PUCRS
Pontifícia Universidade Católica
do Rio Grande do Sul

OLAP Docking

Uma solução OLAP para análise de experimentos de docagem molecular: Aplicação com a enzima InhA.

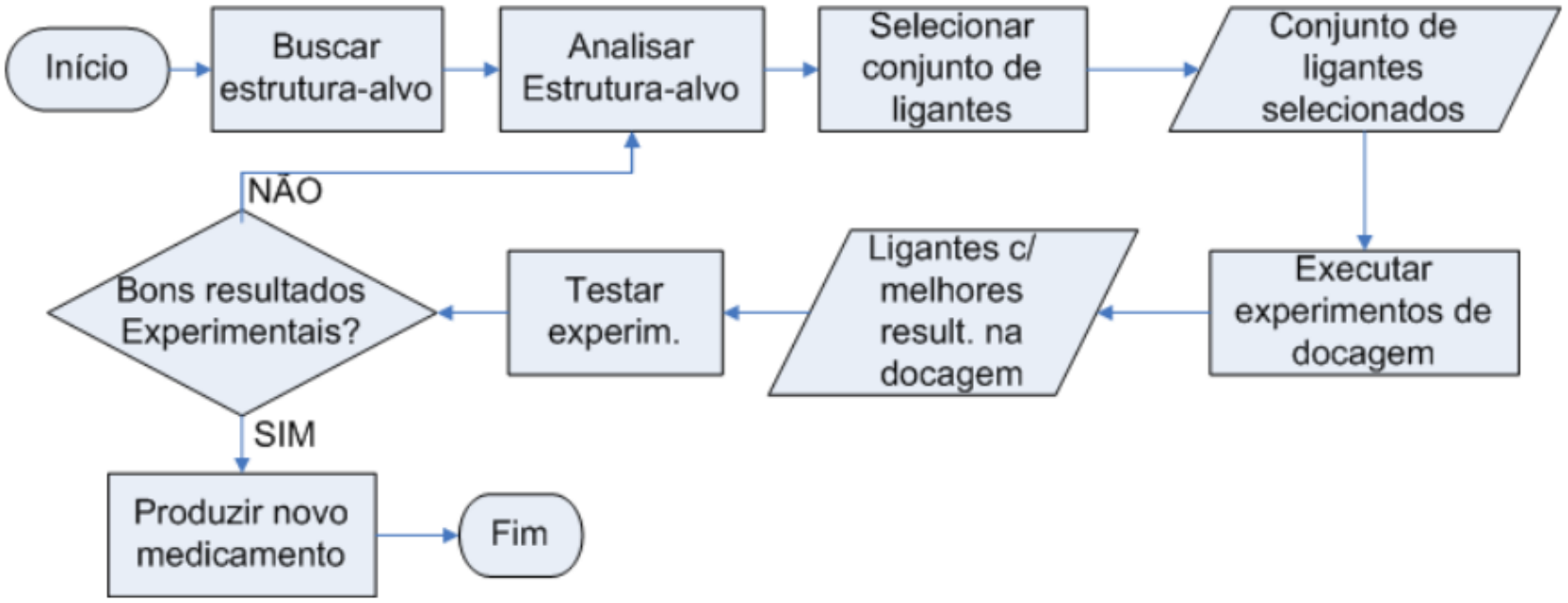
Orientador:
Duncan D. A. Ruiz

Bruno Rubin dos Santos
Matheus Eduardo Bonifácio Morais

Sumário

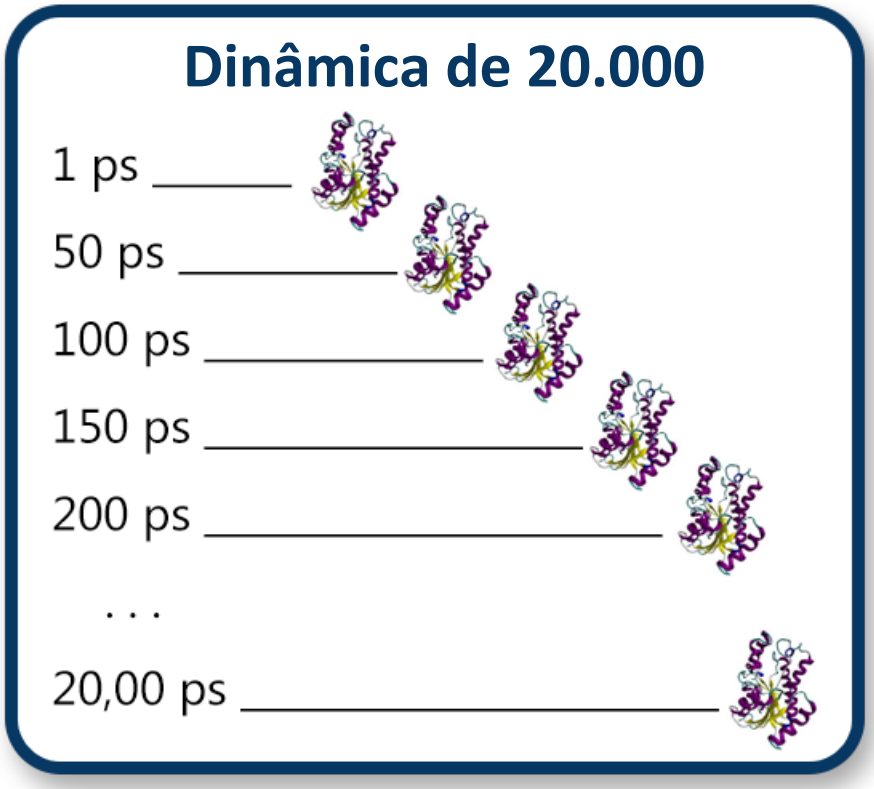
Introdução

- Processo de desenvolvimento racional de novos fármacos (RDD);
- Experimentos in-silico;
- Alto custo (1,2 bilhões de dólares);
- Processo demorado (12 anos).



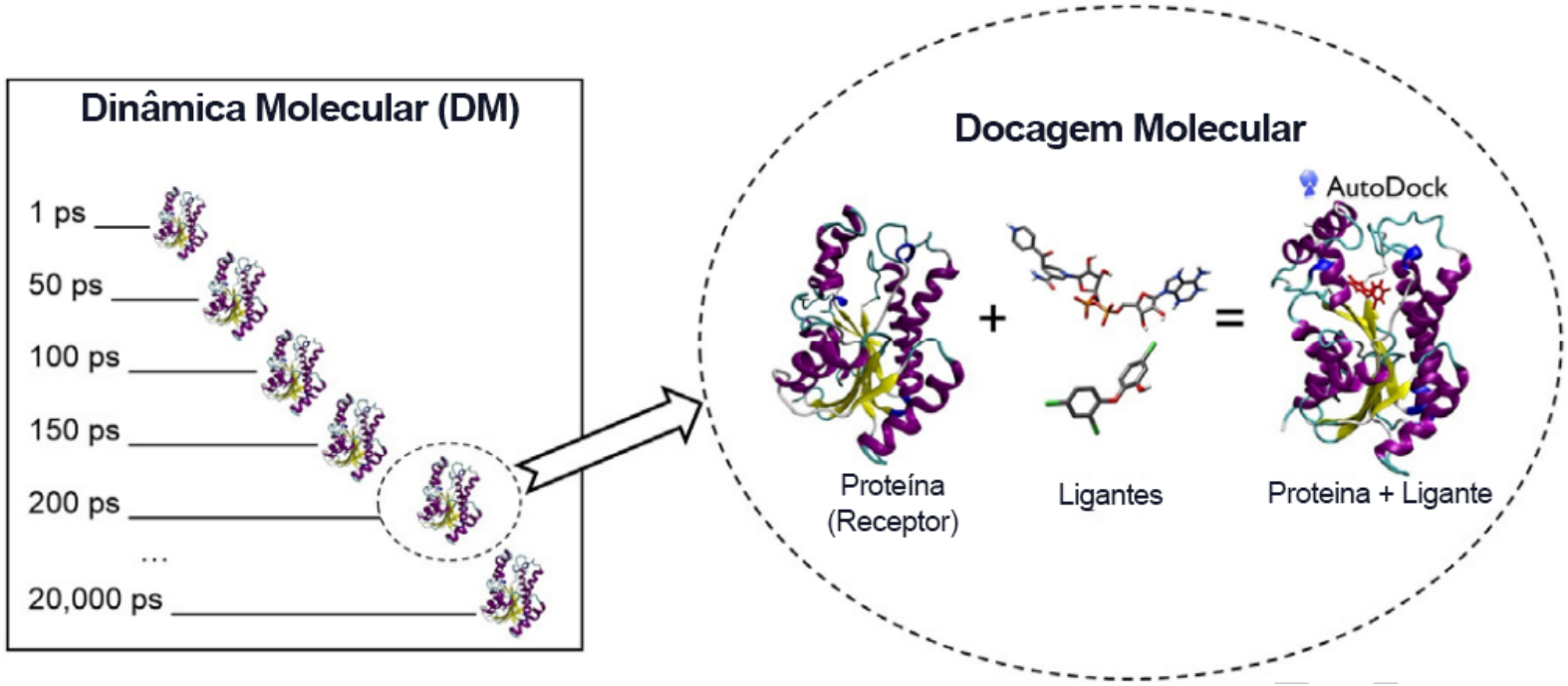
Dinâmica Molecular

- Utilizado para representar a flexibilidade da proteína receptora;
- Conjunto de snapshots (conformação);
- Cada snapshot representa um instante de tempo;
  - Posicionamento dos átomos
  - Coordenadas variáveis



Docagem Molecular

- Objetivo principal é identificar novos candidatos a fármaco ou melhorar compostos já existentes.
  - Potenciais inibidores da proteína receptora
- Simulações entre proteína receptora e ligantes
  - Analisar interações (ligações)
- Docagem Molecular utilizando Dinâmica Molecular:



Docagem Molecular

- Dados resultantes de experimentos são analisados de forma manual.
- Especialista de domínio segue um protocolo de avaliação para identificar as ligações estáveis.
- Dados resultantes crescem de acordo com a dinâmica utilizada.

- **Resultado:**

Data set CSV

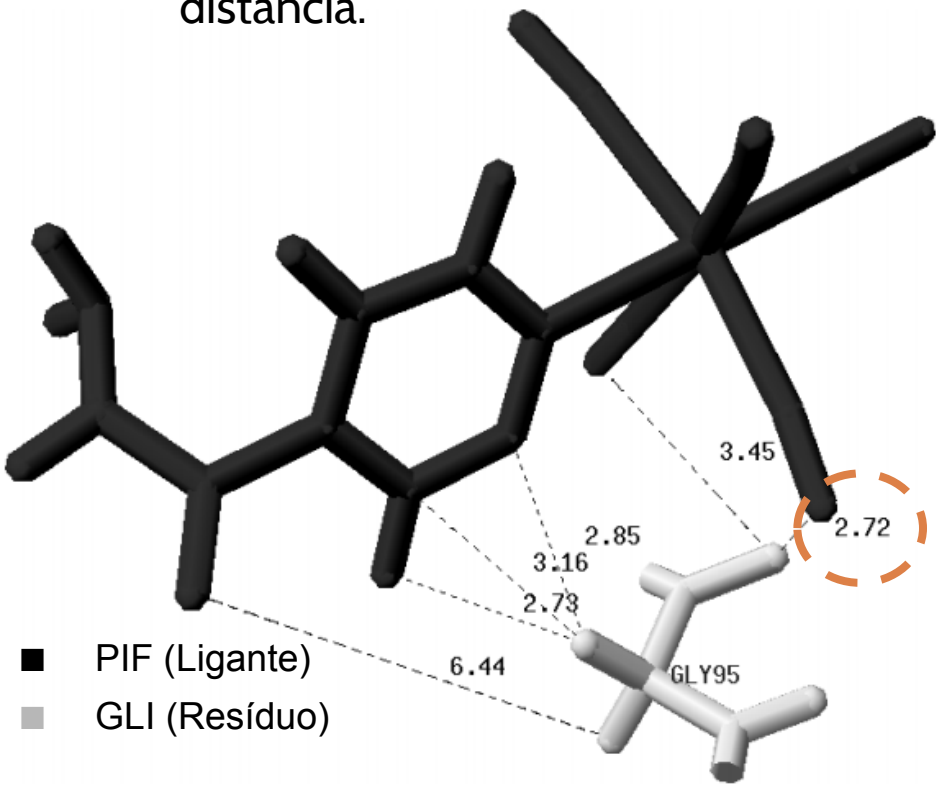
- **Lista de resíduos**
  Lista de átomos
- **Posicionamento por Snapshot**
  Coordenadas X, Y e Z
- **Melhor FEB do ligante**
- **Melhor RMSD do ligante**

Identificação de Métricas

1. **FEB**
   - Energia livre de ligação;
   - Termodinâmica;
   - Quanto menor o valor, mais estável é a ligação.
2. **RMSD**
   - *Root-mean-square Deviation;*
   - Comparar o posicionamento inicial do ligante, proposto pelo especialista de domínio, com o posicionamento após a execução da docagem.
3. **Número de contatos**
   - Ligações estabelecidas entre ligante e os principais resíduos da molécula receptora.

Identificação de Métricas

- Os contatos entre o ligante e o resíduo são medidos através da distância entre seus átomos.
  - Coordenada X, Y e Z de cada átomo
  - **Ångström (Å)** é a medida de comprimento usada para expressar a distância.



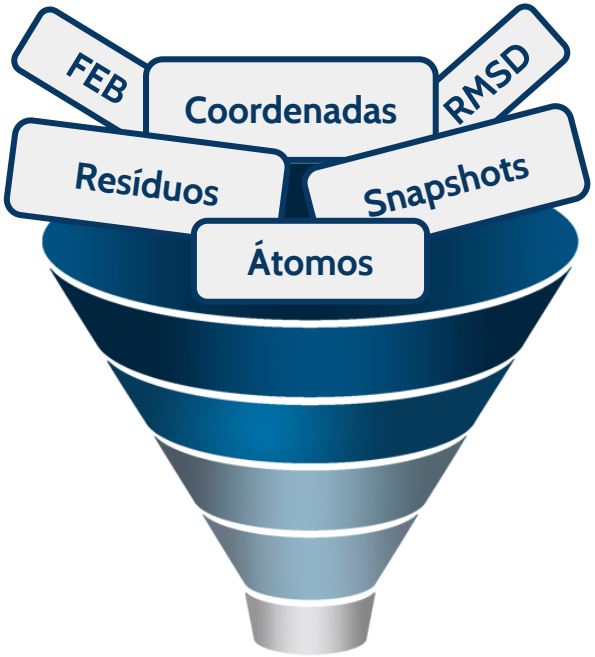
2,72 Å

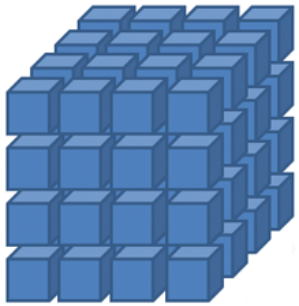
De todas as distâncias, considera-se apenas a menor

Solução Proposta

- **OLAP:**
  - Solução que permite analisar os dados de forma flexível
    - Roll-up (Consolidação)
    - Drill-down
  - Multidimensionalidade;
  - Cruzamento de dados;
  - Cubo de dados
    - Análise de informações
    - Relatórios

Solução Proposta





Cubo de dados

- Utilizado na esfera estratégica das organizações com foco nas áreas financeiras, vendas, marketing e etc;
- Análisar resultados de múltiplos experimentos de docagem;
- Identificar padrões baseados em histórico de dados;
- Responder questões relevantes para o negócio;
- Pela literatura não há registro do uso deste tipo de tecnologia para endereçar questões relacionadas à docagem molecular.

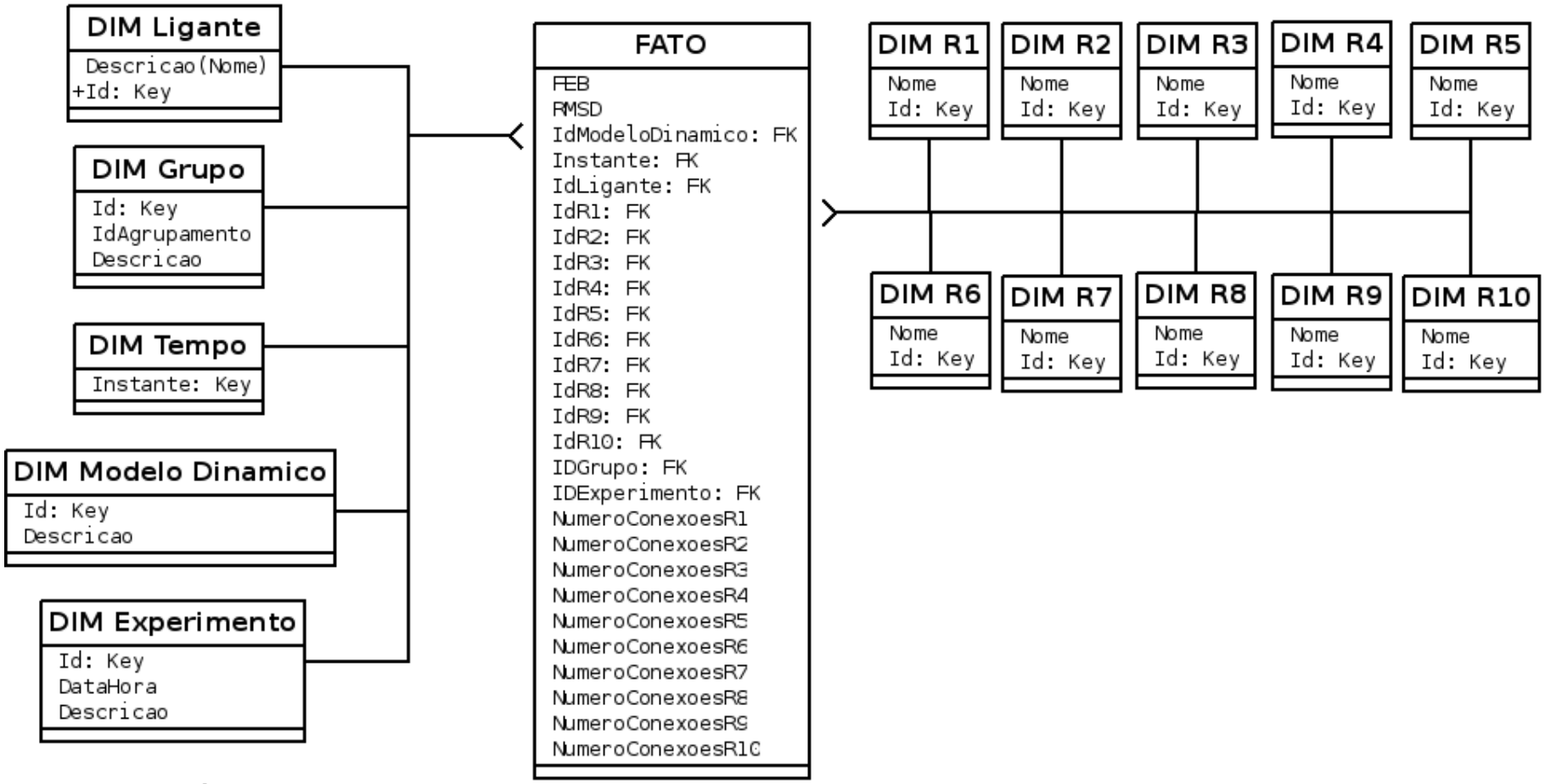
Solução Proposta

- **Definição de questões de negócio**:
  - Entrevistas com os especialistas do LABIO;
  - Identificação de questões relevantes;
  - Flexibilidade para uso em diversos cenários.

- **Questões identificadas:**

1. Associar um grupo para cada conformação.
2. Identificar o comportamento das conformações baseado nas métricas de FEB e RMSD.
3. Identificar conformações/grupos que possuem o maior número de contatos com os ligantes.
4. Com base no item 3, identificar quais são os resíduos mais importantes.
5. Com base no item 3, identificar quais grupos possuem melhores valores de FEB e RMSD.

Solução Proposta



- FATO
  - Compreende as chaves das dimensões e as métricas FEB, RMSD e o número de contatos dos principais ligantes.

Solução Proposta

- **DIM Ligante**
  - Representação dos ligantes utilizados nos experimentos.
- **DIM Grupo**
  - Agrupamento de conformações que o LABIO utiliza em seus experimentos, baseados no posicionamento de cada conformação.
- **DIM Tempo**
  - Representação do tempo em picosegundos.
- **DIM Modelo Dinâmico**
  - Quais dinâmicas foram utilizadas para um determinado experimento.
- **DIM Experimento**
  - Pode possuir um mesmo experimento utilizando diferentes algoritmos de docagem e também versões de software de docagem diferente.
- **DIM R1 ... DIM R10**
  - Resíduos considerados mais relevantes para um experimento.

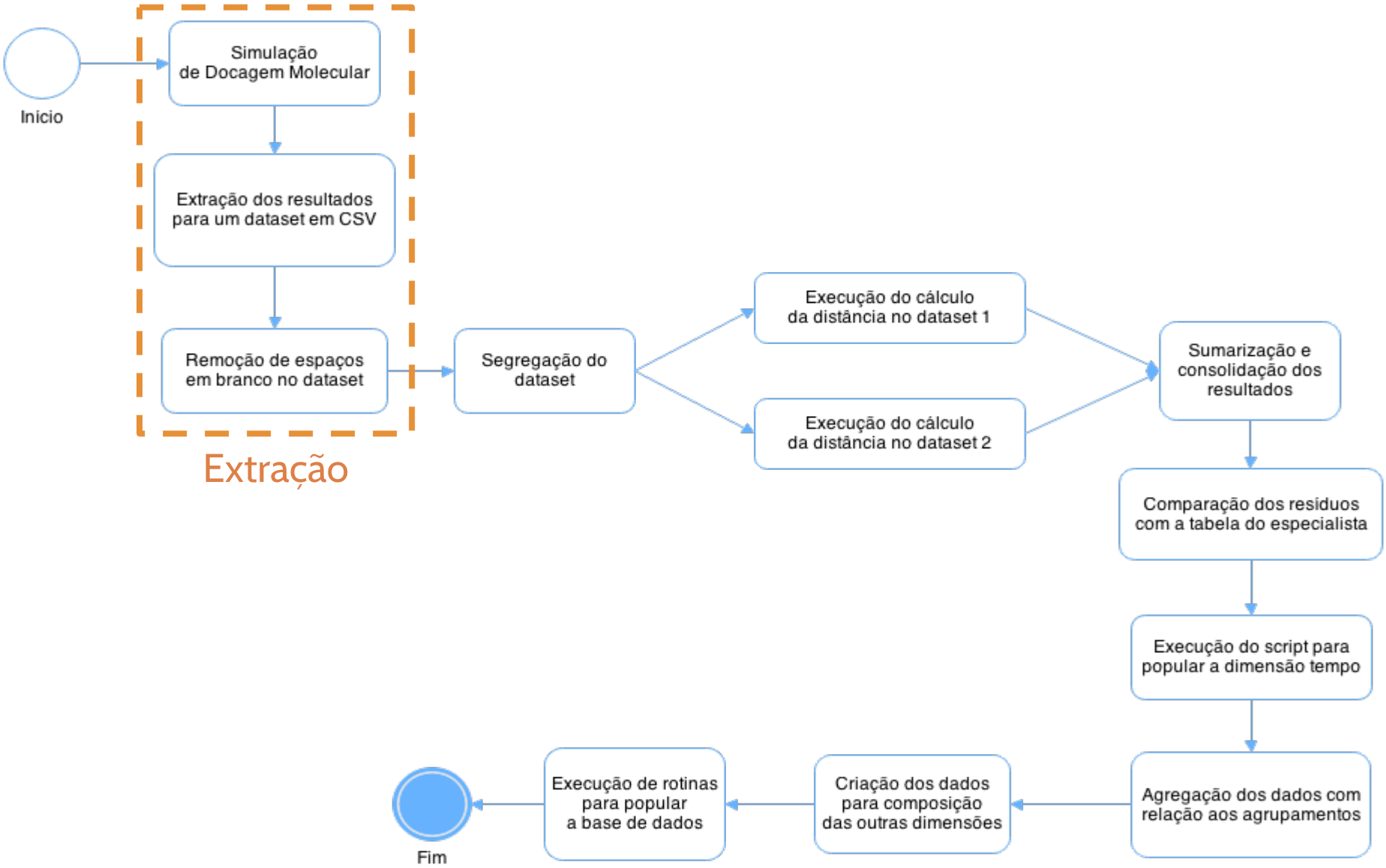
Processo de Extração

Data set CSV

- Data set resultante da simulação de docagem:
  - 3100 linhas
  - 12335 colunas
- Composição:
  - 3.100 Snapshots
  - 268 Resíduos
  - 4.008 Átomos (Cada um com X, Y e Z)
  - Ligantes TCL e ETH
  - FEB e RMSD.

- Script para remover espaços em branco no arquivo para otimizar o tempo de execução dos processos de transformação.
  - Original com 550 MB
  - Redução para 250 MB

Processo de Extração



Resíduos Relevantes

- Critérios definidos:
  - Ligações com distâncias entre 2 e 4 Ångströms são consideradas um contato;
  - Para cada resíduo, somente átomos sem hidrogênio são calculados;
  - Para cada snapshot, todos os resíduos são calculados.

- Distância dos átomos pelo cálculo da distância Euclidiana:

$$d(R, L) = \sqrt{(r_x - l_x)^2 + (r_y - l_y)^2 + (r_z - l_z)^2}$$

- Critérios de mensuração:
  - Quanto mais ligações estabelecer, mais relevante o resíduo se torna para o experimento de docagem;
  - Comparar resíduos com uma lista existente no LABIO.

Processo de Transformação

- Segregação do arquivo gerado no processo de extração em dois:
  - Um arquivo apenas para o resíduo NAH e o ligante TCL;
  - Outro arquivo contendo todo o restante exceto o resíduo NAH.
- Script para identificar os resíduos:
  - Cálcula a distância Euclidiana entre átomos do ligante e do receptor;
  - Ignora o cálculo para átomos de hidrogênio.
- Sumarização do número de contatos de cada resíduo com o ligante em um único arquivo.

- Tempo de execução das rotinas de transformação para o cenário descrito foi de ~35 minutos em um Intel Core i5 (2.5GHz)

Processo de Transformação

- Output do script que calcula as distâncias:

```
$ ./calculaDistancia.py residuos.txt ligantes.txt tmp/Docking_Simul.csv
Snapshot,Residuo,Ligante,Distancia,Classificacao
1,SER_93,ETH,3.94,2-4
1,SER_93,ETH,3.45,2-4
1,SER_93,ETH,3.65,2-4
1,SER_93,ETH,2.79,2-4
1,SER_93,ETH,3.66,2-4
1,SER_93,ETH,3.86,2-4
1,SER_93,ETH,2.81,2-4
1,SER_93,ETH,3.31,2-4
1,SER_93,ETH,2.01,2-4
1,ILE_94,ETH,3.95,2-4
1,ILE_94,ETH,3.94,2-4
1,ILE_94,ETH,3.28,2-4
1,ILE_94,ETH,3.78,2-4
1,ILE_94,ETH,3.77,2-4
1,ILE_94,ETH,3.07,2-4
1,ILE_94,ETH,2.99,2-4
```

- Dados sumarizados:

```
Total,Snapshot,Residuo,Ligante,Classificacao
1,1000,PHE_96,ETH,2-4
1,1000,SER_122,ETH,2-4
1,1001,ALA_190,ETH,2-4
1,1001,LYS_164,ETH,2-4
```

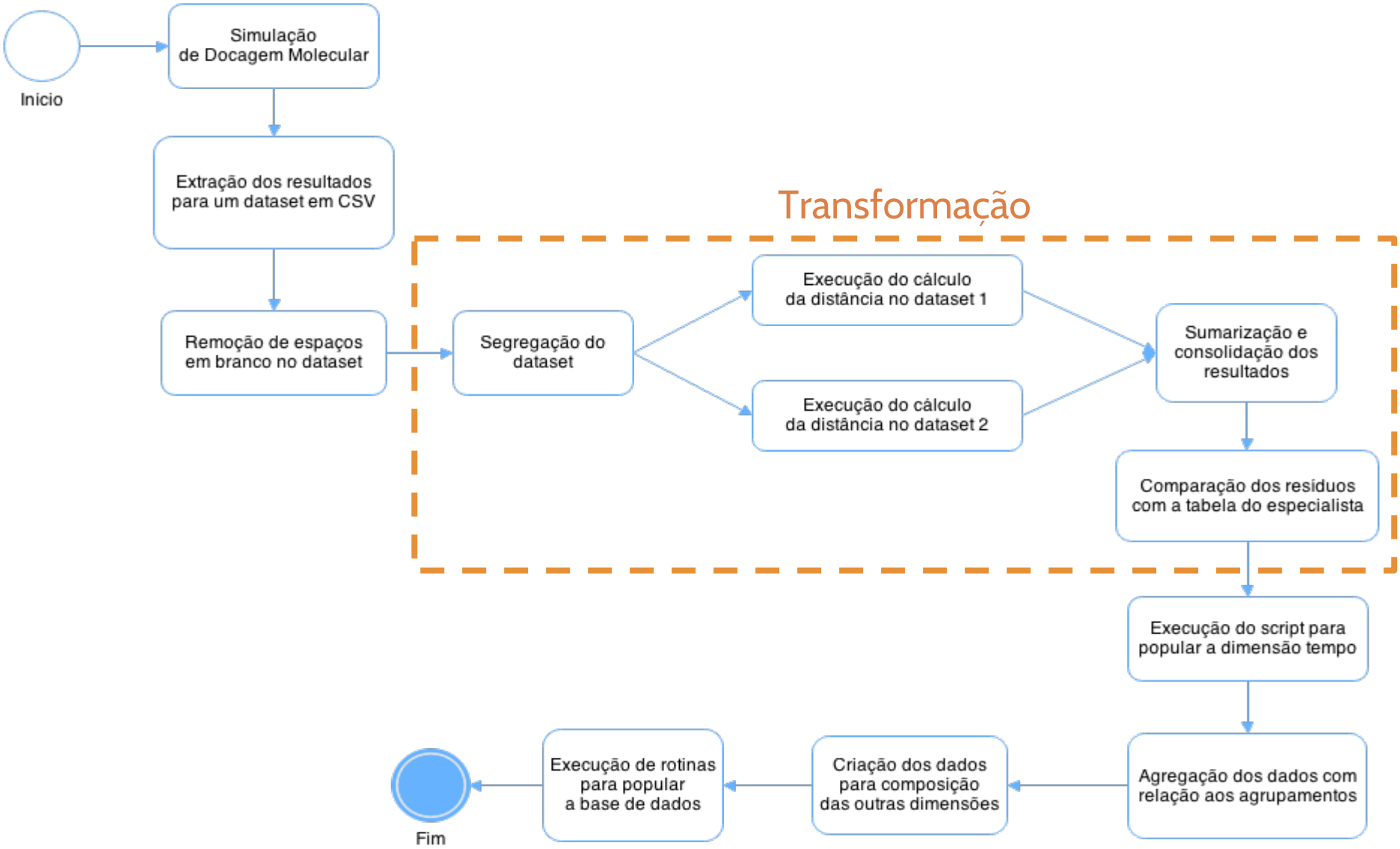
Processo de Transformação

- **Resultado:**
  - Identificados 15 resíduos comuns para os ligantes TCL e ETH.
- Comparação dos resultados com a listagem do LABIO:
  - Resíduos mais relevantes baseados em histórico de experimentos;
  - 10 dos 15 resíduos encontrados estavam na lista;

- Definição dos 10 resíduos relevantes para o experimento:

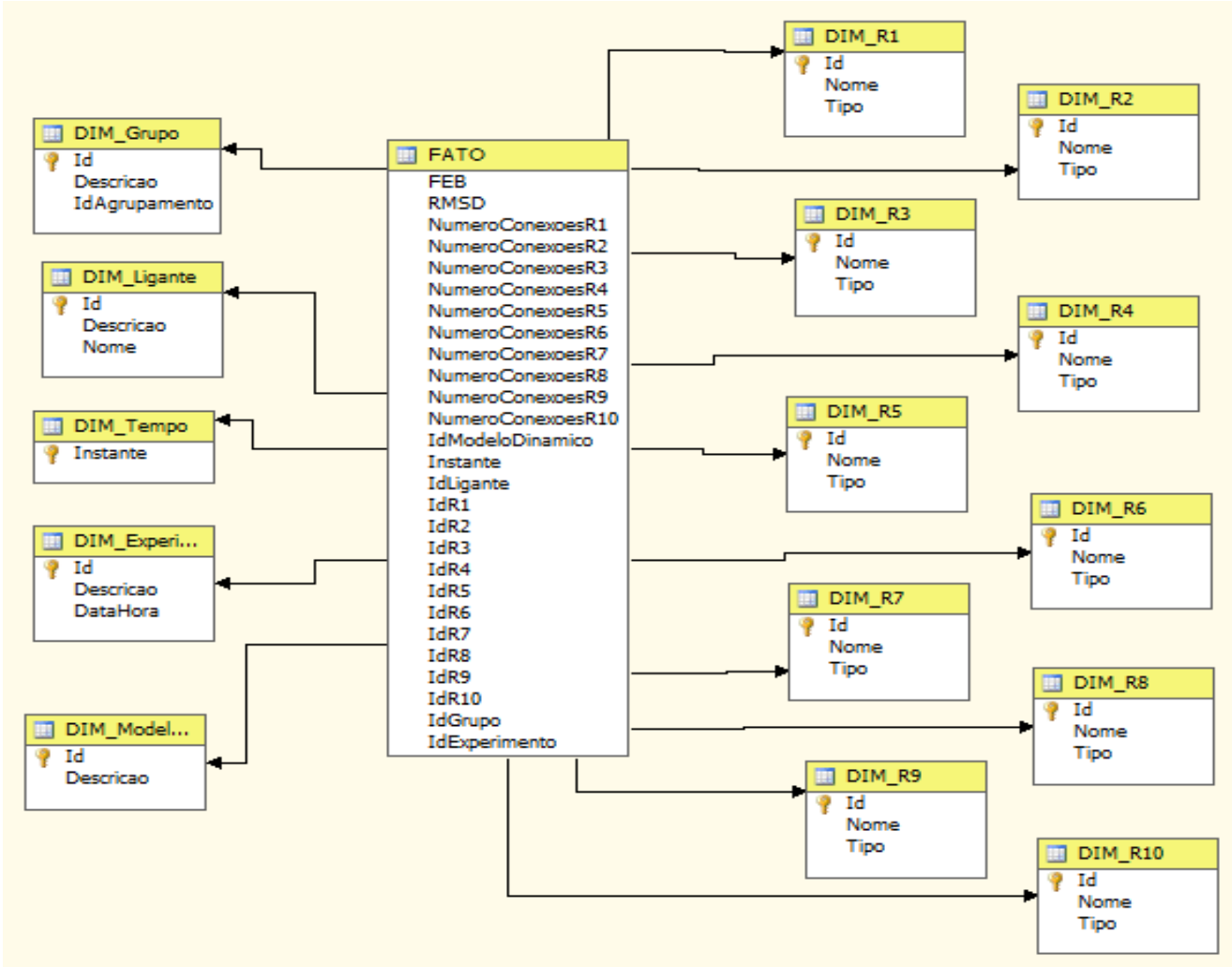
ILE_15	SER_19	ILE_94	GLY_95	PHE_148
TYR_157	MET_160	ILE_193	THR_195	MET_198

Processo de Transformação



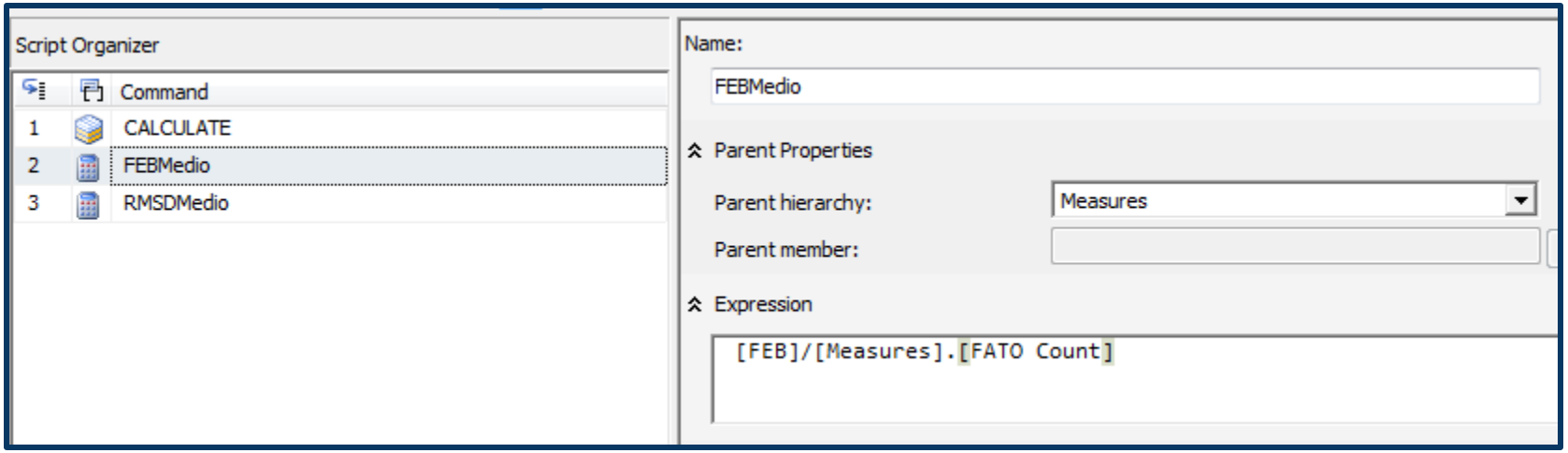
Construção do Modelo

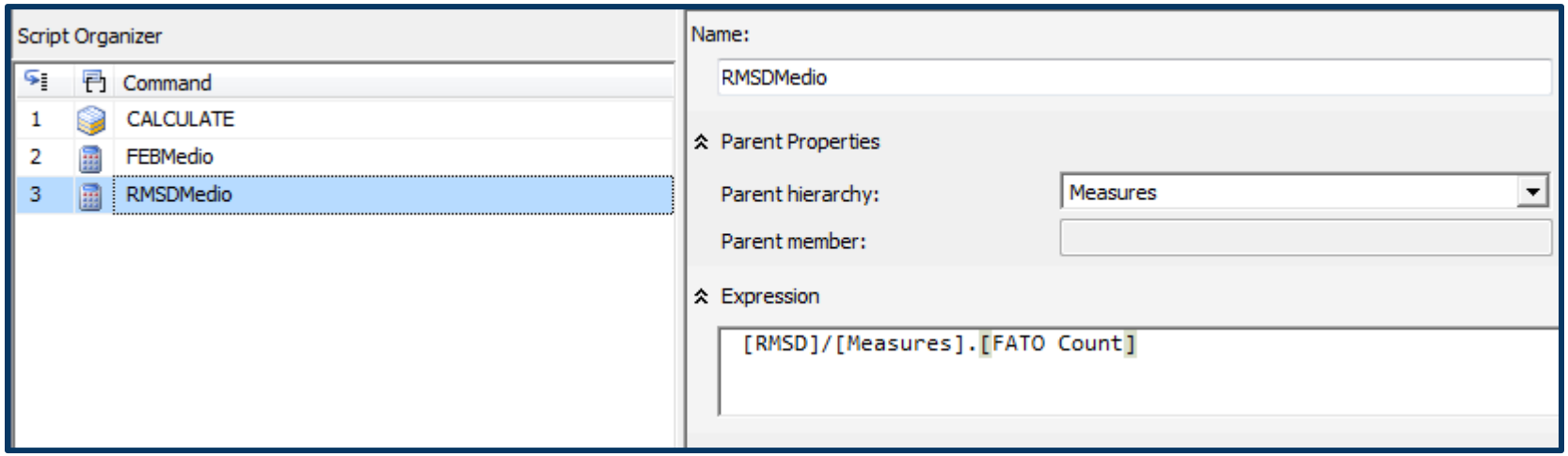
- Modelo criado no Microsoft Analysis Services:



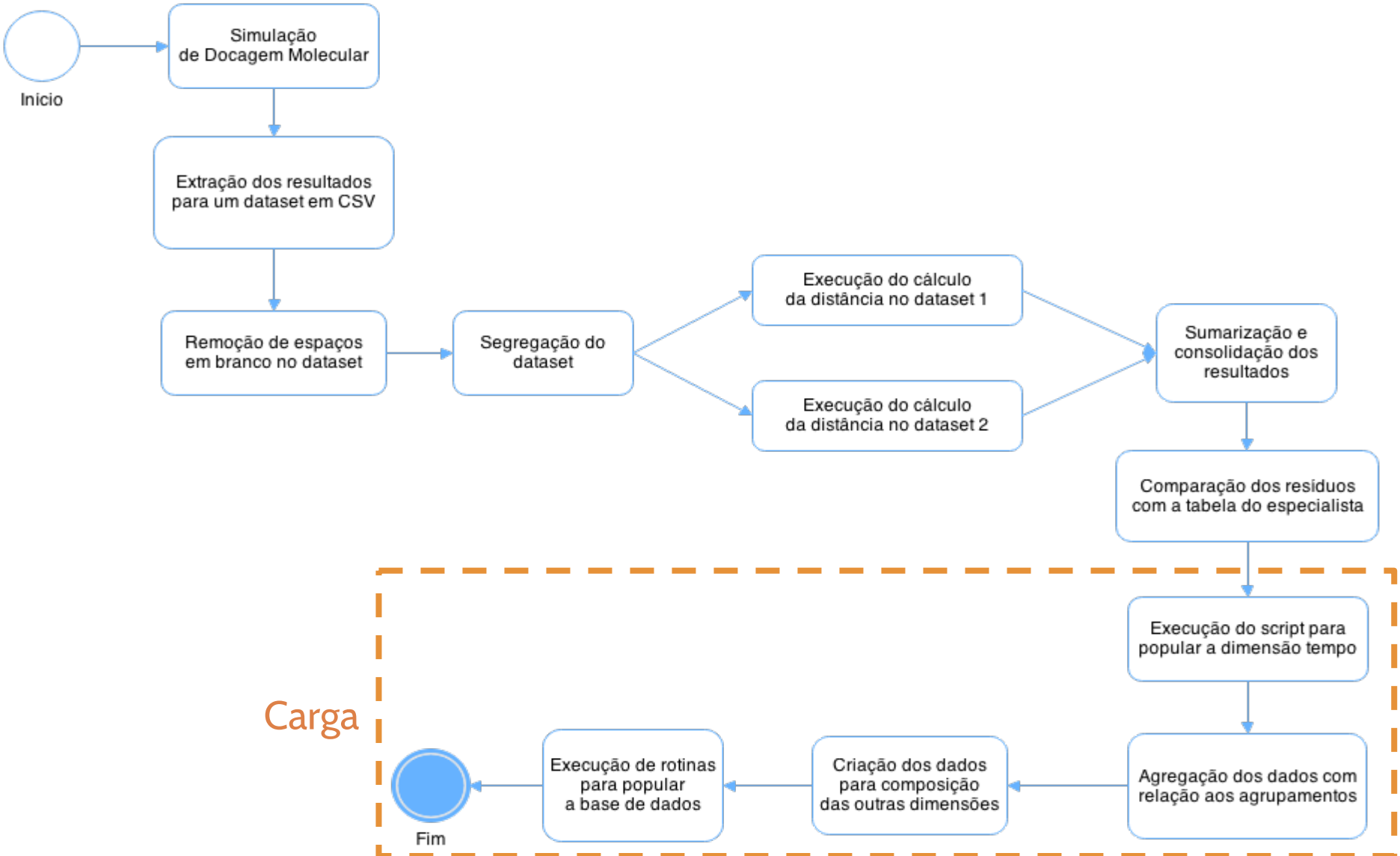
Construção do Modelo

- Métricas 'FEBMedio' e 'RMSDMedio' criadas no cubo:





Processo de Carga



Processo de Carga

- Consolidação das informações para composição das dimensões;
- Identificação de FEB positivas:
  - Estes valores foram descartados (colocando o valor 0 para estes casos)
- Script para popular dimensão 'Tempo';
- Script para carga de dados:
  - Gera os comandos SQL para inserção no banco de dados baseado nos resultados do processo de transformação.

- Tempo de execução do processo de carga para o cenário descrito foi de ~15 minutos em um Intel Core i5 (2.5GHz).
- Os processos de transformação e carga podem ter o tempo de execução otimizados através da implementação de paralelismo (multi-thread) nas rotinas.

Processo de Carga

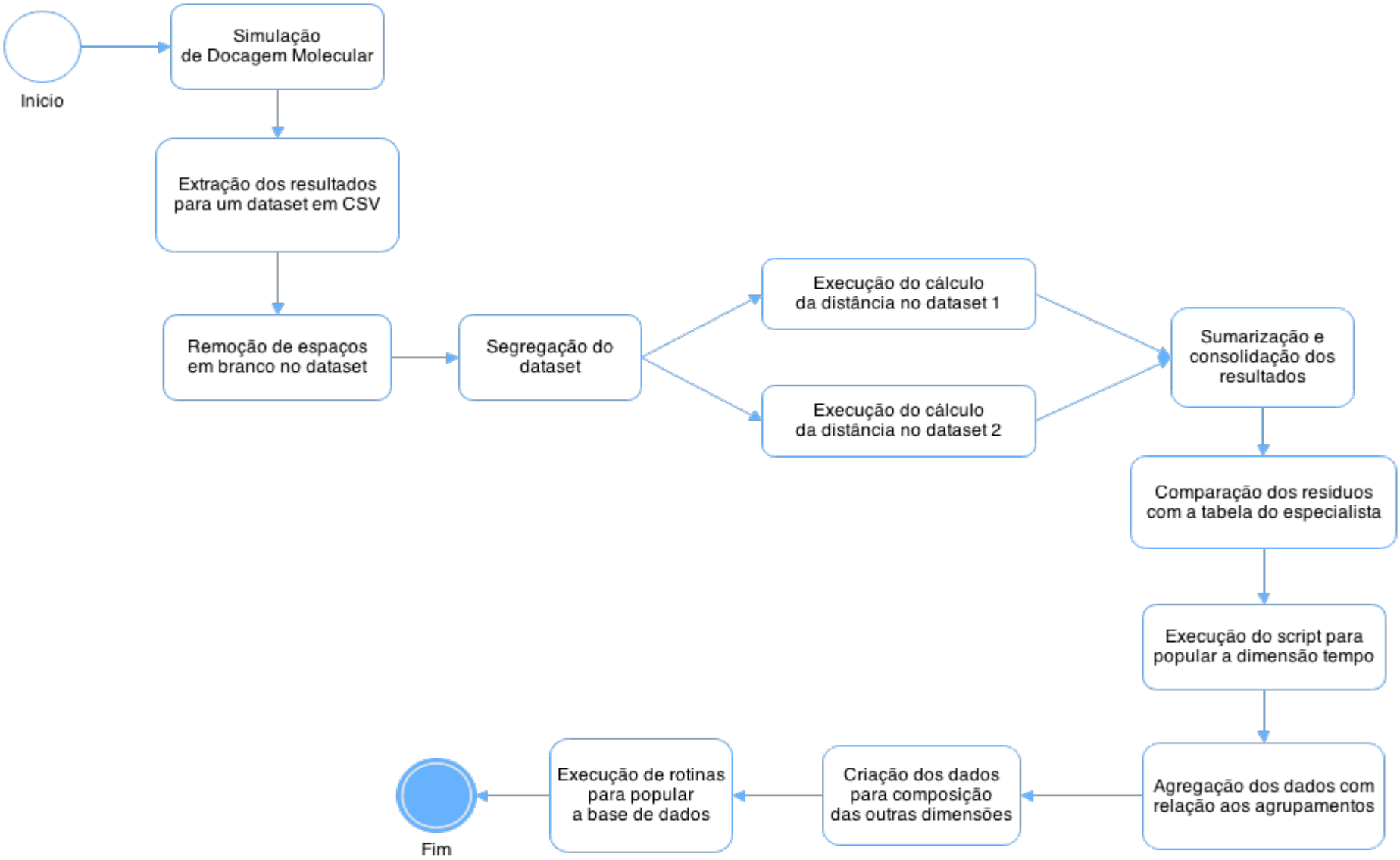
- **Exemplo de entradas e saídas dos scripts de carga:**
- Execução script para popular dimensão 'Tempo':

```
$ ./criaDadosDIMTempo.py 10
INSERT INTO DIM_Tempo VALUES (1), (2), (3), (4), (5), (6), (7), (8), (9), (10);
```

- Execução do script para popular o restante dos dados:

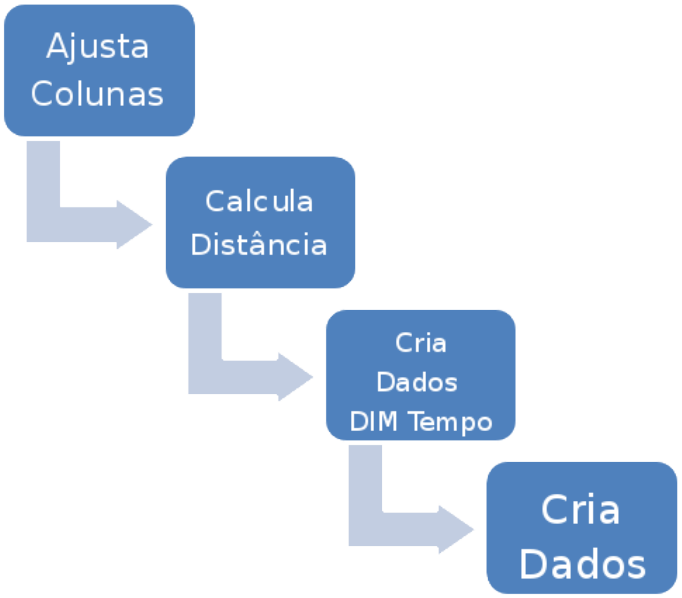
```
$ ./criaDados.py tmp/dim_ligante.csv tmp/dim_grupo.csv tmp/dim_modelo_dinamico.csv tmp/dim_experimento.csv tmp/residuos.csv tmp/arquivos_sumarizados.csv tmp/Docking_Simul.csv tmp/grupos_DM_3100.csv
INSERT INTO DIM_Ligante (Id, Nome, Descricao) VALUES ('1', 'ETH', 'Etionamida'), ('2', 'TCL', 'Triclosan');
INSERT INTO DIM_Grupo (Id, IdAgrupamento, Descricao) VALUES ('1', '0', 'k_means_NADH'), ('2', '1', 'k_means_NADH'), ('3', '2', 'k_means_NADH'), ('4', '3', 'k_means_NADH'), ('5', '4', 'k_means_NADH'), ('6', '5', 'k_means_NADH');
INSERT INTO DIM_Modelo_Dinamico (Id, Descricao) VALUES ('1', 'Teste de Modelo Dinamico');
INSERT INTO DIM_Experimento (Id, DataHora, Descricao) VALUES ('1', '2009/10/12', 'Experimento de docagem para InhA com 3100 conformacoes com ligante Etionamida'), ('2', '2009/10/12', 'Experimento de docagem para InhA com 3100 conformacoes com ligante Triclosan');
INSERT INTO DIM_R1 (Id, Nome, Tipo) VALUES (1, 'PHE_148', 'Fenilalanina');
INSERT INTO DIM_R2 (Id, Nome, Tipo) VALUES (2, 'ILE_193', 'Isoleucina');
INSERT INTO DIM_R3 (Id, Nome, Tipo) VALUES (3, 'GLY_95', 'Glicina');
INSERT INTO DIM_R4 (Id, Nome, Tipo) VALUES (4, 'THR_195', 'Treonina');
INSERT INTO DIM_R5 (Id, Nome, Tipo) VALUES (5, 'ILE_94', 'Isoleucina');
INSERT INTO DIM_R6 (Id, Nome, Tipo) VALUES (6, 'MET_198', 'Metionina');
INSERT INTO DIM_R7 (Id, Nome, Tipo) VALUES (7, 'MET_160', 'Metionina');
INSERT INTO DIM_R8 (Id, Nome, Tipo) VALUES (8, 'SER_19', 'Serina');
INSERT INTO DIM_R9 (Id, Nome, Tipo) VALUES (9, 'ILE_15', 'Isoleucina');
INSERT INTO DIM_R10 (Id, Nome, Tipo) VALUES (10, 'TYR_157', 'Tirosina');
INSERT INTO FATO (FEB, RMSD, NumeroConexoesR1, NumeroConexoesR2, NumeroConexoesR3, NumeroConexoesR4, NumeroConexoesR5, NumeroConexoesR6, NumeroConexoesR7, NumeroConexoesR8, NumeroConexoesR9, NumeroConexoesR10, IdModeloDinamico, Instante, IdLigante, IdR1, IdR2, IdR3, IdR4, IdR5, IdR6, IdR7, IdR8, IdR9, IdR10, IdGrupo, IdExperimento) VALUES (-8.74, 3.79, 1, 0, 6, 0, 12, 0, 1, 0, 0, 0, 1, 1, 1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 1);
```

Processo ETL completo



Processo ETL completo

- **Fluxo de execução dos scripts desenvolvidos:**
- Rotinas escritas em Python e funcionam com entrada de parâmetros.
  - $ ./calculaDistancia.py residuos.txt ligantes.txt Docking_ETH_TCL.csv > resultado.csv
- Parâmetros de entrada e saída são arquivos TXT, CSV ou SQL.



Disponíveis em https://github.com/memorais/scripts_prep_bi

Resultados Obtidos

Demonstração

Resultados Obtidos

- Exemplo do cálculo das métricas 'FEBMedio' e 'RMSDMedio' por experimento:

Data Hora	Descricao	Id	FEBMedio	RMSDMedio	FATO Count
2009-10-12	Experimento de docagem para InhA com 3100 conformacoes com ligante Etionamida	1	-9,4622...	5,08509...	3100
2009-10-12	Experimento de docagem para InhA com 3100 conformacoes com ligante Triclosan	2	-0,5473...	6,95340...	3100

Resultados Obtidos

- Associar um grupo para cada conformação:

	A	B
1	Grupo	Conformações
2	0	582
3	1	948
4	2	1602
5	3	1014
6	4	1044
7	5	1010
8	Total Geral	6200

Resultados Obtidos

- Identificar o comportamento das conformações baseado nas métricas de FEB e RMSD:

	A	B	C	D
1	Grupo	Conformações	FEBMedio	RMSDMedio
2	0	582	-5,3108	5,0949
3	1	948	-4,7957	6,0078
4	2	1602	-5,3095	6,1983
5	3	1014	-5,0178	5,9891
6	4	1044	-4,7345	6,2437
7	5	1010	-4,8076	6,0770
8	**Total Geral**	**6200**	**-5,0048**	**6,0192**

Resultados Obtidos

- Identificar conformações/grupos que possuem o maior número de contatos com os ligantes:

Grupos	Numero Contatos R1	Numero Contatos R2	Numero Contatos R3	Numero Contatos R4	Numero Contatos R5	Numero Contatos R6	Numero Contatos R7
0	**1147**	**737**	**2043**	**643**	**967**	**858**	**412**
ETH	1083	694	314	499	959	283	167
TCL	64	43	1729	144	8	575	245
1	**673**	**547**	**2266**	**390**	**996**	**69**	**669**
ETH	673	547	561	192	983	29	515
TCL	0	0	1705	198	13	40	154
2	**1077**	**1101**	**4267**	**421**	**2169**	**147**	**1154**
ETH	1022	1101	840	301	2160	118	1042
TCL	55	0	3427	120	9	29	112
3	**981**	**1029**	**3218**	**659**	**1074**	**393**	**761**
ETH	926	1029	430	535	1063	287	631
TCL	55	0	2788	124	11	106	130
4	**228**	**248**	**2628**	**446**	**1233**	**108**	**854**
ETH	228	248	829	252	1227	50	681
TCL	0	0	1799	194	6	58	173
5	**714**	**760**	**2169**	**338**	**1230**	**181**	**711**
ETH	714	760	573	262	1219	147	668
TCL	0	0	1596	76	11	34	43
Total Geral	**4820**	**4422**	**16591**	**2897**	**7669**	**1756**	**4561**

Resultados Obtidos

- Com base no item 3, identificar quais são os resíduos mais importantes:

	A	B
1	**Grupos**	
2	**0**	
3	Numero Contatos R1	1147
4	Numero Contatos R2	737
5	Numero Contatos R3	2043
6	Numero Contatos R4	643
7	Numero Contatos R5	967
8	Numero Contatos R6	858
9	Numero Contatos R7	412
10	Numero Contatos R8	78
11	Numero Contatos R9	284
12	Numero Contatos R10	232
13	**1**	
14	Numero Contatos R1	673
15	Numero Contatos R2	547
16	Numero Contatos R3	2266
17	Numero Contatos R4	390
18	Numero Contatos R5	996
19	Numero Contatos R6	69
20	Numero Contatos R7	669
21	Numero Contatos R8	846
22	Numero Contatos R9	721
23	Numero Contatos R10	6
24	**2**	
25	Numero Contatos R1	1077

Resultados Obtidos

- Com base no item 3, identificar quais grupos possuem melhores valores de FEB e RMSD:

	A	B	C
1	Rótulos de Linha	FEBMedio	RMSDMedio
2	**0**	**-5,3108**	**5,0949**
3	k_means_NADH 0	-5,3108	5,0949
4	**1**	**-4,7957**	**6,0078**
5	k_means_NADH 1	-4,7957	6,0078
6	**2**	**-5,3095**	**6,1983**
7	k_means_NADH 2	-5,3095	6,1983
8	**3**	**-5,0178**	**5,9891**
9	k_means_NADH 3	-5,0178	5,9891
10	**4**	**-4,7345**	**6,2437**
11	k_means_NADH 4	-4,7345	6,2437
12	**5**	**-4,8076**	**6,0770**
13	k_means_NADH 5	-4,8076	6,0770
14	**Total Geral**	**-5,0048**	**6,0192**

Conclusões

- Modelo apresentou eficácia na análise dos dados:
  - Explorar dimensões
  - Facilidade na análise
  - Resolução das questões de negócio
- Não foi possível explorar todas as possibilidades de análise devido a baixa quantidade de dados de docagem.
- Uso em longo prazo pode mostrar a real diferença que esta abordagem pode trazer para o processo de análise da docagem.
  - Identificar padrões de comportamento
  - Comparar histórico de experimentos
- Flexibilidade para outros cenários:
  - Permite a inclusão de novas dinâmicas moleculares;
  - Inclusão de diferentes experimentos;
  - Pode ser adaptado para experimentos de docagem.

Dúvidas?

Referências

1. Kimball, R.; Ross, M. "The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling". Wiley, 2013.

2. Machado, K. S. "Um workflow científico para a modelagem do processo de desenvolvimento de fármacos assistido por computador utilizando receptor flexível", 2007.

3. Machado, K. S. "Seleção eficiente de conformações de receptor flexível em simulações de docagem molecular", 2011.

4. Turban, E.; Potter, R. "Administração de tecnologia da informação: teoria e prática". Elsevier, 2005.

Obrigado!