Structure-based Assessment of Homologous Analogues of NATural products -*SAHANA*: A novel approach to screen bioactive compounds in comparison with their synthetic drug counterparts

Aditya Rao (✉ adityaraosj@gmail.com)

CSIR-CFTRI: Central Food Technological Research Institute CSIR <https://orcid.org/0000-0002-2546-5109>

Nandini Shetty

CSIR-CFTRI: Central Food Technological Research Institute CSIR

Research Article

Keywords: Drug design, Synthetic drugs, Pharmacokinetics, Structural scaffolds, Activity cliff, Structural fingerprint

Posted Date: June 14th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-507920/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background The present study describes a novel strategy to screen natural products (NPs) for their therapeutic effects with predicted most-probable mode of action. The method entitled '*Structure-based Assessment of Homologous Analogues of NATural products-SAHANA*' follows a strategy to compare natural products against synthetic drugs based on their molecular properties and 2-dimensional structural similarities. The method is based on a well-established hypothesis that, the molecules with similar structural properties will have similar functions.

Methods The method was validated by implementing it for the screening of NPs against different disease conditions. The biological effect of the identified NPs were evaluated based on their binding affinity with the target of their synthetic drug counterpart, prior to which, their *in-silico* pharmacokinetic and pharmacodynamics properties were assessed. The stability of binding was validated using molecular dynamics simulation studies.

Results The study yielded NPs with significant structural similarities to synthetic drugs and binding interactions stabilizing the macromolecular structures.

Conclusions The results envisage a strong indication that the natural product drugs can work in a manner similar to that of synthetic drugs if administered individually. Further, the results encourage the application of the current strategy to screen competent drugs against any disease condition ad libitum.

Background

Searching and identification of chemical compounds is considered to be one of the important aspect of modern drug design. It is based on the assumption that, structurally similar molecules frequently exhibit similar functions. This assumption has fuelled many drug-designing programs to develop many 2-dimensional and 3-dimensional similarity approaches, which are found to be highly efficient [1]. For any screening strategy to be effective, it should involve a close coupling of search algorithms and database implementation [2]. In this avenue, several strategies have been proposed, giving emphasis on core-structure and sub-structure searching across the databases. This included searching compounds by converting structure information to unique string sequence, like WLN (Wiswesser Line Notation), ROSDAL (Representation of Organic Structures Description Arranged Linearly) or SMILES (Simplified Molecular Input Line Entry System) [3] and perform a simple string search, to using advanced computational methods to develop new search algorithms by implementing modern data structure modules [4–6]. However, these search algorithms give preference to structure of the compounds and not their function; it becomes necessary to filter them facilitating the prediction of probable biological effect.

Molecular similarities are considered to optimize the potency and druglikeness of lead compounds to derive structure-activity relationships [1]. For an orally administered drug, its effectiveness depends on the pharmacokinetic (PK) and pharmacodynamic (PD) properties, which in turn decides its absorption, distribution, metabolism, and excretion (ADME) and its bioavailability. The PK/PD properties of a drug

candidate can be predicted using mathematical models. These models describe the association of a drug with the target and the possible mechanism the drug exerts to evoke a physiological response. Based on these characters the drug's affinity and efficacy are predicted.

The therapeutic effect of a drug is generally associated with a target molecule, playing a critical role in the disease mechanism [7]. Any malfunction in these links results in disease. Validating such links and targeting them to interfere with biological processes has resulted in controlling or managing several disease conditions. The underlying mechanism of action of such target molecules would be the inhibition by a small molecules (reversible or irreversible), as an agonist or antagonist for a receptor molecule, intercalator (binder), modifier (alkylating agent) or substrate mimicking for a DNA molecule, membrane blocker or opener for cell membrane ion channels or uptake inhibitors for a transporter molecule.

Structure resemblance as a basis to ascertain function:

The biological activity of any molecule is attributed by its structural arrangements. If two molecules have similar structure then they most probably will have similar biological effect [1,8,9]. Though there are some converse effects being reported [10,11], the computational chemists are successful in exploiting this principle for the construction of diverse compound libraries and selection of compounds for high-throughput screening experiments [8]. A comparison describing the structural similarities between thymidine, a naturally occurring nucleotide base and zidovudine, a synthetic drug used to treat HIV patients is given in figure 1. Structurally both the molecules share >93% identity. Such similarities has driven the synthesis of many anti-HIV drugs, which share maximum identity with the nucleotide bases [12].

Unlike synthetic compounds, which usually covers a small range of the chemical diversities, natural compounds exhibit more versatile functionality, even with their structural complexity. As they are synthesized from specific enzyme-catalysed reactions, the asymmetric factors play critical role in imparting their pharmacological effect. Hence, by considering the molecular fingerprint of their pharmacophore it is possible to predict their biological behaviour by comparing with structurally similar synthetic drug [9,13]. In this avenue, the present study proceeds to identify the therapeutic effects of natural products against specific disease conditions by comparing against synthetic small molecule drugs based on molecular properties and 2-dimensional structural similarities. As the chemically synthesised drugs exert their biological effects precisely and their action mechanism is well defined, their structurally similar natural products are also expected to exert similar effect. The selected molecules were subjected to different *in-silico* studies to validate the proposed methodology.

Overview of SAHANA:

The present protocol details an approach to screen natural products by comparing structural similarities with synthetic drug counterparts giving emphasis to their therapeutic potentials. The method follows the comparison of NPs with synthetic drugs with known mechanism of action and proposes the action mechanism for natural products, even if their bioactivity is not known. The comparison was made based

on the molecular composition and two-dimensional structural similarities between the natural products and the synthetic drugs. As results demonstrated a significant level of similarities in terms of structure, the study was continued to explore the possibility that identified NPs to demonstrate similar therapeutic effect. The selected molecules showed remarkable binding interactions with the targets of their synthetic drug counterparts. The method was applied to screen investigational drugs currently being used to treat Type 2 diabetes (T2D) and COVID-19 patients. Our findings encourages the application of this strategy for any other disease condition ad libitum. The method can also be used to define new role of natural products with therapeutic benefits reducing the time and cost of lead discovery and validation. However, emphasis has to be given for experimental validations to ascertain effects on the action mechanisms proposed based on this strategy. An overview of methodology followed in *SAHANA* is presented in figure 2.

Method

Dataset collection:

The proposed method-*SAHANA* was implemented on two different disease conditions viz. Type-2 Diabetes (T2D) and COVID-19 infection. For the study, the information on prescribed drugs for T2D condition were collected from Drugs.com (<https://www.drugs.com/condition/diabetes-mellitus-type-ii.html>) and their respective structures were collected from PubChem database (<https://pubchem.ncbi.nlm.nih.gov>). The structures of natural products were collected from NPASS database [14] (<http://bidd.group/NPASS/index.php>), COVID19 experimental drugs from PubChem database [15] (<https://pubchem.ncbi.nlm.nih.gov/#query=covid-19>), COVID19 drug-leads from ClinicalTrials.gov database [16] (<https://www.clinicaltrials.gov/ct2/home>).

Natural products and synthetic drug library construction:

Natural product activity and species source database (NPASS) offers a large library of natural products with experimental-derived quantitative activity data [14]. For T2D drug comparison, based on the literature survey, separate libraries consisting of 166 natural product structures and 84 synthetic drugs being prescribed for T2D condition were constructed. The study was extended to a larger data set by constructing a library of 26,311 structures of natural products derived from their SMILES data (Simplified molecular-input line-entry system) collected from NPASS database. The natural product library was compared by further categorizing as flavans (339), flavones (193) and isoflavonoids (457) and rest of the molecules as a general group. For comparison against COVID 19 conditions the small molecule synthetic drugs were categorised into molecules present in Pubchem Covid19 portal (306) and molecules present at different stages of clinical trials (138) (As on 31st August 2020). Further, the study was also extended for the comparison between most promising investigational drugs, which included Remdesivir, Arbidol, Lopinavir and Ritonavir molecules from which, top 10 results were selected for *in-silico* PK/PD analysis and HTVS studies. To support our study, both T2D prescribed drug library and COVID 19 drug libraries were compared internally to justify the similar biological activity among structurally similar molecules. All

the libraries were constructed using Osiris DataWarrior software in a local machine equipped with AMD Ryzen 5 six-core 3.4 GHz processor, 8GB graphics and 16 GB RAM with Microsoft Windows 10 and Ubuntu 16.04 LTS dual boot operating systems.

Implementation of the method:

Pharmacophore based comparison of natural products with their synthetic drugs counterparts:

The non-redundant natural product libraries were compared against their structurally similar small molecule synthetic drugs currently being prescribed for T2D and being prescribed/studied to treat COVID 19 infection. The molecular properties were used to assign the druglikeness of each molecule by considering their molecular weight, cLogP, hydrogen atom donors, hydrogen atom acceptors and rotatable hydrogen bonds. Based on these molecular properties, activity cliff of each structure was derived. To find structurally similar compounds rather than compounds sharing a specific sub-structure, core fragment based SAR analysis was performed by considering the most central ring structure. The similarities between two fragments were assessed based on the number of fragments that both molecules have in common divided by the number of fragments being found in any of the two structures [17]. The structures were further analysed for structural scaffolds based on plane ring system to determine the substructures and to define the similarity cut-off during the structure comparison. To derive molecular properties, activity cliff, core fragments and structural scaffolds, Osiris DataWarrior V.4.4.3 software [17] was used. For the comparison, a cut-off of more than 60% structural similarities was set to screen natural product library against synthetic drug library based on 2D-pharmacophore core fragment analysis. The graphical representation of the comparison result is given in figure 3.

Similarity score cut-off limit:

Natural products exist in many of their stable analogues forms in nature. The analogous form of natural products even with their minor structural change are capable of exerting unique biological effects. Due to their complex structure at higher cut-off limits such variations will get excluded. Following table details the selection criteria for the inclusion of analogous structures of a parent molecule- Procyanidin. As the cut-off limits decreased, different structures were included which are the analogous forms of the native structure. Based on this a similarity cut-off of 60% was fixed for the comparison.

Method validation:

State 1: *In-silico* pharmacokinetics, pharmacodynamic and toxicology profiling:

The molecular complexity is one of the major characteristics of natural products. The stereo-specificity, chirality, and cis/ trans-configuration contributing to the structural properties of the natural product are very specific. The presence of fewer aromatic rings, larger macrocyclic aliphatic rings, lower nitrogen content, and increased oxygen content contribute majorly to this structural complexity [18]. The oral bioavailability of natural products can be predicted by considering such structural properties. In the pharmaceutical research, Rule of five or Lipinski rule-of-five (RO5) has become one of the widely used

computational approaches for the estimation of solubility and permeability of new drug candidates [19]. Although, these parameters cannot be strictly applied to filter natural products, they can be used as a rule of thumb to describe the molecular properties necessary to filter candidate drug's pharmacokinetic (PK) and pharmacodynamics (PD) [20,21]. And hence, the natural product obtained from structural comparison can be considered as hits for PK/PD assessment *in-silico*.

Oral bioavailability assessment was done using Osiris Data warrior V.4.4.3 to analyse pharmacodynamic properties like mutagenicity, tumorigenicity, reproductive effective, irritant properties, drug likeliness along with basic chemical properties like total molecular weight, ClogP, H-acceptors, H-donors, rotatable bonds as for RO5 filters. Bioactivity scores were predicted using molinspiration [20], and admetSAR [22] server was used to predict the pharmacokinetic properties like solubility, permeability, drug-likeness, GPCR ligand, ion channel modulator, protease inhibitor, kinase inhibitor, enzyme inhibitor, nuclear receptor inhibitor, aqueous solubility, TPSA (Topological polar surface area), blood-brain barrier penetration, human intestinal absorption, Caco-2 permeability, AMES toxicity, carcinogenicity and acute oral toxicity by submitting each molecule individually [21,23]. The PK/PD properties of the natural products were predicted using Osiris DataWarrior V.4.4.3 [17].

Stage 2: High throughput virtual screening to predict Molecular interactions:

The selected molecules were subjected for high throughput virtual screening (HTVS) using automated docking to identify their binding affinity with the selected targets [24,25]. The information on action mechanism of synthetic drugs was collected from Inxight: Drugs database of NIH- National Center for Advancing Translational Sciences (NCATS) server (<https://drugs.ncats.io/>). We used automated docking as a preliminary screening method to predict the probable interactions of selected natural products with their corresponding targets proteins. The targets were selected by considering the action mechanism of the most structurally similar synthetic drug (Table 2A and 3A). The Protein structure files were obtained from protein data bank and were edited by removing hetero atoms and added with Kollmann charges were added as a part of target preparation. For Empagliflozin and Luseogliflozin, and their structurally similar natural products SGLT2 inhibitor with PDB ID 2XQ2 was used for interaction studies. Similarly, for Prednisolone, Glucocorticoid receptor agonist with PDB ID 1M2Z was used. To study the binding effectiveness of natural products against COVID19 infection, Lopinavir and its related natural products were docked against antiretroviral protease inhibitor (I50V isolate) (PDB ID 3OXV), Ritonavir and its related natural products were docked against antiretroviral protease inhibitor (A02 isolate) (PDB ID 4NJV), Remdesivir and its related natural products were docked against antiretroviral protease inhibitor (PDB ID 7BV2) and Arbidol and its related natural products were docked against antiretroviral protease inhibitor (PDB ID 5T6S). For each target, residues forming the binding site were identified using PDBsum server. Grid box was set around the residues forming binding pocket and Broyden-Fletcher-Goldfarb-Shanno algorithm implemented in the AutoDockVina was employed to study appropriate binding modes of the ligand in different conformations [26]. For the ligand molecules, all the torsions were allowed to rotate during docking. The *in-silico* studies were performed on a local machine equipped with AMD Ryzen 5 six-

core 3.4 GHz processor, 8GB graphics and 16 GB RAM with Microsoft Windows 10 and Ubuntu 16.04 LTS dual boot operating systems.

Stage 3: United-atom molecular dynamic simulation studies to predict the protein stability:

United-atom molecular dynamics (MD) simulation studies are used to analyse the physical movement of atoms and molecules. To assess the structural stability upon ligand binding [27,28] MD simulations were run for a time scale of 20 ns. From each docked group, natural product bound targets with lowest binding energy was selected and subjected for MD simulations to study the small molecular interaction and protein stability. For simulation studies GROMOS96 54a7 [29] force field implemented in GROMACS package (Abraham *et al.*, 2015), version 2018 was employed. Briefly, the system was solvated using simple point charge (SPC) model [30] and neutralized using sodium and chloride ions. Energy minimization was done using steepest descent method followed with temperature coupling (300K) using V-rescale thermostat [31] and pressure coupling (10^5 Pa) using Parrinello-Rahman barostat [32]. LINCS algorithm [33] was employed to adjust bond length constraints and particle mesh Ewald method (PME) [34] was used to evaluate electrostatic interactions. Equilibration trajectories were prepared for a time scale of 1ns under periodic boundary conditions and final MD trajectories were prepared for a time scale of 20ns at a time step of 2fs with trajectory coordinates regularly written at 10ps intervals. Thus produced MD trajectories were analysed using *gmx energy*, *gmx rms*, *gmx rmsf*, *gmx gyrate*, *gmx hbond*, *gmx do_dssp*, *gmx covar*, *gmx anaeig* and *gmx sasa* modules of GROMACS along with interaction energies in terms of electrostatic and van der Waals energy between the ligand and the macromolecule.

Binding free energy calculations using g_mmpbsa:

Molecular mechanics/Poisson-Boltzmann surface area (MMPBSA) is one of the most widely used approaches for estimating the binding free energy of a protein-ligand complex. For MMPBSA calculations, a new trajectory file was created from final 10 ns trajectories consisting of frames generated at every 200ps using *gmx trjconv* module. The binding energy components were calculated using *g_mmpbsa* package [35]. For calculation of polar solvation energy, *g_mmpbsa* relies on Adaptive Poisson-Boltzmann Solver (APBS) program. For non-polar solvation energy calculation, Solvent Accessible Surface Area (SASA) model was used. Molecular mechanical (MM) energy consisted of electrostatic (Eelec) and van der Waals (EvdW) contributions. The *g_mmpbsa* package uses the following equation to calculate the binding energy of the protein-ligand complex;

$$\Delta G_{\text{Binding}} = G_{\text{Complex}} - (G_{\text{Protein}} + G_{\text{Ligand}}) \quad (I)$$

The 'G' term can be further decomposed into the following components-

$$\Delta G = \Delta E_{\text{MM}} + \Delta G_{\text{Solvation}} - T\Delta S = \Delta E_{\text{(Bonded + Non-bonded)}} + \Delta G_{\text{(Polar + Non-polar)}} - T\Delta S \quad (II)$$

Where,

G_{Complex} = total free energy of the binding complex,

G_{Protein} and G_{Ligand} = total free energies of protein and ligand, respectively.

E_{MM} = vacuum potential energy; $G_{\text{Solvation}}$ = free energy of solvation

In equation (II), E_{MM} term consists of E_{bonded} and $E_{\text{non-bonded}}$ energies. $E_{\text{non-bonded}}$ energy is the sum of van der Waals (Lennard-Jones potential function) and electrostatic (Coulomb potential function) energies. $G_{\text{solvation}}$ consists of sum of G_{polar} and $G_{\text{non-polar}}$ terms, which represent the electrostatic and the non-electrostatic energies. G_{polar} is calculated based on continuum implicit solvent model using the Poisson-Boltzmann equation [36]. The $G_{\text{non-polar}}$ term consists of sum of G_{cavity} and G_{vdW} . G_{cavity} is the work done by the solute to create a cavity in the solvent and G_{vdW} is the attractive van der Waals energy between solvent and solute. $G_{\text{non-polar}}$ accounts for the hydrophobic effect [37]. The entropic term ($T\Delta S$) denotes the product of the temperature and the entropic contribution.

Results

***In-silico* pharmacokinetics, pharmacodynamic and toxicology profiling:**

The comparison within the T2D prescribed drugs library and COVID 19 drug library supported their similarities in structures and similar mechanisms of action (Table S1 & S2). The comparison of 166 selected natural products against 84 T2D prescribed drugs yielded 15 natural products with more than 60% structural similarity (Table 2C). The druglikeness estimated based on the molecular properties of the selected structures indicated that out of 15 molecules, 09 molecules with positive scores demonstrated druglikeness, while 14 molecules found to penetrate into human intestine, 09 molecules penetrating blood brain barrier and none of them being the Cytochromes P450 substrate indicating a high possibility of their bioavailability (Table 2A). The result of the *in-silico* PK/PD studies indicated that, out of 15 molecules, 02 were predicted to be mutagenic while none of the molecules was found to be tumorigenic, with any reproductive effects or irritants (Table 2B).

The method was extended with a larger dataset to validate the selection of natural products expected to be effective against COVID 19 infection. The complete natural product library consisting of 26,311 structures was screened against local Pubchem Covid19 library and clinical trials drug library. Among the total number of molecules screened, 17,798 natural product structures were found to have more than 60% structural similarities against Pubchem Covid19 library of which, 41 molecules were flavans, 41 were flavones, 272 were isoflavonoids. The comparison against clinical trials drug library yielded 14,689 natural products which have more than 60% structure similarity consisting of 30 flavans, 18 flavones, 78 isoflavonoids. The study was extended to compare the complete natural product library against most promising investigational drugs, viz. Remdesivir, Arbidol, Lopinavir and Ritonavir molecules which yielded 35 natural product structures with more than 60% structural similarity (Table 3C).

The druglikeness estimated based on the molecular properties of the selected structures indicated that out of 35 molecules, 23 molecules with positive scores indicated their potential drug like effects, while 33

molecules found to penetrate into human intestine, 17 molecules penetrating blood brain barrier and none of them being the Cytochromes P450 substrate, indicating a high possibility of their bioavailability (Table 3A). The result of the *in-silico* PK/PD studies indicated that, out of 35 molecules, 34 were predicted to be non-mutagenic and non-tumorigenic and non-irritant, with 10 molecules predicted to have reproductive effects (Table 3B).

High throughput virtual screening to predict Molecular interactions:

The *in-silico* molecular interaction studies resulted in predicting the most effective natural product that could bind to appropriate target [24,25]. The binding was compared with the structurally similar synthetic drug counterpart whose target was used for the interaction studies. The study yielded natural products being effectively bound to their respective targets (Table 2C and 3C). The results were expressed in terms of docking energy (kcal/mol). The effectiveness of binding was validated by considering the protein stability using molecular dynamic simulation studies.

United-atom molecular dynamic simulation studies to predict the protein stability:

In the present study MD simulations were performed to confirm the accuracy of binding resulted from docking studies. The result of MD simulation displayed the conformational changes acquired by target proteins of T2D and COVID 19 conditions after respective ligand binding (fig 4 & 5). The calculated MD parameters for unbound and ligand bound targets detailing RMSD, RMSF, Rg, SASA, Secondary structure, Coul-SR energy and LJ-SR energy of T2D and COVID 19 target proteins along with binding free energies and the contributing energy terms are tabulated in table 2D & 3D respectively. The MD trajectory analysis revealed that the natural products Naringin and Isoquercitrin displayed fewer RMS deviations and RMS fluctuations compared to their structurally similar SGLT2 inhibitors; Empagliflozin and Luseogliflozin. Similarly, Ursolic acid displayed fewer RMS deviation and RMS fluctuation compared to glucocorticoid receptor inhibitor Prednisolone. The binding free energy calculated using g_mmpbsa module displayed better interactions between selected natural products with their respective target proteins when compared to their synthetic drug counterparts.

The average RMS deviations and RMS fluctuations were calculated from the MD trajectories of natural product and synthetic drug bound HIV-1 protease (I50V isolate), Influenza virus hemagglutinin, SARS-CoV NSP 12 polymerase and HIV-1 protease (A02 isolate) and were compared with their respective unbound structures. Lesser RMS deviations were observed in HIV-1 protease (I50V isolate) after the binding of Hexahydropyrrolo Derivative, in SARS-CoV NSP 12 polymerase after the binding of HydroxyManzamin_A and HIV-1 protease (A02 isolate) after the binding of Bionectin_B compared to their respective chemical drug counterparts. RMS deviations was found to be lower in Arbidol bound Influenza virus hemagglutinin compared to natural product Phellibaurin_A bound structure. Lesser RMS fluctuations was observed in the natural product bound structures of HIV-1 protease, Influenza virus hemagglutinin and HIV-1 protease compared to their respective chemical drug bound structures. The binding free energy calculations performed using g_mmpbsa module displayed better binding of natural products with their respective target proteins compared to respective chemical drug counterparts.

Discussion And Conclusion

The present research describes a novel method to screen natural products by primarily considering their molecular composition and 2-dimensional structural arrangements, followed by their effectiveness assessment against specific disease condition. The method relies on the central foundation of medicinal chemistry that structurally similar molecules will have similar biological effect [1,8]. As an initial part of the study, the natural products were screened against synthetic and/or prescribed drugs for a specific disease condition based on their molecular and 2-dimensional structural similarities which resulted in the selection of effective bioactive molecules. The identified natural products were checked for druglikeness, pharmacokinetics and pharmacodynamics properties by comparing against a large group of already reported datasets supporting the screening strategy, which formed the next part of the study. The biological interactions involved in the action mechanism of synthetic drugs was extended to their structurally similar natural products. The interaction studies forming the third part of the study described the binding effectiveness of the selected molecules against suitable drug targets. The effectiveness of binding was studied using MD simulation which formed the final part of the methodology, finally reporting with the most effective drug-like natural product for disease conditions under study. Based on the molecular interactions, target's structural stability induced by natural products, in a way similar to that of their synthetic drug counterparts, it is convincing to state that the selected natural products can also exert similar effects if administered individually.

Several approaches have been proposed for searching and identification of chemical compounds in a dataset which includes conventional string sequence search, to using modern data structure module based search algorithms [3–6]. The similarity measurements generally consider either structural representations involving physiochemical properties, topological indices, molecular graphs, pharmacophore features, molecular shapes, molecular fields or quantitative measures involving Tanimoto coefficient, Dice index, cosine coefficient, Euclidean distance, Tversky index for similarity assessment between two structures [1]. The 3-dimensional confirmation based searching methods include pharmacophore modeling, shape similarity, molecular field-based methods, 3D fingerprints to compare chemical compounds. The popular chemical databases like PubChem uses percent similarity measures employing Tanimoto equation [38] and a dictionary-based fingerprint, analogous to the MACCS (Molecular ACCess System) structure-based keys [39], while the tools like SwissTargetPredictor predicts the target for bioactive molecules based on a combination of 2D and 3D similarity measures. However, it fails to predict the biological effects if the prediction accuracy is significantly lower for molecules with unknown bioactivity [9]. The OSIRIS DataWarrior program, employed in the current study for the assessment of molecular similarity uses chemical similarity based methods which relies on substructure fragments to biological similarity considering 3D-geometry and binding [17]. *SAHANA* by utilizing these calculations uniquely screens natural products against specific disease conditions by targeting a specific drug target, even for molecules whose bioactivity is not known, and proceeds for bioactivity assessment based on molecular interactions and structural stability assessments using molecular docking and MD simulation studies.

Limitations and other considerations:

The methodology described in *SAHANA* relies completely on computational calculations. However, the results have the statistical accuracy; the experimental validations are required to confirm the reliability of the prediction. In this avenue, the prediction made using *SAHANA* can act as a 'rule of thumb' in screening a large dataset of natural products against any specific disease condition. The other limitation is *SAHANA* cannot be applied to conditions, which do not have any prescribed drugs with predefined mechanism of action. As the prediction relies on the structural similarities between natural products and the available drugs, the presence of prescribed drug with well-defined action mechanism is a prerequisite for *SAHANA*.

Abbreviations

ADME- Absorption, Distribution, Metabolism, and Excretion; APBS- Adaptive Poisson- Boltzmann Solver; HTVS- High Throughput Virtual Screening; MACCS- Molecular ACCess System; MD- Molecular Dynamics; MM- Molecular mechanical; MMPBSA- Molecular mechanics/Poisson-Boltzmann surface area; NCATS- National Center for Advancing Translational Sciences; NPASS- Natural product activity and species source database; NPs - Natural products; PD- Pharmacodynamics; PDB- Protein Data Bank; PK – Pharmacokinetics; PME- Particle Mesh Ewald method; Rg- Radius of Gyration; RMSD- Root Mean Square Deviation; RMSF- Root Mean Square Fluctuation; RO5- Rule-of-Five; ROSDAL- Representation of Organic Structures Description Arranged Linearly; SASA- Solvent Accessible Surface Area; SMILES- Simplified Molecular Input Line Entry System; SPC- Simple Point Charge; T2D- Type 2 diabetes; TPSA- Topological polar surface area; WLN- Wiswesser Line Notation.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: All the data used during the current study are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no conflicts of interest.

Funding: This work is supported by the Department of Science and Technology (DST)- Science and Engineering Research Board (SERB), Govt. of India (Grant number: PDF/2018/00237).

Authors' contributions: ARSJ: designed and conceived the study, performed the research, and wrote the manuscript. NPS: participated in the results discussion and technical support. Both the authors read and approved the final manuscript.

Acknowledgement: The authors thank the Department of Science and Technology (DST) - Science and Engineering Research Board (SERB), Govt. of India for their financial support.

References

1. Kumar A, Zhang KYJ. Advances in the development of shape similarity methods and their application in drug discovery. *Front. Chem.* 2018.
2. Wang H, Volarath P, Harrison R. An approach in building a chemical compound search engine in oracle database. *Annu Int Conf IEEE Eng Med Biol - Proc.* 2005. p. 2839–42.
3. Gasteiger J, Engel T. *Chemoinformatics: A Textbook.* Wiley-VCH Verlag GmbH & Co. KGaA; 2003.
4. Volarath P, Wang H, Fu H, Harrison R. Knowledge-based algorithms for chemical structure and property analysis. *Annu Int Conf IEEE Eng Med Biol - Proc.* 2004. p. 3011–4.
5. Ullmann JR. An Algorithm for Subgraph Isomorphism. *J ACM.* 1976;23:31–42.
6. Cordella LP, Foggia P, Sansone C, Vento M. Performance evaluation of the VF graph matching algorithm. *Proc - Int Conf Image Anal Process ICIAP 1999.* 1999. p. 1172–7.
7. Schneider G. Madame Curie Bioscience Database. Predict Drug-Like Prop [Internet]. Austin (TX): Landes Bioscience; 2013. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK6404/>
8. Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity? *J Med Chem.* 2002;45:4350–8.
9. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: A web server for target prediction of bioactive small molecules. *Nucleic Acids Res.* 2014;42.
10. Borman S. SIMILAR MOLECULES, OPPOSITE EFFECTS. *Chem Eng News.* 2010;88:11.
11. Wisén S, Bertelsen EB, Thompson AD, Patury S, Ung P, Chang L, et al. Binding of a small molecule at a protein-protein interface regulates the chaperone activity of Hsp70-Hsp40. *ACS Chem Biol.* 2010;5:611–22.
12. Baker DC, Chu C. *Nucleosides and Nucleotides as Antitumor and Antiviral Agents.* 1st ed. Springer US; 1993.
13. Capecchi A, Probst D, Reymond JL. One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome. *J Cheminform.* 2020;12.
14. Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, et al. NPASS: Natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* 2018;46:D1217–22.

15. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res.* 2019;47:D1102–9.
16. U.S. National Institute of Health (NIH), U.S. National Library of Medicine (NLM). ClinicalTrials.gov database [Internet]. U.S. Natl. Lib. Med. 2008. Available from: <https://clinicaltrials.gov/ct2/home>
17. Sander T, Freyss J, Von Korff M, Rufener C. DataWarrior: An open-source program for chemistry aware data visualization and analysis. *J Chem Inf Model.* 2015;55:460–73.
18. Huigens RW, Morrison KC, Hicklin RW, Timothy TA, Richter MF, Hergenrother PJ. A ring-distortion strategy to construct stereochemically complex and structurally diverse compounds from natural products. *Nat Chem.* 2013;5:195–202.
19. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 1997. p. 3–25.
20. Jarrahpour A, Motamedifar M, Zarei M, Youssoufi MH, Mimouni M, Chohan ZH, et al. Petra, Osiris, and Molinspiration Together as a Guide in Drug Design: Predictions and Correlation Structure/Antibacterial Activity Relationships of New N-Sulfonyl Monocyclic β -Lactams. *Phosphorus Sulfur Silicon Relat Elem* [Internet]. 2010;185:491–7. Available from: <http://www.tandfonline.com/doi/abs/10.1080/10426500902953953>
21. Aditya R, Venugopal T, Jayanna N, Paramesha M, Ramesh C. Bioactive isolates of *Morus* species as antibacterial agents and their insilico profiling. *Lett Drug Des Discov.* 2020;17.
22. Cheng F, Li W, Zhou Y, Shen J, Wu Z, Liu G, et al. AdmetSAR: A comprehensive source and free tool for assessment of chemical ADMET properties. *J Chem Inf Model.* 2012;52:3099–105.
23. Raghavendra S, Aditya Rao SJ, Kumar V, Ramesh CK. Multiple ligand simultaneous docking (MLSD): A novel approach to study the effect of inhibitors on substrate binding to PPO. *Comput Biol Chem.* 2015;59.
24. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol.* 2007;25:197–206.
25. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. *Science (80-).* 2008;321:263–6.
26. Trott O, Olson AJ. Software news and update AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem.* 2010;31:455–61.

27. Lin J-H. Accommodating Protein Flexibility for Structure-Based Drug Design. *Curr Top Med Chem.* 2012;11:171–8.
28. Salsbury FR. Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Curr. Opin. Pharmacol.* 2010. p. 738–44.
29. Schmid N, Eichenberger AP, Choutko A, Riniker S, Winger M, Mark AE, et al. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur Biophys J.* 2011;40:843–56.
30. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J. Interaction Models for Water in Relation to Protein Hydration. 1981. p. 331–42.
31. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. *J Chem Phys.* 2007;126.
32. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys.* 1981;52:7182–90.
33. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A Linear Constraint Solver for molecular simulations. *J Comput Chem.* 1997;18:1463–72.
34. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. *J Chem Phys.* 1995;103:8577–93.
35. Kumari R, Kumar R, Consortium OSDD, Lynn A. g _ mmpbsa - A GROMACS tool for MM-PBSA and its optimization for high-throughput binding energy calculations. *J Chem Inf Model.* 2014;54:1951–62.
36. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc Natl Acad Sci U S A.* 2001;98:10037–41.
37. Richmond TJ. Solvent accessible surface area and excluded volume in proteins. *J Mol Biol.* 1984;178:63–89.
38. Tanimoto T. IBM Internal Report 17th [Internet]. 1957. Available from: http://en.wikipedia.org/wiki/Jaccard_index
39. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci.* 2002;42:1273–80.

Tables

Due to technical limitations the Tables are available as a download in the Supplementary Files.

Figures

Thymidine

Zidovudine

Figure 1

Structural similarity between thymidine and zidovudine with variations highlighted inside the ellipse.

Figure 2

Overview of the methodology of SAHANA

Isoflavonoids	Structure of Isoflavonoid	Similarity (FragFp)	Most Similar Structure	Most similar Generic Drug name
Genistein 8-C-Beta-D-Glucoside		0.78291816 0.6730769		Dapagliflozin Etoposide
Puerarin		0.76156586 0.6565934		Dapagliflozin Etoposide
Grincamycin F		0.78608245 0.7744361 0.70387244 0.67058825 0.6681319 0.65984654 0.65346533 0.63116884 0.62094766		Resiniferatoxin Etoposide Naltrexone Doxycycline Macrolide Ciclesonide Ivermectin Budesonide Esin
Trifolirhizin		0.76512456 0.71590906 0.6386768 0.60928965		Dapagliflozin Etoposide Naltrexone Resiniferatoxin
7-Demethoxy-7-D-(Glucosyloxy)-Homopterocarpin		0.7544484 0.7073864 0.63104326 0.6098901		Dapagliflozin Etoposide Naltrexone Resiniferatoxin

Figure 3

Output of structurally similar molecule comparison results from DataWarrior software.

Figure 4

RMSD, RMSF, Rg and SASA plots obtained from MD trajectories analysis of native, Natural product bound and chemical drug bound structure of SGLT2 receptor (a & b) and glucocorticoid receptor (c).

Figure 5

RMSD, RMSF, Rg and SASA plots obtained from MD trajectories analysis of native, Natural product bound and chemical drug bound structure of HIV-1 protease (a), influenza virus hemagglutinin (b), SARS-CoV NSP12 polymerase (c) and HIV-1 protease (d).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [tables08052021.pdf](#)
- [Supplementarydata08052021.docx](#)