Data in Brief 55 (2024) 110677

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset for discovering new hypertension small molecules using machine learning-aided computational fragment-based design

Odifentse Mapula-e Lehasa*, Uche A.K. Chude-Okonkwo

Institute for Intelligent Systems, University of Johannesburg, 69 Kingsway Avenue, Auckland Park, Johannesburg 2092, Gauteng Province, South Africa

ARTICLE INFO

Article history:
Received 28 May 2024
Revised 17 June 2024
Accepted 19 June 2024
Available online 26 June 2024

Dataset link: Input Data (Original data)
Dataset link: Output Data (Original data)

Keywords:
Fragment-based drug discovery
High blood pressure
Renin-angiotensin-aldosterone system
Angiotensin-converting enzyme inhibitor
Angiotensin II receptor blockers
K-Means clustering

ABSTRACT

This dataset demonstrates the use of computational fragmentation-based and machine learning-aided drug discovery to generate new lead molecules for the treatment of hypertension. Specifically, the focus is on agents targeting the renin-angiotensin-aldosterone system (RAAS), commonly classified as Angiotensin-Converting Enzyme Inhibitors (ACEIs) and Angiotensin II Receptor Blockers (ARBs). The preliminary dataset was a target-specific, user-generated fragment library of 63 molecular fragments of the 26 approved ACEI and ARB molecules obtained from the ChEMBL and DrugBank molecular databases. This fragment library provided the primary input dataset to generate the new lead molecules presented in the dataset. The newly generated molecules were screened to check whether they met the criteria for oral drugs and comprised the ACEI or ARB core functional group criterion. Using unsupervised machine learning, the molecules that met the criterion were divided into clusters of drug classes based on their functional group allocation. This process led to three final output datasets, one containing the new ACEI molecules, another for the new ARB molecules, and the last for the new unassigned class molecules. This data can aid in the timely and efficient design of novel antihypertensive drugs. It can also be used in precision hypertension medicine for patients with treatment

* Corresponding author.
E-mail addresses: odifentselehasa@gmail.com (O.M.-e. Lehasa), ucheo@uj.ac.za (U.A.K. Chude-Okonkwo).
Social media: @Odi_Lehasa (O.M.-e. Lehasa)

https://doi.org/10.1016/j.dib.2024.110677
2352-3409/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

resistance, non-response or co-morbidities. Although this dataset is specific to antihypertensive agents, the model can be reused with minimal changes to produce new lead molecules for other health conditions.

© 2024 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Specifications Table

Subject	Bioinformatics.
Specific subject area	Applying machine learning and computational fragment-based drug discovery to design new hypertension small molecules.
Type of data	Tables, Images, and Figures. Raw, Analysed, Filtered.
Data collection	The preliminary data on existing FDA-approved ACE Inhibitors and ARBs was collected from ChEMBL, a chemical database of bioactive molecules, and DrugBank, a drug database. This data was extracted in SDF format, which allows the molecule's structure, atoms and bonds to be analysed. This analysis was conducted computationally on Jupyter Notebook using Python version 3.9.12 with the RDKit library, on a Conda version 23.1.0 environment.
Data source location	Institution: University of Johannesburg, Institute for Intelligent Systems City: Johannesburg Country: South Africa.
Data accessibility	The data can be accessed from the data repository in the location specified below. Repository name: Mendeley Data Data identification number: 10.17632/brgzpd5wj4.1 Direct URL to data: https://data.mendeley.com/datasets/brgzpd5wj4/1 The instructions for accessing these data are contained in the data repository, alongside the data files. These instructions also contain the steps for reproducing the data, which can be accessed from: https://doi.org/10.5281/zenodo.11636007
Related research article	*None.*

1. Value of the Data

- This data is useful for understanding the computational generation of new small molecules for the treatment of hypertension, using existing hypertension small molecules as a building block.
- The data demonstrates a pool of newly generated ACEIs and ARBs that can aid in the future development of novel hypertension drugs. These drugs can then be personalised to individual hypertensive patients and their treatment response, increasing therapeutic outcomes, and reducing side effects.
- The new lead molecules displayed in the dataset can be considered for the treatment of hypertension through the renin-angiotensin-aldosterone system (RAAS).
- Computational chemists and bioinformaticians can conduct molecular docking simulations on the new hypertension lead ACEIs and ARBs to determine their binding affinity to the target site. Subsequently, these molecules can be synthesised for further screening.
- The fragment library dataset can be used to generate new molecules that treat other health conditions that are not specified in this dataset.
- Other researchers can consider incorporating generative machine learning techniques using this dataset to produce additional candidate ACEI and ARB molecules, containing the required core functional group structure and molecular properties. This AI-enhanced approach can also be applied to other drug discovery studies.

2. Background

With over 1 billion affected adults [1], hypertension (or high blood pressure) is one of the most critical public health challenges worldwide. This condition is one of the main risk factors for developing cardiovascular, diabetes, kidney, and other related diseases [2]. Although effective antihypertensive agents exist, only 21 % of hypertensive patients have their condition under control [3], with the rest failing to adhere to their treatment prescription due to adverse effects and others suffering from treatment non-response (resistant hypertension) [4].

By augmenting the conventional drug discovery process, using computational methods such as machine learning and fragment-based drug design, new drugs can be designed in shorter timelines and at lower costs [5]. This study aims to use fragmentation-based and machine learning-aided drug discovery techniques to generate a dataset of new lead drug molecules for the treatment of hypertension. These molecules are generated using fragments of existing hypertension molecules and subsequently clustered and screened to confirm that they possess the properties of ACEI and ARB hypertension drugs.

The value of this dataset is to increase the pool of available treatment options, especially for those hypertensive individuals who may not respond well to the existing treatment.

3. Data Description

This dataset [6] consists of eight (8) data files for the input data and nine (9) files from the output dataset. These files, their locations and contents are described in Tables 1 and 2 below.

Fig. 1 contains a two-dimensional (2D) illustration of the original set of ACE Inhibitors obtained from the ChEMBL and DrugBank databases. In contrast, Fig. 2 illustrates the original set of ARB molecules used for the preliminary dataset obtained from the same databases. These original molecules were fragmented to generate the fragment library, the primary input dataset.

Table 3 shows the physicochemical properties of the original molecules (15 ACEIs and 10 ARBs) and the newly generated 492 ACEIs, 681 ARBs and 363 unassigned new lead molecules. Here, a side-by-side comparison can be drawn to ascertain how well the new molecules performed against the original molecules and the criteria for oral drugs.

4. Experimental Design, Materials and Methods

4.1. Materials

The experiments were conducted using Python 3 (version 3.9.12) on Jupyter Notebook (version 6.4.8) with the Anaconda (Conda v23.1.0) package manager and environment on a MacBook Pro with an M1 max chip. The code workbooks comprising the experiments, Conda libraries used, and their dependencies are located in the code repository [7]. Additionally, tools provided by RDKit (an open-source cheminformatics and machine learning software) were used to compute the chemistry-specific functions.

4.2. Data collection method

"The best way to find a new drug is to start with an existing one" [8]. Therefore, the objective of developing a dataset of new drugs for the treatment of hypertension requires the use of existing hypertension drugs. Publicly available (existing) hypertension drugs were sourced from ChEMBL, a chemical database of bioactive molecules, and DrugBank, a drug database, to form the initial input dataset. Considering that the focus is on antihypertensive agents acting on the renin-angiotensin-aldosterone system (RAAS), Angiotensin-Converting Enzyme Inhibitors (ACEIs)

Table 1
Raw input data.

No.	Title	Location	File Type	Description
1	Chembl_C09.sdf	Input Data	Structured Data File (.sdf)	List of FDA-approved C09 drugs (agents acting on the renin-angiotensin-aldosterone system) from the ChEMBL database.
2	Drugbank_C09.sdf	Input Data	Structured Data File (.sdf)	List of FDA-approved C09 drugs (agents acting on the renin-angiotensin-aldosterone system) from the DrugBank database.
3	ACE Inhibitors.csv	Input Data > ACE Inhibitors	Comma-separated values (.csv)	List of the 15 FDA-approved ACE Inhibitors sourced from both ChEMBL and DrugBank databases. The dataset contains the following columns: - Molecule Name - ATC Code - Class - Canonical SMILES
4	Original ACE Inhibitors (Molecules in 2D format).png	Input Data > ACE Inhibitors	Portable Network Graphic (.png)	Two-dimensional (2D) images of the 15 ACEI molecules from the ACE Inhibitors.csv file.
5	ARBs.csv	Input Data > ARBs	Comma-separated values (.csv)	List of the 10 FDA-approved ARBs sourced from both ChEMBL and DrugBank databases. The dataset contains the following columns: - Molecule Name - ATC Code - Class - Canonical SMILES
6	Original ARBs (Molecules in 2D format).png	Input Data > ARBs	Portable Network Graphic (.png)	Two-dimensional (2D) images of the 10 ARB molecules from the ARBs.csv file.
7	Fragment_library.csv	Input Data > Fragment Library	Comma-separated values (.csv)	List of 63 fragments used in generating the new C09 molecules. The dataset contains the following columns: - Fragment - Molecular Weight - LogP - Hydrogen Donors - Hydrogen Acceptors - Ro3 Pass - Frequency
8	Fragments as 2D images.png	Input Data > Fragment Library	Portable Network Graphic (.png)	Two-dimensional (2D) images of the 63 molecule fragments from the *Fragment_library.csv* file.

Table 2
Analysed and filtered output data.

No.	Title	Location	File type	Description
1	New ACEI lead molecules (with properties).xlsx	Output Data > New ACE Inhibitors	Microsoft Excel Spreadsheet (.xlsx)	492 newly generated lead ACEI molecules with their scores for the following properties: Aromatic Rings (No.); Aliphatic Rings (No.); AVG Molecular weight; Exact Molecular weight; LogP; Hdonors; Hacceptors; Rotatable bonds; Heavy Atoms (No.); QED; Property Forecast Index; PSA; SAscore.
2	New lead ACE Inhibitors (492 molecules in 2D format).png	Output Data > New ACE Inhibitors	Portable Network Graphic (.png)	Two-dimensional (2D) images of the 492 new ACEI molecules from the *New ACEI lead molecules (with properties).xlsx* file.
3	New ARB lead molecules (with properties).xlsx	Output Data > New ARBs	Microsoft Excel Spreadsheet (.xlsx)	681 newly generated lead ARB molecules with their scores for the following properties: Aromatic Rings (No.); Aliphatic Rings (No.); AVG Molecular weight; Exact Molecular weight; LogP; Hdonors; Hacceptors; Rotatable bonds; Heavy Atoms (No.); QED; Property Forecast Index; PSA; SAscore.
4	New lead ARBs (681 molecules in 2D format).png	Output Data > New ARBs	Portable Network Graphic (.png)	Two-dimensional (2D) images of the 681 new ARB molecules from the *New ARB lead molecules (with properties).xlsx* file.
5	New unassigned molecules_cluster 1 (with properties).xlsx	Output Data > New unassigned molecules	Microsoft Excel Spreadsheet (.xlsx)	363 newly generated lead molecules that were neither classified as ARB or ACEI, along with their scores for the following properties: Aromatic Rings (No.); Aliphatic Rings (No.); AVG Molecular weight; Exact Molecular weight; LogP; Hdonors; Hacceptors; Rotatable bonds; Heavy Atoms (No.); QED; Property Forecast Index; PSA; SAscore.
6	New unassigned cluster 1 molecules (2D format).png	Output Data > New unassigned molecules	Portable Network Graphic (.png)	Two-dimensional (2D) images of the 363 new unassigned molecules from the *New unassigned molecules_cluster 1 (with properties).xlsx* file.
7	Original ACEIs (with properties).xlsx	Output Data > Properties of original molecules	Microsoft Excel Spreadsheet (.xlsx)	15 original ACEI molecules with their scores for the following properties: Aromatic Rings (No.); Aliphatic Rings (No.); AVG Molecular weight; Exact Molecular weight; LogP; Hdonors; Hacceptors; Rotatable bonds; Heavy Atoms (No.); QED; Property Forecast Index; PSA; SAscore.
8	Original ARBs (with properties).xlsx	Output Data > Properties of original molecules	Microsoft Excel Spreadsheet (.xlsx)	10 original ARB molecules with their scores for the following properties: Aromatic Rings (No.); Aliphatic Rings (No.); AVG Molecular weight; Exact Molecular weight; LogP; Hdonors; Hacceptors; Rotatable bonds; Heavy Atoms (No.); QED; Property Forecast Index; PSA; SAscore.
9	Comparison of physicochemical properties for all molecules.xlsx	Output Data > Properties of new molecules	Microsoft Excel Spreadsheet (.xlsx)	A side-by-side comparison of the physicochemical properties of the new and original molecules.

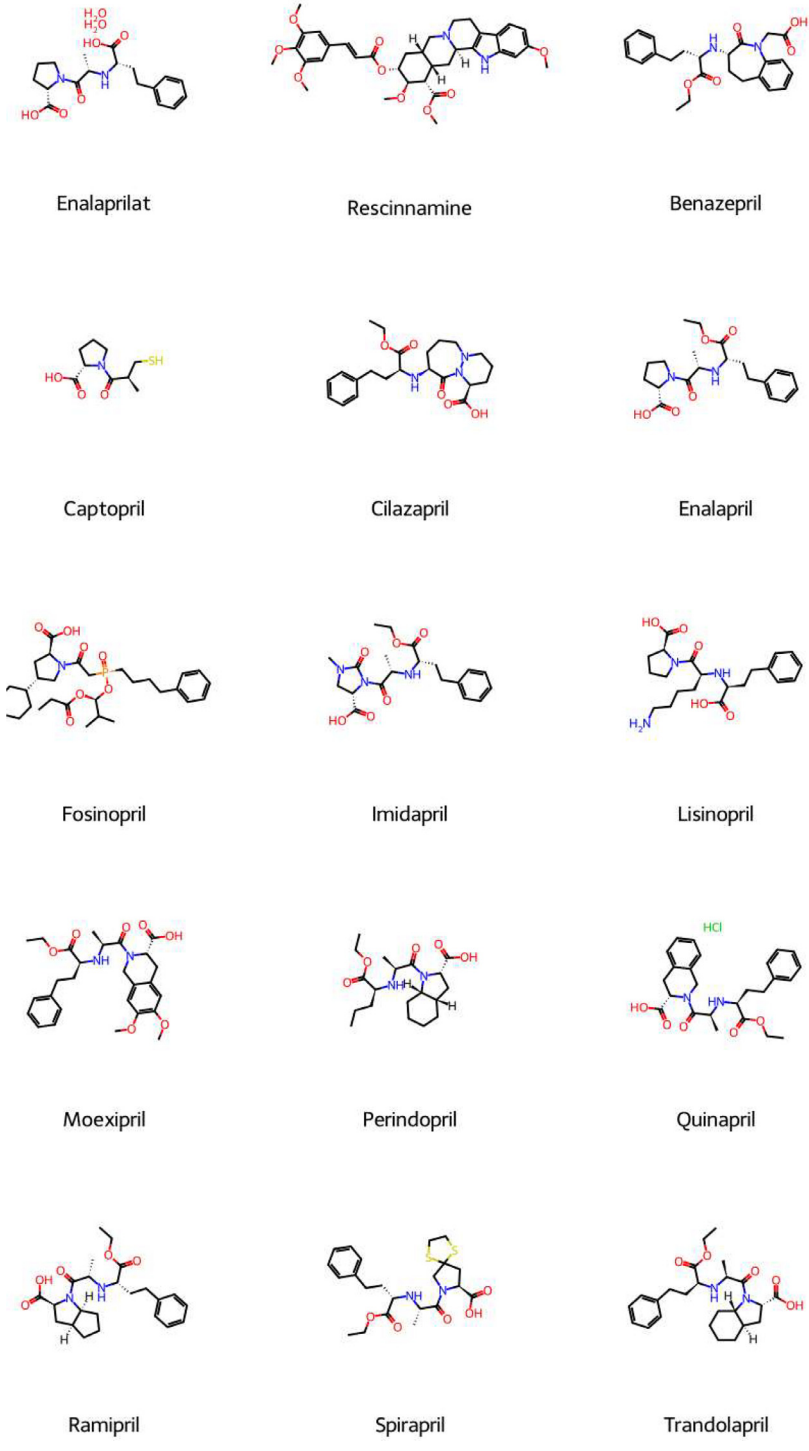
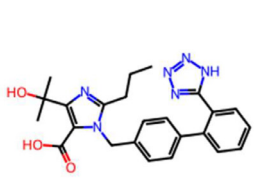
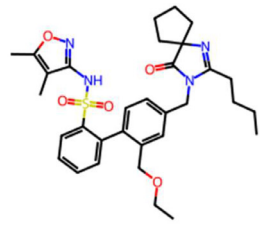


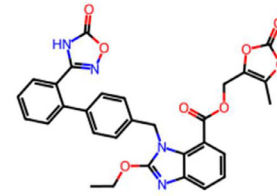
Fig. 1. The original ACEIs used to create the fragment library (reference drugs) to inform the generation of new lead ACEI molecules.



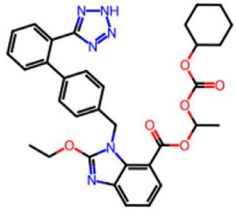
Olmesartan



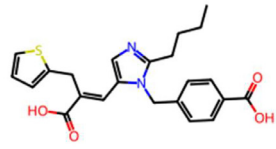
Sparsentan



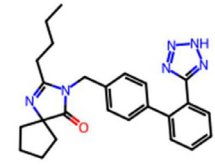
Azilsartan medoxomil



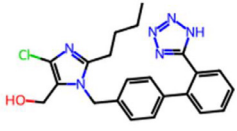
Candesartan Cilexetil



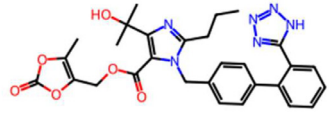
Eprosartan



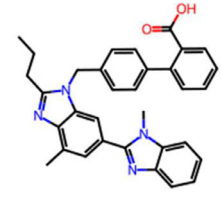
Irbesartan



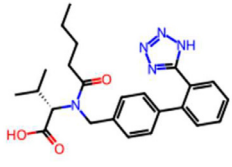
Losartan



Olmesartan Medoxomil



Telmisartan



Valsartan

Fig. 2. The original ARBs used to create the fragment library (reference drugs) to inform the generation of new lead ARB molecules.

Table 3
Comparison of physicochemical properties of the original versus new molecules.

		Lipinski's Rule of Five				Other Physicochemical Properties						Synthetic Accessibility Score (SAscore)
		Molecular Weight (g/Mol)	Lipophilicity (LogP)	Hydrogen Bond Donors (HBD)	Hydrogen Bond Acceptors (HBA)	Rotatable Bonds	Heavy Atoms	Aromatic Rings	Quantitative Estimate of Drug-Likeness (QED)	Property Forecast Index (PFI)	Topological Polar Surface Area (tPSA)	
Criteria		<=500	<=5	<=5	<=10	<=10	<38	<4	0.5–1	<7	<140	N/A
Original	**Count**	15	15	15	15	15	15	15	15	15	15	15
ACEIs	**Mean**	432.01	2.25	2.33	7.67	9.07	30.60	1.20	0.52	3.45	105.99	3.51
	Std. Dev.	93.78	1.60	1.59	1.54	2.34	6.93	0.77	0.15	2.07	24.27	0.44
	Min.	217.08	−0.52	1	4	3	14	0	0.13	0.48	57.61	3.00
	Max.	634.29	6.12	7	11	14	46	3	0.68	7.57	169.94	4.46
New Lead	**Count**	492	492	492	492	492	492	492	492	492	492	492
ACEIS	**Mean**	394.07	2.68	1.40	6.66	6.46	28.33	1.17	0.72	3.85	80.47	3.67
	Std. Dev.	60.76	1.09	0.61	1.35	1.37	4.67	0.85	0.09	1.67	18.43	0.44
	Min.	143.09	−0.36	0	3	2	10	0	0.35	-0.18	40.54	2.44
	Max.	495.22	4.99	4	10	10	35	3	0.92	6.99	136.81	4.64
Original	**Count**	10	10	10	10	10	10	10	10	10	10	10
ARBs	**Mean**	499.91	5.06	1.6	8.8	9.1	36.6	4.2	0.35	9.26	116.21	3.03
	Std. Dev.	76.71	1.21	0.70	2.44	1.60	5.74	1.23	0.14	1.97	30.64	0.37
	Min.	422.16	3.66	1	6	7	30	3	0.14	7.16	72.94	2.52
	Max.	610.25	7.26	3	12	12	45	6	0.59	13.26	162.16	3.62
New Lead	**Count**	681	681	681	681	681	681	681	681	681	681	681
ARBs	**Mean**	335.11	1.94	1.40	7.82	4.89	24.11	2.09	0.81	4.02	91.20	3.73
	Std. Dev.	53.40	1.11	0.54	1.45	1.28	3.95	0.69	0.09	1.49	19.85	0.48
	Min.	197.13	−0.58	0	4	3	14	1	0.54	0.42	41.29	2.24
	Max.	481.28	4.72	4	10	10	35	3	0.94	6.94	138.35	4.77
New	**Count**	363	363	363	363	363	363	363	363	363	363	363
Unassigned	**Mean**	378.84	2.92	0.81	5.29	4.90	24.51	1.05	0.78	3.97	55.61	4.66
Molecules	**Std. Dev.**	42.17	1.15	0.80	1.95	1.20	3.11	0.78	0.07	1.39	28.02	0.31
(Cluster 1)	**Min.**	260.14	−0.05	0	2	3	16	0	0.56	1.24	6.48	3.79
	Max.	497.18	4.99	4	10	9	34	3	0.91	6.99	124.26	5.35

and Angiotensin II Receptor Blockers (ARBs) were the drug classes extracted from the ChEMBL and DrugBank databases.

The drugs of the classes ACEIs and ARBs (or agents acting on the renin-angiotensin-aldosterone system) were identified using the ATC classification code C09. These drugs were identified by filtering the 'Level 1 ATC Code Description' to 'C – cardiovascular system'. We then filtered the 'Level 2 ATC Code Description' to 'C09 – agents acting on the renin-angiotensin system'. This search yielded a list of 29 small molecules from ChEMBL and 47 from DrugBank, a total of 76 C09 drugs.

However, after removing the duplicates, experimental and investigational drugs from the list, there were 26 FDA-approved C09 drugs. The drug data of these 26 drugs was extracted from these databases in SDF format in order to retrieve each molecule's chemical structure and details of its atoms, bonds and connectivity [9].

As a result, the dataset ($\mathcal{D}$) used for the initial preliminary work (input data) was a list of all 26 FDA-approved C09 drugs from the chemical databases, where 15 were ACEIs, 10 were ARBs, and 1 was labelled as 'other'. The one drug that was neither ACEI nor ARB was removed from the list, to focus only on the ACEI and ARB drug classes. Ultimately, our input dataset comprised 25 C09 small molecules, of which 10 were ARBs and 15 were ACEIs. This dataset can be represented as

$$\mathcal{D} = \{d_1, d_2, d_2, \ \ldots, d_{25}\}$$

This preliminary dataset $\mathcal{D}$ was used as the input into the fragmentation and generation process to develop the new C09 small molecule output dataset.

4.3. Data processing method

The proposed framework illustrated in Fig. 3 below outlines the steps to transform the original C09 drug data into the fragment library of fragment molecules (input dataset) that were used to generate the new set of C09 lead molecules (final output dataset). This framework entails designing new C09 molecules using computational fragmentation-based drug design aided by machine learning.

The pool of 26 original C09 molecules were used as the preliminary data to inform the discovery of new C09 lead molecules. This preliminary data was fragmented using the RDKit BRICS module. A set of 64 unique fragments were produced from this process. These fragments were each tested to determine whether they met the fragment Rule of Three (Ro3) criteria. This criterion states that fragments should have a molecular weight lower than or equal to 300 Da, a LogP, hydrogen bond donor, and hydrogen bond acceptor scores of three (3) or less. Following the application of these criteria, 1 fragment was eliminated as it did not pass this criterion. The remaining 63 fragments and their Ro3 scores formed the fragment library.

Using the RDKit BRICS module, a sample of 10,000 from a possible 39,711 new molecules were generated from the fragment library, forming the new molecule dataset. The properties of the new molecules were screened to determine whether the new molecules met the Lipinski Rule of Five (Ro5) and physicochemical property criterion for oral drugs.

The criterion recommends that molecules meet the following rules:

1. Lipinski Ro5 [10]:
   a. Molecular Weight <= 500
   b. LogP <= 5
   c. Hydrogen Bond Donors <= 5
   d. Hydrogen Bond Acceptors <= 10
2. Physicochemical properties:
   a. Number of Heavy Atoms < 38 [11]
   b. Number of Rotatable Bonds <= 10 [12]
   c. Number of Aromatic Rings < 4 [13]

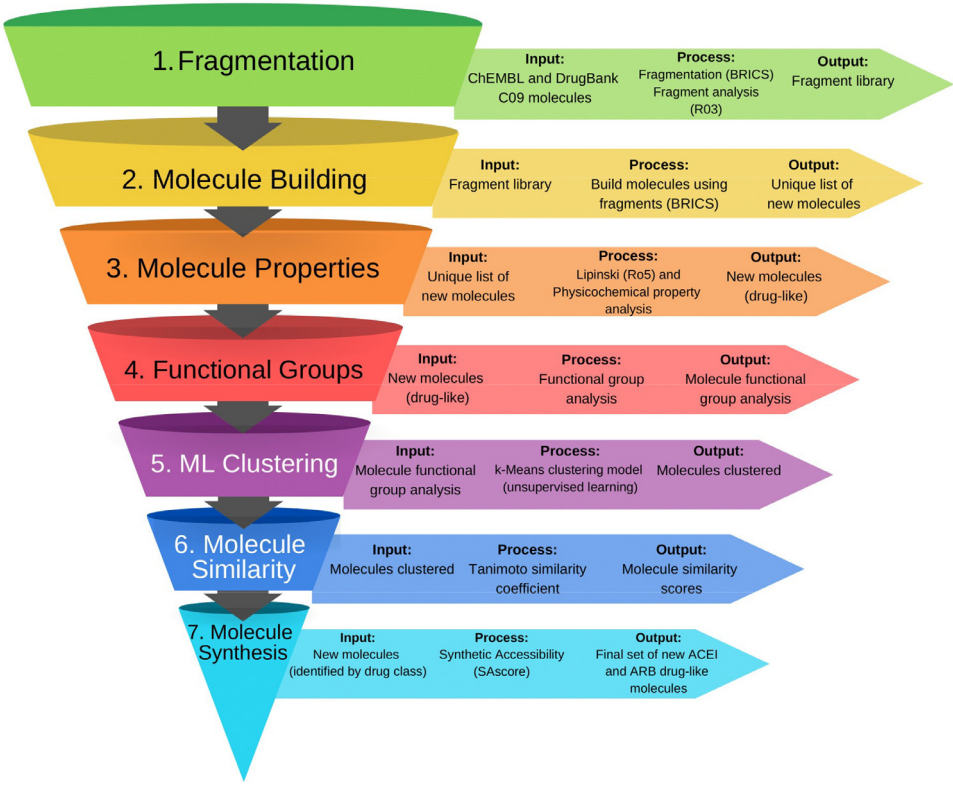


Fig. 3. Proposed framework of the steps followed to produce the datasets of the new lead C09 molecules.

d. PSA (Polar Surface Area) < 140 [12]
e. QED (Quantitative Estimate of Drug-Likeness) 0.5 – 1 [14]
f. PFI (Property Forecast Index) < 7 [15].

The scores for each of these metrics were calculated using the RDKit.Chem Descriptors, rdMolDescriptors, Lipinski and QED modules.

They were then screened to check whether they contained the core functional groups required for ACEIs (carboxyl, sulfhydryl, and phosphinyl [16,17]) and ARBs (tetrazole, biphenyl, benzimidazole [18,19]). The new molecules that did not meet the Ro5, physicochemical properties, and core functional groups criterion were eliminated from the dataset.

Using the k-means unsupervised machine learning algorithm, the remaining new molecules in the dataset were grouped into three (3) clusters, the *k* value recommended by the elbow and silhouette methods. The k-means clustering, and molecular similarity screening steps are important for determining which of the new molecules are in the ACEI or ARBs drug classes. The allocation of new molecules into each cluster was based on the functional groups present in each molecule. In order to ascertain which cluster contained ACEIs or ARBs, molecular similarity tests were conducted. The Tanimoto molecular similarity metric was used to determine the similarity scores of the new molecules against the original ACEI molecules and the original ARB molecules. The new and original molecules were translated to MACCS Keys Fingerprints to apply the FingerprintSimilarity (Tanimoto equivalent) function from the RDKit DataStructs library.

The similarity scores demonstrate which clusters of new molecules are more ACE-like or ARB-like based on their similarity scores to the original ACEI and ARB molecules. Cluster 1 obtained the lowest similarity score with the original ACEI and ARB datasets. Therefore, this cluster

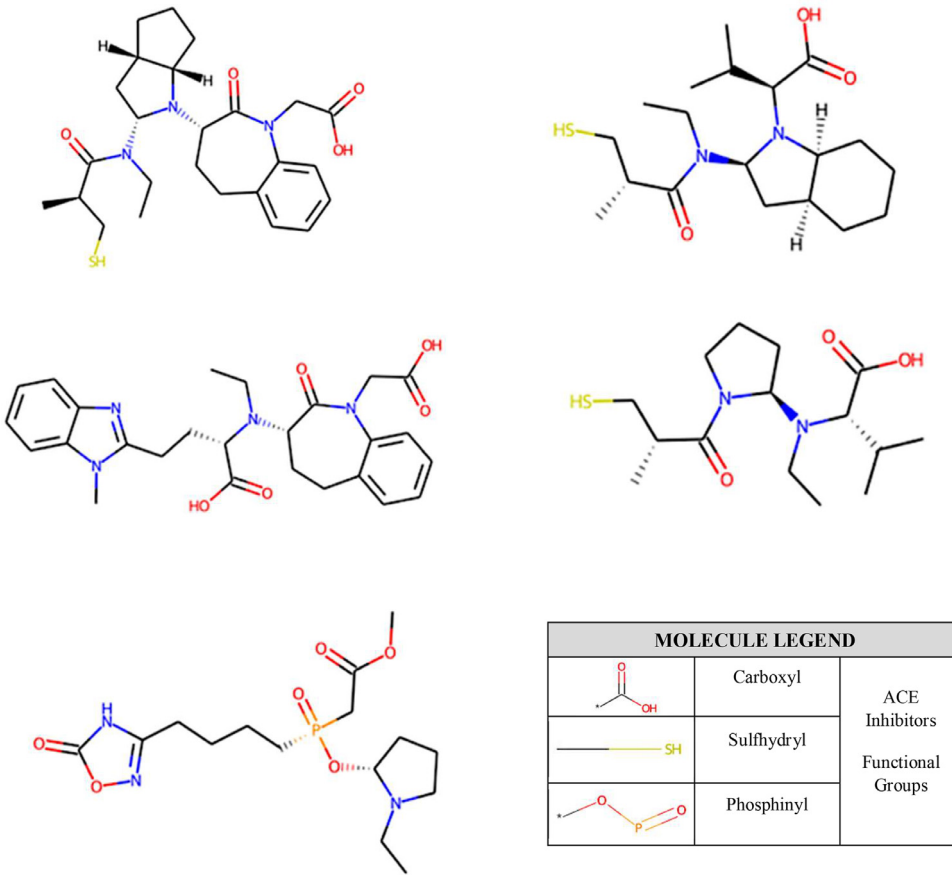
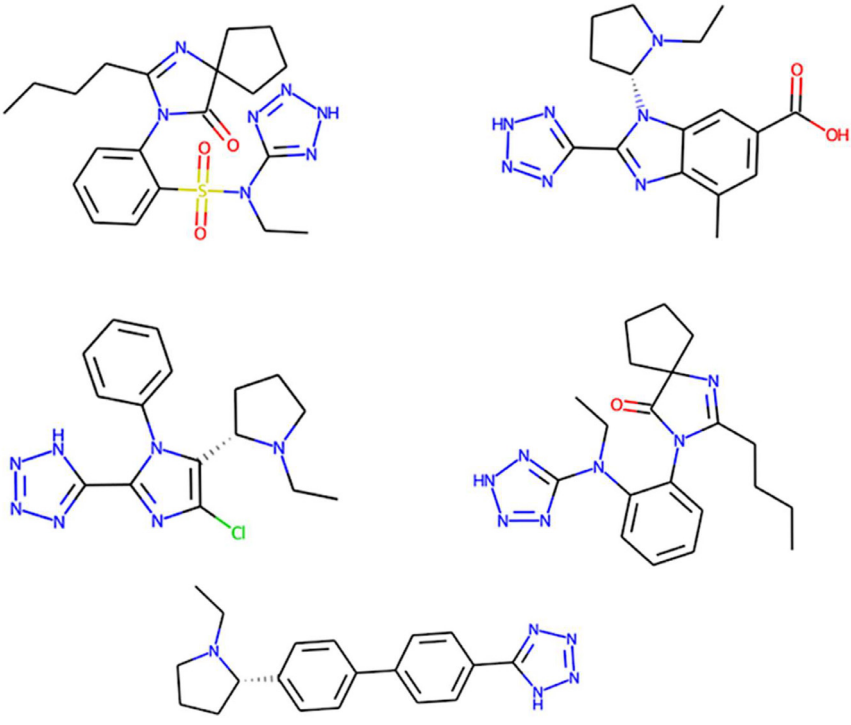


Fig. 4. Sample of newly generated ACEI lead molecules.

and its molecules were not labelled ACEI or ARB, and instead remained uncategorised. Only the molecules with a similarity score greater than or equal to 0.7 (out of 1) were considered for the new datasets of ACEI and ARB lead molecules [20], while the rest were eliminated.

Finally, the datasets of new lead ACEI and ARB molecules were assessed for their synthetic accessibility score (SAscore). This metric, which determines the ease of synthesizing a molecule, was used to compare the original and new lead molecules, as indicated in Table 3. The final datasets of the newly generated ACEI and ARB lead molecules contain 492 new ACEIs and 681 new ARBs. A sample of the new lead ACEI, ARB and uncategorised lead molecules is provided in Figs. 4, 5, and 6. Below, along with a legend indicating the core functional group structures of each drug class, to confirm that the new lead molecules are indeed ACEIs and ARBs.

The comprehensive list of final datasets of the new C09 molecules are listed in Table 2 and can be accessed through the data repository, together with the step-by-step code used to generate this data. A Jupyter Notebook is provided for each step of the framework, indicating the input data used, the process required to transform that data, and the output data generated from that step. These input and output datasets can also be accessed from the data repository, under the folders “Input Data” and “Output Data”.



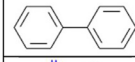
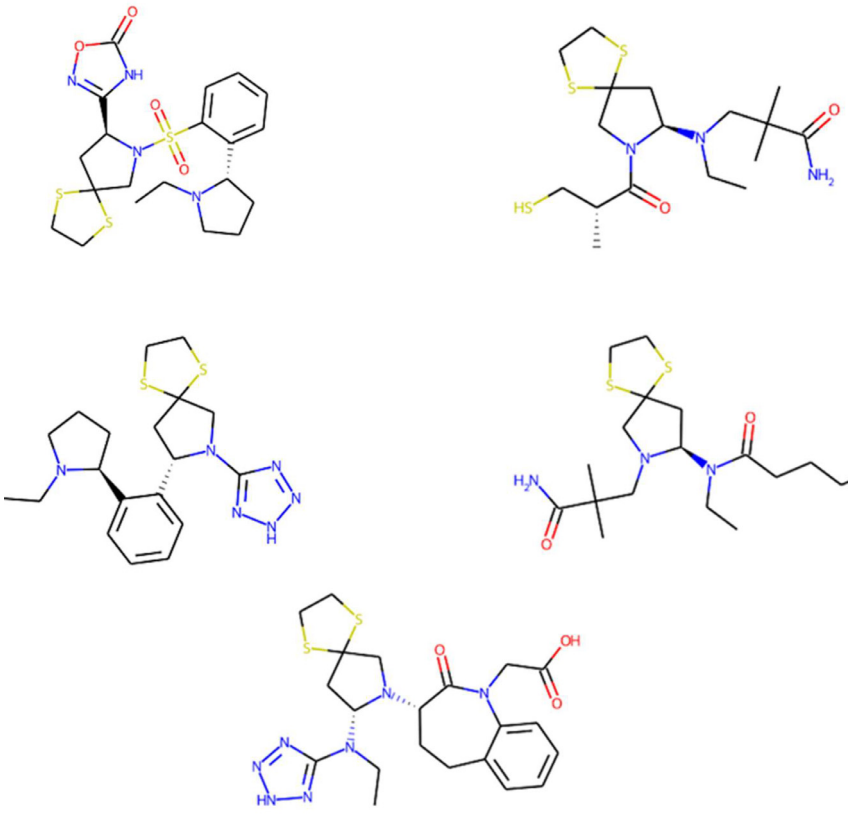
MOLECULE LEGEND		
	Biphenyl	ARBs Functional Groups
	Benzimidazole	
	Tetrazole	

Fig. 5. Sample of newly generated ARB lead molecules.



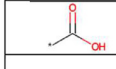
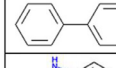
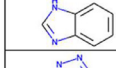
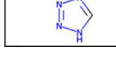
MOLECULE LEGEND		
	Carboxyl	ACE Inhibitors Functional Groups
	Sulfhydryl	
	Phosphinyl	
	Biphenyl	ARBs Functional Groups
	Benzimidazole	
	Tetrazole	

Fig. 6. Sample of newly generated unassigned new lead molecules.

Limitations

Although the preliminary dataset of original C09 molecules was relatively small, thousands of new molecules could be generated using the framework proposed in the previous section. However, the limitation of the final output datasets is that they contain molecules that are specif-

ically targeting the RAAS to treat hypertension. These new molecules were not examined for their ability to treat hypertension through other therapeutic target sites.

The dataset provided can be further enhanced by conducting molecular docking screening to determine those new lead molecules with high binding affinity to their respective target sites, and subsequently compare this data against that of the original molecules.

Ethics Statement

The authors have read and followed the ethical requirements for publication in Data in Brief and confirm that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

CRediT Author Statement

Odifentse Mapula-e Lehasa: Conceptualisation, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Funding acquisition.

Uche A.K. Chude-Okonkwo: Conceptualisation, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Review & Editing, Supervision, Funding acquisition.

Data Availability

Input Data (Original data) (Mendeley Data).
Output Data (Original data) (Mendeley Data).

Acknowledgments

The University of Johannesburg funded this work under the Global Excellence Stature (GES 4.0) grant (grant number: 201212831).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] B.C. Rossier, M. Bochud, O. Devuyst, The hypertension pandemic: an evolutionary perspective, Physiology 32 (2) (2017) 112–125, doi:10.1152/physiol.00026.2016.
[2] A. Mannan, et al., Association between comorbidity and health-related quality of life in a hypertensive population: a hospital-based study in Bangladesh, BMC Public Health 22 (1) (2022) 1–12, doi:10.1186/s12889-022-12562-w.
[3] World Health Organization (WHO), "Hypertension" [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension.
[4] R.M. Carey, et al., Resistant hypertension: detection, evaluation, and management a scientific statement from the American heart association, Hypertension 72 (5) (2018) E53–E90, doi:10.1161/HYP.0000000000000084.
[5] D.H. Freedman, Hunting for new drugs with AI, Nature 576 (7787) (2019) S49–S53, doi:10.1038/d41586-019-03846-0.
[6] O. Lehasa, U.A.K. Chude-Okonkwo, Data for machine learning-aided computational fragment-based design of small molecules for hypertension treatment, Mendeley Data 1 (2024), doi:10.17632/brgzpd5wj4.1.
[7] O.M. Lehasa and U.A.K. Chude-Okonkwo, "Computational-FBDD-for-hypertension: version 2." Zenodo, 2024. 10.5281/zenodo.11636007.

[8] W. Jahnke, D.A. Erlanson, Fragment-based Approaches in Drug Discovery, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2006.
[9] USA Food and Drug Administration (FDA), "Quick guide to creating a structure-data file (SD File) for DMF submissions disclaimer." [Online]. Available: https://www.fda.gov/media/151718/download#:~:text=sdf"%2C"SDFile", bonds%2Cconnectivity%2Candcoordinates.
[10] C.A. Lipinski, Lead- and drug-like compounds: the rule-of-five revolution, Drug Discov. Today Technol. 1 (4) (2004) 337–341, doi:10.1016/j.ddtec.2004.11.007.
[11] A.L. Hopkins, G.M. Keserü, P.D. Leeson, D.C. Rees, C.H. Reynolds, The role of ligand efficiency metrics in drug discovery, Nat. Rev. Drug Discov. 13 (2) (2014) 105–121, doi:10.1038/nrd4163.
[12] S. Kralj, M. Jukič, U. Bren, Molecular filters in medicinal chemistry, Encyclopedia 3 (2) (2023) 501–511, doi:10.3390/encyclopedia3020035.
[13] T.J. Ritchie, S.J.F. MacDonald, R.J. Young, S.D. Pickett, The impact of aromatic ring count on compound developability: further insights by examining carbo- and hetero-aromatic and -aliphatic ring types, Drug Discov. Today 16 (3–4) (2011) 164–171, doi:10.1016/j.drudis.2010.11.014.
[14] G.R. Bickerton, G.V. Paolini, J. Besnard, S. Muresan, A.L. Hopkins, Quantifying the chemical beauty of drugs, Nat. Chem. 4 (2) (2012) 90–98, doi:10.1038/nchem.1243.
[15] S.B. Bunally, C.N. Luscombe, R.J. Young, Using physicochemical measurements to influence better compound design, SLAS Discov. 24 (8) (2019) 791–801 Society for Laboratory Automation and Screening, doi:10.1177/2472555219859845.
[16] W. Zheng, et al., Small molecule angiotensin converting enzyme inhibitors: a medicinal chemistry perspective, Front. Pharmacol. 13 (2022), doi:10.3389/fphar.2022.968104.
[17] R.W. Piepho, Overview of the angiotensin-converting-enzyme inhibitors, Am. Soc. Health-Syst. Pharm. 57 (2000) 3–7 1079-2082/00/1001-00S3$06.00.
[18] T. Takezako, H. Unal, S.S. Karnik, K. Node, The non-biphenyl-tetrazole angiotensin AT 1 receptor antagonist eprosartan is a unique and robust inverse agonist of the active state of the AT 1 receptor, Br. J. Pharmacol. 175 (12) (2018) 2454–2469, doi:10.1111/bph.14213.
[19] M. Fujino, et al., A small difference in the molecular structure of angiotensin II receptor blockers induces at 1 receptor-dependent and-independent beneficial effects, Hypertens. Res. 33 (10) (2010) 1044–1052, doi:10.1038/hr.2010.135.
[20] E. Gandini, G. Marcou, F. Bonachera, A. Varnek, S. Pieraccini, M. Sironi, Molecular similarity perception based on machine-learning models, Int. J. Mol. Sci. 23 (11) (2022), doi:10.3390/ijms23116114.