CADMA-py
Documentation

Guzmán-López Eduardo Gabriel, Hernandez-Ayala Luis Felipe; Miguel Reina; Annia Galano
Universidad Autónoma Metropolitana - Facultad de Química - UNAM

June 2024

Contents

1 Contact Information

For further assistance, please contact:

- Eduardo Gabriel Guzmán-López (eggl.quimica@gmail.com)
- Theoretical and Applied Chemistry Group of Dr. Annia Galano (agal@xanum.uam.mx)
- Departamento de Química, UAM-Iztapalapa

2 How to Cite CADMA-py

When referencing this tool in your research, please cite the following articles:

- Guzman-Lopez, E.G.; Reina, M.; Perez-Gonzalez, A.; Francisco-Marquez, M.; Hernandez-Ayala, L.F.; Castañeda-Arriaga, R.; Galano, A. *Int. J. Mol. Sci.* 2022, 23, 13246. https://doi.org/10.3390/ijms232113246
- Guzmán-López, E.G.; Reina, M.; Hernández-Ayala, L.F.; Galano, A. *Antioxidants* 2023, 12, 1256. https://doi.org/10.3390/antiox12061256

3 Introduction

CADMA-Chem (Computer-Assisted Design of Multifunctional Antioxidants based on Chemical Properties) is a comprehensive computational protocol aimed at designing new multifunctional antioxidants with versatile behavior. This protocol integrates several software tools, including CADMA-py, to facilitate the drug design process. This document outlines the steps involved in using the CADMA-Chem protocol and provides detailed instructions for using CADMA-py, which is a key step in the overall protocol.

4 CADMA-Chem Protocol

The CADMA-Chem protocol aims to modify existing drugs, whether currently approved or in preclinical phases, to enhance their activity by incorporating potential chemical and biological antioxidant properties while preserving the interactions responsible for their previously reported pharmacological activity. This is achieved through the following steps:

1. **Select a disease of interest**
   A disease can be selected and treated comprehensively, but we operate under the assumption that the oral drugs used do not necessarily target the same receptors. Therefore, it is crucial to identify specific targets to compare the designed derivatives against the original drugs. Without this specific information, it is difficult to ensure that the biological activity will be maintained. Instead, we can only design bioavailable compounds with antioxidant capacity.

2. **Select a reference set of drugs**
   The reference set is built from oral drugs currently available to treat or mitigate the side effects of the disease. This protocol is not designed to create de novo compounds for treating diseases. Instead, it uses compounds with reported uses and aims to improve their properties.

3. **Determine the ADMETSA properties of the reference drugs**
   ADMETSA properties include Absorption, Distribution, Metabolism, Excretion, Toxicological properties, and Synthetic Accessibility. ADME properties ensure the bioavailability of orally administered drugs, toxicological properties ensure their safe consumption, and synthetic accessibility ensures that production costs remain reasonable.

   - **Physicochemical properties**: Assess molecular weight, logP, hydrogen bond donors/acceptors, number of non-hydrogen atoms, topological polar surface area, and molar refractivity.
   - **Toxicological properties**: Evaluate oral rat LD50, mutagenicity, and developmental toxicity.

- **Synthetic properties**: Analyze synthetic accessibility.

4. **Calculate mean and standard deviation values for the reference drugs**
   Estimating the statistics of the reference set allows us to use them as a basis for comparing our derivatives and determining which properties have improved and which may have worsened. These statistical comparisons also enable us to identify which modifications have been detrimental and which have been beneficial to the structural frameworks.

5. **Propose a lead compound**
   We must select a compound that will serve as the lead compound and whose properties we wish to modify. The primary objective is to add multifunctional properties while preserving its biological functionality.

6. **Generate derivatives with Smile-It**
   Derivatives are generated by selecting the sites to be modified on the structure and choosing the functional groups or structural frameworks to be incorporated into it. Modifications can be made at a single site or multiple sites simultaneously, a functionality provided by the SMILE-It tool, which was previously developed by our research group.

7. **Evaluate the generated derivatives**
   For each derivative, determine the toxicological and synthetic properties:
   - Oral rat LD50
   - Mutagenicity
   - Developmental toxicity
   - Synthetic accessibility

8. **Determine selection scores**
   Determine the selection scores for the proposed derivatives and statistically evaluate whether these derivatives present any significant differences that should be considered. Since these are scores, there is a possibility that excellent performance in certain properties could mask deficiencies in others. Therefore, it is necessary to consider variations in the reference sets and evaluate which properties deviate from acceptable ranges.

9. **Evaluating polygenic neuroprotection**
   (a) Use molecular docking to evaluate interactions with known enzymes related to the disease.
   (b) Compare binding energies and inhibition constants with known inhibitors.
   (c) Select candidates with strong interactions and potential polygenic neuroprotective effects.

10. **Evaluating multifunctional antioxidant behavior**
    (a) **Free radical scavenging (AOX-I)**:
    - Investigate thermochemistry and kinetics of reactions with radicals.
    - Calculate overall rate coefficients considering all reaction sites and environmental conditions.

    (b) **OIL behavior (AOX-II)**:
    - Study metal chelation and its impact on reducing metal-induced oxidation.
    - Analyze Gibbs free energies of reaction pathways.

    (c) **Repairing biological molecules (AOX-III)**:
    - Evaluate the capability to repair lipids, proteins, and DNA using models of oxidative damage.
    - Calculate Gibbs free energies and rate constants for repair mechanisms.

5 What exactly is CADMA-py?

CADMA-py is a tool for analyzing chemical compounds and predicting their feasibility as multifunctional compounds. It provides a user-friendly interface to load and process SMILES files and CSV files containing toxicity and synthetic accessibility data. CADMA-py is capable of executing the first four steps of the CADMA-Chem protocol and step eight to select the best candidates for use in the subsequent stages of the protocol.

6 Creating a reference set (RefSet) of drugs with CADMA-py

CADMA-py allows you to create a clean reference set of drugs by filtering out compounds with missing information and calculating their physicochemical properties. This ensures that only compounds with complete data are included for further analysis. Follow these steps:

1. **Prepare the necessary data**
   Gather the required information for each drug and ensure this data is saved in the following CSV files:
   - **SMILES file**: This file should have two columns: one with the names of the reference drugs and another with their corresponding SMILES strings. This ensures that the data is stored clearly as a reference set of drugs.
   - **Toxicity files**: You will need three separate files for toxicity data, all obtained using the TEST software. It is essential to maintain the headers as the program will look for them to identify the column of interest in each file:
     - **LD50 file**: Contains data on oral rat LD50 values.
     - **Developmental toxicity file**: Contains data on developmental toxicity.
     - **Mutagenicity file**: Contains data on mutagenicity.
   - **Synthetic accessibility file**: This file is generated using the AMBIT software. It is necessary to clean this file manually, removing any non-data rows at the beginning and end, ensuring that only the headers and relevant molecular information remain.
2. **Open CADMA-py and navigate to the 'CREATE A REFERENCE SET' Button.**
   Launch CADMA-py and go to the appropriate section in the graphical user interface.
3. **Enter the necessary information**
   Fill in the 'Disease Name' field with the name of the disease you are targeting.
4. **Load the required files**
   Click on the 'SMILES file' button to select and load your SMILES file.
   Click on the 'Toxicity development' button to select and load the developmental toxicity data file.
   Click on the 'LD50 file' button to select and load the LD50 data file.
   Click on the 'Mutagenicity file' button to select and load the mutagenicity data file.
   Click on the 'Synthetic accessibility' button to select and load the synthetic accessibility data file.
5. **Create the Reference Set**
   Click the 'CREATE A REFERENCE SET' button. CADMA-py will load and process the data, calculating various molecular properties for each drug and discarding any compounds with missing information.
6. **Save the results**
   Click the 'Process data -¿' button. CADMA-py will prompt you to provide a name for the output file. The cleaned data will be saved as `{name}_RefSet.csv` and the statistical summaries as `{name}_SummaryStatistics.csv`.

6.1 Output Information

After processing, CADMA-py will provide the following information for each drug in the reference set:

- Molecular Weight **(MW)**: The mass of the molecule.
- **LogP**: The partition coefficient, indicating lipophilicity.
- Molar Refractivity **(MR)**: A measure of the volume occupied by an atom or group of atoms.
- Heavy Atom Count **(AtX)**: The number of non-hydrogen atoms.
- Lipinski Hydrogen Bond Acceptors **(HBLA)**: Number of hydrogen bond acceptors.
- Lipinski Hydrogen Bond Donors **(HBLD)**: Number of hydrogen bond donors.

- Rotatable Bonds **(RB)**: Number of rotatable bonds.
- Topological Polar Surface Area **(PSA)**: The surface area of polar atoms.

Additionally, CADMA-py will generate statistical summaries (mean, standard deviation, min, max, median, variance, skewness, and kurtosis) for each property. These summaries are essential for subsequent steps in the CADMA-Chem protocol.

6.2 Graphical Interface (results_text)

The `results_text` section in the graphical interface displays the following:

- **Summary Statistics**: A statistical summary of each property, including mean, standard deviation, min, max, median, variance, skewness, and kurtosis.
- **Cleaned Data**: The data frame after removing rows with any missing values. This ensures that only complete data sets are used for further analysis.
- **Rows with at least one N/A**: A listing of rows that were excluded from the cleaned data due to missing values in one or more properties. This helps in identifying and understanding the extent of missing data in the initial dataset.

7 Calculating SELECTION SCORES and Plotting Derivatives

This section provides detailed instructions on using the `SELECTION SCORES` button in the CADMA-py interface. This feature is crucial for analyzing compound derivatives and calculating their selection scores.

7.1 Loading and Processing SMILES Files for Derivatives

1. **Enter the necessary information**
   Fill in the 'Name' and 'Acronym' fields with the appropriate values for your study. These fields are mandatory. Optionally, if you want to see the plots or include the parent/lead molecule, you will need to fill in the 'Original SMILE' field.
2. **Load the SMILES file**
   Click on the 'Smiles file' button to select and load your SMILES file containing the derivatives you wish to analyze. This file should contain a single column with the SMILES strings of the derivatives and have a .smi extension.

7.2 Processing Data from CSV Files

1. **Load the required CSV files**
   You will need to load several CSV files containing information on various properties:
   - **Developmental Toxicity:** Contains data on developmental toxicity.
   - **Mutagenicity:** Contains data on mutagenicity.
   - **Oral Rat LD50:** Contains data on oral rat LD50 values.
   - **Synthetic Accessibility:** Generated using the AMBIT software, this file should be manually cleaned to remove any non-data rows at the beginning and end.
2. **Click 'CALCULATE SCORES'**
   CADMA-py will load and process the data from these files, calculating various molecular properties for each derivative and performing statistical analyses.

7.3 Selecting a Disease

The `Select Disease` dropdown menu provides pre-set values studied by our group for each disease. These values are editable, allowing you to modify any parameter. The modified value will be used for all score and counter calculations. If you click the left button "👁✎", the database will be shown in a new window, allowing you to review the information on all the drugs being considered (for those within the applicability domain of the employed software and for which property estimation was successful).

7.4 CADMA-Chem Intervals

The `CADMA-Chem intervals` are the intervals proposed by our group. These fields are editable, allowing you to adjust the intervals as needed to suit your specific requirements.

7.5 Saving Results

- After processing the data, CADMA-py will prompt you to save the results.
- The cleaned data will be saved as {`name`}`_SS.csv`, and the statistical summaries will be saved as {`name`}`_SS_SummaryStatistics.csv`.

7.6 Graphical Interface (results_text)

The `results_text` section in the graphical interface displays the following:

- **Summary Statistics:** A statistical summary of each property, including mean, standard deviation, min, max, median, variance, skewness, and kurtosis.
- **Cleaned Data:** The data frame after removing rows with any missing values. This ensures that only complete data sets are used for further analysis.
- **Rows with at least one N/A:** A listing of rows that were excluded from the cleaned data due to missing values in one or more properties. This helps in identifying and understanding the extent of missing data in the initial dataset.

7.7 Plotting Results

- **Click 'PLOT RESULTS 📊'**
  This button opens a new window displaying plots of the selection scores.
- **Generated Plots:**
  - **S_S vs Derivative:** Plot of the selection score (S_S) against each derivative.
  - **S_ADME vs Derivative:** Plot of the ADME properties score against each derivative.
  - **S_ADMET vs Derivative:** Plot of the ADMET score against each derivative.
- **Reference Lines:**
  - The red horizontal line represents the values of the reference drugs.
  - The green line represents the values associated with the SMILE placed in the `Original SMILE` field. Note that the SMILE value must be within the set as its properties are needed for comparison.

7.8 Loading and Processing SMILES Files for Derivatives

1. **Enter the necessary information**
   Fill in the "Name" and "Acronym" fields with the appropriate values for your study. These fields are mandatory. Optionally, if you want to see the PLOTS or to include the parent/lead molecule you will need to fill the "Original SMILE" field.
2. **Load the SMILES file**
   Click on the "Smiles file" button to select and load your SMILES file containing the derivatives you wish to analyze. This file should contain a single column with the SMILES strings of the derivatives and has .smi extension.

8 Generating 3D conformers from SMILES

This section explains how to use the CONFORMERS OF SMILES feature in CADMA-py, which allows you to generate 3D conformers in SDF format. These conformers can be opened in software such as Gaussian16, DataWarrior, or similar tools for further analysis.

8.1 Using the Interface

1. **Open the "CONFORMERS GENERATED FROM SMILES" Interface**
   Launch CADMA-py and click on the 'CONFORMERS OF SMILES' button to open the interface.

2. **Input Options**
   There are two main input options:

   - **Processing a CSV file**:
     (a) **Load the SS CSV File**: Click on the 'Load CSV file' button to select and load the CSV file containing the selection scores. The file should contain the selection scores (`S_S`) for each derivative.
     (b) **Specify a Compound (Optional)**: If there is a specific compound you wish to include regardless of its position in the selection scores, enter its identifier of the column "Compound" in the provided field.
     (c) **Number of SMILES to Convert**: Enter the number of top-scoring SMILES you want to process. This limits the number of SMILES strings processed based on their selection scores.
     (d) **Process SS file**: Click on this button to start processing the CSV file, generating 3D conformers for the selected SMILES strings.
   - **Processing a List of SMILES Strings**:
     (a) **Enter SMILES Strings**: In the provided text area, input the list of SMILES strings you wish to process. Each SMILE should be on a new line.
     (b) **Run SMILES**: Click on this button to start processing the entered SMILES strings, generating 3D conformers for each.

8.2 Conformer Generation and Output

- **Generating 3D Conformations**:
  - CADMA-py uses the RDKit library to generate 3D conformers for each SMILES string. It generates and optimizes fifty conformers per method using the MMFF94 and ETKDG methods, selecting the best conformers from each method, resulting in two final conformers. This ensures that at least two sufficiently different starting structures are available for subsequent calculations. To guarantee that the two force fields do not generate the same or very similar structures, a condition is applied requiring a minimum RMSD difference of 0.1 between the best conformers, using the GetConformerRMS function from RDKit. Therefore, two structures are saved for each selected SMILE. Note that these structures do not account for solvent effects and are merely starting points for further optimizations.
- **Saving Results**:
  - After processing, CADMA-py will prompt you to save the results. The conformers are saved in an SDF file format, which is compatible with various molecular visualization tools.
  - The file is saved as `estructuras_moleculares_optimizadas.sdf`.
- **Generating 2D Representations**:
  - CADMA-py also generates 2D representations of the molecules. These representations are saved as a grid image in PNG format for easy visualization.
  - The file is saved as `structures_grid.png`.

8.3 Conditions and Constraints

- **Data Integrity**: Verify that the input files do not contain missing or corrupted data to avoid processing errors.
- **Performance Considerations**: Processing a large number of SMILES strings can be computationally intensive. Adjust the number of SMILES to process based on your system's capabilities.

By following these instructions, users can effectively utilize CADMA-py to generate molecular representations and conformers from SMILES strings, facilitating further analysis and visualization of potential drug candidates.

9 Instalación en una terminal de Linux, macOS y Windows

Para instalar y configurar todo lo necesario para usar CADMA-py en una terminal de Linux, macOS o Windows, siga estos pasos:

9.1 Instalar Conda

Primero, necesita instalar Conda, un sistema de gestión de paquetes y entornos. Recomiendo instalar Miniconda, una distribución mínima de Conda:

1. Descargue el script de instalación de Miniconda:
   - Para Linux:

     `wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh`
   - Para macOS:

     `wget https://repo.anaconda.com/miniconda/Miniconda3-latest-MacOSX-x86_64.sh`
   - Para Windows, descargue el instalador desde el siguiente enlace:

     `https://repo.anaconda.com/miniconda/Miniconda3-latest-Windows-x86_64.exe`
2. Ejecute el script de instalación:
   - En Linux:

     `bash Miniconda3-latest-Linux-x86_64.sh`
   - En macOS:

     `bash Miniconda3-latest-MacOSX-x86_64.sh`
   - En Windows, ejecute el instalador descargado y siga las instrucciones en pantalla.
3. Siga las instrucciones en pantalla. Acepte el acuerdo de licencia, elija el directorio de instalación y permita que el instalador inicialice Conda.
4. Una vez finalizada la instalación, cierre y vuelva a abrir la terminal o ejecute:
   - En Linux:

     `source ~/.bashrc`
   - En macOS, es posible que necesite ejecutar:

     `source ~/.zshrc`

     si está usando el shell zsh por defecto.
   - En Windows, abra una nueva ventana de Anaconda Prompt desde el menú de inicio.

9.2 Crear un Entorno Virtual

Después de instalar Conda, cree un nuevo entorno virtual donde instalará las dependencias necesarias para CADMA-py:

```
conda create --name cadma-py python=3.10.13

conda activate cadma-py
```

9.3 Instalar las Dependencias

Con el entorno virtual activado, instale las dependencias necesarias usando Conda y el canal conda-forge:

```
conda install -c conda-forge pandas=2.1.4 rdkit=2023.09.4 py3dmol=2.1.0 numpy matplotlib tk pillow
```

Esto instalará las siguientes bibliotecas:

- pandas (2.1.4)
- rdkit (2023.09.4)
- py3dmol (2.1.0)
- numpy
- matplotlib
- tk (para tkinter)
- pillow

9.4 Verificar la Instalación

Para asegurarse de que todas las dependencias se han instalado correctamente, ejecute:

```
conda list
```

Este comando mostrará una lista de todos los paquetes instalados en el entorno activo, junto con sus versiones y canales de origen.

9.5 Ejecutar CADMA-py

Con todas las dependencias instaladas, ya puede ejecutar CADMA-py. Navegue al directorio donde está guardado el script `CADMA-py.py` y ejecútelo:

```
python CADMA-py.py
```

Esto lanzará la interfaz gráfica de CADMA-py y permitirá comenzar a utilizar todas sus funcionalidades.

10 Acknowledgements

This work was supported by the National Council of Humanities, Science and Technology (CONAHCYT) of Mexico. We express our gratitude for the funding provided:

- "Fronteras de la ciencia 2023-2024" No. [CBF2023-2024-1141], "Diseño de fármacos multifuncionales para abordar enfermedades multifactoriales: Una aproximación computacional y experimental hacia la mejora de la terapia".
- Post-Doc Luis Felipe detalles de su beca
- Proyecto UNAM Miguel detalles (¿hay algún proyecto que agradecer aparte del de fronteras?) .

We also extend our thanks to all the researchers and institutions involved for their valuable contributions and support throughout the development of this work.

Special thanks to CONAHCYT for the doctoral scholarship 649128 awarded to Eduardo Gabriel Guzmán-López, which has been instrumental in supporting this research.