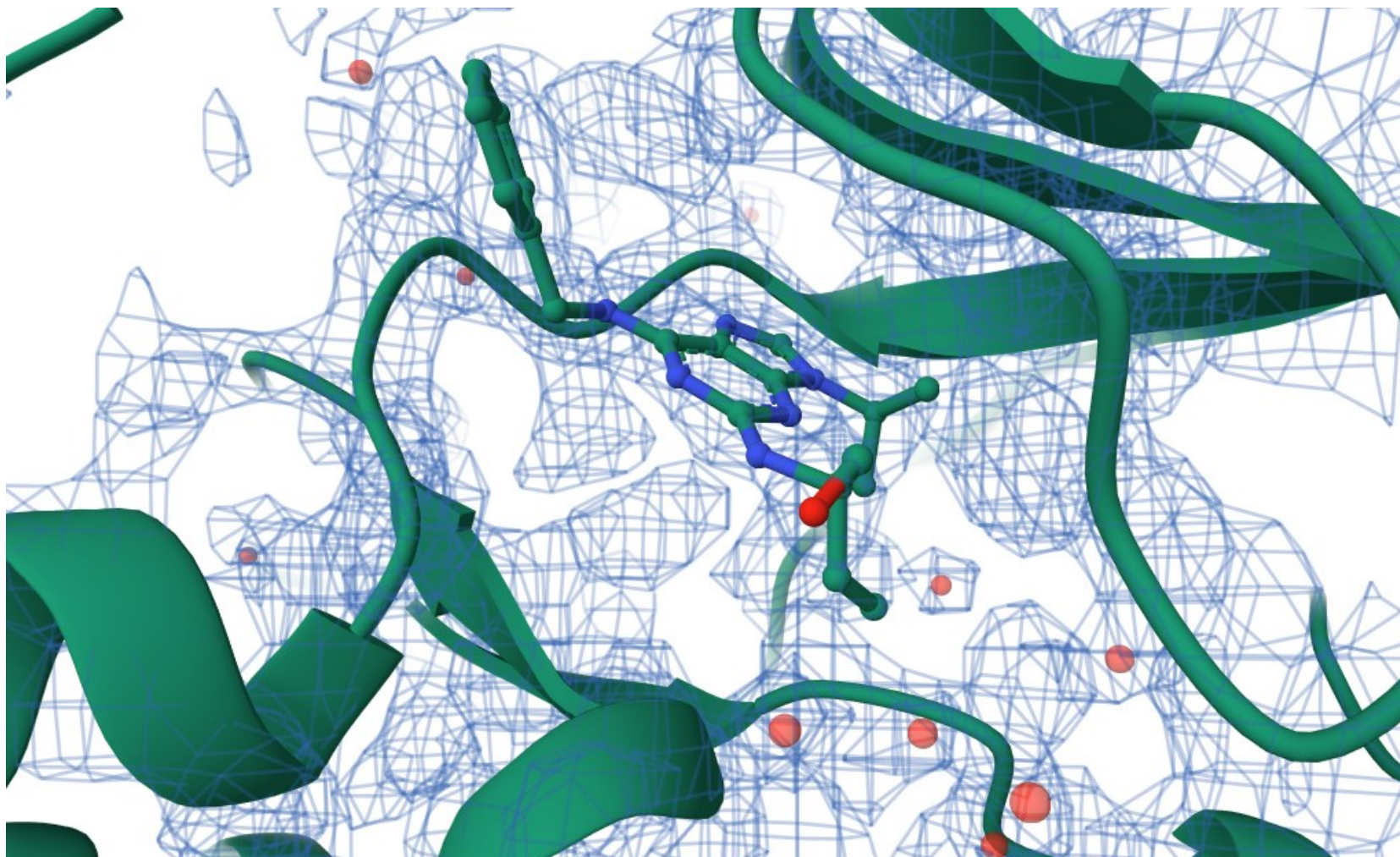


Clique na  
estrutura para  
acessar o [3D-  
View](#)



## Sistema Proteína-Ligante

Abaixo temos o zoom para a região da proteína onde encaixam os inibidores. Especificamente, a figura mostra a região de encaixe da molécula Roscovitina. Estamos vendo um zoom da fechadura (parte da proteína) com a chave ligada (molécula de Roscovitina). O gradeado azul indica a densidade eletrônica.



## Estruturas no PDB

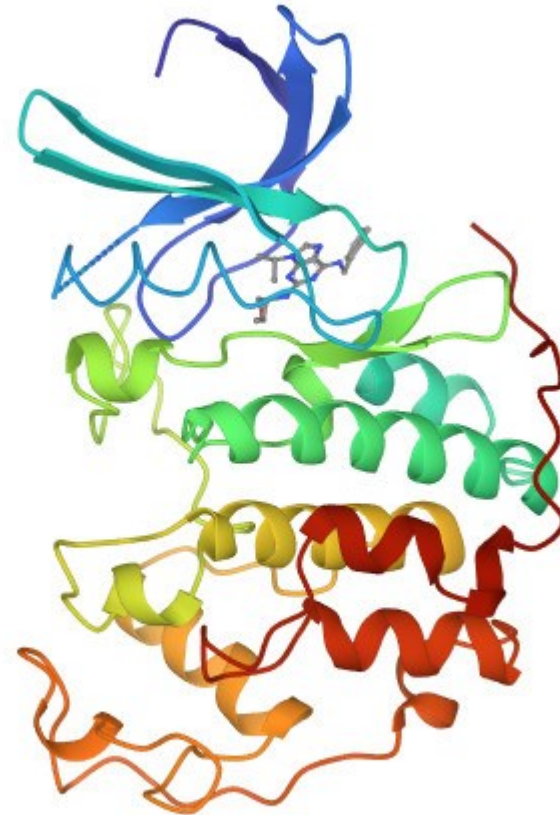
A base de dados Protein Data Bank (PDB) (<https://www.rcsb.org/>) é o maior repositório de informações sobre a estrutura tridimensional de macromoléculas biológicas. Essa informação está armazenada na forma de arquivos de texto, onde cada linha tem a informação sobre as coordenadas x, y e z de cada átomo da proteína. Ao lado temos a página de entrada do Protein Data Bank (imagem capturada em Abril de 2024). Nesta aula iremos olhar como é o formato de um arquivo PDB.



Página de entrada do Protein Data Bank (<https://www.rcsb.org/>) como vista em Abril de 2024. Todo mês o PDB muda a molécula do mês. Essa parte do PDB visa fornecer para comunidade interessada uma fonte de informação sobre o papel biológico das moléculas que foram depositadas no PDB.

## Estruturas no PDB

Na maioria dos artigos científicos que descrevem as estruturas tridimensionais de proteínas, os autores indicam que as coordenadas atômicas da moléculas estão disponíveis no site do PDB (<https://www.rcsb.org/>). Além disso, os autores indicam um código alfanumérico de quatro dígitos que traz a identificação da proteína depositada. Destacamos a CDK2, veremos como fazer download da CDK2 em complexo com o inibidor Roscovitina.





## Estruturas no PDB

Acesse o site do PDB (<https://www.rcsb.org/>). Você terá a página abaixo.

The screenshot shows the RCSB PDB homepage in a web browser. The browser's address bar displays <https://www.rcsb.org>. The page features a dark blue navigation bar with the following menu items: RCSB PDB, Deposit, Search, Visualize, Analyze, Download, Learn, About, Documentation, Careers, and COVID-19. There are also buttons for 'MyPDB' and 'Contact us'. Below the navigation bar, the RCSB PDB logo is displayed alongside statistics: '217,966 Structures from the PDB' and '1,068,577 Computed Structure Models (CSM)'. A search bar is present with the placeholder text 'Enter search term(s), Entry ID(s), or sequence' and a search button. The search bar also includes a '3D Structures' dropdown and an 'Include CSM' toggle. Below the search bar, there are links for 'Advanced Search' and 'Browse Annotations', and a 'Help' link. A banner below the search bar promotes 'Access Computed Structure Models (CSMs) of all available model organisms' with a 'Learn more' button. The main content area is divided into three sections: a 'Welcome' section, a 'Deposit' section, and an 'April Molecule of the Month' section. The 'Welcome' section contains a navigation menu with 'Welcome', 'Deposit', 'Search', 'Visualize', and 'Analyze'. The 'Deposit' section features a 'Deposit' icon and text: 'RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis.' The 'April Molecule of the Month' section features a large 3D protein structure visualization.

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB PROTEIN DATA BANK 217,966 Structures from the PDB 1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entry ID(s), or sequence Include CSM

Advanced Search | Browse Annotations Help

PDB-101 PDB EMDataResource NAKB wwPDB Foundation PDB-Dev

Access Computed Structure Models (CSMs) of all available model organisms Learn more

Welcome

Deposit

Search

Visualize

Analyze

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis.

3D structures from the Protein Data Bank

CSM from AlphaFold DB and

These data can be explored in context of external annotations providing a structural view of biology.

April Molecule of the Month

## Estruturas no PDB

Digite 2A4L no campo indicado e pressione *Enter*.

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB PROTEIN DATA BANK 217,966 Structures from the PDB 1,068,577 Computed Structure Models (CSM)

3D Structures 2A4L Include CSM  Help

in Entry ID(s)  
2A4L

PDB-101 PDB EMDataResource NAKB wwPDB Foundation PDB-Dev

Access Computed Structure Models (CSMs) of all available model organisms [Learn more](#)

Welcome

Deposit

Search

Visualize

Analyze

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive
- Computed Structure Models (CSM)** from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

April Molecule of the Month

## Estruturas no PDB

Veremos a estrutura da CDK2 em complexo com Roscovitina. A partir da página abaixo, temos acesso a diversos dados sobre a estrutura da proteína. Veremos alguns deles, por exemplo a estrutura primária da proteína.

The screenshot shows the RCSB PDB website interface. The browser address bar displays <https://www.rcsb.org/structure/2A4L>. The navigation menu includes: RCSB PDB, Deposit, Search, Visualize, Analyze, Download, Learn, More, Documentation, Careers, and MyPDB. The main header features the RCSB PDB logo and the text: "194011 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education". A search bar is present with a "PDB Archive" dropdown and a search icon. Below the search bar are links for "Advanced Search" and "Browse Annotations", and a "Help" link. The footer of the header includes logos for PDB-101, Worldwide PDB, EMDataResource, Nucleic Acid Database, and Worldwide Protein Data Bank Foundation, along with social media icons for Facebook, Twitter, YouTube, and LinkedIn.

The main content area has a navigation bar with tabs: Structure Summary (selected), 3D View, Annotations, Experiment, Sequence, Genome, Ligands, and Versions. Below this, there are buttons for "Display Files" and "Download Files".

The entry details for **2A4L** are as follows:

- Human cyclin-dependent kinase 2 in complex with roscovitine**
- PDB DOI:** [10.2210/pdb2A4L/pdb](https://doi.org/10.2210/pdb2A4L/pdb)
- Classification:** TRANSFERASE
- Organism(s):** Homo sapiens
- Expression System:** Spodoptera frugiperda
- Mutation(s):** No
- Deposited:** 2005-06-29 **Released:** 2006-10-03

On the left side of the main content area, there is a "Biological Assembly 1" section with a 3D ribbon diagram of the protein structure. On the right side, there is a vertical "Contact Us" button.



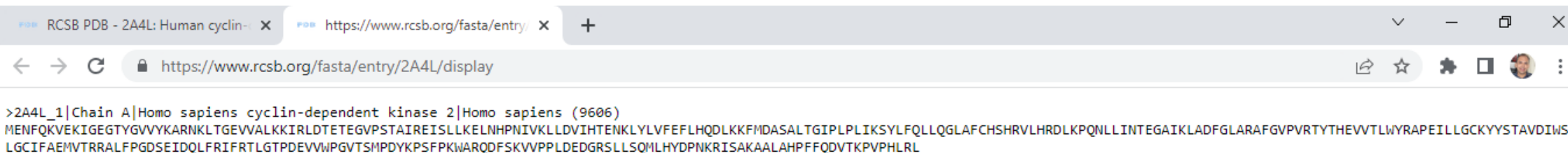
## Estruturas no PDB

Clicamos em *Display Files*->*FASTA Sequence*.

The screenshot shows the RCSB PDB website interface. The browser address bar displays <https://www.rcsb.org/structure/2A4L>. The main navigation bar includes links for Deposit, Search, Visualize, Analyze, Download, Learn, More, Documentation, and Careers, along with a MyPDB button. The PDB logo and tagline "194011 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education" are visible. A search bar with "PDB Archive" and a search icon is present. Below the navigation, there are logos for PDB-101, Worldwide PDB, EMDataResource, Nucleic Acid Database, and Worldwide Protein Data Bank Foundation. The main content area features a tabbed interface with "Structure Summary" selected. The entry details for 2A4L are shown, including the title "Human cyclin-dependent kinase 2 in complex with ro...", PDB DOI: 10.2210/pdb2A4L/pdb, Classification: TRANSFERASE, Organism(s): Homo sapiens, Expression System: Spodoptera frugiperda, and Mutation(s): No. A "Display Files" dropdown menu is open, showing options: FASTA Sequence, mmCIF Format, mmCIF Format (Header), PDB Format, and PDB Format (Header). A blue arrow points from the text above to the "FASTA Sequence" option in the dropdown menu.

## Estruturas no PDB

Abaixo temos o arquivo no formato FASTA para a CDK2. Os arquivos no formato FASTA trazem a estrutura primária da proteína. A primeira linha do arquivo FASTA mostra a identificação da proteína. Da segunda linha em diante, temos a sequência de aminoácidos com o código de uma letra.



```
>2A4L_1|Chain A|Homo sapiens cyclin-dependent kinase 2|Homo sapiens (9606)
MENFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDTETEGVPSTAIRESISLLKELNHPNIVKLLDVIHTENKLYLVFEFLHQDLKKFMDASALTGIPLPLIKSYLFQLLQGLAFCHSHRVLHRDLKPQNLLINTEGAIKLADFGLARAFGVPVRTYTHEWVTLWYRAPEILLGCKYYSTAVDIWS
LGCIFAEMVTRRALFPGDSEIDQLFRIFRTLGTPEVVWPGVTSMPDYKPSFPKWARQDFSKVPPPLDEDGRSLLSQMLHYDPNKRISAKAALAHPPFQDVTKPVPHLRL
```

## Estruturas no PDB

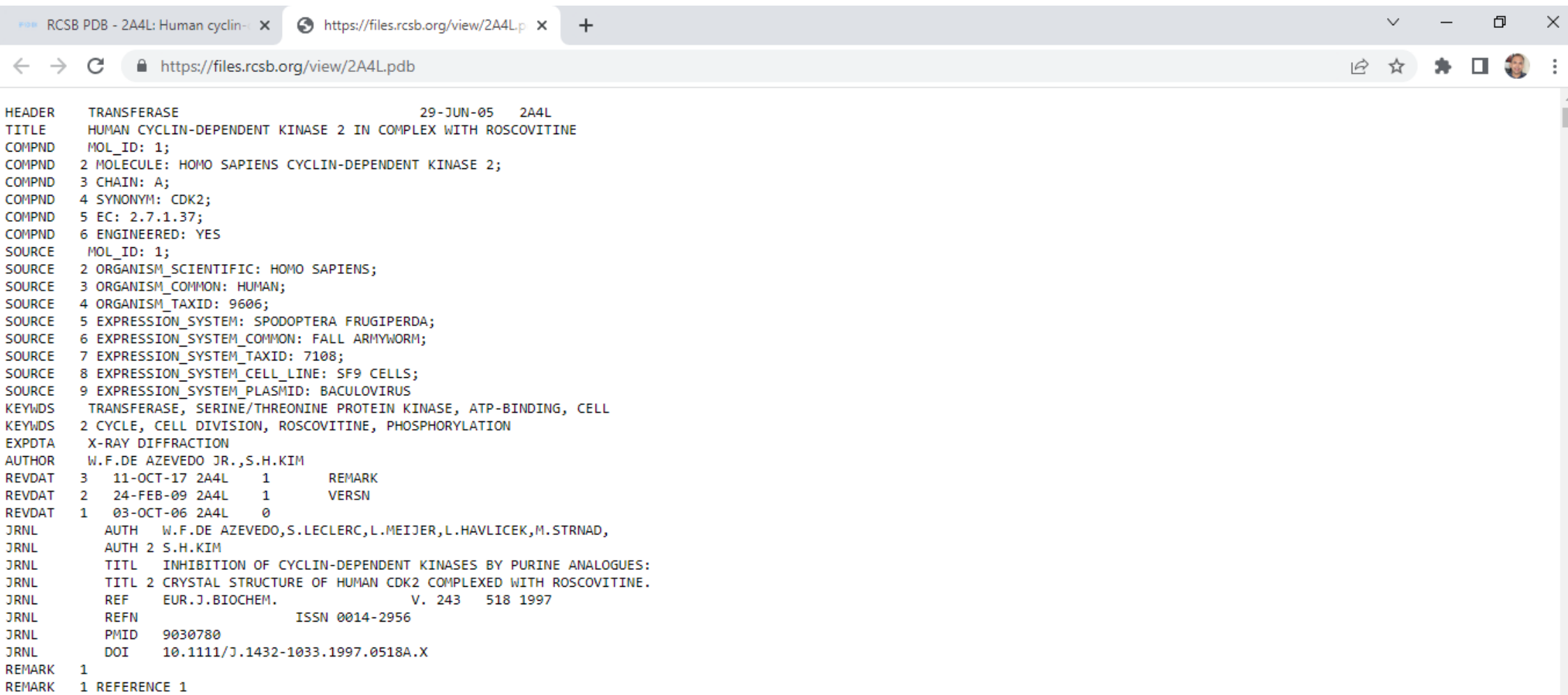
Para termos acesso às coordenadas atômicas, clicamos *Display Files*->*PDB Format*.

The screenshot shows the RCSB PDB website interface. The browser address bar displays <https://www.rcsb.org/structure/2A4L>. The main navigation bar includes links for Deposit, Search, Visualize, Analyze, Download, Learn, More, Documentation, and Careers, along with a MyPDB button. The header features the RCSB PDB logo and the text "194011 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education". A search bar is present with a "PDB Archive" dropdown and a search icon. Below the header, there are logos for PDB-101, Worldwide PDB, EMDataResource, Nucleic Acid Database, and Worldwide Protein Data Bank Foundation. The main content area has tabs for Structure Summary, 3D View, Annotations, Experiment, Sequence, Genome, Ligands, and Versions. The "Structure Summary" tab is active, showing a 3D ribbon diagram of the protein structure. To the right of the diagram, the entry ID "2A4L" is displayed, along with the title "Human cyclin-dependent kinase 2 in complex with ro". Below the title, the PDB DOI is "10.2210/pdb2A4L/pdb". The classification is "TRANSFERASE", the organism is "Homo sapiens", the expression system is "Spodoptera frugiperda", and there are no mutations. The deposition date is "2005-06-29" and the release date is "2006-10-03". A "Display Files" dropdown menu is open, showing options: FASTA Sequence, mmCIF Format, mmCIF Format (Header), PDB Format (highlighted), and PDB Format (Header). A blue arrow points from the text above to the "PDB Format" option in the dropdown menu.



## Estruturas no PDB

Abaixo temos as primeiras linhas do arquivo no PDB para a CDK2 com Roscovitina. Essas linhas trazem informações sobre a identificação da proteínas, os autores seguidas de detalhes sobre a técnica experimental usada para a resolução da estrutura da proteína. As coordenadas atômicas estão mais abaixo no arquivo.



```

HEADER      TRANSFERASE                      29-JUN-05  2A4L
TITLE      HUMAN CYCLIN-DEPENDENT KINASE 2 IN COMPLEX WITH ROSCOVITINE
COMPND     MOL_ID: 1;
COMPND     2 MOLECULE: HOMO SAPIENS CYCLIN-DEPENDENT KINASE 2;
COMPND     3 CHAIN: A;
COMPND     4 SYNONYM: CDK2;
COMPND     5 EC: 2.7.1.37;
COMPND     6 ENGINEERED: YES
SOURCE     MOL_ID: 1;
SOURCE     2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE     3 ORGANISM_COMMON: HUMAN;
SOURCE     4 ORGANISM_TAXID: 9606;
SOURCE     5 EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA;
SOURCE     6 EXPRESSION_SYSTEM_COMMON: FALL ARMYWORM;
SOURCE     7 EXPRESSION_SYSTEM_TAXID: 7108;
SOURCE     8 EXPRESSION_SYSTEM_CELL_LINE: SF9 CELLS;
SOURCE     9 EXPRESSION_SYSTEM_PLASMID: BACULOVIRUS
KEYWDS     TRANSFERASE, SERINE/THREONINE PROTEIN KINASE, ATP-BINDING, CELL
KEYWDS     2 CYCLE, CELL DIVISION, ROSCOVITINE, PHOSPHORYLATION
EXPDTA     X-RAY DIFFRACTION
AUTHOR     W.F.DE AZEVEDO JR.,S.H.KIM
REVDAT     3 11-OCT-17 2A4L 1      REMARK
REVDAT     2 24-FEB-09 2A4L 1      VERSN
REVDAT     1 03-OCT-06 2A4L 0
JRNL       AUTH  W.F.DE AZEVEDO,S.LECLERC,L.MEIJER,L.HAVLICEK,M.STRNAD,
JRNL       AUTH  2 S.H.KIM
JRNL       TITL  INHIBITION OF CYCLIN-DEPENDENT KINASES BY PURINE ANALOGUES:
JRNL       TITL  2 CRYSTAL STRUCTURE OF HUMAN CDK2 COMPLEXED WITH ROSCOVITINE.
JRNL       REF   EUR.J.BIOCHEM.                V. 243  518 1997
JRNL       REFN  ISSN 0014-2956
JRNL       PMID  9030780
JRNL       DOI   10.1111/J.1432-1033.1997.0518A.X
REMARK     1
REMARK     1 REFERENCE 1
  
```

## Estruturas no PDB

No PDB as coordenadas atômicas são identificadas com a palavra-chave ATOM, como destacado abaixo.

Browser tabs: RCSB PDB - 2A4L: Human cyclin- x | https://files.rcsb.org/view/2A4L.p x

Address bar: https://files.rcsb.org/view/2A4L.pdb

ORIGX3	0.000000	0.000000	1.000000	0.000000							
SCALE1	0.013830	0.000000	0.000000	0.000000							
SCALE2	0.000000	0.013686	0.000000	0.000000							
SCALE3	0.000000	0.000000	0.018422	0.000000							
ATOM	1	N	MET	A	1	101.710	112.330	93.759	1.00	48.54	N
ATOM	2	CA	MET	A	1	102.732	113.140	94.479	1.00	47.79	C
ATOM	3	C	MET	A	1	103.199	114.420	93.762	1.00	47.20	C
ATOM	4	O	MET	A	1	102.995	114.577	92.561	1.00	51.55	O
ATOM	5	CB	MET	A	1	103.933	112.272	94.785	1.00	50.37	C
ATOM	6	CG	MET	A	1	104.548	112.540	96.126	1.00	55.72	C
ATOM	7	SD	MET	A	1	106.336	112.671	95.934	1.00	62.79	S
ATOM	8	CE	MET	A	1	106.542	114.250	95.159	1.00	54.71	C
ATOM	9	N	GLU	A	2	103.906	115.275	94.503	1.00	44.44	N
ATOM	10	CA	GLU	A	2	104.085	116.695	94.178	1.00	40.49	C
ATOM	11	C	GLU	A	2	105.065	117.015	93.046	1.00	35.47	C
ATOM	12	O	GLU	A	2	104.918	118.030	92.386	1.00	35.53	O
ATOM	13	CB	GLU	A	2	104.531	117.459	95.428	1.00	43.49	C
ATOM	14	CG	GLU	A	2	103.464	117.597	96.515	1.00	52.62	C
ATOM	15	CD	GLU	A	2	103.286	116.347	97.374	1.00	53.08	C
ATOM	16	OE1	GLU	A	2	102.216	115.703	97.266	1.00	57.29	O
ATOM	17	OE2	GLU	A	2	104.183	116.042	98.197	1.00	54.12	O
ATOM	18	N	ASN	A	3	106.112	116.203	92.926	1.00	34.83	N
ATOM	19	CA	ASN	A	3	107.148	116.396	91.921	1.00	34.19	C
ATOM	20	C	ASN	A	3	106.657	116.065	90.504	1.00	31.01	C
ATOM	21	O	ASN	A	3	107.411	116.198	89.539	1.00	31.95	O
ATOM	22	CB	ASN	A	3	108.367	115.508	92.233	1.00	36.44	C
ATOM	23	CG	ASN	A	3	109.066	115.890	93.527	1.00	36.39	C
ATOM	24	OD1	ASN	A	3	109.003	115.176	94.514	1.00	34.23	O
ATOM	25	ND2	ASN	A	3	109.942	116.909	93.449	1.00	37.86	N
ATOM	26	N	PHE	A	4	105.407	115.612	90.401	1.00	29.45	N
ATOM	27	CA	PHE	A	4	104.807	115.143	89.150	1.00	27.72	C
ATOM	28	C	PHE	A	4	103.454	115.819	88.841	1.00	31.06	C
ATOM	29	O	PHE	A	4	102.491	115.724	89.620	1.00	31.86	O
ATOM	30	CB	PHE	A	4	104.593	113.612	89.178	1.00	26.30	C
ATOM	31	CG	PHE	A	4	105.855	112.802	89.435	1.00	27.14	C

## Estruturas no PDB

A seguir temos a descrição de cada campo de uma linha de coordenadas atômicas de um arquivo PDB.

Identificador de linha com coordenadas atômicas

Número do átomo

Tipo do átomo

Código de três letras do resíduo de aminoácido

Identificador de cadeia

Número do resíduo de aminoácido

ATOM	1	N	MET A	1	101.710	112.330	93.759	1.00	48.54		N
ATOM	2	CA	MET A	1	102.732	113.140	94.479	1.00	47.79		C
ATOM	3	C	MET A	1	103.199	114.420	93.762	1.00	47.20		C
ATOM	4	O	MET A	1	102.995	114.577	92.561	1.00	51.55		O
ATOM	5	CB	MET A	1	103.933	112.272	94.785	1.00	50.37		C
ATOM	6	CG	MET A	1	104.548	112.540	96.126	1.00	55.72		C
ATOM	7	SD	MET A	1	106.336	112.671	95.934	1.00	62.79		S
ATOM	8	CE	MET A	1	106.542	114.250	95.159	1.00	54.71		C
ATOM	9	N	GLU A	2	103.906	115.275	94.503	1.00	44.44		N
ATOM	10	CA	GLU A	2	104.085	116.695	94.178	1.00	40.49		C

.....



## Estruturas no PDB

Coordenadas atômicas (x,y,z) em Å ( $1 \text{ \AA} = 10^{-10}\text{m}$ )

Fator de ocupação

B-factor ( $\text{\AA}^2$ )

Tipo do átomo

					x	y	z				
ATOM	1	N	MET	A	1	101.710	112.330	93.759	1.00	48.54	N
ATOM	2	CA	MET	A	1	102.732	113.140	94.479	1.00	47.79	C
ATOM	3	C	MET	A	1	103.199	114.420	93.762	1.00	47.20	C
ATOM	4	O	MET	A	1	102.995	114.577	92.561	1.00	51.55	O
ATOM	5	CB	MET	A	1	103.933	112.272	94.785	1.00	50.37	C
ATOM	6	CG	MET	A	1	104.548	112.540	96.126	1.00	55.72	C
ATOM	7	SD	MET	A	1	106.336	112.671	95.934	1.00	62.79	S
ATOM	8	CE	MET	A	1	106.542	114.250	95.159	1.00	54.71	C
ATOM	9	N	GLU	A	2	103.906	115.275	94.503	1.00	44.44	N
ATOM	10	CA	GLU	A	2	104.085	116.695	94.178	1.00	40.49	C

.....

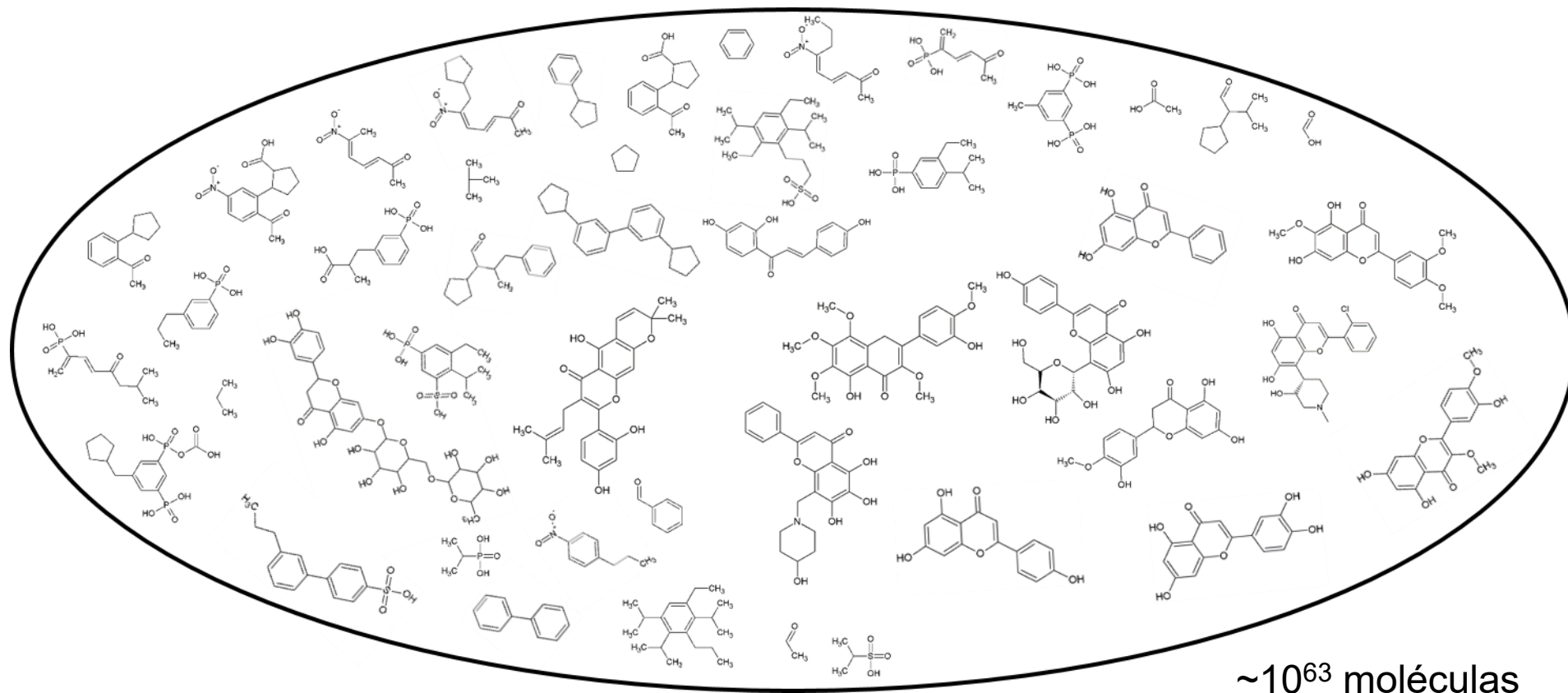
## Estruturas no PDB

Os programas de docagem como o Molegro Virtual Docker usam estas coordenadas atômicas para calcular a distância geométrica entre átomos. Os programas de docagem consideram os átomos como esferas centradas nestas coordenadas. Todos os cálculos realizados para o cálculo dos termos energéticos da interação entre um ligante e a proteína começam com as coordenadas atômicas.

ATOM	1	N	MET	A	1	101.710	112.330	93.759	1.00	48.54	N
ATOM	2	CA	MET	A	1	102.732	113.140	94.479	1.00	47.79	C
ATOM	3	C	MET	A	1	103.199	114.420	93.762	1.00	47.20	C
ATOM	4	O	MET	A	1	102.995	114.577	92.561	1.00	51.55	O
ATOM	5	CB	MET	A	1	103.933	112.272	94.785	1.00	50.37	C
ATOM	6	CG	MET	A	1	104.548	112.540	96.126	1.00	55.72	C
ATOM	7	SD	MET	A	1	106.336	112.671	95.934	1.00	62.79	S
ATOM	8	CE	MET	A	1	106.542	114.250	95.159	1.00	54.71	C
ATOM	9	N	GLU	A	2	103.906	115.275	94.503	1.00	44.44	N
ATOM	10	CA	GLU	A	2	104.085	116.695	94.178	1.00	40.49	C

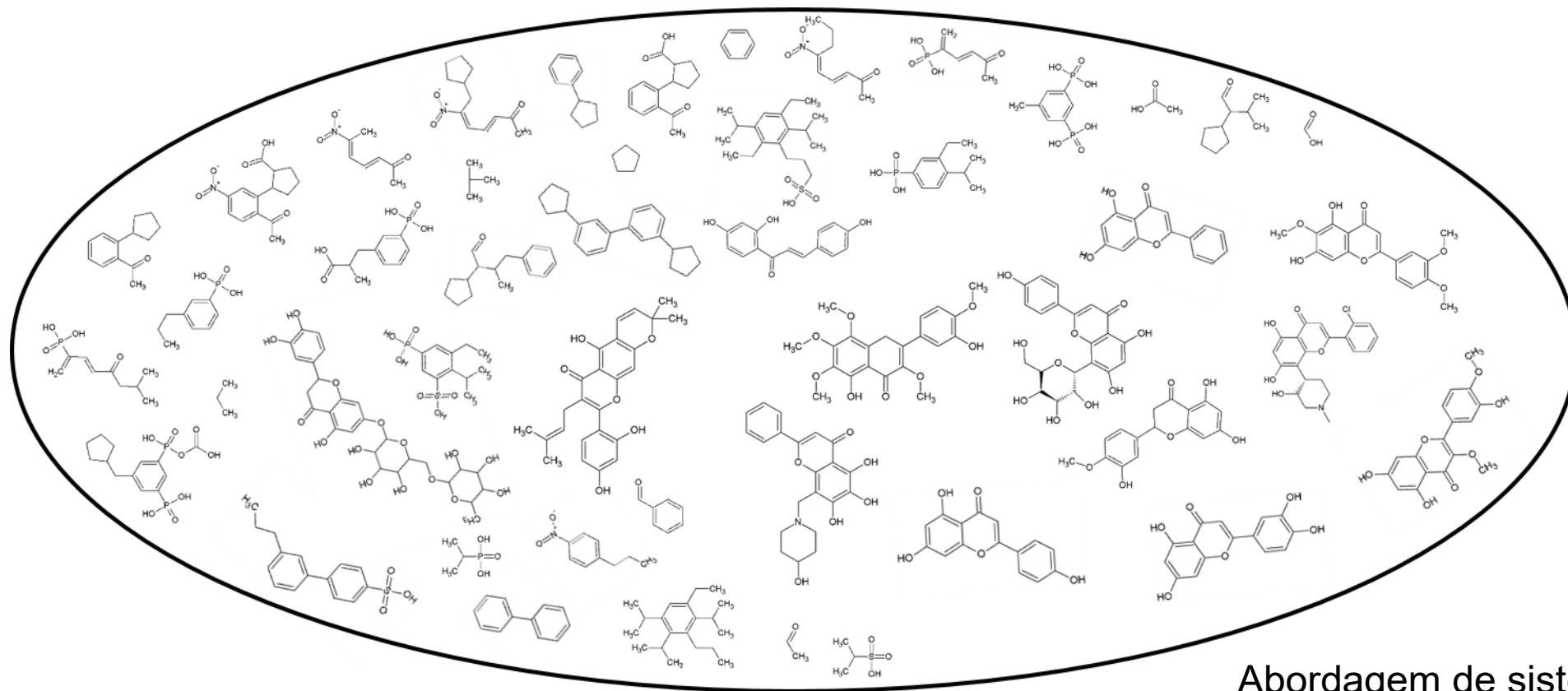
.....

## Espaço Químico



Fonte: Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996; 16(1):3–50.

## Espaço Químico - Vantagens

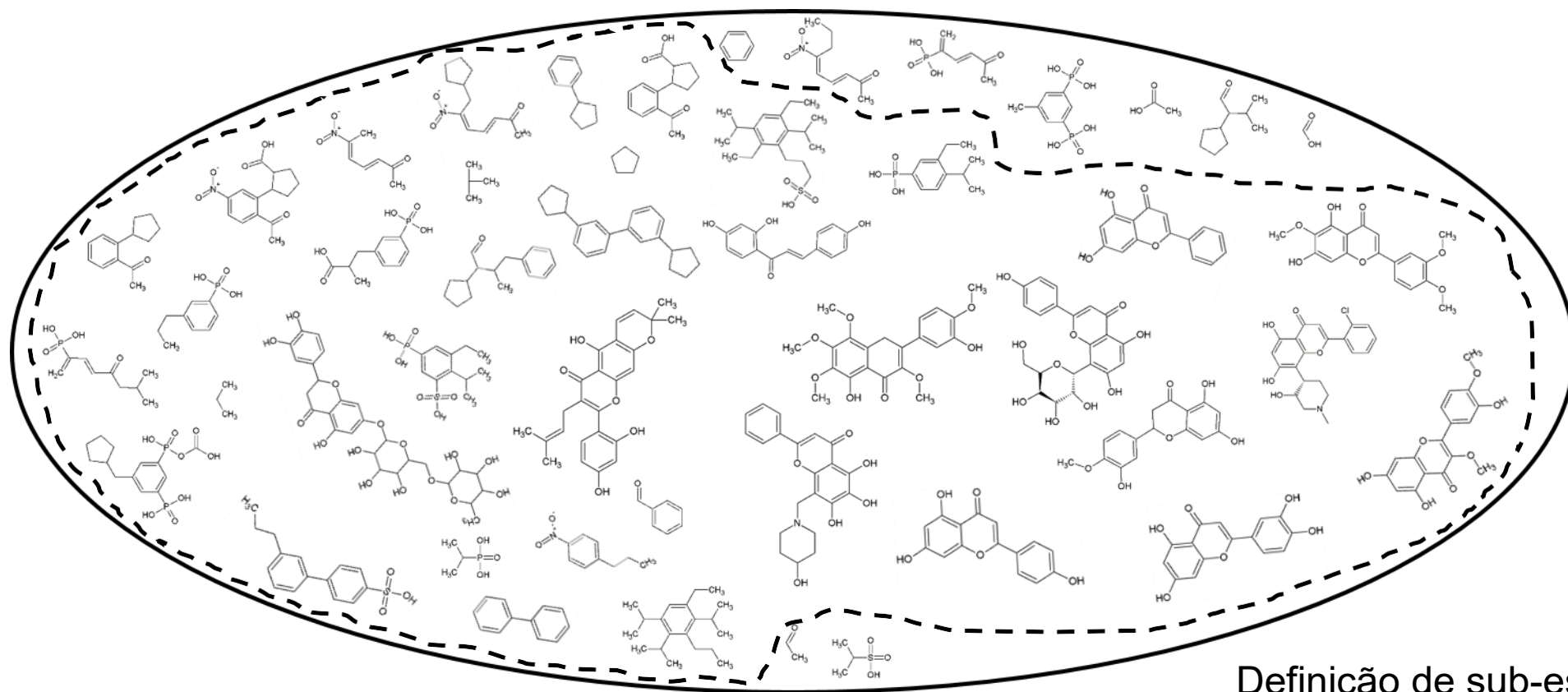


Abordagem de sistemas

Fonte: Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996; 16(1):3–50.



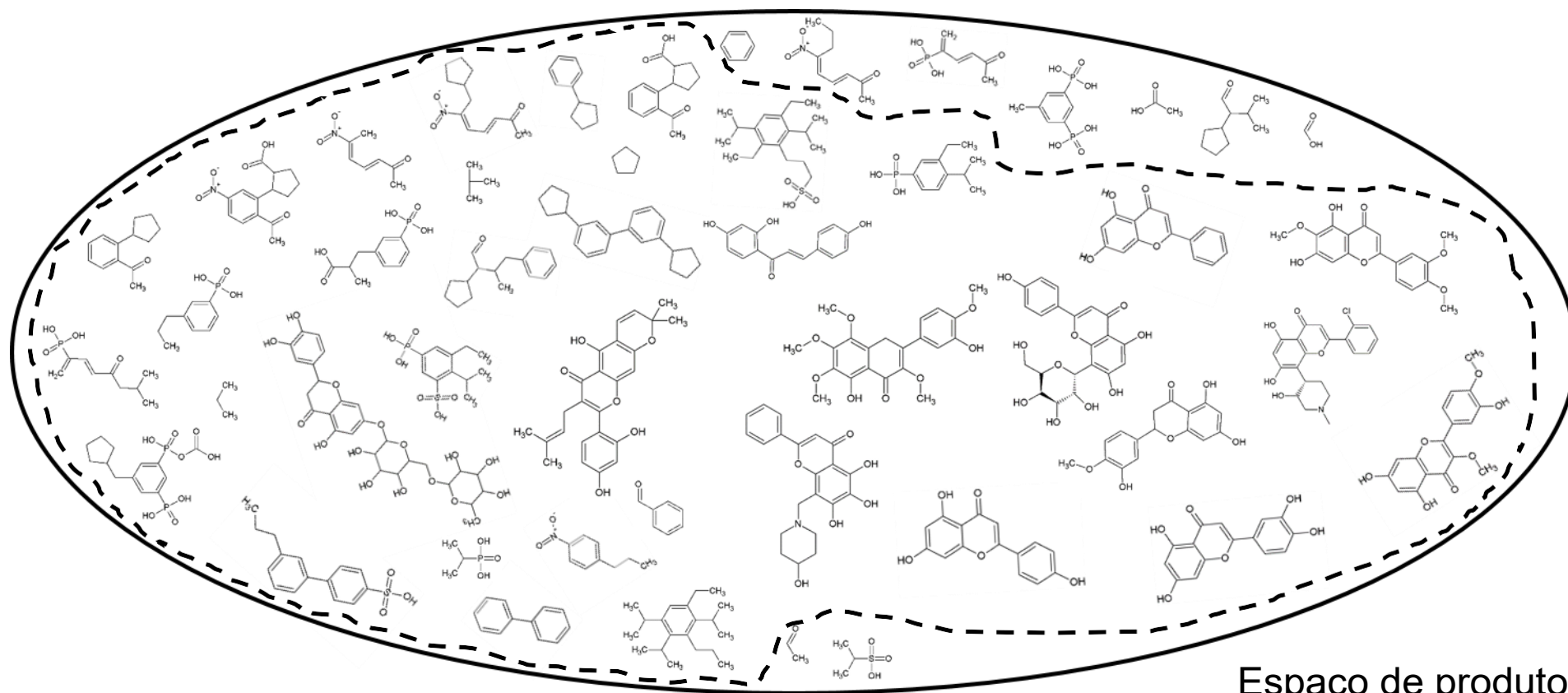
## Espaço Químico - Vantagens



Definição de sub-espacos

Fonte: Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996; 16(1):3–50.

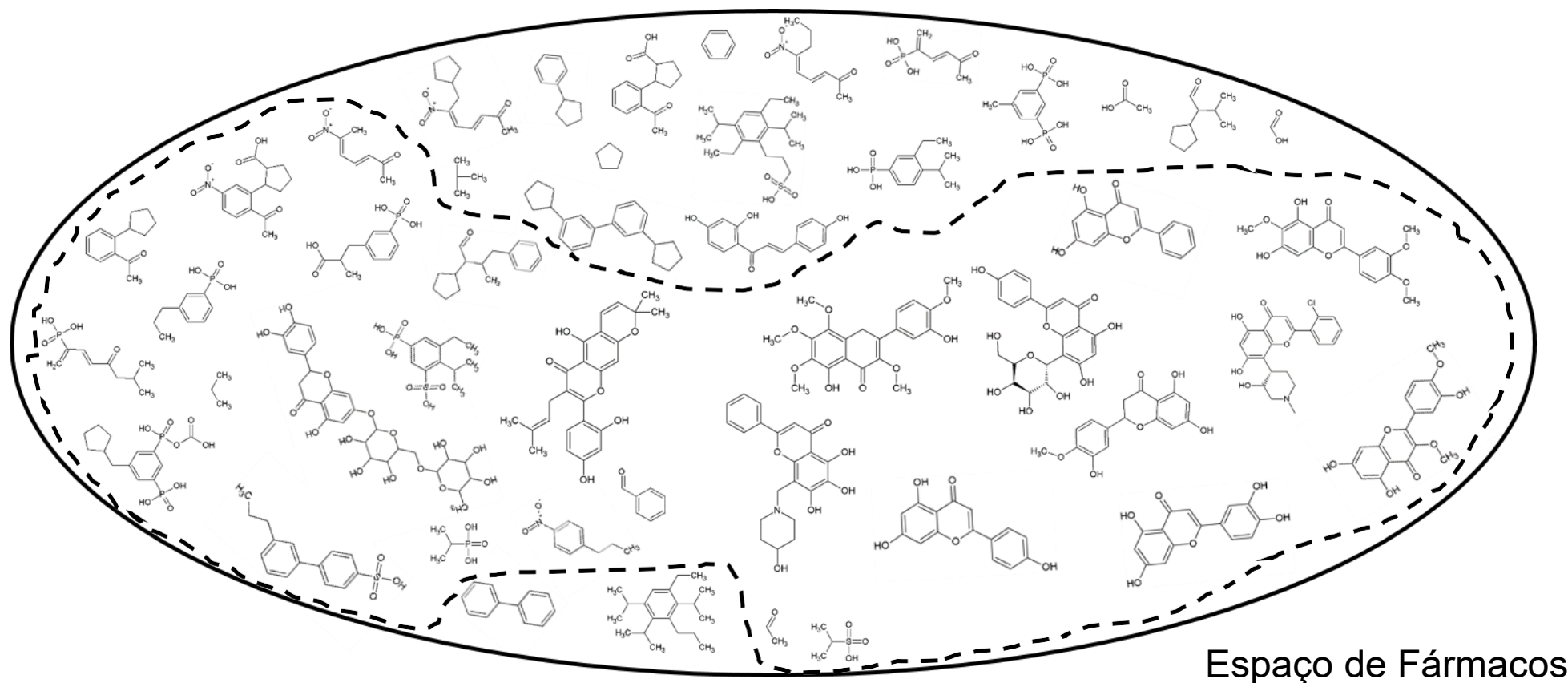
## Espaço Químico - Vantagens



Espaço de produtos naturais

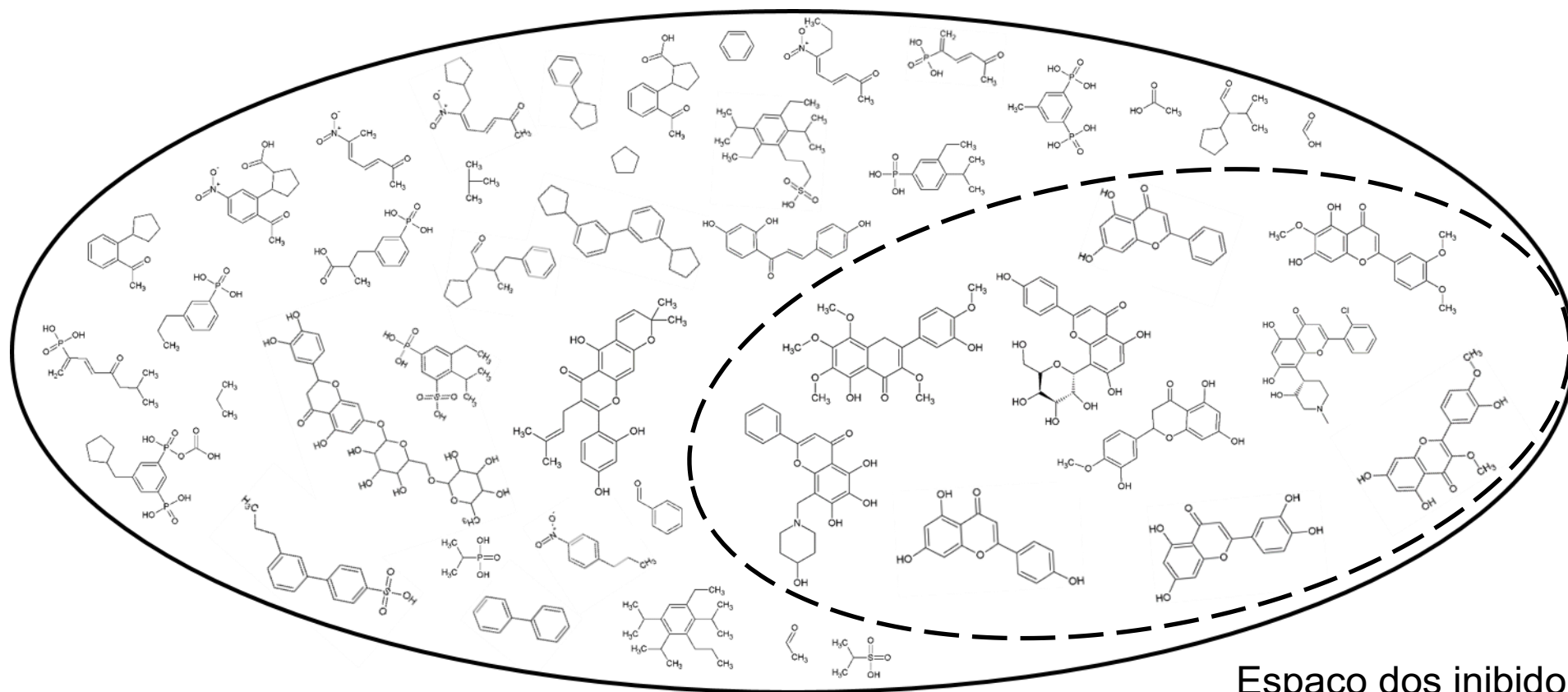
Fonte: Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996; 16(1):3–50.

## Espaço Químico - Vantagens



Fonte: Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996; 16(1):3–50.

## Espaço Químico - Vantagens

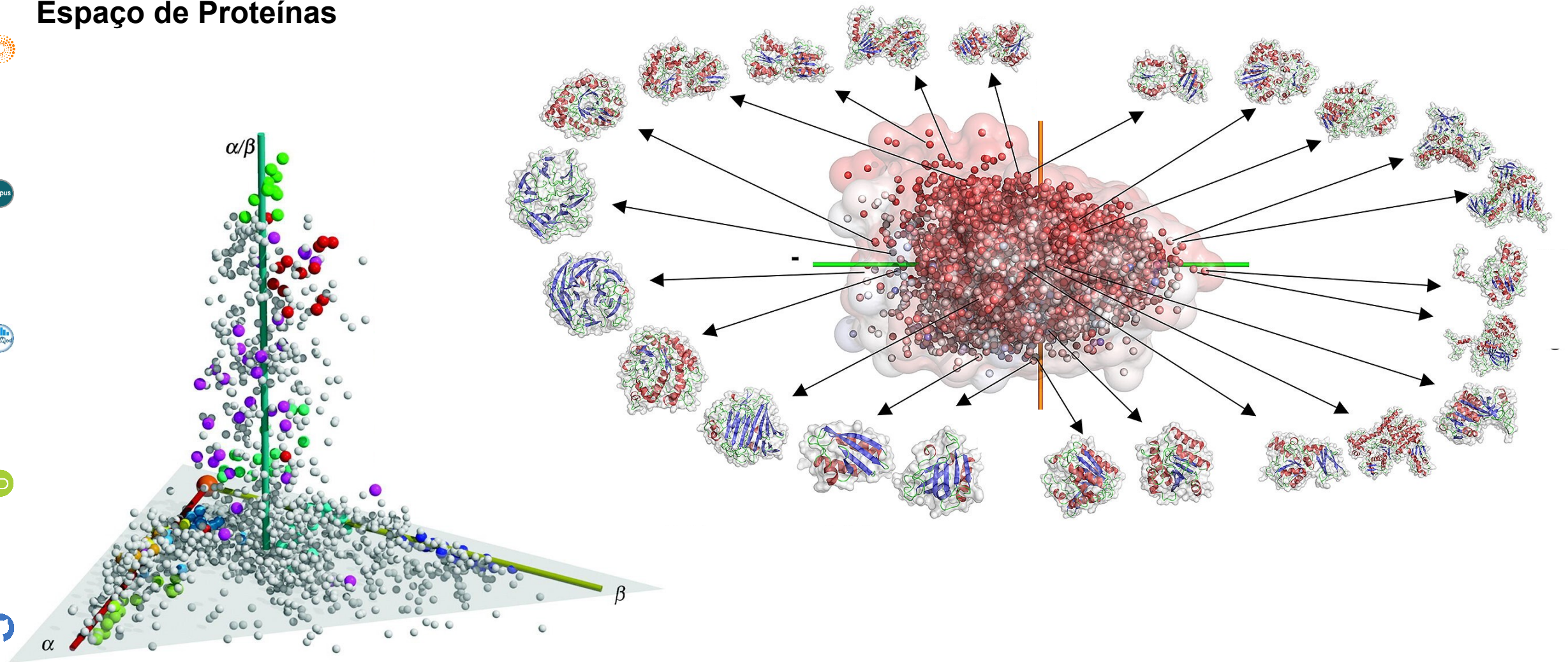


Espaço dos inibidores de CDK2

Fonte: Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996; 16(1):3–50.



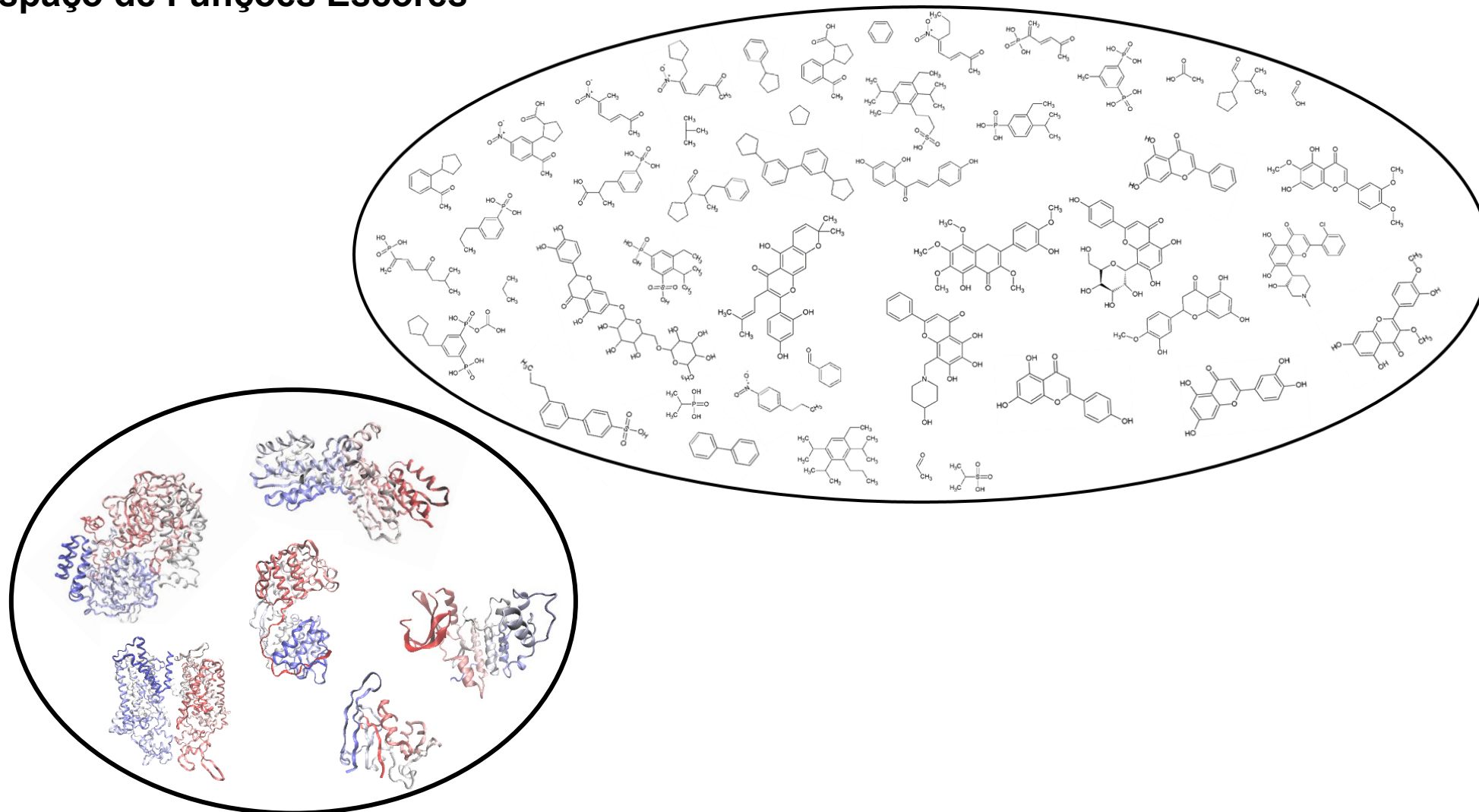
# Espaço de Proteínas



Fonte: Han X, Sit A, Christoffer C, Chen S, Kihara D. A global map of the protein shape universe. PLoS Comput Biol. 2019; 15(4):e1006969.

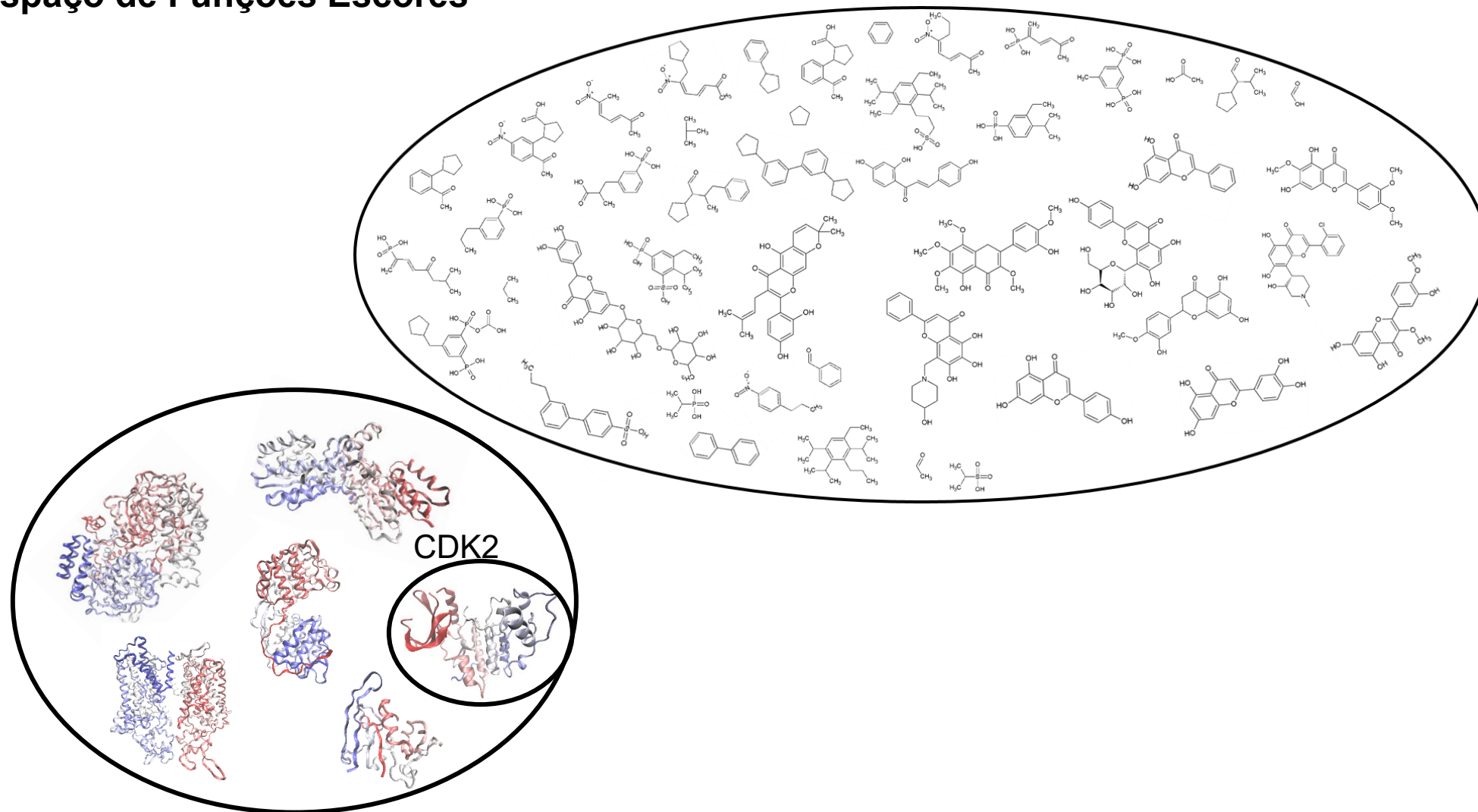


## Espaço de Funções Escores



Fonte: Heck GS, Pintro VO, Pereira RR, de Ávila MB, Levin NMB, de Azevedo WF. Supervised Machine Learning Methods Applied to Predict Ligand- Binding Affinity. *Curr Med Chem.* 2017; 24(23):2459-2470.

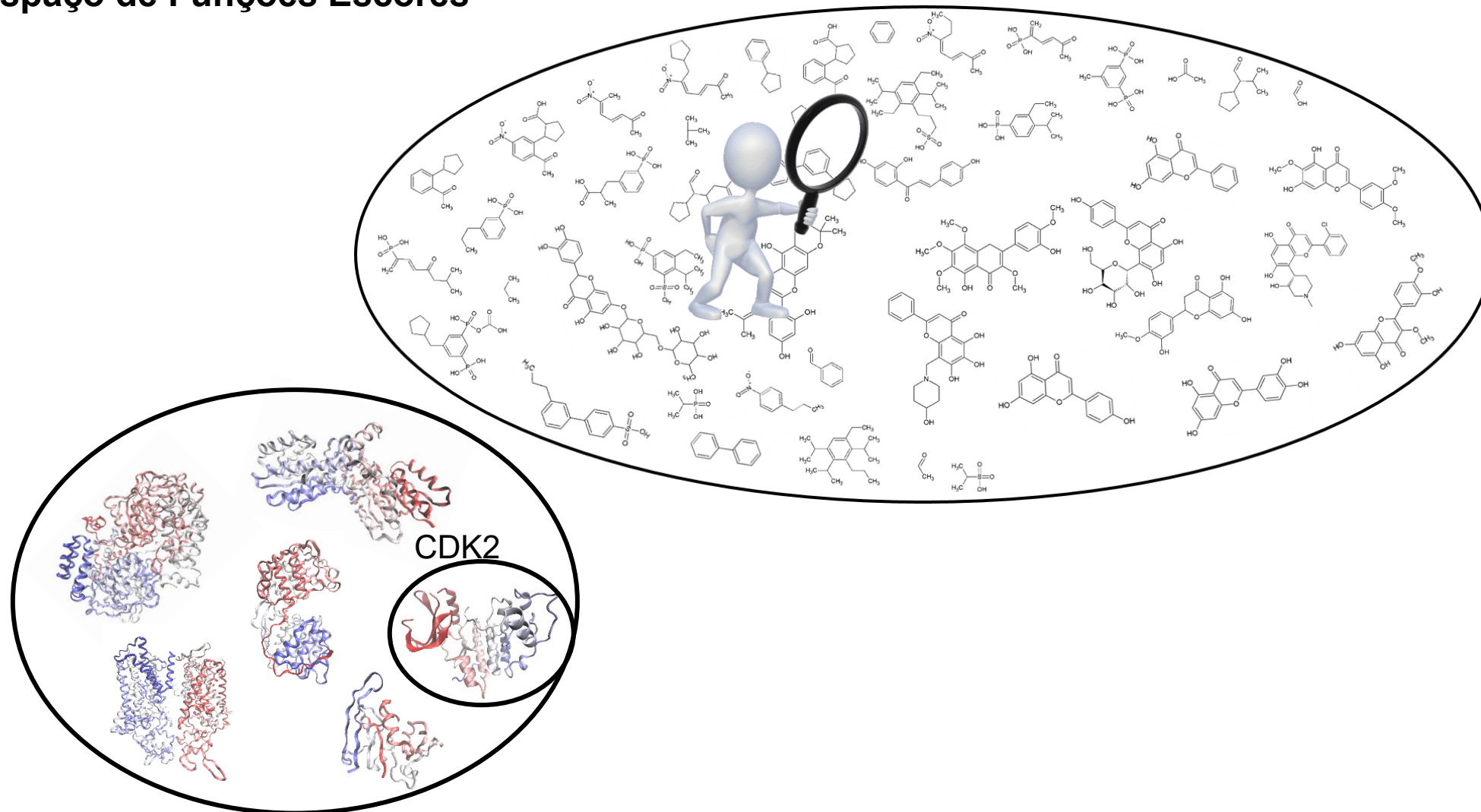
## Espaço de Funções Escores



Fonte: Heck GS, Pintro VO, Pereira RR, de Ávila MB, Levin NMB, de Azevedo WF. Supervised Machine Learning Methods Applied to Predict Ligand- Binding Affinity. *Curr Med Chem.* 2017; 24(23):2459-2470.

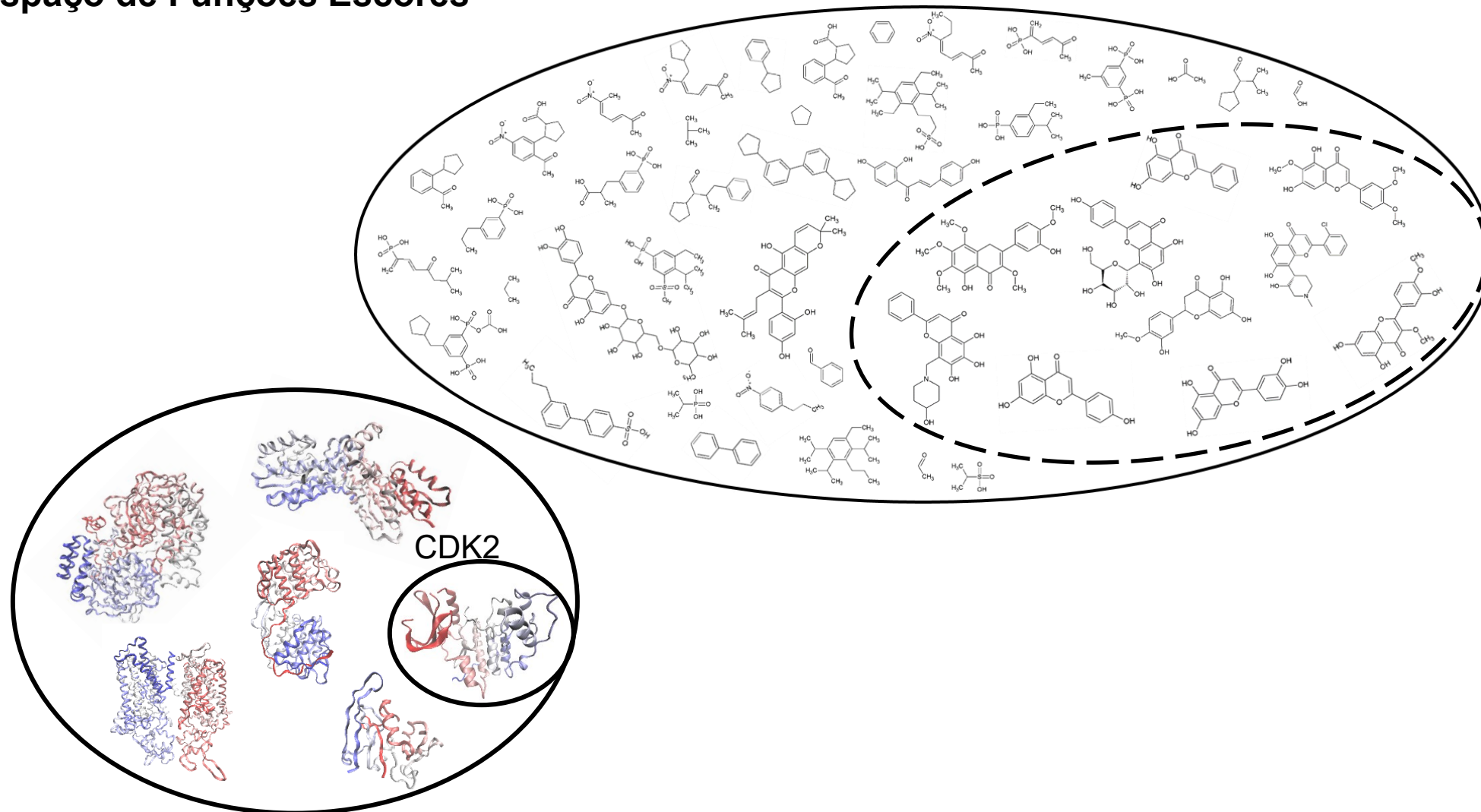


## Espaço de Funções Escores



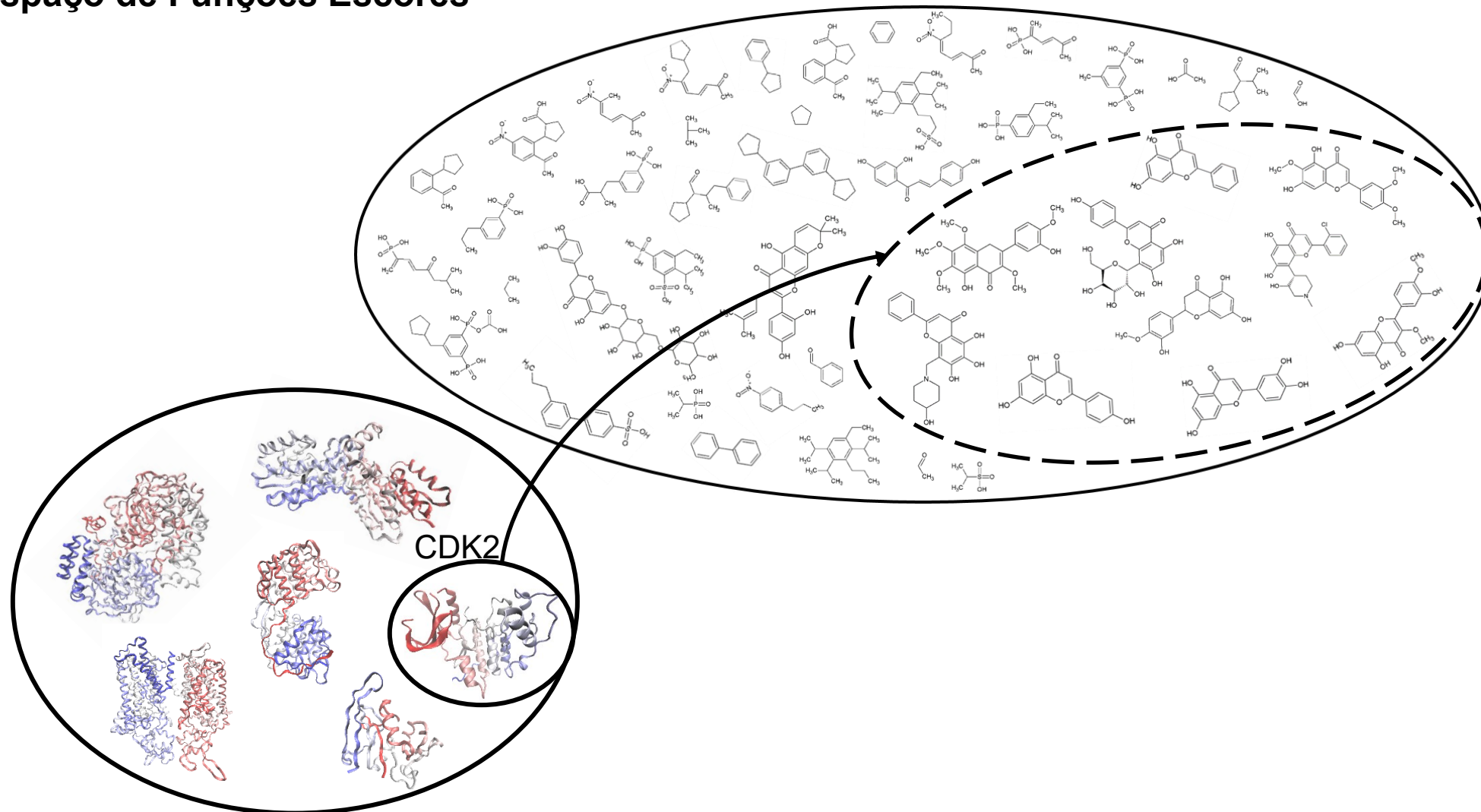
Fonte: Heck GS, Pintro VO, Pereira RR, de Ávila MB, Levin NMB, de Azevedo WF. Supervised Machine Learning Methods Applied to Predict Ligand- Binding Affinity. *Curr Med Chem.* 2017; 24(23):2459-2470.

## Espaço de Funções Escores



Fonte: Heck GS, Pintro VO, Pereira RR, de Ávila MB, Levin NMB, de Azevedo WF. Supervised Machine Learning Methods Applied to Predict Ligand- Binding Affinity. *Curr Med Chem.* 2017; 24(23):2459-2470.

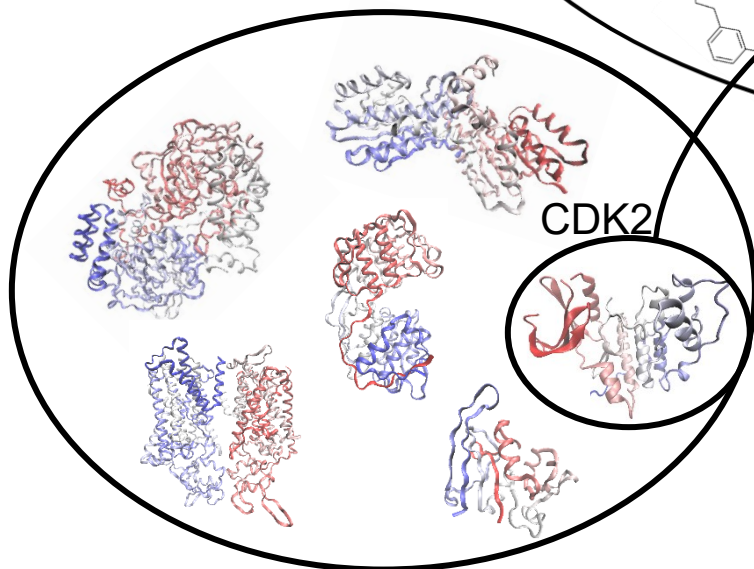
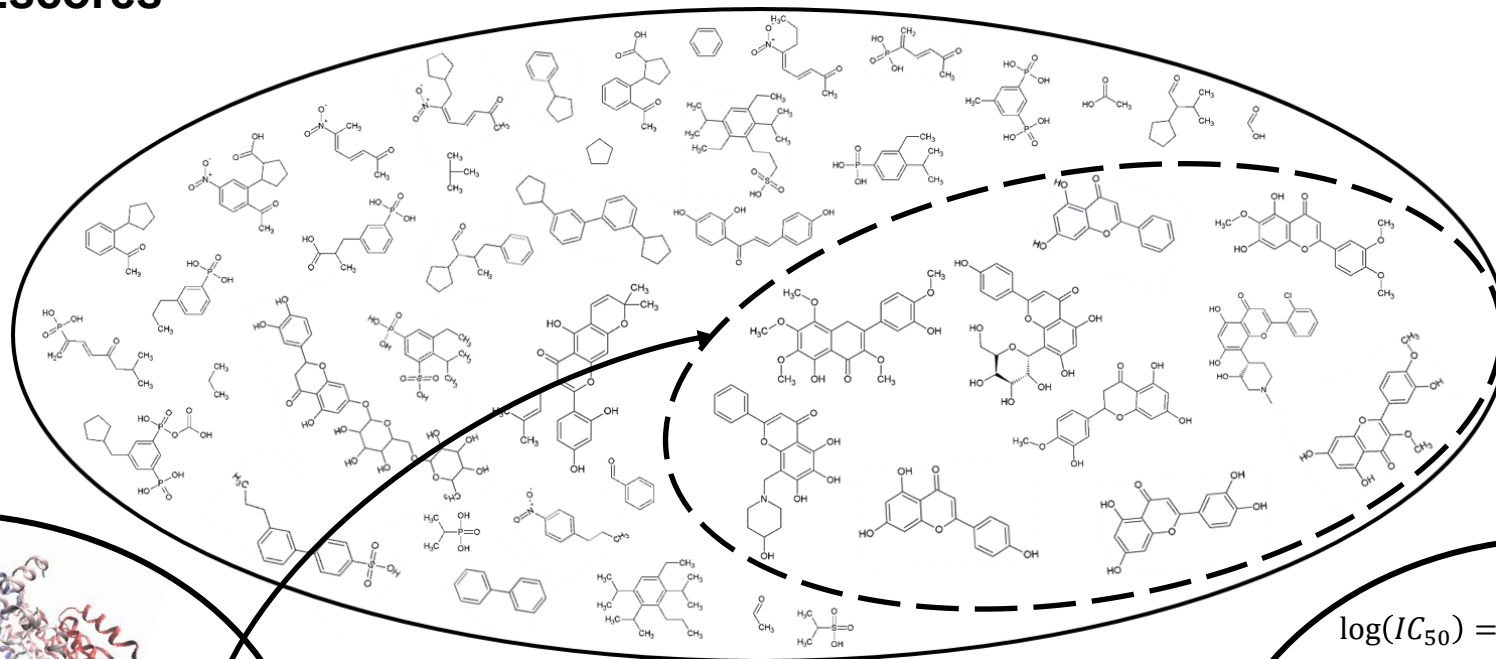
## Espaço de Funções Escores



Fonte: Heck GS, Pintro VO, Pereira RR, de Ávila MB, Levin NMB, de Azevedo WF. Supervised Machine Learning Methods Applied to Predict Ligand- Binding Affinity. *Curr Med Chem.* 2017; 24(23):2459-2470.



# Espaço de Funções Escores



**TABA**



$$\log(IC_{50}) = \sum_{i=0}^N \omega_i x_i + \sum_{j=0}^N \alpha_j x_j^i$$

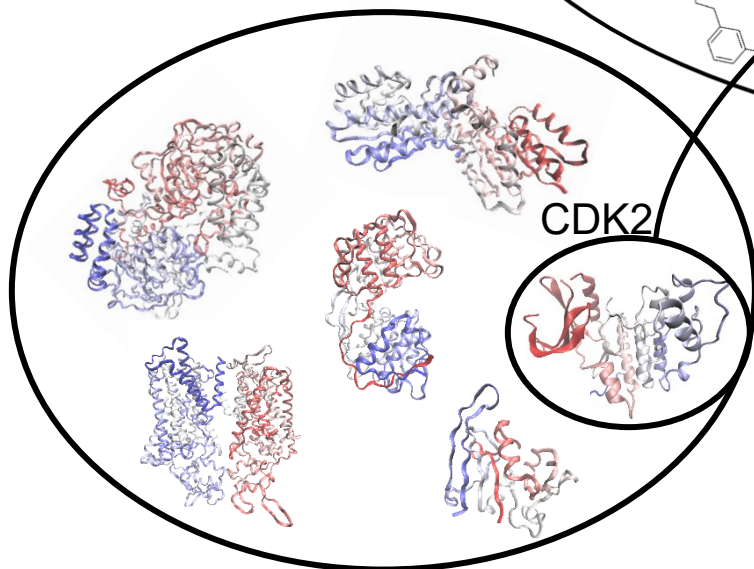
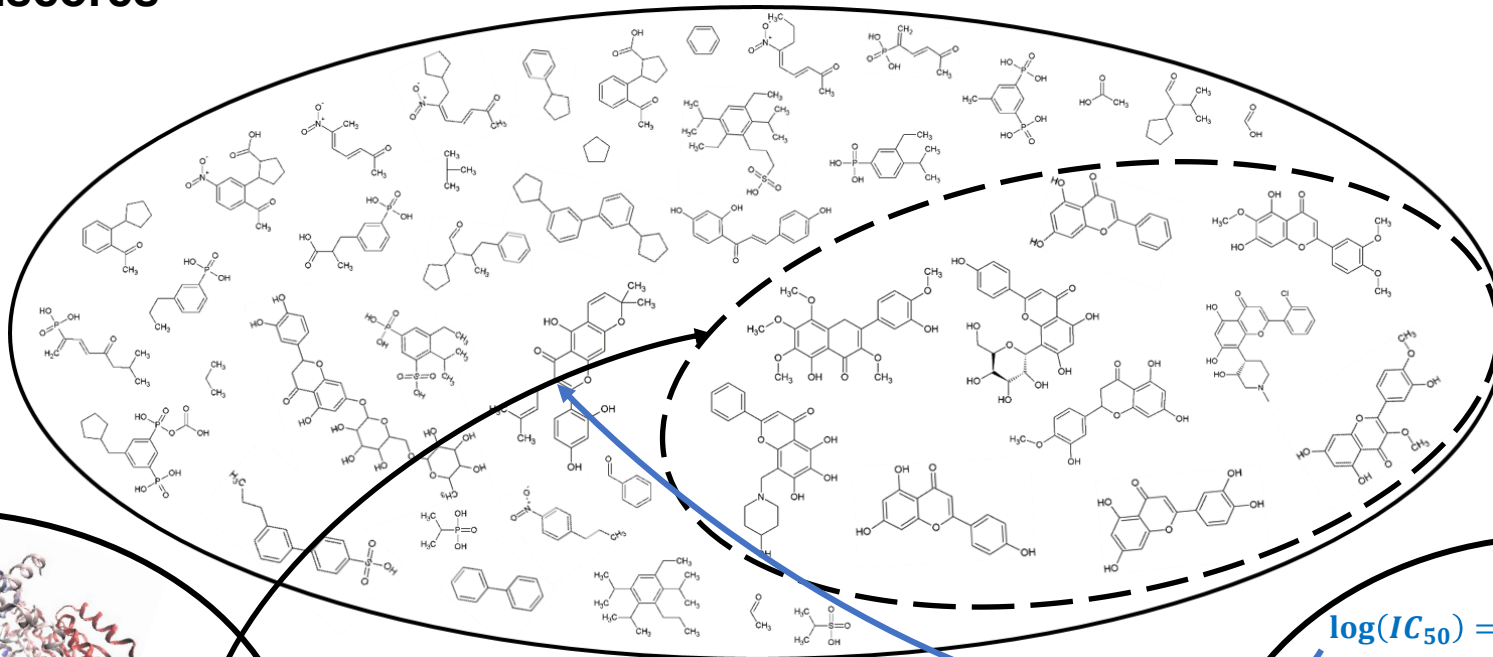
$$\Delta G = \sum_{i=0}^N \omega_i x_i \quad f = \sum_{i=1}^N \alpha_i x_i - x_j^{-3} + \sum_{j=1}^M x$$

$$\Delta S = \alpha_j - x_i \sum_{i=1}^N x_i y_j \quad f = \alpha_j \beta_i + x$$

$$\log(K_I) = \sum_{i=0}^N \omega_i x_i + \sum_{j=1}^M \sum_{l=1}^N \lambda$$

Fonte: Heck GS, Pintro VO, Pereira RR, de Ávila MB, Levin NMB, de Azevedo WF. Supervised Machine Learning Methods Applied to Predict Ligand- Binding Affinity. *Curr Med Chem.* 2017; 24(23):2459-2470.

# Espaço de Funções Escores



**TABA**



$$\log(IC_{50}) = \sum_{i=0}^N \omega_i x_i + \sum_{j=0}^N \alpha_j x_j^i$$

$$\Delta G = \sum_{i=0}^N \omega_i x_i \quad f = \sum_{i=1}^N \alpha_i x_i - x_j^{-3} + \sum_{j=1}^M x$$

$$\Delta S = \alpha_j - x_i \sum_{i=1}^N x_i y_j \quad f = \alpha_j \beta_i + x$$

$$\log(K_I) = \sum_{i=0}^N \omega_i x_i + \sum_{j=1}^M \sum_{i=1}^N \lambda$$

Fonte: Heck GS, Pintro VO, Pereira RR, de Ávila MB, Levin NMB, de Azevedo WF. Supervised Machine Learning Methods Applied to Predict Ligand- Binding Affinity. *Curr Med Chem.* 2017; 24(23):2459-2470.

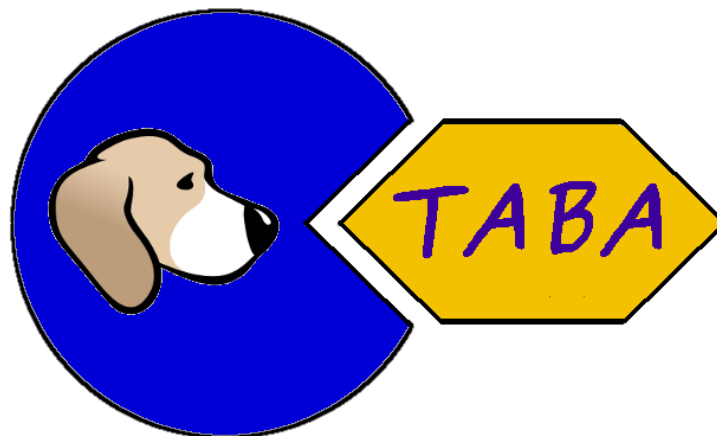
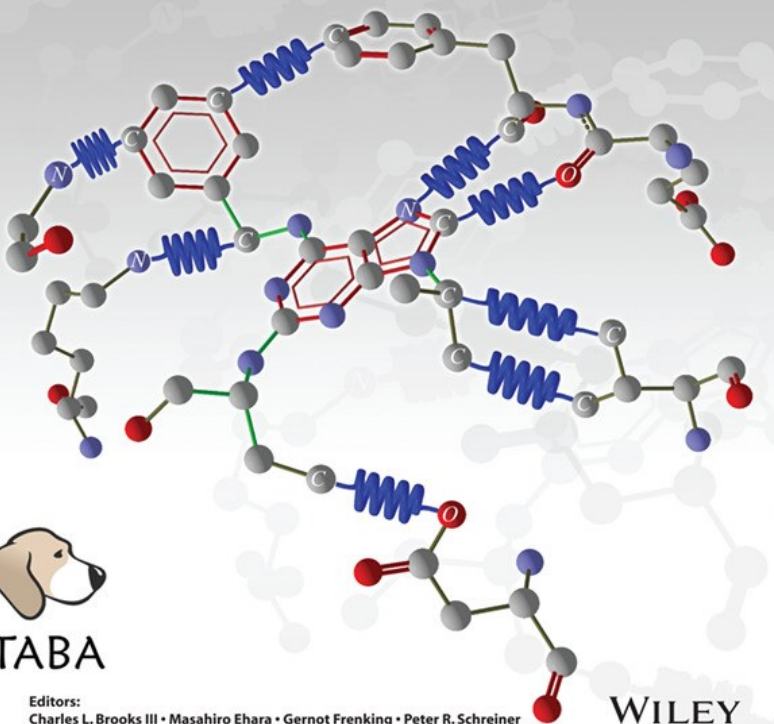


## Taba-Tool to Analyze the Binding Affinity

Volume 41 | Issues 1-2 | 2020  
Included in this print edition:  
Issue 1 (January 5, 2020)  
Issue 2 (January 15, 2020)

Journal of  
**COMPUTATIONAL  
CHEMISTRY**  
Organic • Inorganic • Physical  
Biological • Materials

www.c-chem.org



python  
powered

NumPy

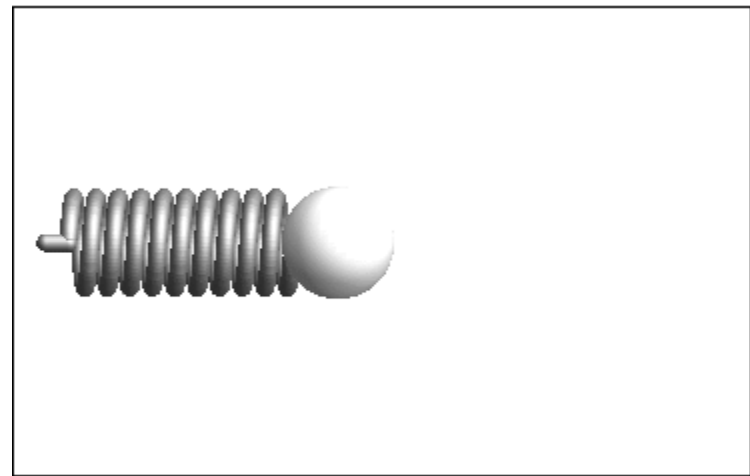
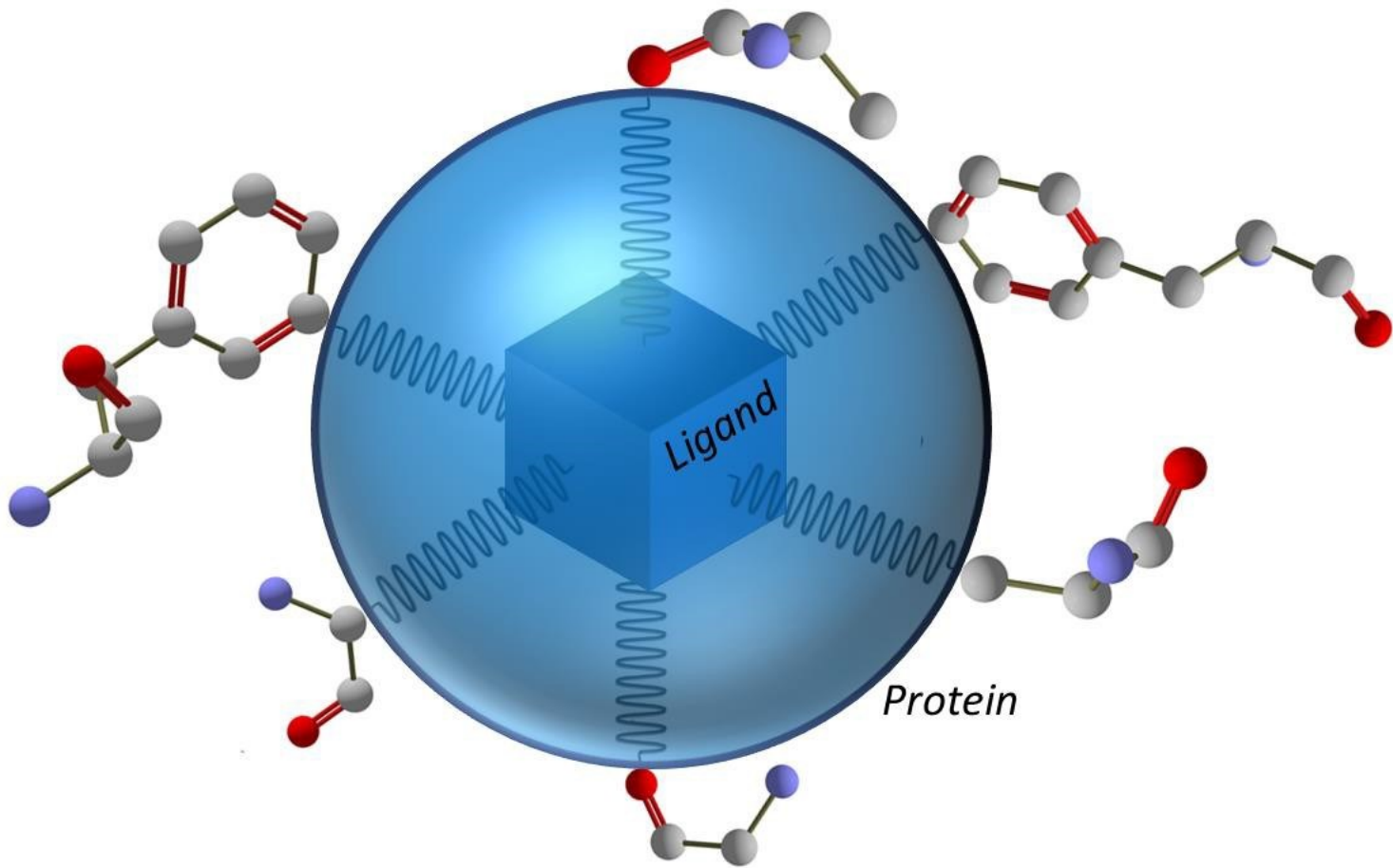
SciPy.org  
Sponsored By  
ENTHOUGHT

scikit  
learn

Available at: <https://github.com/azevedolab>

Fonte: da Silva AD, Bitencourt-Ferreira G, de Azevedo WF Jr. Taba: A Tool to Analyze the Binding Affinity. J Comput Chem. 2020; 41(1):69-73.

# Taba-Tool to Analyze the Binding Affinity



Fonte: da Silva AD, Bitencourt-Ferreira G, de Azevedo WF Jr. Taba: A Tool to Analyze the Binding Affinity. J Comput Chem. 2020; 41(1):69-73.

## Taba-Tool to Analyze the Binding Affinity

$$PBA = \alpha_0 + \sum_i \sum_j \alpha_{i,j} (d_{i,j} - d_{0,i,i})^2$$

Pesos do modelo de regressão:

$$\alpha_0 = -6,581356;$$

$$\alpha_{C,N} = -0,111232;$$

$$\alpha_{C,O} = -0,406456;$$

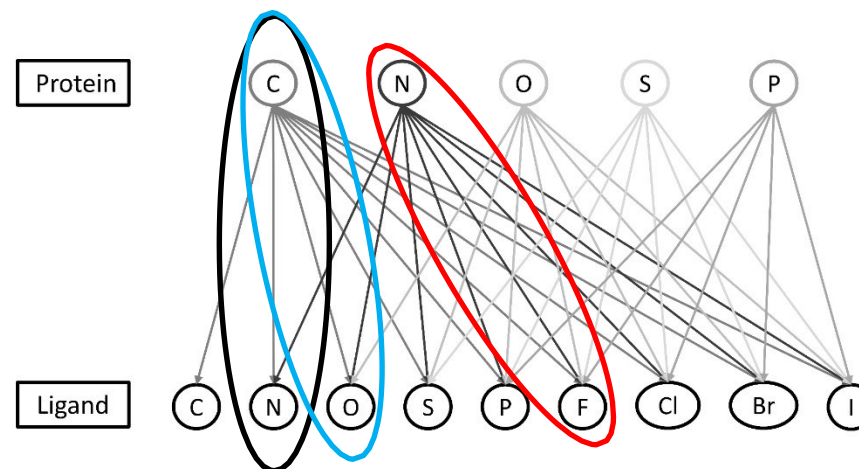
$$\alpha_{N,F} = -0,353717.$$

Distâncias de equilíbrio:

$$d_{0,C,N} = 3,99463;$$

$$d_{0,C,O} = 3,88190;$$

$$d_{0,N,F} = 4,21672 \text{ \AA}.$$



45 combinações

## Tabo-Tool to Analyze the Binding Affinity

Scoring Functions	$\rho$	p-value1	R <sup>2</sup>	p-value2
Free Energy <sup>a</sup>	-0.133	0.7324	0.204	0.2227
Final Intermolecular Energy <sup>a</sup>	0.133	0.7324	0.204	0.2228
vdW+Hbond+desolv Energy <sup>a</sup>	0.133	0.7324	0.204	0.2228
Electrostatic Energy <sup>a</sup>	0.533	0.1392	0.376	0.0789
Final Total Internal Energy <sup>a</sup>	-0.133	0.7324	0.089	0.4365
Torsional Free Energy <sup>a</sup>	0.068	0.8630	0.000	0.9792
Plants Score <sup>b</sup>	0.183	0.6368	0.001	0.9348
MolDock Score <sup>b</sup>	0.217	0.5755	0.010	0.7950
Rerank Score <sup>b</sup>	0.333	0.3807	0.007	0.8336
Interaction Score <sup>b</sup>	0.367	0.3317	0.013	0.7698
Protein Score <sup>b</sup>	0.367	0.3317	0.025	0.6839
Water Score <sup>b</sup>	-0.569	0.1098	0.395	0.0699
Internal Score <sup>b</sup>	0.033	0.9322	0.001	0.9369
Electrostatic Score <sup>b</sup>	0.548	0.1269	0.204	0.2218
Electrostatic Long Score <sup>b</sup>	-0.548	0.1269	0.204	0.2218
H-Bond Score <sup>b</sup>	0.650	0.0581	0.512	0.0301
Ligand Efficiency 1 Score <sup>b</sup>	0.150	0.7001	0.024	0.6935
Ligand Efficiency 3 Score <sup>b</sup>	0.283	0.4600	0.023	0.6968
Affinity Score <sup>c</sup>	-0.067	0.8647	0.117	0.3669
Gauss1 Score <sup>c</sup>	-0.367	0.3317	0.120	0.3603
Gauss2 Score <sup>c</sup>	-0.283	0.4600	0.018	0.7297
Repulsion Score <sup>c</sup>	-0.700	0.0358	0.240	0.1804
Hydrophobic Score <sup>c</sup>	0.100	0.7980	0.002	0.9157
Hydrogen Score <sup>c</sup>	-0.583	0.0992	0.340	0.0993
Tabo (3 variables, d ≤ 4.5 Å)	0.783	0.01252	0.794	0.0107

Predictive performance of scoring functions (test set). <sup>a</sup>AutoDock 4, <sup>b</sup>Molegro Virtual Docker (MVD), <sup>c</sup>AutoDock Vina. p-value1 and p-value2 are related to  $\rho$  and R, respectively.

Fonte: da Silva AD, Bitencourt-Ferreira G, de Azevedo WF Jr. Tabo: A Tool to Analyze the Binding Affinity. J Comput Chem. 2020; 41(1):69-73.

## SAnDReS-Statistical Analysis of Docking Results and Scoring functions



- 54 métodos de aprendizaje de máquina
- AutoDock Vina 1.2.3
- Software libre

Disponível: <https://github.com/azevedolab>



## Rede Internacional de Colaboradores



Dr. Olga Tarasova  
Prof. Vladimir Poroikov  
Institute of Biomedical Chemistry, Moscow  
Russia



Prof. Rodrigo Quiroga  
Prof. Marcos A. Villarreal  
Instituto de Investigaciones en Físicoquímica de Córdoba  
(INFIQC), CONICET-Departamento de Matemática y Física,  
Facultad de Ciencias Químicas, Universidad Nacional de  
Córdoba, Ciudad Universitaria, Córdoba,  
Argentina



Prof. Fernanda Canduri  
Instituto de Química de São Carlos  
Universidade de São Paulo-São Carlos-SP.  
Brazil



Dr. Stéphanie Baud  
Dr. Angelo Steffanel  
Prof. Manuel Dauchez  
Université de Reims Champagne Ardenne, Reims.  
France

Dr. José Henrique Pereira  
Molecular Biophysics and Integrated Bioimaging Division,  
Lawrence Berkeley National Laboratory, Berkeley, CA, 94720,  
USA

Prof. Marco Tutone  
Department of Biological, Chemical and Pharmaceutical Sciences  
and Technologies (STEBICEF)  
University of Palermo  
Palermo  
Italy

## Exemplos de Projetos Propostos (Center: X: -8.66 Å, Y: 9.96 Å, Z: 13.37 Å) (RANDOM 1123581321)

#	Title	Parameter Settings	Docking Results	ML Method
01	Differential Evolution and Gradient Boosting Regression to Predict Inhibition of Cyclin-Dependent Kinase 2	Score: MolDock Score Radius: 10 Å Search algorithm: Differential Evolution Number of runs: 1 Population size: 40 Max iterations: 200 Scaling factor: 0.5 Crossover rate: 0.8 Offspring: Scheme 1 Termination: Variance-based	RMSD: 0.557591 Å MolDock Score: -164.555 au	Gradient Boosting Regression

## Exemplos de Projetos Propostos (Center: X: -8.66 Å, Y: 9.96 Å, Z: 13.37 Å) (RANDOM 1123581321)

#	Title	Parameter Settings	Docking Results	ML Method
02	Combining Ant Colony Optimization and AdaBoost Regression to Address Protein-Ligand Interactions for Cyclin-Dependent Kinase 2	Score: Plants Score Radius: 15 Å Search algorithm: Ant Colony Optimization Number of runs: 1 Max iterations: 50 Population size: 20 <b>Simplex local search</b> Maximum steps: 2000 Tolerance: 0.01 Tolerance (iteration best solution): 0.0001 <b>Adaptative Sampling (ACO)</b> Evaporation rate: 0.15 Probability of best ant (pBest): 0.50	RMSD: 0.553042 Å Plants Score: -94.0287 au MolDock Score: -159.108 au	AdaBoost Regression

## Exemplos de Projetos Propostos (Center: X: -8.66 Å, Y: 9.96 Å, Z: 13.37 Å) (RANDOM 1123581321)

#	Title	Parameter Settings	Docking Results	ML Method
03	Voting Regression Meets Differential Evolution: A Computational Model to Predict Inhibition of Cyclin-Dependent Kinase 2	Score: MolDock Score Radius: 10 Å Search algorithm: Differential Evolution Number of runs: 1 Population size: 40 Max iterations: 200 Scaling factor: 0.5 Crossover rate: 0.8 Offspring: Scheme 1 Termination: Variance-based	RMSD: 0.557591 Å MolDock Score: - 164.555 au	Voting Regression

## Exemplos de Projetos Propostos (Center: X: -8.66 Å, Y: 9.96 Å, Z: 13.37 Å) (RANDOM 1123581321)

#	Title	Parameter Settings	Docking Results	ML Method
04	Shaking the Trees: A Machine-Learning Approach to Predict Binding Affinity for Cyclin-Dependent Kinase 2	Score: Plants Score Radius: 15 Å Search algorithm: Ant Colony Optimization Number of runs: 1 Max iterations: 50 Population size: 20 <b>Simplex local search</b> Maximum steps: 2000 Tolerance: 0.01 Tolerance (iteration best solution): 0.0001 <b>Adaptative Sampling (ACO)</b> Evaporation rate: 0.15 Probability of best ant (pBest): 0.50	RMSD: 0.553042 Å Plants Score: -94.0287 au MolDock Score: -159.108 au	Extra-Trees Regression



## Exemplos de Projetos Propostos (Center: X: -8.66 Å, Y: 9.96 Å, Z: 13.37 Å) (RANDOM 1123581321)

#	Title	Parameter Settings	Docking Results	ML Method
05	Combining Bagging Regressor and Ant Colony Optimization to Predict Binding Affinity for the Cyclin-Dependent Kinase 2	Score: Plants Score Radius: 15 Å Search algorithm: Ant Colony Optimization Number of runs: 1 Max iterations: 50 Population size: 20 <b>Simplex local search</b> Maximum steps: 2000 Tolerance: 0.01 Tolerance (iteration best solution): 0.0001 <b>Adaptative Sampling (ACO)</b> Evaporation rate: 0.15 Probability of best ant (pBest): 0.50	RMSD: 0.553042 Å Plants Score: -94.0287 au MolDock Score: -159.108 au	Bagging Regression



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