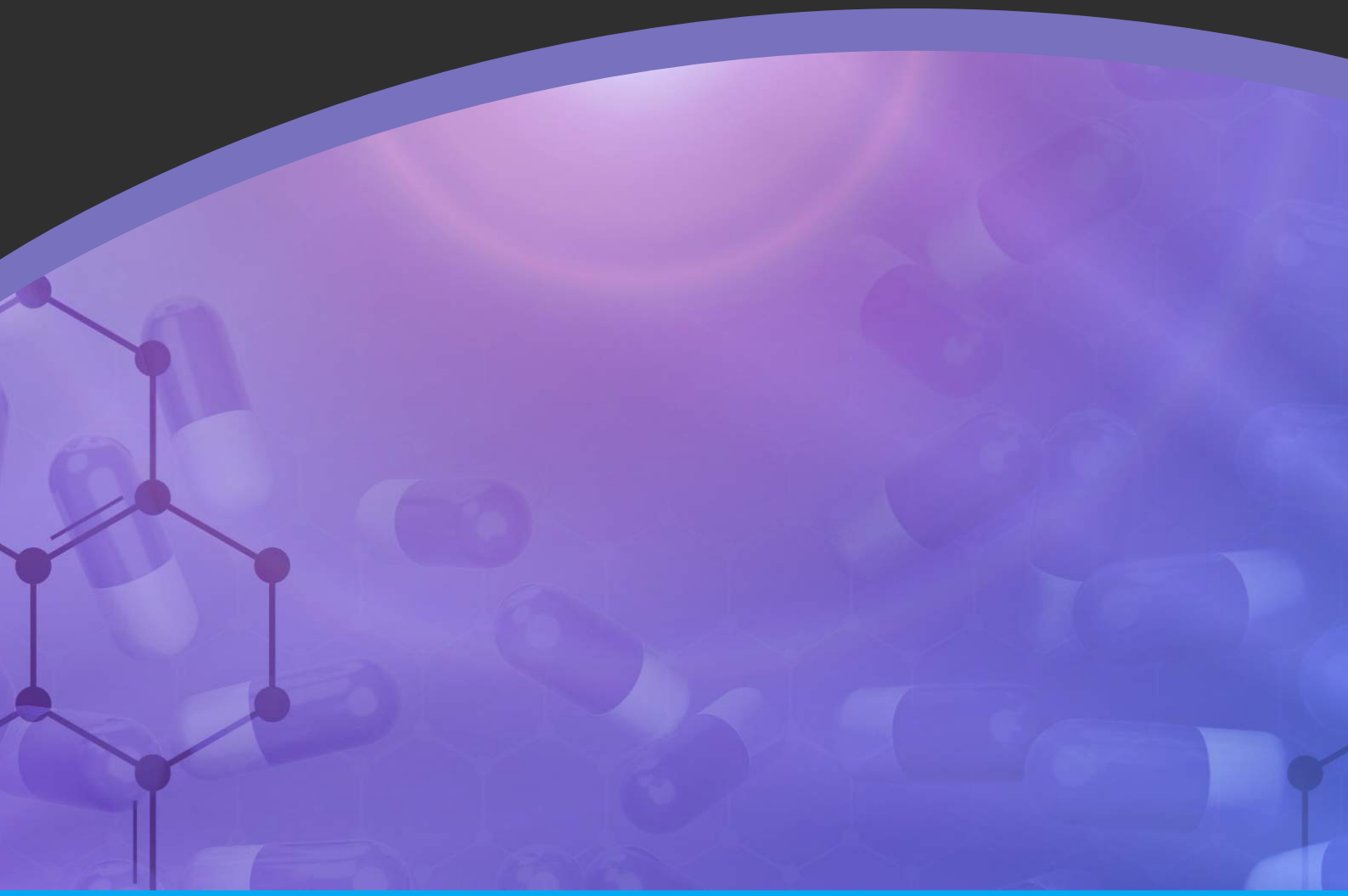
SPECIAL ISSUE

ORGANIC CHEMISTRY FOR HEALTH

Discovery, synthesis, characterisation, and evaluation of new physiologically active compounds



PeerJ Organic Chemistry

PeerJ Life & Environment

A *PeerJ Life & Environment* and *PeerJ Organic Chemistry* cross-journal Special Issue

PeerJ Publishing

PeerJ is a modern and streamlined publisher, built for the internet age. Our mission is to give researchers the publishing tools and services they want, with a unique and exciting experience. All of our seven journals are Gold Open Access and are widely read and cited, with over 500,000 monthly views and 48,500 content alert subscribers. We have published 12,844 peer-reviewed articles since 2013.

Prestigious Editorial Board

High-Impact Research

Quality Peer Review

Rapid Publishing

Optimum Discoverability

High-quality, developmental peer review, coupled with industry leading customer service and an award-winning submission system, means PeerJ journals are the optimal choice for your research

PeerJ Life & Environment

PeerJ Organic Chemistry

PeerJ Organic Chemistry

GC-MS analysis of phytochemical compounds of *Opuntia megarrhiza* (Cactaceae), an endangered plant of Mexico

Madeleyne Cupido[1], Arturo De-Nova[1,2], María L. Guerrero-González[1], Francisco Javier Pérez-Vázquez[3], Karen Beatriz Méndez-Rodríguez[3] and Pablo Delgado-Sánchez[1]

[1] Facultad de Agronomía y Veterinaria, Universidad Autónoma de San Luis Potosí, Soledad de Graciano Sánchez, San Luis Potosí, Mexico
[2] Instituto de Investigación de Zonas Desérticas, Universidad Autónoma de San Luis Potosí, San Luis Potosí, San Luis Potosí, Mexico
[3] Coordinacion para la Innovación y Aplicación de la Ciencia y la Tecnología, Universidad Autónoma de San Luis Potosí, San Luis Potosí, San Luis Potosí, Mexico

ABSTRACT

Opuntia megarrhiza is an endemic plant used in Mexican traditional medicine for the treatment of bones fractures in humans and domestic animals. One of the most used technique for the detection and characterization of the structure of phytochemical compounds is the Gas Chromatography Coupled to Mass Spectrometry. The goals of the present study were to identify and characterize the phytochemical compounds present in wild individuals of *O. megarrhiza* using this analysis. We used chloroform and methanol extracts from cladodes, and they were analyzed by gas chromatography-electron impact-mass spectrometry. We obtained 53 phytochemical compounds, 19 have been previously identified with some biological activity. Most of these compounds are alkanes, alkenes, aromatic hydrocarbons, fatty acids, and ketones. We detected some fragmentation patterns that are described for the first time for this species. The variety of metabolites presents in *O. megarrhiza* justifies the medicinal use of this plant in traditional medicine and highlight it as a source of phytochemical compounds with potential in medicine and biotechnology.

Subjects Organic Chemistry (other), Organic Compounds, Photochemistry
Keywords Bioprospection, Endemic, Metabolites, Phytochemicals screening

Submitted 12 August 2021
Accepted 22 February 2022
Published 22 March 2022

Corresponding authors
Arturo De-Nova,
arturo.denova@uaslp.mx
Pablo Delgado-Sánchez,
pablo.delgado@uaslp.mx

Academic editor
Eder Lenardao

Additional Information and Declarations can be found on page 15

DOI **10.7717/peerj-ochem.5**

INTRODUCTION

The members of Cactaceae represent a diverse evolutionary lineage endemic to America, over 1,450 species belonging to ca. 127 genera (*Barthlott & Hunt, 1993*; *Hunt, Taylor & Charles, 2006*; *Hernández-Hernández et al., 2011*). They are successful plants adapted to arid and semiarid environments, where the conditions imply a constant stress, so they have developed different phytochemical compounds with an important biological activity such as alkaloids, amino acids, antioxidant phenol components (betalains and flavonoids), carotenoids, coumarins, esters, fibers, phytosterols, tannins, terpenes,

tocopherols, and vitamins C and E (*Piattelli, Minale & Prota, 1965*; *Stintzing, Schieber & Carle, 2001*; *Strack, Vogt & Schliemann, 2003*; *Paiz et al., 2010*; *Sim et al., 2010*; *Osorio-Esquivel et al., 2011*; *Harlev et al., 2012*; *Aruwa, Amoo & Kudanga, 2018*; *Araújo et al., 2021*). Bioactive phytochemical compounds are of great interest since their possible applications in biotechnology and industry, and they are usually categorized into phenolic and non-phenolic compounds and pigments (*Martins et al., 2011*; *Aruwa, Amoo & Kudanga, 2018*; *Araújo et al., 2021*). Some of them have nutritional benefits (*Kris-Etherton et al., 2002*; *Kudanga, Nemadziva & Le Roes-Hill, 2017*; *Araújo et al., 2021*; *Yu et al., 2021*), pharmaceutical applications (*Aruwa, Amoo & Kudanga, 2018*; *Araújo et al., 2021*; *Patra et al., 2021*; *Yu et al., 2021*), and are used in the production of nutraceuticals (*Gil-Chávez et al., 2013*; *Aruwa, Amoo & Kudanga, 2018*; *Yu et al., 2021*), in novel food formulations (*Gurrieri et al., 2000*; *Pawar, Killedar & Dhuri, 2017*; *Aruwa, Amoo & Kudanga, 2018*; *Araújo et al., 2021*), and for animal feed supplementation (*Ennouri et al., 2006*; *Aruwa, Amoo & Kudanga, 2018*; *Araújo et al., 2021*).

The species of genus *Opuntia* (Cactaceae) are native of Mexico, where they originated and diversified (*Barthlott & Hunt, 1993*; *Reyes-Agüero, Aguirre & Carlín, 2004*). Different cultures, ancient and modern, have used them as fuel, forage, fences, food, and particularly in traditional medicine (*González-Durán, Riojas & Arreola, 2001*; *Reyes-Agüero, Aguirre-Rivera & Hernández, 2005*; *Andrade-Cetto & Wiedenfeld, 2011*). Several *Opuntia* species have the ability to synthesize molecules with unique and complex structures with therapeutical potential (*Shedbalkar et al., 2010*; *Bargougui, Le Pape & Triki, 2013*; *Weli et al., 2019*). For example, *Opuntia dillenii* (Ker Gawl.) Haw. has beneficial effects for the human health as anti-inflammatory, analgesic, hypoglycemiant, hypocholesterolemiant, and antioxidant (*Perfumi & Tacconi, 1996*; *Park et al., 2001*; *Ghasemzadeh & Ghasemzadeh, 2011*). Other edible species like *Opuntia ficus-indica* (L.) Mill. present antioxidant and antiproliferative activities useful in colon cancer (*Serra et al., 2013*; *Yeddes et al., 2013a*; *Yeddes et al., 2013b*), and have nutraceutical, anticarcinogen, and antiviral properties useful in the digestive processes, reducing the risks of obesity, gastrointestinal suffering and high levels of cholesterol (*Feugang, 2006*; *Bensadón et al., 2010*). Usually, the used parts are the fruit, stem, cladode, and root (*Estrada-Castillón et al., 2012*). It has been estimated that plants have ca. 200,000 different metabolites, primary and secondary (*Pichersky & Gang, 2000*; *Fiehn, 2001*; *Fiehn, 2002*), like aminoacids, fatty acids, carbohydrates, lipids, and more (*Velmurugan & Anand, 2017*; *Banakar & Jayaraj, 2018*). Several studies in *Opuntia* have focused on the analysis of phytochemical compounds as alkaloids, carotenoids, flavonoids, phenols, and vitamins C and E in different plant species in order to discover and produce new drugs for several illnesses (*Yahia & Mondragon-Jacobo, 2011*; *Weli et al., 2019*), as well as nutritional supplements since they provide metabolites, mineral, and vitamins essential for the human organism (*Caballero-Gutiérrez & Gonzáles, 2016*). In the agriculture, some of these compounds are used for the control of phytopathogen microorganisms (*De Corato et al., 2010*). Some other, are used industrially to produce detergents, cosmetics and dermatological products, solvents, lubricants, textiles, and others (*Kim et al., 2019*).

Figure 1 ***Opuntia megarrhiza*** **from wild populations.** (A) Herbarium specimen from the studied locality; (B) Flowering adult plant; (C) Adult plant showing part of its characteristic massive roots. Photos (A) and (B) by J.A. De Nova, (C) by P. Delgado. Full-size DOI: 10.7717/peerj-ochem.5/fig-1

The main chemical analysis for the detection and characterization of the structure of phytochemical compounds are the Thin-Layer Chromatography, the UV-Vis spectrophotometry, the Nuclear Magnetic Resonance, the liquid chromatography-mass spectrometry (LC-MS), and the Gas Chromatography Coupled to Mass Spectrometric (GC-MS) (*Robertson, 2005*; *Marquet, 2012*). The last one is the most used in metabolomics research since it is a very selective technique for the detection and characterization of metabolites (*Fiehn, 2016*). The LC-MS is a robust technique for general unknown screening, however its major drawback is the lack of universal reference libraries obtained with different instrument types (*Marquet, 2012*), as in GC-MS. The GC-MS together with the metabolomic analysis are a key for the profiling of metabolites in plants since they perform the qualitative and quantitative characterization of all the molecules (metabolites) present in their cellules (*Harrigan & Goodacre, 2012*).

Opuntia megarrhiza Rose is a species endemic to Mexico, locally known as "nopal camote" (Fig. 1). It is restricted to some regions of the Chihuahuan Desert, particularly in the State of San Luis Potosi (*Hernández, Gómez-Hinostrosa & Bárcenas, 2001*), and it is listed as endangered in the IUCN Red List (*Hernández et al., 2013*). It grows in different habitats as xerophytic scrubs, oak forest, and other mountain forests (*Hernández, Gómez-Hinostrosa & Bárcenas, 2001*; *Segura-Venegas & Rendón-Aguilar, 2016*). This species is characterized by its massive roots, which are succulent, gross, and deeply buried in ground, 30 to 60 cm long and 5 to 10 cm diameter (*Bravo-Hollis & Sánchez-Mejorada, 1991*;

Hernández & Godínez, 1994). The cladodes are relatively small contrasting with other *Opuntia* species. The flowers are yellowish-green to pink, 3 to 5.5 cm long and 2.5 to 6 cm diameter at anthesis. Fruits are ovoid, 2.5 cm long, and the seeds are ca. 4 mm diameter (*Bravo-Hollis & Scheinvar, 1999*; *Hernández, Gómez-Hinostrosa & Bárcenas, 2001*).

Opuntia megarrhiza is used by locals in the treatment of bones fractures, both animals and humans (*Hernández, Gómez-Hinostrosa & Bárcenas, 2001*). In Cerro El Borrego (Guadalcazar, San Luis Potosi), the root is applied directly in the fractures, but in other localities like Xoconoxtle (Zaragoza, San Luis Potosi) people use cladodes or the whole plant, and sometimes is mixed with parts of *Cylindropuntia* spp., to make a paste that is applied with bandages in the injury (*Segura-Venegas & Rendón-Aguilar, 2016*). Given the ethnopharmacological value of *O. megarrhiza*, previously highlighted by the empirical traditional medicine, the goals of the present study were to identify and characterize the phytochemical compounds present in cladodes from wild individuals of the species, using GC-MS. There are not previous studies about the bioactive phytochemical compounds in this species, so its phytochemical characterization could contribute to increase the knowledge of the species and its potential biotechnological applications, but also improving bio-valorization and environment.

MATERIALS AND METHODS

Sampling

All the protocols involving plants were adhered to relevant ethical guidelines and permissions for plant sampling from herbaria JAAA and SLPM (SGPA/DGGFS/712/0501/18) were used. Non-lethal samples of cladodes were obtained from ten wild individuals, sampling randomly, of *Opuntia megarrhiza* located at Cerro San Pedro (San Luis Potosi). We do not collect whole plants and the identification was conducted in field by taxonomists Arturo De Nova and Eleazar Carranza form herbarium SLPM and registered with photographs then confirmed with Research Grade in iNaturalist (46236747, 46978331). The Fig. 1A show a reference herbarium voucher from the studied locality for identification (SLPM 22132).

Extracts

The extractions were conducted at Laboratorio de Biotecnologia de la Facultad de Agronomia y Veterinaria, UASLP and Centro Regional de Biociencias, UASLP. Cladode fragments were cleaned using brushes and distillated water to eliminate possible associated microorganisms. We used 95 g of sample, which was macerated to make a semisolid paste, then either 25 ml of methanol (MeOH) for extraction 1 or 25 ml of chloroform ($CHCl_3$) for extraction 2. These solutions were vortex mixed one hour to prevent conglomeration and sedimentation of small particles. The extracts were filtered three times using Whatman paper, grade 1, 5 and 6 in order, in a vacuum chamber. The volumes were adjusted to 10 mg/ml for all extracts. Subsequently, we made a dilution for each solution using acetone as solvent (1:1): (1) acetone-chloroform ($(CH_3)_2CO/CHCl_3$), and (2) acetone-methanol ($(CH_3)_2CO$/MeOH). Before depositing in a vial, extracts were filtered through a polytetrafluoroethylene (PTFE) or polyvinylidene (PVDF) membranes

(with different hydrophobic adsorption ranges and size exclusion pores), and transferred to a vial for analysis in the GC-MS.

GC-MS analysis scan mode

This process was conducted at Coordinacion para la Innovacion y la Aplicacion de la Ciencia y la Tecnologia, UASLP. We used the Hewlett Packard gas chromatograph HP 6890, coupled to mass spectrometry detector with electronic impact HP 5973 (Agilent Technologies, Palo Alto, CA, USA). The column exerted was an absorbed silica capillary column of 95% methyl-poly-siloxane and 5% phenyl (HP-5MS; length: 60 m, diameter: 0.25 mm and film 0.25 µm). Helium was used as carrier gas, and the flow rate was 1 ml/min. The GC oven temperature gradient was: 60 °C (hold for 3 min) initially, then increased 5 °C each minute until 300 °C, this final temperature was hold 5 minutes. The transfer line temperature was 280 °C. The ion source temperature was 230 °C and the MS was scanned at 50 to 550 mass range. The essays were processed in the ChemStations software (Houston, TX, USA) to generate the chromatograms for the interpretation of the spectra.

Identification of phytochemical compounds

The identification was performed by comparing the spectrum of unknown compounds with the spectrum of known compounds in the National Institute of Standards and Technology (NIST98) mass spectral library. Compound name, synonyms, molecular weight, and the mass spectrum for each compound were obtained from NIST Standard Reference 69 and PubChem databases to confirm the GC-MS results. Nomenclature for all compound names was standardized with IUPAC rules. Relative quantification of the compounds present in each sample was obtained from the relative area of the peaks in the chromatograms. Biological activity for each identified compound was obtained with an exhaustive search in scientific publications and from the Dr. Duke phytochemical and ethnobotanical databases. The identity of five compounds showing similarity above 90% (as recommended in *Mangas-Marín et al., 2018*) with phytochemical compounds previously reported with biological activity was verified by using the commercial pure standards: The used standards were 1,3-benzothiazole (Purity: minimum 97.0%), henicosane (Purity: minimum 99.5%), hentriacontane (Purity: minimum 98.0%), methyl hexadecanoate (Purity: minimum 98.5%), triacontane (Purity: minimum 98.0%), were purchased from Sigma (St. Louis, MO, USA). All standards were diluted in acetone as the final solvent at a concentration of 5 ppm and were analyzed in the GC-MS using the same parameters than in the samples.

RESULTS

A total of 53 phytochemical compounds were detected based on the analyses of the obtained chromatograms (Tables 1–4). The $(CH_3)_2CO$/MeOH extract showed 11 peaks with the PVDF membrane filter and 12 with the PTFE. The $(CH_3)_2CO/CHCl_3$ extract showed 23 peaks with PVDF and seven with PTFE (Fig. 2).

Table 1 Phytochemical compounds identified by GC-MS from MeOH extract with PVDF membrane filter.

Peak No.	RT (min)	Name of the compound	Molecular weight (g/mol)	Peak area (%)	Similarity* (%)	Molecular formula	Compound nature
1	8.21	2,3,6,7-tetramethyloctane	170.33	1.17	64	$C_{12}H_{26}$	Alkane
2	8.33	2,3,5,8-tetramethyldecane	198.38	1.78	64	$C_{14}H_{30}$	Alkane
3	15.35	1,3-di*tert*-butylbenzene	190.32	4.19	95	$C_{14}H_{22}$	Aromatic Hydrocarbon
4	16.74	methyl (2*R*)-5-oxo-2-propan-2-ylhexanoate	186.25	2.78	53	$C_{10}H_{18}O$	Ester
5	17.20	4-propan-2-ylcyclohexane-1,3-dione	154.21	2.19	60	$C_9H_{14}O_2$	Ketone
6	17.28	heptadecane	240.46	1.03	59	$C_{17}H_{36}$	Alkane
7	21.25	7-methylhexadecane	240.46	1.00	72	$C_{17}H_{36}$	Alkane
8	22.69	2,3,6-trimethyldecane	184.36	1.80	64	$C_{13}H_{28}$	Alkane
9	31.75	hexadecanoic acid	256.42	4.67	97	$C_{16}H_{32}O_2$	Fatty acid
10	35.43	octadecanoic acid	284.47	2.47	93	$C_{18}H_{36}O_2$	Fatty acid
11	46.72	1,54-dibromotetrapentacontane	917.2	1.41	76	$C_{54}H_{108}Br_2$	Halogenated hydrocarborn

Notes:
* Percentage of similarity to the reference spectrum of the NIST library.
RT, retention time.

Table 2 Phytochemical compounds identified by GC-MS from MeOH extract with PTFE membrane filter.

Peak No.	RT (min)	Name of the compound	Molecular weight (g/mol)	Peak area (%)	Similarity* (%)	Molecular formula	Compound nature
1	8.33	4-methyldecane	156.30	1.44	64	$C_{11}H_{24}$	Alkane
2	15.35	1,3-di*tert*-butylbenzene	190.32	3.53	95	$C_{14}H_{22}$	Aromatic Hydrocarbon
3	17.20	4-propan-2-ylcyclohexane-1,3-dione	154.21	1.84	53	$C_9H_{14}O_2$	Ketone
4	21.61	heptacosane	380.73	0.98	83	$C_{27}H_{56}$	Alkane
5	22.05	2,4-di*tert*-butylphenol	206.32	2.13	97	$C_{14}H_{22}O$	Aromatic Hydrocarbon
6	22.69	pentacosane	352.68	1.66	64	$C_{25}H_{52}$	Alkane
7	32.89	henicosane	296.57	0.98	90	$C_{21}H_{44}$	Alkane
8	46.73	nonacosane	408.8	2.28	95	$C_{29}H_{60}$	Alkane
9	48.04	triacontane	422.81	1.4	96	$C_{30}H_{62}$	Alkane
10	49.33	hentriacontane	436.83	1.63	93	$C_{31}H_{64}$	Alkane
11	50.57	octacosane	394.76	1.28	95	$C_{28}H_{58}$	Alkane
12	52.22	(3*S*,8*S*,9*S*,10*R*,13*R*,14*S*,17*R*)-17-[(2*R*,5*S*)-5-ethyl-6-methylheptan-2-yl]-10,13-dimethyl-2,3,4,7,8,9,11,12,14,15,16,17-dodecahydro-1*H*-cyclopenta[a]phenanthren-3-ol	414.70	1.76	90	$C_{29}H_{50}O$	Lipids

Notes:
* Percentage of similarity to the reference spectrum of the NIST library.
RT, retention time.

The analysis from $(CH_3)_2CO$/MeOH extract with PVDF membrane filter showed the presence of 11 phytochemical constituents. Five alkanes: 2,3,6,7-tetramethyloctane (1.17%), 2,3,5,8-tetramethyldecane (1.78%), heptadecane (1.03%), 7-methylhexadecane (1%), 2,3,6-trimethyldecane (1.80%). One aromatic hydrocarbon: 1,3-di*tert*-butylbenzene (4.19%). One ester: methyl (2*R*)-5-oxo-2-propan-2-ylhexanoate (2.78%). One ketone: 4-propan-2-ylcyclohexane-1,3-dione (2.19%). One Halogenated hydrocarbons: 1,54-dibromotetrapentacontane (1.41%). Two fatty acids: hexadecanoic acid (4.67%), octadecanoic acid (2.47%) (Fig. 2A, Table 1).

Table 3 Phytochemical compounds identified by GC-MS from $CHCl_3$ extract with PVDF membrane filter.

Peak No.	RT (min)	Name of the compound	Molecular weight (g/mol)	Peak area (%)	Similarity* (%)	Molecular formula	Compound nature
1	10.45	4-ethyl-1,2-dimethylbenzene	134.21	2.27	93	$C_{10}H_{14}$	Aromatic Hydrocarbon
2	11.34	1,2,4,5-tetramethylbenzene	134.21	0.51	81	$C_{10}H_{14}$	Aromatic Hydrocarbon
3	11.47	1,2,3,4-tetramethyl-5-methylidenecyclopenta-1,3-diene	134.21	0.82	95	$C_{10}H_{14}$	Alkene
4	14.53	1,3-benzothiazole	135.18	5.33	95	C_7H_5NS	Aromatic Hydrocarbon
5	18.59	(*E*,7*R*,11*R*)-3,7,11,15-tetramethylhexadec-2-en-1-ol	296.53	0.45	42	$C_{20}H_{40}O$	Alcohol
6	22.70	6-hexyloxan-2-one	184.27	0.58	59	$C_{11}H_{20}O_2$	Ketone
7	30.09	(*E*)-octadec-5-ene	252.47	0.67	78	$C_{18}H_3$	Alkene
8	30.95	7,9-di*tert*-butyl-1-oxaspiro[4.5]deca-6,9-diene-2,8-dione	276.37	1.61	50	$C_{17}H_{24}O_3$	Ketone
9	31.01	methyl hexadecanoate	270.45	1.1	93	$C_{17}H_{34}O_2$	Fatty acid
10	33.98	(*E*)-octadec-9-ene	252.5	1.65	89	$C_{18}H_{36}$	Alkene
11	34.20	methyl (9*Z*,12*Z*)-octadeca-9,12-dienoate	294.47	0.65	96	$C_{19}H_{34}O_2$	Fatty acid
12	34.31	(3*Z*,13*Z*)-2-methyloctadeca-3,13-dien-1-ol	280.5	0.45	53	$C_{19}H_{36}O$	Alcohol
13	34.78	methyl octadecanoate	298.50	0.97	93	$C_{19}H_{38}O_2$	Fatty acid
14	34.88	tridecanedial	212.33	0.58	62	$C_{13}H_{24}O_2$	Aldehyde
15	34.97	1-(7-hydroxy-8,9-dimethoxy-17-oxa-5,15-diazahexacyclo[13.4.3.$0^{1,16}$.$0^{4,12}$.$0^{6,11}$.$0^{12,16}$]docosa-6,8,10-trien-5-yl)ethanone	414.5	0.84	52	$C_{23}H_{30}N_2O_5$	Ketone
16	35.84	butyl hexadecanoate	312.53	4.83	87	$C_{20}H_{40}O_2$	Fatty acid
17	36.24	dioctadecoxy(oxo)phosphanium	585.9	0.55	53	$C_{36}H_{74}O_3P^+$	Fatty acid
18	39.21	2-methylpropyl octadecanoate	340.58	3.10	87	$C_{22}H_{44}O_2$	Fatty acid
19	39.39	tetratriacontane	478.91	0.92	90	$C_{34}H_{70}$	Alkane
20	40.96	tetrapentacontane	759.45	0.93	80	$C_{54}H_{110}$	Alkane
21	42.48	icosane	282	1.34	68	$C_{20}H_{42}$	Alkane
22	45.81	(6*E*,10*E*,14*E*,18*E*)-2,6,10,14,18-pentamethylicosa-2,6,10,14,18-pentaene	350.6	4.80	74	$C_{25}H_{50}$	Alkene
23	52.21	17-(5-ethyl-6-methylheptan-2-yl)-10,13-dimethyl-2,7,8,9,11,12,14,15,16,17-decahydro-1*H*-cyclopenta[a]phenanthrene	396.7	13.80	60	$C_{29}H_{48}$	Alkene

Notes:
* Percentage of similarity to the reference spectrum of the NIST library.
RT, retention time.

The analysis from the $(CH_3)_2CO$/MeOH extract with PTFE membrane filter detected 12 phytochemical constituents. Eight alkanes: 4-methyldecane (1.44%), heptacosane (0.98%), pentacosane (1.66%), henicosane (0.98%), nonacosane (2.28%), triacontane (1.4%), hentriacontane (1.63%), octacosane (1.28%). Two aromatic hydrocarbons: 1,3-di*tert*-butylbenzene (3.53%), 2,4-di*tert*-butylphenol (2.13%). One ketone: 4-propan-2-ylcyclohexane-1,3-dione (1.84%). One lipid: (3*S*,8*S*,9*S*, 10*R*,13*R*, 14*S*,17*R*)-17-[(2*R*,5*S*)-5-ethyl-6-methylheptan-2-yl]-10,13-dimethyl-2,3,4,7,8,9,11,12, 14,15,16,17-dodecahydro-1*H*-cyclopenta[a]phenanthren-3-ol (1.76%) (Fig. 2B, Table 2).

Table 4 Phytochemical compounds identified by GC-MS from $CHCl_3$ extract with PTFE membrane filter.

Peak No.	RT (min)	Name of the compound	Molecular Weight (g/mol)	Peak area (%)	Similarity* (%)	Molecular formula	Compound nature
1	10.25	1-ethyl-2,4-dimethylbenzene	134.21	2.51	64	$C_{10}H_{14}$	Aromatic Hydrocarbon
2	10.45	4-ethyl-1,2-dimethylbenzene	134.21	4.24	80	$C_{10}H_{14}$	Aromatic Hydrocarbon
3	11.48	2-ethyl-1,3-dimethylbenzene	134.21	0.83	74	$C_{10}H_{14}$	Aromatic Hydrocarbon
4	22.07	2,4-di*tert*-butylphenol	206.32	5.28	83	$C_{14}H_{22}O$	Aromatic Hydrocarbon
5	36.26	1-(6-methylheptan-2-yl)-4-(4-methylpentyl)cyclohexane	280.53	1.20	53	$C_{20}H_{40}$	Alkane
6	39.22	2-methylpropyl octadecanoate	340.58	10.83	90	$C_{22}H_{44}O_2$	Fatty acid
7	41.77	2-(2-ethylhexoxycarbonyl)benzoic acid	278.34	7.55	53	$C_{16}H_{22}O_4$	Ester

Notes:
* Percentage of similarity to the reference spectrum of the NIST library.
RT, retention time.

The analysis from the $(CH_3)_2CO/CHCl_3$ extract with PVDF membrane filter showed the presence of 23 phytochemical constituents. Three alkanes: tetratriacontane (0.92%), tetrapentacontane (0.93%), icosane (1.34%). Three aromatic hydrocarbons: 4-ethyl-1,2-dimethylbenzene (2.27%), 1,2,4,5-tetramethylbenzene (0.51%), 1,3-benzothiazole (5.33%). Three ketones: 6-hexyloxan-2-one (0.58%), 7,9-di*tert*-butyl-1-oxaspiro[4.5]deca-6,9-diene-2,8-dione (1.61%), 1-(7-hydroxy-8,9-dimethoxy-17-oxa-5,15-diazahexacyclo [13.4.3.$0^{1,16}$.$0^{4,12}$.$0^{6,11}$.$0^{12,16}$]docosa-6,8,10-trien-5-yl)ethanone (0.84%). Two alcohols: (*E*,7*R*,11*R*)-3,7,11,15-tetramethylhexadec-2-en-1-ol (0.45%), (3*Z*,13*Z*)-2-methyloctadeca-3,13-dien-1-ol (0.45%). One aldehyde: tridecanedial (0.58%). Five alkenes: 1,2,3,4-tetramethyl-5-methylidenecyclopenta-1,3-diene (0.82%), (*E*)-octadec-5-ene (0.67%), (*E*)-octadec-9-ene (1.65%), (6*E*,10*E*,14*E*,18*E*)-2,6,10,14,18-pentamethylicosa-2,6,10,14,18-pentaene (4.80%), 17-(5-ethyl-6-methylheptan-2-yl)-10,13-dimethyl-2,7,8,9,11,12,14,15,16,17-decahydro-1*H*-cyclopenta[a]phenanthrene (13.80%). Six fatty acids: methyl hexadecanoate (1.1%), methyl (9*Z*,12*Z*)-octadeca-9,12-dienoate (0.65%), methyl octadecanoate (0.97%), butyl hexadecanoate (4.83%), dioctadecoxy(oxo)phosphanium (0.55%), 2-methylpropyl octadecanoate (3.10%) (Fig. 2C, Table 3).

The analysis from $(CH_3)_2CO/CHCl_3$ extract with PTFE membrane filter seven compounds were observed. One alkane: 1-(6-methylheptan-2-yl)-4-(4-methylpentyl) cyclohexane (1.20%). Four aromatic hydrocarbons: 1-ethyl-2,4-dimethylbenzene (2.51%), 4-ethyl-1,2-dimethylbenzene (4.24%), 2-ethyl-1,3-dimethylbenzene (0.83%), 2,4-di*tert*-butylphenol (5.28%). One ester: 2-(2-ethylhexoxycarbonyl)benzoic acid (7.55%). One fatty acid: 2-methylpropyl octadecanoate (10.83%) (Fig. 2D, Table 4).

The analyses reveled different nature kinds for the identified compounds such as alkanes, aromatic hydrocarbons, esters, ketones halogenated hydrocarbons, alcohols, aldehydes, alkenes, lipids, and fatty acids, some of them with a biological activity previously reported (Tables 5–7). From the identified compounds, 19 shown similarities to

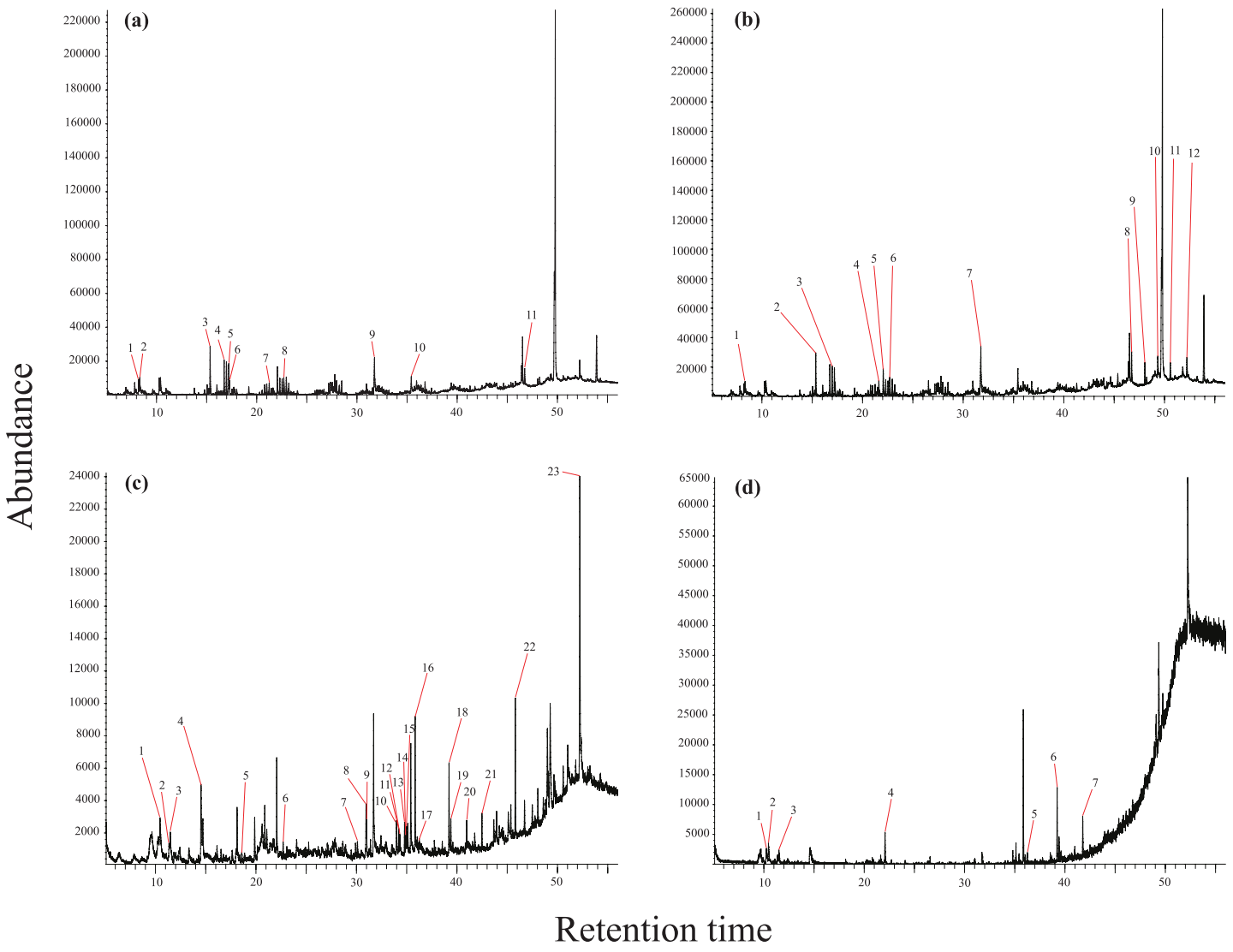


Figure 2 GC-MS chromatograms. (A) MeOH extract with PVDF membrane filter, (B) MeOH extract with PTFE membrane filter, (C) $CHCl_3$ extract with PVDF membrane filter, and (D) $CHCl_3$ extract with PTFE membrane filter. Full-size DOI: 10.7717/peerj-ochem.5/fig-2

Table 5 Number and type of phytochemical compounds in *Opuntia megarrhiza* by extract and membrane filter.

Extract/ Filter	A	Ah	Fa	K	Ak	Al	E	Ad	L	Hh	Total
MeOH/PVDF	5	1	2	1	0	0	1	0	0	1	11
MeOH/PTFE	8	2	0	1	0	0	0	0	1	0	12
$CHCl_3$/PVDF	3	3	6	3	5	2	0	1	0	0	23
$CHCl_3$/PTFE	1	4	1	0	0	0	1	0	0	0	7
Total	17	10	9	5	5	2	2	1	1	1	53

Note:
Alkanes (A), Aromatic hydrocarbons (Ah), Esters (E), Ketones (K), Halogenated hydrocarbons (Hh), Alcohols (Al), Aldehydes (Ad), Alkenes (Ak), Lipids (L), Fatty acids (Fa).

Table 6 Phytochemical compounds with biological activity identified by GC-MS in MeOH extract.

Peak No.	Name of the compound	Molecular weight (g/mol)	Peak area (%)	Molecular formula	Ions (*m/z*)	Compound nature	Biological activity
PVDF filter							
1	heptadecane	240.46	1.03	$C_{17}H_{36}$	57, 71, 85	Alkane	Antifungal (*Adeleye, Daniels & Omadime, 2010*), antimicrobial (*Rahbar, Shafagha & Salimi, 2012*), anti-inflammatory and antioxidative (*Kim et al., 2013*)
2	hexadecanoic acid	256.42	4.67	$C_{16}H_{32}O_2$	57, 73, 129	Fatty acid	Anti-inflammatory (*Aparna et al., 2012*) antialopecic, anti-androgenic, antifibrinolytic, antioxidant, antipsychotic, hemolytic, hypocholesterolemic, nematicide, pesticide and 5-Alpha reductase inhibitor (*USDA, 1992*–2016 [U.S. Department of Agriculture, Agricultural Research Service]). Dr. Duke's)
3	octadecanoic acid	284.47	2.14	$C_{18}H_{36}O_2$	73, 55, 129	Fatty acid	Antibacterial, antifungal and antitumoral (*Gehan et al., 2009*; *Hsouna et al., 2011*)
PTFE filter							
1	heptacosane	380.73	0.98	$C_{27}H_{56}$	57, 71, 85	Alkane	Antioxidant (*Marrufo et al., 2013*)
2	2,4-di*tert*-butylphenol	206.32	2.13	$C_{14}H_{22}O$	191, 57, 206	Aromatic hydrocarbon	Anti-inflammatory, antimicrobial and antioxidant *USDA, 1992*–2016 [U.S. Department of Agriculture, Agricultural Research Service]). Dr. Duke's)
3	pentacosane	352.68	1.66	$C_{25}H_{52}$	57, 71, 85	Alkane	Antimicrobial and antioxidant (*Marrufo et al., 2013*)
4	henicosane	296.57	0.96	$C_{21}H_{44}$	85,71,57	Alkane	Antiasthmatics, urine acidifiers and antimicrobial (*Usha Nandhini, Sangareshwari & Lata, 2015*)
5	triacontane	422.81	1.4	$C_{30}H_{62}$	57, 85, 113	Alkane	Antimicrobial and cytotoxic (*Hsouna et al., 2011*), antidiabetic, antitumor and antibacterial (*Tiagy & Agarwal, 2017*)
6	hentriacontane	436.83	1.63	$C_{31}H_{64}$	57, 85, 113	Alkane	Antibacterial activity (*Olubunmi et al., 2009*), and anti-inflammatory (*Kim et al., 2011*)
7	octacosane	394.76	1.28	$C_{28}H_{58}$	57, 141, 239	Alkane	Antimicrobial and antioxidant (*Jun et al., 2018*)

phytochemical compounds with biological activities previously reported (Tables 6 and 7); their mass spectra resulting from the GC-MS analyses and chemical structures are presented in Figs. S1–S4.

Ten phytochemical compounds shown in Fig. 3 were the most prevailing in the two extracts ($CHCl_3$ and MeOH) and both membrane filters (PVDF and PTFE): Benzene, 1,3-bis(1,1-dimethylethyl) (4.19%) in MeOH/PVDF and (3.53%) in (MeOH/PTFE), hexadecanoic acid (4.67%) in MeOH/PVDF; 1,3-benzothiazole (5.33%), butyl hexadecanoate (4.83%), 2-methylpropyl octadecanoate (3.10%), (6*E*,10*E*,14*E*,18*E*)-2,6,10,14,18-pentamethylicosa-2,6,10,14,18-pentaene (4.80%), and 17-(5-ethyl-6-methylheptan-2-yl)-10,13-dimethyl-2,7,8,9,11,12,14,15,16,17-decahydro-1*H*-cyclopenta[a]phenanthrene (13.80%) in $CHCl_3$/PVDF; and 4-ethyl-1,2-dimethylbenzene (4.24%), 2,4-di*tert*-butylphenol (5.28%) and 2-(2-ethylhexoxycarbonyl)benzoic acid (7.55%) in $CHCl_3$/PTFE (Fig. 3).

Table 7 Phytochemical compounds with biological activity identified by GC-MS in $CHCl_3$ extract.

Peak No.	Name of the compound	Molecular weight (g/mol)	Peak area (%)	Molecular formula	Ions (*m/z*)	Compound nature	Biological activity
PVDF filter							
1	1,3-benzothiazole	135.18	5.33	C_7H_5NS	135, 108, 69	Aromatic Hydrocarbon	Anticonvulsant, anti-inflammatory, antileishmanial, antimicrobial and antitumor (*Siddiqui, Khan & Rana, 2007*)
2	(*E*,7*R*,11*R*)-3,7,11,15-tetramethylhexadec-2-en-1-ol	296.53	0.45	$C_{20}H_{40}O$	55, 71, 81	Alcohol	Antispasmodic (*Pongprayoon et al., 1992*), anticarcinogen *USDA, 1992*–2016 [U.S. Department of Agriculture, Agricultural Research Service]). Dr. Duke's; *Lee, Lee & Park, 1999*; *Hema, Kumaravel & Alagusundaram, 2011*), antitubercular (*Saikia et al., 2010*), antibacterial, antifungal, antimalaria, analgesic and stimulant (*Hema, Kumaravel & Alagusundaram, 2011*), anticonvulsant (*Costa et al., 2012*), anti-inflammatory, antinociceptive (*Okiei et al., 2009*; *Silva et al., 2014*; *Islam et al., 2018*), anxiolytic, cell autophagy and apoptosis inducer metabolism-modulating, cytotoxic and immune-modulating (*Islam et al., 2018*), antimicrobial (*Islam et al., 2018*), and antioxidant (*Mohammad, Omran & Hussein, 2016*; *Islam et al., 2018*)
3	7,9-di*tert*-butyl-1-oxaspiro[4.5]deca-6,9-diene-2,8-dione	276.37	1.61	$C_{17}H_{24}O_3$	57, 175, 217	Ketone	Antioxidant (*USDA, 1992*–2016 [U.S. Department of Agriculture, Agricultural Research Service]). Dr. Duke's)
4	methyl hexadecanoate	270.45	1.1	$C_{17}H_{34}O_2$	74, 143, 227 74, 87, 43, 55, 143	Fatty Acid	Antibacterial (*Agoramoorthy et al., 2007*), antifungal, anti-inflammatory, blood cholesterol decrease (*Hema, Kumaravel & Alagusundaram, 2011*), antioxidant (*Agoramoorthy et al., 2007*; *Hema, Kumaravel & Alagusundaram, 2011*)
5	methyl octadecanoate	298.50	0.97	$C_{19}H_{38}O_2$	143, 74, 55	Fatty Acid	Antimicrobial (*Abubakar & Majinda, 2016*)
6	butyl hexadecanoate	312.53	4.83	$C_{20}H_{40}O_2$	257, 129, 56	Fatty Acid	Antioxidant (*Prakash, Gondwal & Pant, 2011*), and antimicrobial (*Sujatha et al., 2014*)
7	icosane	282	1.34	$C_{20}H_{42}$	113, 85, 57	Alkane	Antibacterial (*Boussaada et al., 2008*; *Hsouna et al., 2011*), antifungal, antitumor and cytotoxic (*Hsouna et al., 2011*)
PTFE filter							
1	2,4-di*tert*-butylphenol	206.32	5.28	$C_{14}H_{22}O$	57, 191, 206	Aromatic hydrocarbon	Anti-inflammatory, antimicrobial and antioxidant *USDA, 1992*–2016 [U.S. Department of Agriculture, Agricultural Research Service]). Dr. Duke's)
2	2-(2-ethylhexoxycarbonyl) benzoic acid	278.34	7.55	$C_{16}H_{22}O_4$	149, 167, 112	Ester	Cytotoxic (*Krishnan, Mani & Jasmine, 2014*)

Identity from five compounds that showed a similarity percentage above 95%, was supported by comparison of their retention times with pure commercial standards (Figs. S5–S9). 1,3-benzothiazole was found at RT of 14.61 min, with ions (*m/z*) of 135 and

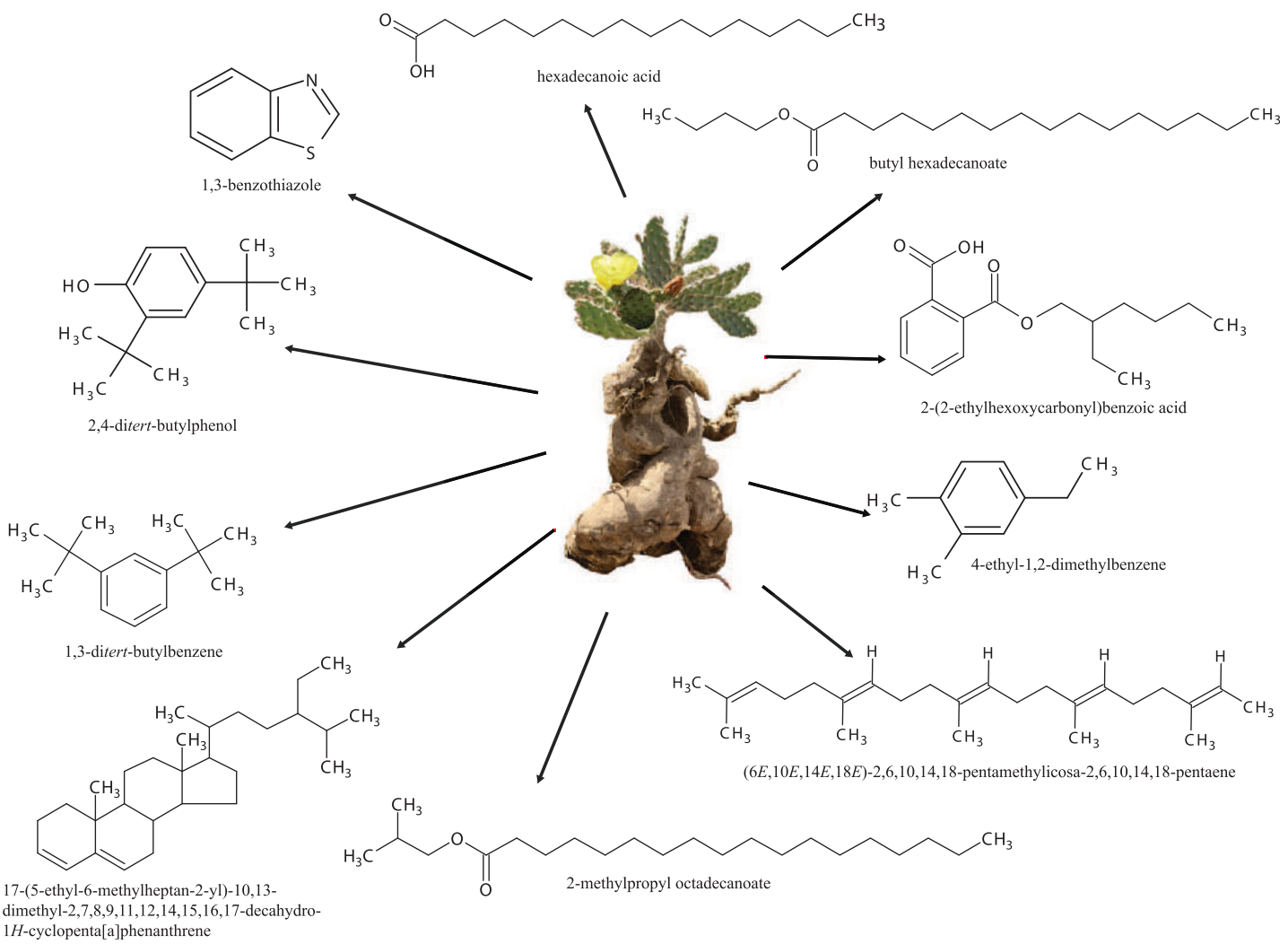


Figure 3 Chemical structures. Prevailing phytochemical compounds of *Opuntia megarrhiza* identified in GC-MS analysis in $CHCl_3$ and MeOH extracts with PVDF and PTFE membrane filters. Full-size DOI: 10.7717/peerj-ochem.5/fig-3

107.9. henicosane at RT of 33.5 min, with ions (*m/z*) of 57, 113. hentriacontane at RT of 48.7 min, with ions (*m/z*) of 57, 85, and 113. methyl hexadecanoate at RT of 30.5 min, with ions (*m/z*) of 74, 143, 227, and 55. triacontane was detected at RT of 33.5 min, with ions (*m/z*) 57, 85 and 113.

Finally, 34 compounds with no identified biological activity were found, eight in the $(CH_3)_2CO$/MeOH extract with PVDF membrane filter (MeOH/PVDF), five in the $(CH_3)_2CO$/MeOH extract with PTFE (MeOH/PTFE), 16 in the $(CH_3)_2CO/CHCl_3$ extract with PVDF ($CHCl_3$/PVDF), and five in the $(CH_3)_2CO/CHCl_3$ extract with PTFE ($CHCl_3$/PTFE).

DISCUSSION

The use of *Opuntia megarrhiza* in traditional medicine in Mexico has been reported previously (*Segura-Venegas & Rendón-Aguilar, 2016*), however, this is the first study that

demonstrate the presence of phytochemical compounds with biological activities. *Opuntia* species are used in the world as local medicinal interventions for chronic diseases and as food sources, mainly because they possess nutritional properties and biological activities that has been recently reviewed (*Aruwa, Amoo & Kudanga, 2018*). Here we report for the first time, the identification of several phytochemical compounds in *O. megarrhiza* with biological activities. Our findings highlight the relevance of this species in developing of new drugs, trough future chemical studies, and encourage of planting this species once this one is listed as endangered in the IUCN Red List.

Biotechnological methods are reliable and provide continuous sources of raw material and natural products for food, pharmaceutical, and cosmetic industries (*Rao & Ravishankar, 2002*; *Nalawade et al., 2003*; *Julsing, Quax & Kayser, 2007*). Previously, it has been indicated that more than 50,000 plant species are used in phytotherapy and medicine, and around 66% of them are harvested from nature leading to local extinction of many species or degradation of their habitats (*Tasheva & Kosturkova, 2012*). Alternatives to protect these useful plants, should be directed to both preservation of the plant populations and elevating the level of knowledge for sustainable utilization of these plants in medicine have been previously indicated (WHO, 2010, http://www.who.int/mediacentre/factsheets/fs134/en/). Biotechnological methods offer possibilities not only for faster cloning and conservation of the genotype of the plants (*Verpoorte, Contin & Memelink, 2002*; *Tripathi & Tripathi, 2003*) but for modification of their gene information, regulation, and expression for production of valuable substances in higher amounts or with better properties (*Rao & Ravishankar, 2002*; *Khan et al., 2009*).

GC-MS is one of the most precise methods to identify various metabolites present in plant extracts (*Fiehn et al., 2000*; *Roessner et al., 2000*; *Roessner et al., 2001*; *Kopka, 2006a*; *Kopka, 2006b*; *Fiehn, 2006*; *Fernie, 2007*; *Saito & Matsuda, 2010*; *Tiago et al., 2016*; *Dinesh-Kumar & Rajakumar, 2018*) since some of these chromatographs include preloaded libraries or databases (NIST and WILEY) that allows to know the possible identity of the metabolites by comparing the resulting mass spectra with those found as reference in these libraries (*Kim et al., 2019*; *Wei et al., 2014*). Several studies indicate that *Opuntia* plants contain different phytochemical groups such as phenolic acids, sterols, esters, coumarins, terpenoids, and alkaloids with several health benefits (*Piattelli, Minale & Prota, 1965*; *Stintzing, Schieber & Carle, 2001*; *Strack, Vogt & Schliemann, 2003*; *Paiz et al., 2010*; *Osorio-Esquivel et al., 2011*; *Aruwa, Amoo & Kudanga, 2018*). However, the nature of the compound extracted depends largely on their solubility in the extraction solvent, the degree of polymerization of the phenols, and the interaction of the phenols with other constituents of the plant (*Choi et al., 2002*; *El Cadi et al., 2020*). But the use of different membrane filters allows to identified chemical compounds with different hydrophobicity and molecule sizes. Previously, it has been indicated that PTFE has less hydrophobic adsorption but more size exclusion (*Xiao et al., 2014*).

In addition, identity of five of the compounds found was corroborated using pure commercial standards. The ions obtained from each of the standards corresponded to those found in the extracts according to the NIST base of the equipment. GC-MS has a library of mass spectra, which makes it easy to obtain compounds that have the most

similar mass to the library spectrum (*Kim et al., 2019*). However, the attribution of a GC-MS chromatographic peak should be confirmed whenever possible by comparison with a standard compound analyzed under the same experimental conditions (*Sturaro, Parvoli & Doretti, 1994*). We identified five compounds in the extracts performed through the use of standards. In this context, the analytical standard is used as a reference in the qualitative, quantitative and identity determinations of a compound, it must also have high purity and stability (*Sun et al., 2015*).

On the other hand, the major phytochemical compounds found in our study have been reported to possess several biological activities. Some alkanes like hentriacontane and triacontane have antibacterial activity (*Boussaada et al., 2008*; *Olubunmi et al., 2009*; *Hsouna et al., 2011*; *Tiagy & Agarwal, 2017*). Heptadecane have antifungal activity (*Adeleye, Daniels & Omadime, 2010*). Icosane has both antibacterial and antifungal activity (*Hsouna et al., 2011*). Henicosane, heptadecane, octacosane, and pentacosane have antimicrobial activity (*Rahbar, Shafagha & Salimi, 2012*; *Marrufo et al., 2013*; *Usha Nandhini, Sangareshwari & Lata, 2015*; *Jun et al., 2018*). Heptadecane and hentriacontane have anti-inflammatory activity (*Kim et al., 2011*; *Kim et al., 2013*). Heptacosane, heptadecane, octacosane, and pentacosane have antioxidant activity (*Kim et al., 2013*; *Marrufo et al., 2013*; *Jun et al., 2018*). Icosane and triacontane have antitumor activity (*Hsouna et al., 2011*; *Tiagy & Agarwal, 2017*). Triacontane has antidiabetic activity (*Tiagy & Agarwal, 2017*). Henicosane has been reported as an antiasthmatic, urine acidifier (*Usha Nandhini, Sangareshwari & Lata, 2015*). Icosane, octadecane, and hexadecanoic acid has been previously identified in *Opuntia stricta* (*Izuegbuna, Otunola & Bradley, 2019*).

Fatty acids like octadecanoic acid have antibacterial or antifungal activity (*Gehan et al., 2009*; *Hsouna et al., 2011*), and it has previously identified in *Opuntia dillenii* (*Ben-Lataief et al., 2020*). The hexadecanoic acid have antialopecic, anti-androgenic, antifibrinolytic, antioxidant, antipsychotic, hemolytic, hypocholesterolemic, nematicide, pesticide, 5-Alpha reductase inhibitor (*USDA, 1992*–2016 [U.S. Department of Agriculture, Agricultural Research Service]), and anti-inflammatory (*Aparna et al., 2012*). Octadecanoic acid have been reported as anticarcinogen or antitumoral (*Hsouna et al., 2011*; *Gehan et al., 2009*). Fatty acid like butyl hexadecanoate, methyl hexadecanoate, and methyl octadecanoate, has antimicrobial activity (*Sujatha et al., 2014*; *Abubakar & Majinda, 2016*). Benzenoids like 2-(2-ethylhexoxycarbonyl)benzoic acid ester has been reported as cytotoxic (*Krishnan, Mani & Jasmine, 2014*). The diterpene (*E*,7*R*,11*R*)-3,7,11,15-tetramethylhexadec-2-en-1-ol has been reported to have multiple activities like anticarcinogen, anticonvulsant, antifungal, anti-inflammatory, antimalaria, antimicrobial, antinociceptive, antioxidant, antitubercular, antispasmodic, anxiolytic, autophagy and apoptosis inducing, cytotoxic, immune-modulating, metabolism-modulating, resistant to gonorrhea, and stimulant (*USDA, 1992*–2016 [U.S. Department of Agriculture, Agricultural Research Service]; *Pongprayoon et al., 1992*; *Lee, Lee & Park, 1999*; *Okiei et al., 2009*; *Saikia et al., 2010*; *Hema, Kumaravel & Alagusundaram, 2011*; *Costa et al., 2012*; *Silva et al., 2014*; *Mohammad, Omran & Hussein, 2016*; *Islam et al., 2018*), and the 2-(2-ethylhexoxycarbonyl)benzoic

acid has anti-inflammatory, antimicrobial, antioxidant, antiviral, and cytotoxicity activities (*Krishnan, Mani & Jasmine, 2014*).

Additionally, phytochemical compounds we found in *Opuntia megarrhiza* with no reports of biological activity, have been previously identified in other *Opuntia* species. For example, (*Z*)-octadec-9-enoic acid and 17-(5-ethyl-6-methylheptan-2-yl)-10,13-dimethyl-2,7,8,9,11,12,14,15,16,17-decahydro-1*H*-cyclopenta[a]phenanthrene was identified in *O. dillenii* (*Ben-Lataief et al., 2020*). Additionally, β-Sitosterol is the major sterol extracted from different parts of the fruit oils of *Opuntia ficus-indica* (*Ramadan & Mörsel, 2003a*, *2003b*). Herein, we identify (3*S*,8*S*,9*S*,10*R*,13*R*,14*S*,17*R*)-17-[(2*R*,5*S*)-5-ethyl-6-methylheptan-2-yl]-10,13-dimethyl-2,3,4,7,8,9,11,12,14,15,16,17-dodecahydro-1*H*-cyclopenta[a]phenanthren-3-ol in *O. megarrhiza*.

CONCLUSIONS

The GC-MS analysis of cladode extracts of *Opuntia megarrhiza* conducted here proves the presence of several phytochemical compounds responsible for biological activities previously reported support the medicinal use of this plant in traditional medicine. In particular, the anti-inflammatory activity in compounds with a high similarity percentage in our results (*e.g.*, hexadecanoic acid, 2,4-di*tert*-butylphenol, hentriacontane, 1,3-benzothiazole, and methyl hexadecanoate) supports its use for treating bone fractures. Hence, *O. megarrhiza* represents a source for finding phytochemical compounds with potential use in medicine and biotechnology. Our results represent an advance in the knowledge of an endangered plant, not previously studied, and with ethnical uses, and support future target studies through the identification of compounds with biotechnological potential using certified standard and additional tools.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Madeleyne Cupido received a grant of CONACYT (Graduate Studies Scholarship 1007054). This research was funded by the international cooperative research of Rural Development Administration (RDA) from Republic of Korea, (Grant PJ012429012016 to Pablo Delgado-Sánchez) and CONACyT (Grant 2014-243454 to JADN). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
CONACYT: 1007054.
International Cooperative Research of Rural Development Administration (RDA), Republic of Korea: PJ012429012016.
CONACyT: 2014-243454.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Madeleyne Cupido conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Arturo De-Nova conceived and designed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, contributed materials and reagents, and approved the final draft.
- María L. Guerrero-González conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, contributed analysis tools, and approved the final draft.
- Francisco Javier Pérez-Vázquez conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, contributed materials and reagents, and approved the final draft.
- Karen Beatriz Méndez-Rodríguez performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Pablo Delgado-Sánchez conceived and designed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, contributed materials and reagents, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data is available in the Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-ochem.5#supplemental-information.

REFERENCES

Abubakar M, Majinda R. 2016. GC-MS analysis and preliminary antimicrobial activity of *Albizia adianthifolia Schumach* and *Pterocarpus angolensis* DC. *Medicines* **3(1)**:1–9 DOI 10.3390/medicines3010003.

Adeleye IA, Daniels FV, Omadime M. 2010. Characterization of volatile components of epa-ijebu: a native wonder cure recipe. *Journal of Pharmacology and Toxicology* **6(1)**:97–100 DOI 10.3923/jpt.2011.97.100.

Agoramoorthy G, Chandrasekaran M, Venkatesalu V, Hsu MJ. 2007. Antibacterial and antifungal activities of fatty acid methyl esters of the blind-your-eye mangrove from India. *Brazilian Journal of Microbiology* **38(4)**:739–742 DOI 10.1590/S1517-83822007000400028.

Andrade-Cetto A, Wiedenfeld H. 2011. Anti-hyperglycemic effect of *Opuntia streptacantha* Lem. *Journal of Ethnopharmacology* **133(2)**:940–943 DOI 10.1016/j.jep.2010.11.022.

Aparna V, Dileep KV, Mandal PK, Karthe P, Sadasivan C, Haridas M. 2012. Anti-inflammatory property of n-hexadecanoic acid: structural evidence and kinetic assessment. *Chemical Biology and Drug Design* **80(3)**:434–439 DOI 10.1111/j.1747-0285.2012.01418.x.

Araújo F, Farias D, Neri-Numa I, Pastore G. 2021. Underutilized plants of the Cactaceae family: nutritional aspects and technological applications. *Food Chemistry* **(362)**:130196 DOI 10.1016/j.foodchem.2021.130196.

Aruwa CE, Amoo SO, Kudanga T. 2018. Opuntia (Cactaceae) plant compounds, biological activities and prospects-A comprehensive review. *Food Research International* **112**:328–344 DOI 10.1016/j.foodres.2018.06.047.

Banakar P, Jayaraj M. 2018. GC-MS analysis of bioactive compounds from ethanolic leaf extract of *Waltheria indica* Linn. and their pharmacological activities. *International Journal of Pharmaceutical Sciences and Research* **9(5)**:2005–2010 DOI 10.13040/IJPSR.0975-8232.9(5).2005-10.

Bargougui A, Le Pape P, Triki S. 2013. Antiplasmodial efficacy of fruit extracts and cladodes of *Opuntia ficus-indica*. *Journal of Natural Sciences Research* **3**:31–37 DOI 10.7176/JNSR.

Barthlott W, Hunt D. 1993. Cactaceae. In: Kubitzki K, Rohmer JG, Bittridi V, eds. *The Families and Genera of Vascular Plants*. Berlin: Springer, 161–197.

Ben-Lataief S, Zourgui MN, Rahmani R, Najjaa H, Gharsallah N, Zourgui L. 2020. Chemical composition, antioxidant, antimicrobial and cytotoxic activities of bioactive compounds extracted from Opuntia dillenii cladodes. *Journal of Food Measurement and Characterization* **15(1)**:782–794 DOI 10.1007/s11694-020-00671-2.

Bensadón S, Hervert-Hernández D, Sáyao-Ayerdi SG, Goñi I. 2010. By-products of *Opuntia ficus-indica* as a source of antioxidant dietary fiber. *Plant Foods for Human Nutrition* **65(3)**:210–216 DOI 10.1007/s11130-010-0176-2.

Boussaada O, Ammar S, Saidana D, Chriaa J, Chraif I, Daami M, Helal AN, Mighri Z. 2008. Chemical composition and antimicrobial activity of volatile components from capitula and aerial parts of *Rhaponticum acaule* DC growing wild in Tunisia. *Microbiological Research* **16(1)**:87–95 DOI 10.1016/j.micres.2007.02.010.

Bravo-Hollis H, Sánchez-Mejorada H. 1991. *Las Cactáceas de México*. México: Universidad Nacional Autónoma de México.

Bravo-Hollis H, Scheinvar L. 1999. *El interesante mundo de las cactáceas*. México D.F.: Fondo de Cultura Económica.

Caballero-Gutiérrez L, Gonzáles G. 2016. Alimentos con efecto anti-inflamatorio. *Acta Medica Peruana* **33(1)**:1–50 DOI 10.35663/amp.2016.331.18.

Choi CW, Kim SC, Hwang SS, Choi BK, Ahn HJ, Lee MY, Park SH, Kim SK. 2002. Antioxidant activity and free radical scavenging capacity between Korean medicinal plants and flavonoids by assay-guided comparison. *Plant Science* **163(6)**:1161–1168 DOI 10.1016/S0168-9452(02)00332-1.

Costa JP, Ferreira PB, Sousa DP, Jordan J, Freitas RM. 2012. Anticonvulsant effect of phytol in a pilocarpine model in mice. *Neuroscience Letters* **523(2)**:115–118 DOI 10.1016/j.neulet.2012.06.055.

De Corato U, Maccioni O, Trupo M, Di Sanzo G. 2010. Use of essential oil of *Laurus nobilis* obtained by means of a supercritical carbon dioxide technique against post harvest spoilage fungi. *Journal of Crop Protection* **29(2)**:142–147 DOI 10.1016/j.cropro.2009.10.012.

Dinesh-Kumar G, Rajakumar R. 2018. GC-MS analysis of bioactive compounds from ethanolic leaves extract of *Eichhornia crassipes* (Mart) Solms. and their pharmacological activities. *Journal of Pharmaceutical Innovation* **7**:459–462 DOI 10.22271/tpi.

El Cadi H, El Cadi A, Kounnoun A, Oulad El Majdoub Y, Lovillo MP, Brigui J, Dugo P, Mondello L, Cacciola F. 2020. Wild strawberry (Arbutus unedo): phytochemical screening and antioxidant properties of fruits collected in northern Morocco. *Arabian Journal of Chemistry* **13(8)**:6299–6311 DOI 10.1016/j.arabjc.2020.05.022.

Ennouri M, Fetoui H, Bourret E, Zeghal N, Attia H. 2006. Evaluation of some biological parameters of *Opuntia ficus indica*. 1. Influence of a seed oil supplemented diet on rats. *Bioresource Technology* **97(12)**:1382–1386 DOI 10.1016/j.biortech.2005.07.010.

Estrada-Castillón E, Soto-Mata B, Garza-López M, Villarreal-Quintanilla J, Jiménez-Pérez J, Pando-Moreno M, Sánchez-Salas J, Scott-Morales L, Cotera-Correa M. 2012. Medicinal plants in the southern region of the State of Nuevo León, México. *Journal of Ethnobiology and Ethnomedicine* **11(1)**:8–45 DOI 10.1186/1746-4269-8-45.

Fernie AR. 2007. The future of metabolic phytochemistry: larger numbers of metabolites, higher resolution, greater understanding. *Phytochemistry* **68(22–24)**:2861–2880 DOI 10.1016/j.phytochem.2007.07.010.

Feugang JM. 2006. Nutritional and medicinal use of Cactus pear (*Opuntia* spp.) cladodes and fruits. *Frontiers in Bioscience* **11(1)**:2574–2589 DOI 10.2741/1992.

Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L. 2000. Metabolite profiling for plant functional genomics. *Nature Biotechnology* **18(11)**:1157–1161 DOI 10.1038/81137.

Fiehn O. 2001. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics* **2(3)**:155–168 DOI 10.1002/cfg.82.

Fiehn O. 2002. Metabolomics the link between genotypes and phenotypes. *Plant Molecular Biology* **48(1/2)**:155–171 DOI 10.1023/A:1013713905833.

Fiehn O. 2006. Metabolite profiling in *Arabidopsis*. In: Totowa NJ, Salinas J, Sanchez-Serrano JJ, eds. *Arabidopsis Protocols: Methods in Molecular Biology*. Totowa: Humana Press, 439–447.

Fiehn O. 2016. Metabolomics by gas chromatography-mass spectrometry: combined targeted and untargeted profiling. *Current Protocols in Molecular Biology* **(114)**:30.4.1–30.4.32 DOI 10.1002/0471142727.mb3004s114.

Gehan MA, Hanan AE, Hassan AH, Okbah MA. 2009. Marine natural products and their potential applications as antiinfective agents. *World Applied Sciences Journal* **7(7)**:872–880.

Ghasemzadeh A, Ghasemzadeh N. 2011. Flavonoids and phenolic acids: role and biochemical activity in plants and human. *Journal of Medicinal Plant Research* **5(31)**:6696–6703 DOI 10.5897/jmpr11.1404.

Gil-Chávez JG, Villa JA, Ayala-Zavala FJ, Heredia JB, Sepulveda D, Yahia EM, González-Aguilar GA. 2013. Technologies for extraction and production of bioactive compounds to be used as nutraceuticals and food ingredients: an overview. *Comprehensive Reviews in Food Science and Food Safety* **12(1)**:5–23 DOI 10.1111/1541-4337.12005.

González-Durán A, Riojas L, Arreola N. 2001. *El género Opuntia en Jalisco. Guía de campo*. México: Universidad de Guadalajara-Comisión Nacional para el Conocimiento y Uso de la Biodiversidad.

Gurrieri S, Miceli L, Lanza CM, Tomaselli F, Bonomo RP, Rizzarelli E. 2000. Chemical characterization of Sicilian prickly pear (*Opuntia ficus indica*) and perspectives for the storage of its juice. *Journal of Agricultural and Food Chemistry* **48(11)**:5424–5431 DOI 10.1021/jf9907844.

Harlev E, Nevo E, Lansky EP, Lansky S, Bishayee A. 2012. Anticancer attributes of desert plants. *Anti-Cancer Drugs* **23(3)**:255–271 DOI 10.1097/cad.0b013e32834f968c.

Harrigan G, Goodacre R. 2012. *Metabolic profiling: its role in biomarker discovery and gene function analysis*. New York: Springer.

Hema R, Kumaravel S, Alagusundaram K. 2011. GC/MS determination of bioactive components of *Murraya koenigii*. *American Journal of Science* **7**:80–83.

Hernández H, Godínez AH. 1994. Contribución al conocimiento de las cactáceas mexicanas amenazada. *Acta Botánica Mexicana* **26(26)**:33–52 DOI 10.21829/abm26.1994.690.

Hernández H, Gómez-Hinostrosa C, Bárcenas R. 2001. Studies on Mexican Cactaceae. II. *Opuntia megarrhiza*, a poorly known endemic from San Luis Potosí, Mexico. *Brittonia* **53(4)**:528–533 DOI 10.1007/bf02809653.

Hernández HM, Gómez-Hinostrosa C, Goettsch BK, Sotomayor M. 2013. Opuntia megarrhiza. *The IUCN Red List of Threatened Species 2013* e.T41219A2952324 DOI 10.2305/IUCN.UK.2013-1.RLTS.T41219A2952324.

Hernández-Hernández T, Hernández HM, De-Nova JA, Puente R, Eguiarte L, Magallón S. 2011. Phylogenetic relationships and evolution lineages within Cactaceae. *American Journal of Botany* **98(1)**:44–61 DOI 10.3732/ajb.1000129.

Hsouna AB, Trigui M, Mansour RB, Jarraya RM, Damak M, Jaoua S. 2011. Chemical composition, cytotoxicity effect and antimicrobial activity of Ceratonia siliqua essential oil with preservative effects against Listeria inoculated in minced beef meat. *International Journal of Food Microbiology* **148(1)**:66–72 DOI 10.1016/j.ijfoodmicro.2011.04.028.

Hunt D, Taylor NP, Charles G. 2006. *The new cactus lexicon*. Milborne Port: DH Books.

Islam MT, Ali ES, Uddin SJ, Shaw S, Islam MA, Ahmed MI, Chandra Shill M, Karmakar UK, Yarla NS, Khan IN, Billah MM, Pieczynska MD, Zengin G, Malainer C, Nicoletti F, Gulei D, Berindan-Neagoe I, Apostolov A, Banach M, Yeung A, El-Demerdash A, Xiao J, Dey P, Yele S, Jóźwik A, Strzałkowska N, Marchewka J, Rengasamy K, Horbańczuk J, Amjad-Kamal M, Mubarak M, Mishra S, Shilpi J, Atanasov A. 2018. Phytol: a review of biomedical activities. *Food and Chemical Toxicology* **121(6)**:82–94 DOI 10.1016/j.fct.2018.08.032.

Izuegbuna O, Otunola G, Bradley G. 2019. Chemical composition, antioxidant, anti-inflammatory, and cytotoxic activities of *Opuntia stricta* cladodes. *PLOS ONE* **14(1)**:e0209682 DOI 10.1371/journal.pone.0209682.

Julsing KM, Quax WJ, Kayser O. 2007. The engineering of medicinal plants: prospects and limitations of medicinal plant biotechnology. In: Kayser O, Quax WJ, eds. *Medicinal Plant Biotechnology: from Basic Research to Industrial Applications*. New York: Wiley, 1–8.

Jun M, Rui-Rui X, Yao L, Di-Feng R, Jun L. 2018. Composition, antimicrobial and antioxidant activity of supercritical fluid extract of *Elsholtzia ciliata*. *Journal of Essential Oil Bearing Plants* **21(2)**:556–562 DOI 10.1080/0972060X.2017.1409657.

Khan MY, Aliabbas S, Kumar V, Rajkumar S. 2009. Resent advances in medcinal plant biotehology. *Indian Journal of Biotechnology* **8**:9–22.

Kim DH, Park MH, Choi YJ, Chung KW, Park CH, Jang EJ, An HJ, Yu BP, Chung HY. 2013. Molecular study of dietary heptadecane for the anti-inflammatory modulation of NF-kB in the aged kidney. *PLOS ONE* **8(3)**:1–10 DOI 10.1371/journal.pone.0059316.

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. 2019. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research* **49**:D1388–D1395 DOI 10.1093/nar/gkaa971.

Kim S, Chung W, Kim S, Ko S, Um J. 2011. Antiinflammatory effect of *Oldenlandia diffusa* and its constituent, hentriacontane, through suppression of Caspase-1 activation in mouse peritoneal macrophages. *Phytotherapy Research* **25(10)**:1537–1546 DOI 10.1002/ptr.3443.

Kopka J. 2006a. Current challenges and developments in GC-MS based metabolite profiling technology. *Journal of Biotechnology* **124(1)**:312–322 DOI 10.1016/j.jbiotec.2005.12.012.

Kopka J. 2006b. Gas chromatography mass spectrometry. In: Saito K, Dixon RA, Willmitzer L, eds. *Plant Metabolomics.* Vol. 57. Berlin: Springer, 3–20.

Kris-Etherton PM, Hecker KD, Bonanome A, Coval SM, Binkoski AE, Hilpert KF, Etherton TD. 2002. Bioactive compounds in foods: their role in the prevention of cardiovascular disease and cancer. *The American Journal of Medicine* **113**:71–88 DOI 10.1016/s0002-9343(01)00995-0.

Krishnan K, Mani A, Jasmine S. 2014. Cytotoxic activity of bioactive Compound 1, 2-benzene dicarboxylic acid, mono 2-ethylhexyl ester extracted from a marine derived Streptomyces sp. VITSJK8. *International Journal of Molecular and Cellular Medicine* **3**:246–254.

Kudanga T, Nemadziva B, Le Roes-Hill M. 2017. Laccase catalysis for the synthesis of bioactive compounds. *Applied Microbiology and Biotechnology* **101(1)**:13–33 DOI 10.1007/s00253-016-7987-5.

Lee KL, Lee SH, Park K. 1999. Anticancer activity of phytol and eicosatrienoic acid identified from Perilla leaves. *Journal of Ethnopharmacology* **28**:1107–1112.

Mangas-Marín R, Montes de Oca PR, Herrera PM, Bello AA, Hernández BI, Menéndez S, Lopez M, Paz T, Rodeiro GI. 2018. GC/MS analysis and bioactive properties of extracts obtained from Clusia minor L. leaves. *Journal of the Mexican Chemical Society* **62(4)**:177–188 DOI 10.29356/jmcs.v62i4.544.

Marquet P. 2012. LC-MS vs. GC-MS, online extraction systems, advantages of technology for drug screening assays. In: Langman L, Snozek C, eds. *LC-MS in drug analysis.* Vol. 902. Totowa: Humana Press, 15–27.

Marrufo T, Nazzaro F, Mancini E, Fratianni F, Coppola R, De Martino L, Agostinho AB, De Feo V. 2013. Chemical composition and biological activity of the essential oil from leaves of *Moringa oleifera* Lam. cultivated in Mozambique. *Molecules* **18(9)**:10989–11000 DOI 10.3390/molecules180910989.

Martins S, Mussatto SI, Martínez-Avila G, Montañez-Saenz J, Aguilar CN, Teixeira JA. 2011. Bioactive phenolic compounds: production and extraction by solid-state fermentation. *Biotechnology Advances* **29(3)**:365–373 DOI 10.1016/j.biotechadv.2011.01.008.

Mohammad G, Omran AM, Hussein H. 2016. Antibacterial and phytochemical analysis of *Piper nigrum* using gas chromatography mass spectrum and Fourier-transform infrared spectroscopy. *International Journal of Pharmacognosy and Phytochemical Research* **8**:977–996.

Nalawade SM, Sagare AP, Lee CY, Kao CL, Tsay HS. 2003. Studies on tissue culture of Chinese medicinal plant resources in Taiwan and their sustainable utilization. *Botanical Bulletin of Academia Sinica* **44(2)**:79–98 DOI 10.1079/IVP2003504.

Okiei W, Ogunlesi M, Ofor E, Osibote EA. 2009. Analysis of essential oil constituents in hydro-distillates of *Calotropis procera* (Ait.). *Research Journal of Phytochemistry* **33(3)**:44–53 DOI 10.3923/rjphyto.2009.44.53.

Olubunmi A, Gabriel OA, Stephen AO, Scott FO. 2009. Antioxidant and antimicrobial activity of cuticular wax from *Kigelia africana*. *Fabad Journal of Pharmaceutical Sciences* **34**:187–194.

Osorio-Esquivel O, Ortiz-Moreno A, Álvarez V, Dorantes-Álvarez L, Giusti M. 2011. Phenolics, betacyanins and antioxidant activity in *Opuntia joconostle* fruits. *International Food Research Journal* **44(7)**:2160–2168 DOI 10.1016/j.foodres.2011.02.011.

Paiz R, Juárez-Flores B, Rogelio J, Rivera JR, Cecilia N, Ortega C, Reyes-Agüero JA, Garcia E, Alvarez G. 2010. Glucose-lowering effect of xoconostle (*Opuntia joconostle* A. Web., Cactaceae) in diabetic rats. *Journal of Medicinal Plants Research* **4**:2326–2333 DOI 10.5897/JMPR.

Park EH, Kahng JH, Lee SH, Shin KH. 2001. An anti-inflammatory principle from cactus. *Fitoterapia* **72(3)**:288–290 DOI 10.1016/s0367-326x(00)00287-2.

Patra S, Nayak R, Patro S, Pradhan B, Sahu B, Behera C, Bhutia S, Jena M. 2021. Chemical diversity of dietary phytochemicals and their mode of chemoprevention. *Biotechnology Reports* **(30)**:e00633 DOI 10.1016/j.btre.2021.e00633.

Pawar AV, Killedar SG, Dhuri VG. 2017. *Opuntia*: medicinal plant. *International Journal of Advance Research, Ideas and Innovations in Technology* **3**:148–154.

Perfumi M, Tacconi R. 1996. Antihyperglycemic effect of fresh *Opuntia dillenii* fruit from tenerife (Canary Islands). *International Journal of Pharmacognosy* **34(1)**:41–47 DOI 10.1076/phbi.34.1.41.13186.

Piattelli M, Minale L, Prota G. 1965. Pigments of Centrospermae III: Betaxanthins from *Beta vulgaris* L. *Phytochemistry* **4(1)**:121–125 DOI 10.1016/S0031-9422(00)86153-1.

Pichersky E, Gang D. 2000. Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends in Plant Science* **5(10)**:439–445 DOI 10.1016/s1360-1385(00)01741-6.

Pongprayoon U, Baeckström P, Jacobsson U, Lindström M, Bohlin L. 1992. Antispasmodic activity of beta-damascenone and e-phytol isolated from *Ipomoea pes-caprae*. *Planta Medica* **58(01)**:19–21 DOI 10.1055/s-2006-961381.

Prakash O, Gondwal M, Pant AK. 2011. Essential oils composition and antioxidant activity of water extract from seeds and fruit pulp of *Skimmia anquetilia* N.P. Taylor & Airy Shaw. *Journal of Asian Natural Products Research* **2**:435–441.

Rahbar N, Shafagha A, Salimi F. 2012. Antimicrobial activity and constituents of the hexane extracts from leaf and stem of *Origanum vulgare* L. sp. *viride* (Boiss.) Hayek. Growing wild in Northwest Iran. *Journal of Medicinal Plants Research* **6**:2681–2685 DOI 10.5897/JMPR11.1768.

Ramadan MF, Mörsel JT. 2003a. Oil cactus pear (Opuntia ficus-indica L.). *Food Chemistry* **82(3)**:339–345 DOI 10.1016/s0308-8146(02)00550-2.

Ramadan MF, Möersel JT. 2003b. Lipid profile of prickly pear pulp fractions. *Journal of Food, Agriculture and Environment* **1**:66–70.

Rao SR, Ravishankar GA. 2002. Plant cell cultures: chemical factories of secondary metabolites. *Biotechnology Advances* **20(2)**:101–153 DOI 10.1016/s0734-9750(02)00007-1.

Reyes-Agüero JA, Aguirre R, Carlín F. 2004. Análisis preliminar de la variación morfológica de 38 variantes mexicanas de Opuntia ficus-indica (L.). In: Miller en Esparza G, Valdez R, Méndez G, eds. *El nopal, Tópicos de actualidad*. Chapingo: Univ. Autónoma Chapingo - Colegio de Postgraduados, 21–47.

Reyes-Agüero JA, Aguirre-Rivera J, Hernández M. 2005. Notas sistemáticas y descripción detallada de *Opuntia ficus-indica* (L.) Mill. (Cactaceae). *Agrociencia* **39**:395–408.

Robertson DG. 2005. Metabonomics in toxicology: a review. *Toxicological Sciences* **85(2)**:809–822 DOI 10.1093/toxsci/kfi102.

Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie A. 2001. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13(1)**:11–29 DOI 10.1105/tpc.13.1.11.

Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L. 2000. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant Journal* **23(1)**:131–142 DOI 10.1046/j.1365-313x.2000.00774.x.

Saikia D, Parihar S, Chanda D, Ojha S, Kumar JK, Chanotiya CS, Shanker K, Negi AS. 2010. Antitubercular potential of some semisynthetic analogues of phytol. *Bioorganic & Medicinal Chemistry Letters* **20(2)**:508–512 DOI 10.1016/j.bmcl.2009.11.107.

Saito K, Matsuda F. 2010. Metabolomics for functional genomics, systems biology, and biotechnology. *Annual Review of Plant Biology* **61(1)**:463–489 DOI 10.1146/annurev.arplant.043008.092035.

Segura-Venegas D, Rendón-Aguilar B. 2016. *Opuntia megarrhiza* Rose (Cactaceae) en San Luis Potosí, México: Uso tradicional y distribución de nuevas poblaciones. *Cactáceas y Suculentas Mexicanas* **62**:36–47.

Serra A, Poejo T, Matias J, Bronze A, Duarte C. 2013. Evaluation of *Opuntia* spp. derived products as antiproliferative agents in human colon cancer cell line (HT29). *Food Research International* **54(1)**:892–901 DOI 10.1016/j.foodres.2013.08.043.

Shedbalkar UU, Adki VS, Jadhav JP, Bapat VA. 2010. *Opuntia* and other cacti: applications and biotechnological insights. *Tropical Plant Biology* **3(3)**:136–150 DOI 10.1007/s12042-010-9055-0.

Siddiqui N, Khan SA, Rana A. 2007. Benzothiazoles: a new profile of biological activities. *Indian Journal of Pharmaceutical Sciences* **69**:10 DOI 10.4103/0250-474x.32100.

Silva R, Sousa F, Damasceno S, Carvalho N, Silva V, Oliveira F, Sousa D, Aragão K, Barbosa A, Freitas R, Medeiros JV. 2014. Phytol a diterpene alcohol, inhibits the inflammatory response by reducing cytokine production and oxidative stress. *Fundamental & Clinical Pharmacology* **28(4)**:455–464 DOI 10.1111/fcp.12049.

Sim KS, Sri-Nurestri AM, Sinniah SK, Kim KH, Norhanom AW. 2010. Acute oral toxicity of *Pereskia bleo* and *Pereskia grandifolia* in mice. *Pharmacognosy Magazine* **6(21)**:67–70 DOI 10.4103/0973-1296.59969.

Stintzing F, Schieber A, Carle R. 2001. Phytochemical and nutritional significance of cactus pear. *European Food Research and Technology* **212(4)**:396–407 DOI 10.1007/s002170000219.

Strack D, Vogt T, Schliemann W. 2003. Recent advances in betalain research. *Phytochemistry* **62(3)**:247–269 DOI 10.1016/S0031-9422(02)00564-2.

Sturaro A, Parvoli G, Doretti L. 1994. Standards and GC-MS analysis: an answer to the requirement of compound confirmation. *Chromatographia* **38(3–4)**:239–241 DOI 10.1007/BF02290344.

Sujatha M, Karthika K, Sivakamasundari S, Mariajancyrani J, Chandramohan G. 2014. GC-MS analysis of phytocomponents and total antioxidant activity of hexane extract of Sinapis alba. *International Journal of Pharmaceutical, Chemical and Biological Sciences* **4**:112–117.

Sun W, Tong L, Li D, Huang J, Zhou S, Sun H, Bi K. 2015. Selection of reference standard during method development using the analytical hierarchy process. *Journal of Pharmaceutical and Biomedical Analysis* **107**:280–289 DOI 10.1016/j.jpba.2015.01.006.

Tasheva K, Kosturkova G. 2012. The role of biotechnology for conservation and biologically active substances production of *Rhodiola rosea*: endangered medicinal species. *The Scientific World Journal* **2012**:274942 DOI 10.1100/2012/274942.

Tiago FJ, Rodrigues JA, Caldana C, Schmidt R, Van Dongen JT, Thomas-Oates J, António C. 2016. Massspectrometry-based plant metabolomics: metabolite responses to abiotic stress. *Mass Spectrometry Reviews* **35(5)**:620–649 DOI 10.1002/mas.21449.

Tiagy T, Agarwal M. 2017. Phytochemical screening and GC-MS analysis of bioactive constituents in the ethanolic extract of *Pistia stratiotes* L. and *Eichhornia crassipes* (Mart.) solms. *Journal of Pharmacognosy and Phytochemistry* **6**:195–206 DOI 10.22271/phyto.

Tripathi L, Tripathi JN. 2003. Role of biotechnology in medecinal plants. *Tropical Journal of Pharmaceutical Research* **2(2)**:243–253.

USDA. 1992. [U.S. Department of Agriculture, Agricultural Research Service] Dr. Duke's Phytochemical and Ethnobotanical Databases. *Available at https://phytochem.nal.usda.gov/*.

Usha Nandhini S, Sangareshwari S, Lata K. 2015. Gas chromatography-mass spectrometry analysis of bioactive constituents from the marine Streptomyces. *Asian Journal of Pharmaceutical and Clinical Research* **8**:244–246.

Velmurugan G, Anand P. 2017. GC-MS analysis of bioactive compounds on ethanolic leaf extract of *Phyllodium pulchellum* L. desv. *International Journal of Pharmacognosy and Phytochemical Research* **1(1)**:114–118 DOI 10.25258/ijpapr.v9i1.8051.

Verpoorte R, Contin A, Memelink J. 2002. Biotechnology for the production of plant secondary metabolites. *Phytochemistry Reviews* **1(1)**:13–25 DOI 10.1023/A:1015871916833.

Wei X, Koo I, Kim S, Zhang X. 2014. Compound identification in GC-MS by simultaneously evaluating the mass spectrum and retention index. *The Analyst* **139(10)**:2507–2514 DOI 10.1039/C3AN02171H.

Weli A, Al-Kaabi A, Al-Sabahi J, Said S, Hossain MA, Al-Riyami S. 2019. Chemical composition and biological activities of the essential oils of Psidium guajava leaf. *Journal of King Saud University - Science* **31(4)**:993–998 DOI 10.1016/j.jksus.2018.07.021.

Xiao K, Sun J, Mo Y, Fang Z, Liang P, Huang X, Ma B. 2014. Effect of membrane pore morphology on microfiltration organic fouling: PTFE/PVDF blend membranes compared with PVDF membranes. *Desalination* **343**:217–225 DOI 10.1016/j.desal.2013.09.026.

Yahia E, Mondragon-Jacobo C. 2011. Nutritional components and anti-oxidant capacity of ten cultivars and lines of cactus pear fruit (*Opuntia* spp.). *Food Research International* **44(7)**:2311–2318 DOI 10.1016/j.foodres.2011.02.042.

Yeddes N, Chérif J, Guyot S, Baron A, Trabelsi-Ayadi M. 2013a. Phenolic profile of Tunisian *Opuntia ficus-indica* thornless form flowers via chromatographic and spectral analysis by reversed phase-high performance liquid chromatography-UV-photodiode array and electrospray ionization-mass spectrometer. *International Journal of Food Properties* **17**:741–751 DOI 10.1080/10942912.2012.665404.

Yeddes N, Chérif J, Guyot S, Sotin H, Ayadi M. 2013b. Comparative study of antioxidant power, polyphenols, flavonoids and Betacyanins of the peel and pulp of three Tunisian *Opuntia* forms. *Antioxidants* **2(2)**:37–51 DOI 10.3390/antiox2020037.

Yu M, Gouvinhas I, Rocha J, Barros A. 2021. Phytochemical and antioxidant analysis of medicinal and food plants towards bioactive food and pharmaceutical resources. *Scientific Reports* **(11)**:10041 DOI 10.1038/s41598-021-89437-4.

Abubidentin A, New Oleanane-type Triterpene Ester from *Abutilon bidentatum* and its antioxidant, cholinesterase and antimicrobial activities

Gadah A. Al-Hamoud[1] Nawal M. Al-Musayeib[1] Musarat Amina[1]
Sabrin R.M. Ibrahim[2]

[1] Pharmacognosy, College of Pharmacy, King Saud University, Riyadh, Saudi Arabia
[2] Pharmacognosy, Faculty of Pharmacy, Assiut University, Assiut, Egypt

ABSTRACT

Background. This work describes the phytochemical and biological investigation of aerial parts of *Abutilon bidentatum* Hochst. Of Saudi origin.

Methodology. Petroleum ether fraction of ethanolic extract *A. bidentatum* was fractionated on a silica gel column and further purified with different chromatographic procedures for the isolation of chemical compounds. The chemical structures of all the pure isolated compounds were elucidated by the interpretation of their spectral data using IR, UV, ^{1}H, ^{13}C NMR, and MS spectroscopy and chemical methods (alkaline hydrolysis) as well as comparison with data reported in the literature. The extract and isolated compounds were evaluated for antioxidant, cholinesterase inhibitory, and antimicrobial activities.

Results. A new oleanane-type triterpene ester, namely abubidentin A (**3**) (α, 3β, 30-trihydroxy-29-carboxy-olean-9(11), 12-diene-3-dotriacontanoate), along with two known compounds: 2-hydroxydocosanoic acid (**1**) and stigmasta-22-ene-3-β-ol (**2**) were isolated from the aerial parts of *Abutilon bidentatum* Hochst. (Malvaceae). Concerning the biological potential, the abubidentinA displayed antioxidant, cholinesterase inhibitory and antimicrobial activities. AbubidentinA possessed strong antioxidant activity against DPPH and $ABTS^{+}$ radical scavenging assays. This new triterpene exhibited high inhibition against acetylcholinesterase (IC_{50} 38.13 ± 0.07 μgmL^{-1}) and butyrylcholinesterase (IC_{50} 32.68 ± 0.37 μgmL^{-1}). Abubidentin A displayed promising antimicrobial activity against *Escherichia coli*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus* (125–150 μgmL^{-1}).

Conclusion. These findings suggest *A. bidentatum* can contribute as a source of new biologically active compounds, especially antioxidants and antimicrobial agents.

Subjects Biochemistry, Biomolecules, Bioorganic Chemistry, Natural Products
Keywords *Abutilon bidentatum*, Malvaceae, Abubidentin A, Oleanane-type triterpene, Antioxidant, Cholinesterase, Antimicrobial

Submitted 25 October 2021
Accepted 9 February 2022
Published 8 March 2022

Corresponding authors
Gadah A. Al-Hamoud, galhamoud@ksu.edu.sa
Nawal M. Al-Musayeib, nalmusayeib@ksu.edu.sa

Academic editor
Carlos Fernández Marcos

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj.13040**

How to cite this article A. Al-Hamoud G, Al-Musayeib NM, Amina M, Ibrahim SRM. 2022. Abubidentin A, New Oleanane-type Triterpene Ester from *Abutilon bidentatum* and its antioxidant, cholinesterase and antimicrobial activities. *PeerJ* **10**:e13040
http://doi.org/10.7717/peerj.13040

INTRODUCTION

Abutilon Miller is the extensive genus of family Malvaceae that contains around 150 species of annual or perennial herbs, shrubs or undershrubs, and small trees. This plant specie is native to temperate, subtropical, and tropical areas of Africa, Asia, America and Australia (*Arbat, 2012*). The word Abutilon is the Greek ancient name of the mulberry tree and has been assigned to genus due to its close similarity to the morphology of leaves. The genus Abutilon showed the existence of precious insoluble fibers isolated from different species of this genus and has noteworthy importance (*Gomaa et al., 2016*). The species of Abutilon genus have been used to cure specific health issues including; rheumatoid arthritis, diuretic and demulcent due to the occurrence of high mucilage content (*Ali et al., 2014*; *Baquar, 1989*). Numerous important pharmacological attributes such as hepatoprotective, antioxidant, analgesic, antipyretic, anti-inflammatory, antimicrobial, anticancer, diuretic, anti-hyperglycemic, CNS activity, immunostimulant, anti-hyperlipidemias, anti-hypertensive, antidiarrheal, anti-urolithiatic, and wound healing activities have been assigned to the plant species of this genus (*Khadabadi & Bhajipale, 2010*; *Agrawal, 2017*). Phytochemical studies claimed that the genus contain different secondary metabolites including phenolic acids, flavonoids, triterpenes, sterols, coumarins, quinones, alkaloids, anthocyanins, iridoids, saponins, megastigmanes, and fatty acids (*Gomaa et al., 2018*). In Saudi Arabia, it is represented by five species, namely *A. bidentatum*, *A. figarianum*, *A. fruticosum*, *A. hirtum*, and *A. pannosum* (*Migahid, 1978*).

Abutilon bidentatum Hochst. locally known as Ren-Umbro is an erect shrub, 1–2 m in height with stellate pubescent branches mixed with long simple hairs. Leaves broadly ovate, long-petiolate blade up to c. $12(-17) \times 10(-13)$ cm, deeply cordate at base, sparse, stellate tomentose, serrate, acuminate at apex, velvety on both sides with serrate-dentate margins. Flowers are present in the leaf axils or on a shoot axillary. Seeds are papillose, 2.5 mm long stellately spreading, and black. Different parts of *A.bidentatum* have been utilized to treat several aliments in ethno-medicine, especially its leaves are used to cure infections. Root powder is used in folk-traditional medicine to treat dysentery, colitis, and diarrhea (*Al-Shanwani, 1996*). Literature scrutiny revealed that this plant remained unexplored for phytochemical and biological properties. Only three studies could be traced concerning the chemical constituents and pharmacological activities till date. These studies reported the antibacterial, antioxidant (*Survase, Jamdhade & Chavan, 2012*; *Shahwar et al., 2010*) and hepatoprotective potential of aerial parts of *A. bidentatum* (*Yasmin, Akram Kashmiri & Anwar, 2011*). However, a single study with respect to its chemical constituents, reporting the isolation of cholestane derivative from the *A. bidentatum* has been reported in the literature (*Jain, Jain & Arora, 1996*). The present study reports the isolation and structural characterization of a new oleanane-type triterpene ester: abubidentin A (**3**, Fig. 1) and two known compounds (**1** and **2**) from the aerial parts of *A. bidentatum*.

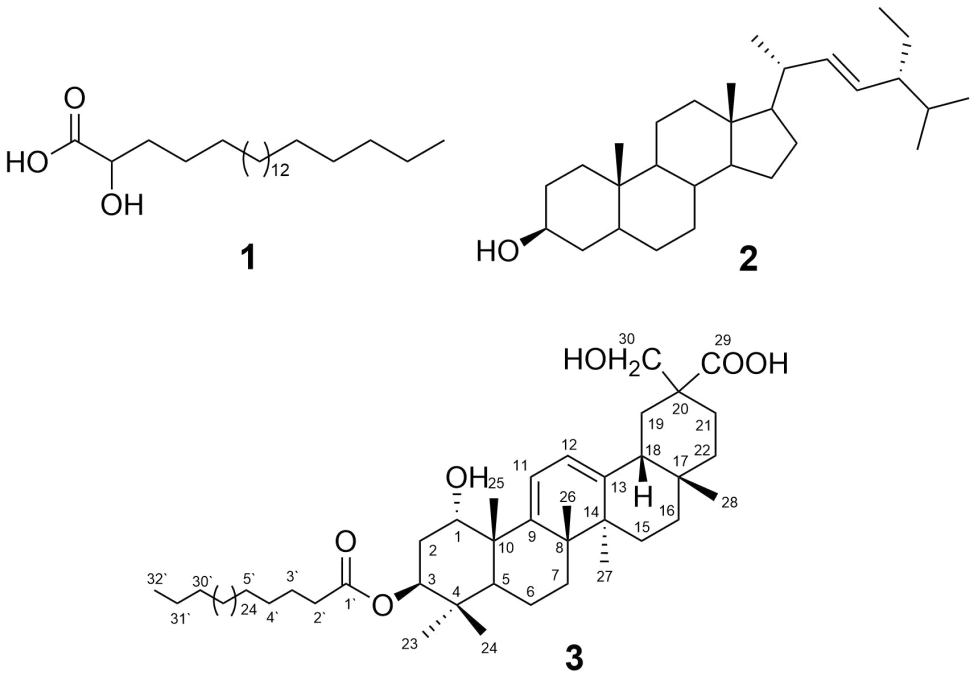


Figure 1 **Structures of isolated compounds 1–3.**

Full-size DOI: 10.7717/peerj.13040/fig-1

MATERIALS & METHODS

Chemicals and reagents

All the chemicals and reagents used for this study including; ethanol (96%), methanol (99.8%), petroleum ether (40–60 °C), dichloromethane (≥99.8%), chloroform (99.9%), n-butanol (99.8%), dimethyl sulfoxide (DMSO), hydrochloric acid (HCl), potassium hydroxide (KOH, ≥85%), 1,1-diphenyl2-picrylhydrazyl (DPPH, 95%), 2, 20-azino-bis[3-ethyl benzo thiazoline-6-sulphonic acid] (ABTS), potassium persulfate ($K_2S_2O_8$), acetylthiocholine iodide (≥99.0%), S-butyrylthiochoilne iodide (≥98.0%), sodium phosphate (Na_3PO_4, 96%), ascorbic acid, resazurin, 2-nitrobenzoic acid and were obtained from Sigma Aldrich (Hamburg, Germany). Muller Hinton agar and Muller Hinton broth were procured from HiMedia (Himedia Laboratories Pvt Ltd., Mumbai, India). Donepezil, galantamine and chloramphenicol were purchased from the local drug store.

Instrumentation

Optical rotation: Perkin-Elmer Model 341 LC polarimeter (Perkin-Elmer, Waltham, MA, USA). UV spectra: in MeOH using a Perkin-Elmer Lambda 25 UV/VIS spectrophotometer (Perkin-Elmer, Waltham, MA, USA). IR spectra: Shimadzu Infrared-400 spectrophotometer (Shimadzu, Kyoto, Japan). ESIMS spectra: Agilent 6320 Ion trap mass spectrometer (Agilent Technologies, Santa Clara, CA, USA) equipped with an electrospray ionization interface. NMR spectra: Bruker DRX 700 spectrometer (Bruker, Rheinstetten, Germany). Column chromatographic separations: silica gel 60 (0.04–0.063

mm, Merck, Darmstadt, Germany). TLC analyses: pre-coated silica gel F254 aluminum sheets (Merck, Darmstadt, Germany).

Enzymes and pathogenic strains

The acetylcholinesterase (AChE) and butyrylcholinesterase (BChE) were obtained from the mouse brain and human blood at the Pharmacology Department of King Saud University, SaudiArabia. *Escherichia coli* (ATCC 25922), *Pseudomonas aeruginosa* (ATCC 27853), and *Staphylococcus aureus* were provided by Microbiology Department, King Khaled University Hospital (KKUH), Saudi Arabia.

Plant material

The fresh aerial parts of *Abutilon bidentatum* Hochst. were collected from Jazan city, Aseer region, Saudi Arabia in March 2009. The plant was kindly identified by Dr. Mohamed Yusuf at the Pharmacognosy Department, College of Pharmacy, King Saud University. A voucher specimen (#16022) was deposited in the herbarium of the Pharmacognosy Department.

Extraction and isolation

Air dried-powdered aerial parts of *A. bidentatum* (900 g) were exhaustively extracted by maceration with 96% EtOH (3L × 7) with shaking. The combined EtOH extract was dried under reduced pressure at 45 °C to give a dry total extract weighing 89 g (EE). A part of the dried residue (87 g) was suspended in 450 mL distilled water and fractionated by shaking with petroleum ether (500 mL × 6), dichloromethane (500 mL × 4), ethyl acetate (500 mL × 5), and n-BuOH (500 mL × 6), respectively. Each combined extract was individually evaporated under reduced pressure to give the petroleum ether (PEF, 13 g), dichloromethane (DCF, 4 g), ethyl acetate (EAF, 6 g), n-BuOH (BF, 9 g), and aqueous soluble (AF, 55 g) fractions. Therefore, part of PEF (11.5 g) was chromatographed on SiO_2 column (330 g, 5 × 90 cm). Elution was started with petroleum ether and the polarity was gradually increased with CH2Cl2, followed by CH_2Cl_2: MeOH mixtures up to 100% MeOH to afford seven subfractions: *AB-1* to *AB-7*. TLC pattern of subfractions *AB-2* and *AB-5* showed the presence of prominent compounds and were selected for further column chromatography. Subfraction *AB-2* (33 mg) was chromatographed over silica gel using petroleum ether:CH_2Cl_2 gradient to give **1** (6.7 mg, white powder). Subfraction *AB-5* (311 mg) was applied on SiO_2 column (15 g, 70 × 1 cm), using petroleum ether:CH_2Cl_2 (95:5 to 80:20) to give two subfractions *AB-5.1* and *AB-5.2* containing two major spots on TLC. Subfraction *AB-5.1* (43 mg) was purified on SiO_2 column, eluted with petroleum ether:CH_2Cl_2 gradient to yield **2** (10 mg, white amorphous powder). Sub fraction *AB-5.2* was similarly treated as subfraction *AB-5.1* to give **3** (12 mg, white amorphous powder).

Abubidentin A (1α,3β,30-Trihydroxy-29-carboxy-olean-9(11), 12-diene-3-dotriacontanoate; 3)

White amorphous powder. $[\alpha]_D$ +45 (c = 0.1, MeOH). UV (MeOH) λ_{max} (log ε): 212 (3.10), 275 (2.21) nm. IR (KBr): 3,420, 2,947, 1,732, 1,712, 1,636, 1,245 cm^{-1}. NMR data ($CDCl_3$, 700 and 176 MHz) see Table 1. (+) ESI-MS *m/z:* 971.4 $[M+Na]^+$; (-) ESI-MS *m/z* 947.8 $[M-H]^-$.

Table 1 NMR spectral data of compound 3 (500 and 125 MHz).

Position	δ_H [mult., *J* (Hz)]	δ_C (mult.)	HMBC
1	4.17 (*d*, *J* = 2.8)	71.6 CH	2, 3, 5, 10, 25
2	1.65–1.66 (m) 1.42–1.43 m	30.4 CH_2	3, 4, 10
3	5.09 (*dd*, *J* = 12.0, 5.0)	74.3 CH	1′, 1, 4, 23, 24
4	–	39.6 C	–
5	1.48–1.50 (m)	44.0 CH	3, 4, 10, 24, 25
6	1.61–1.63 (m) 1.45–1.47 (m)	16.2 CH_2	–
7	2.05–2.07 (m) 1.81–1.82 (m)	30.9 CH_2	–
8	–	37.1 C	–
9	–	149.5 C	–
10	–	43.4 C	–
11	5.73 (*d*, *J* = 6.0)	116.0 CH	8, 9, 10, 13
12	5.58 (*d*, *J* = 6.0)	119.6 CH	9, 14, 18
13	–	146.0 C	–
14	–	39.9 C	–
15	1.68–1.70 (m) 1.25–1.28 (m)	24.1 CH_2	13
16	1.88–1.91 (m) 0.98–1.00 (m)	24.6 CH_2	14
17	–	32.8 C	–
18	2.18 (*dd*, *J* = 12.8, 5.6)	41.4 CH	12, 13, 19, 22, 29, 30
19	1.63–1.65 (m) 1.05–1.08 (m)	45.1 CH_2	17, 21, 29, 30
20	–	30.2 C	
21	1.15–1.17 (m) 1.00–1.02 (m)	33.5 CH_2	29, 30
22	1.48–1.50 (m) 1.32–1.34 (m)	33.8 CH_2	
23	0.97 (s)	24.6 CH_3	3, 4, 5, 24
24	0.91 (s)	15.8 CH_3	3, 4, 5, 23
25	1.26 (s)	22.5 CH_3	1, 5, 9
26	1.17 (s)	19.4 CH_3	9, 14
27	1.07 (s)	19.2 CH_3	8, 13
28	0.92 (s)	27.2 CH_3	17, 18, 22
29	–	173.0 C	–
30	4.06 (*d*, *J* = 10.2) 3.81 (*d*, *J* = 10.2)	69.6 CH_2	19, 20, 21, 29
1′	–	172.3 C	–
2′	2.33 (*t*, *J* = 6.8)	32.0 CH_2	1′
3′	1.62–1.65 (m)	24.1 CH_2	
$(CH_2)_{26}$	1.31–1.27 (m)	28.7–28.1 CH_2	

(*continued on next page*)

Table 1 (*continued*)

Position	δ_H [mult., *J* (Hz)]	δ_C (mult.)	HMBC
30	1.44–1.46 (m)	31.9 CH_2	32′
31′	1.12–1.14 (m)	21.7 CH_2	32′
32′	0.89 (*t*, *J* = 6.8)	13.1 CH_3	31′, 30′

Alkaline hydrolysis of compound 3

Compound **3** (4 mg) was dissolved in five mL of 3% KOH/MeOH solution and kept undisturbed for 15 min at ambient temperature. After 15 min, 1 N HCl/MeOH was added to the solution for the neutralization. The solution was then partitioned with CHCl3, CHCl3 layer was separated, evaporated and the obtained residue was taken up for column chromatography on SiO2 (mesh 60–120) using as eluent hexane: EtOAc gradient to provide methyl ester of dotriacontanoic acid, which was verified by GC/MS (*Ibrahim, Mohamed & Ross, 2016*; *El-Shanawany et al., 2015*; *Al-Musayeib et al., 2013*). A 500-Clarus Perkin-Elmer GC/MS (Waltham, MA, USA) was applied for GC/MS analysis. The integrator combined with software (4.5.0.007 version) controller was turbo mass. A 5 MS/GC elite capillary (30 × 0.25 mm × 0.5 μm) column and helium (He) a carrier gas at 2 ml/min flow rate (55.8 cm/s flow initial with 32 p.s.i., split; 1:40) were applied. Temperature conditions including; source temperature, inlet line temperature, emission trap and electron energy were adjusted at 150 °C, 200 °C, 100 °C and 70 eV, respectively. The injector temperature at 220 °C was maintained. Whereas, the temperature of the column was set at 50 °C for 5 min, raised to 220 °C at the 20 C/min rate. MS was scanned from 50 to 650 m/z.

Antioxidant activity

DPPH radical scavenging activity

The ability of the samples to scavenge DPPH radical was determined by *Wang, Chen & Hou (2019)* method with slight modification (*Wang, Chen & Hou, 2019*). A total of 20 μL of different sample concentrations (5.25–50 mg mL−1) solutions was reacted with 180 μL of DPPH• (6–5 M) dissolved in methanol (80%) in a 96-well plate. Samples were incubated at ambient temperature under dark conditions for 30 min and the absorbance wavelength was measured at 517 nm against a blank sample. DPPH• solution and ascorbic acid were used as blank samples and positive control, respectively. The inhibitory concentration (IC_{50}) was expressed as the concentration that inhibits the 50% of DPPH and calculated using the following equation

$$\text{Inhibition percentage } (\%) = \frac{(A_b - A_s)}{A_b} \times 100$$

where A_b and A_s are the absorbance values of the blank and test samples, respectively.

ABTS radical scavenging activity

ABTS radical cation decolorization assay was performed to measure the total antioxidant activity of the samples by obeying the previously described method (*Re et al., 1999*). In brief, $ABTS^{\bullet+}$ radicals were generated by treating 7 mM $ABTS^{+}$ aqueous solution with 2.4 mM potassium persulfate for 12-16 h at room temperature in the dark. Prior to use, this solution was diluted with ethanol (approx. 1:89 v/v) and equilibrated to 0. 7,000 ±

0.02 at 300 °C absorbance at 734 nm to obtain the working solution. Afterward, a 1.0 mL of working solution was reacted with 20 μL (1 mg mL^{-1}) samples and incubated for 30 min. After incubation, samples were scanned at 734 nm absorbance wavelength and the percentage of inhibition was determined. Ascorbic acid (AA) was applied reference standard.

Evaluation of cholinesterase inhibitory activities

The cholinesterase inhibitory effect of test sample was investigated on two enzymes (acetylcholinesterase and butyrylcholinesterase) by spectrophotometric method by following Ellman et al. procedure with little modification (*Ellman et al., 1961*). Crude enzymes, AChE (acetylcholinesterase) and BChE (butyrylcholinesterase) were collected from brain of mice and blood of humans, respectively, by obeying earlier described procedure (*Asaduzzaman et al., 2014*; *Uddin et al., 2015*). The AChE and BChE assays were tested by using two chemical substrates acetylthiocholine iodide and S-butyrylthiochoilne iodide, respectively. In brief, 10 μL of each enzyme was reacted individually with equal volume (10 μL) of different concentration (25–400 μgmL^{-1}) test sample and reference standard followed by incubation at 37 °C for 15 min for the complete interaction. Afterwards, 2-nitrobenzoic acid (1 mM, 62 μL), sodium phosphate buffer (50 mM, 25 μL, pH 8) provided with bovine albumin serum (0.1%) and acetylcholine iodide (0.5 mm, 13 μL) were added separately into each reaction mixture. Each reaction mixture was individually incubated further for 15 mis at 37 °C and absorbance at 405 nm was immediately noted against the blank. Donepezil and galantamine were used as reference compounds for AChE and BChE activity, respectively. All the experiments were performed in triplicates to avoid error and the results were estimated through the two-tailed Student's *t*-test at a $p < 0.05$ significance. The inhibition percentage of cholinesterase activity was calculated using the following equation

$$I\% = \frac{A_C - A_s}{A_C} \times 100$$

where A_c and A_s is the absorbance of control and sample or reference compound. IC50 values could be determined from the dose response curve obtained by plotting the percent inhibition values against test concentrations of each compound.

Antimicrobial activity

Minimum inhibitory concentration (MIC) assay was performed to evaluate the *in vitro* antimicrobial activities using a broth microdilution method (*Nascente, 2009*). Three bacterial strains *Escherichia coli* (ATCC 25922), *Pseudomonas aeruginosa* (ATCC 27853), and *Staphylococcus aureus* were applied to examine the antimicrobial activity of samples. Briefly, the microbial strains were transferred to Muller Hinton agar (MHB, HiMedia, India) and 24-h colonies were individually suspended in 10 mL Muller Hinton broth (MHB, HiMedia, India). The suspension of each microbial strain was standardized at 575 nm wavelength using a spectrophotometer (CRAIC Technologies, CA, USA), to match the McFarland scale (1.5 × 108 CFU mL^{-1}). The standardized suspension was further diluted to obtain a final concentration of 5 × 105 CFU mL^{-1}. The samples prepared in DMSO

at 1 mgmL^{-1} and different concentrations (30–500 μgmL^{-1}) were attained after dilution in Mueller Hinton broth. Three inoculated wells containing different concentrations of DMSO (4% to 1% range), one non-inoculated well without an antimicrobial agent and negative controls were included. The inoculated well was used to monitor whether the broth was sufficient for the microbial strain to grow. Chloramphenicol (500 to 30 mgmL^{-1}) was applied as a positive control. The sample treated 96-well microplates were sealed and incubated for 24 h at 37 °C After a 1-day incubation, 30 ml of 0.02% resazurin solution was added into each well to examine the viability of the microbial strain (*Palomino et al., 2002*). The minimum used concentration of test sample that inhibited the microorganism growth (MIC value) was calculated as the minimum concentration of the test samples required to prevent the color change of the resazurin solution. All the assays were performed in triplicates.

Statistical analysis

The experiments were conducted in triplicates and results were expressed as mean ± standard deviation. Statistical and graphical analysis were performed on Graph Pad Prism (version 8.0.1) and Microsoft Excel 2010. *T*-test was carried out to determine the statistical significance between the average values and ($p < 0.05$) was considered significant.

RESULTS & DISCUSSION

Compound **3** was isolated as a white amorphous powder. It gave a positive Lieberman-Burchard reaction, indicating its triterpenoidal nature (*Al-Musayeib et al., 2013*; *Ibrahim et al., 2012*). The molecular formula of **3** was determined as $C_{62}H_{108}O_6$ on the basis of the ESI-MS pseudo-molecular ion peaks at *m/z* 971.4 $[M+Na]^+$ and 947.8 $[M-H]^-$. The IR spectrum showed the presence of hydroxyl (3,420 cm^{-1}), ester carbonyl (1,732 cm^{-1}), acid carbonyl (1,712 cm^{-1}), and double bond (1,636 cm^{-1}) functionalities. The ^{13}C and DEPT NMR spectra of **3** displayed resonances for 62 carbons: six tertiary methyl, one primary methyl, 39 methylenes, one of them for an oxymethylene δ(C) 69.6 (C-30), six methines, and ten quaternary carbon signals, including one ester carbonyl at δ(C) 172.3 (C-1′), acid carbonyl δ(C) 173.0 (C-29), and two quaternary olefinic carbons at δ(C) 149.5 (C-9) and 146.0 (C-13). The ^{1}H NMR spectrum showed six methyl singlets at δ(H) 0.97 H-C(23), 0.91 H-C(24), 1.26 H-C(25), 1.17 H-C(26), 1.07 H-C(27), and 0.92 H-C(28), characteristic for an oleanane-type triterpenoid (*Ibrahim, Mohamed & Ross, 2016*; *Al Musayeib et al., 2014*; *Mahato & Kundu, 1994*). They correlated to the carbon signals, resonating at δ(C) 24.6, 15.8, 22.5, 19.4, 19.2, and 27.2, respectively in the HSQC spectrum. Moreover, two coupled olefinic protons at δ(H) 5.73 (*d*, $J = 6.0$, H-C(11)) and 5.58 (*d*, $J = 6.0$, H-C(12)) were observed in the ^{1}H NMR and 1H-1H COSY spectra, indicating the presence of two tri-substituted olefinic double bonds (Fig. 2). They showed HSQC cross peaks to the carbons at δ(C) 116.0 C(11) and 119.6 C(12). The placement of the double bonds at C9-C11 and C12-C13 was established based on the HMBC cross peaks of H-C(11) to C(8), C(9), C(10), and C(13), H-C(12) to C(9), C(13), and C(14), H-C(18) to C(12) and C(13), H-C(27) to C(13), and H-C(25) and H-C(26) to C(9), indicating that **3** had an olean-9(11),12-diene skeleton (*Caceres-Castillo et al., 2008*;

Mahato & Kundu, 1994) (Fig. 2). The ^{1}H and ^{13}C NMR spectra displayed two oxymethine signals at δ(H) 4.17 (*d*, $J = 2.8$, H-C(1))/δ(C) 71.6 (C-1) and 5.09 (dd, $J = 12.0$, 5.0 Hz, H-C(3))/74.3 (C-3). Their assignment was established based on the observed 1H-1H COSY of H-C(2) to H-C(1) and H-C(3) as well as the HMBC cross peaks of H-C(1), H-C(23), and H-C(24) to C(3) and H-C(3) and H-C(25) to C(1). The coupling constants of H-C(1) ($J_{1,2} = 2.8$ Hz) and H-C(3) ($J_{3,2} = 12.0$, 5.0 Hz) revealed that the hydroxyl groups at C(1) and C(3) were α- and β-configured, respectively, based on comparison with the previously reported triterpenoids [$J_{1,2} = 2.5$–3.0 Hz and $J_{3,2} = 10.0$–12.8, 4.2–7.2 Hz] (*Ibrahim et al., 2019*; *Litaudon et al., 2009*; *Rogers & Subramony, 1988*). The signals at δ(H) 4.06 and 3.81 (2H, each *d*, $J = 10.2$, H-C(30))/ δ(C) 69.6 (C-30) and δ(C) 173.0 (C-29) revealed the presence of an oxymethylene and carboxyl functionalities in **3**. Their placement at (C-20) was established by the HMBC cross peaks of H-C(19) and H-C(21) to (C-29) and (C-30) and H-C(30) to (C-29). The multiple aliphatic protons at δ(H) 1.31–1.27 were attributed to the presence of a long chain aliphatic moiety. The ^{13}C NMR signal at δ(C) 172.3 (C-1′) was assigned to an ester carbonyl, which was confirmed by the IR absorption band at 1,732 cm^{-1}. While the signals at δ(H) 0.89 (t, $J = 6.8$)/ δ(C) 13.1 were assigned to the terminal methyl group of a straight-chain fatty acid (*Al-Musayeib et al., 2013*; *El-Shanawany et al., 2015*; *Ibrahim, Mohamed & Ross, 2016*). Upon alkaline hydrolysis of 3, it gave a methyl ester of dotriacontanoic acid, which was identified by GC-MS molecular ion peak at *m/z* 494 $[M]^{+}$ andconfirmed by the ESI-MS fragment ion peak at *m/z* 486 $[M+H-(\text{dotriacontanoyl moiety})]^{+}$. The attachment of the fatty acyl moiety at (C-3) was established by the HMBC cross peak of H-C(3) to (C-1′) and confirmed by the downfield shift of HC-(3) δ(H) 5.09. The relative stereochemistry at stereocenters was assigned by comparing the *J* values and ^{1}H and ^{13}C chemical shifts with those of related triterpenes (*Ibrahim et al., 2019*; *Litaudon et al., 2009*). Based on these findings, **3** was assigned as 1 α,3 β,30-trihydroxy-29-carboxy-olean-9(11),12-diene-3-dotriacontanoate and named abubidentin A.

The known compounds were identified as 2-hydroxydocosanoic acid (**1**) (*Inagaki et al., 2001*) and stigmasta-22-ene-3-β-ol (**2**) (*Connor et al., 2006*) by comparing their NMR spectral and physical data in the literature.

Antioxidant activity assessments

DPPH and $ABTS^{+}$ assays are widely applied to measure the compound's ability to determine its antioxidant potential. Both the spectrophotometric methods used for evaluating antioxidant activity are based on electron transfer reactions and visually rely on the reduction of a colored oxidant. The obtained results from these two assays expressed good correlation. Fig. 3A and Table 2 represents the free radical inhibition of isolated compounds (**1**, **2**, and **3**) and ascorbic acid (AA, standard) at different concentrations. The results showed that the DPPH scavenging potential of compounds **1**, **2**, and **3** incrementally increased with the increase in concentration of compounds. The IC_{50} values obtained were 10.82 ± 0.24, 7.60 ± 0.42, and 4.67 ± 0.28 μgmL^{-1} for compounds **1**, **2**, and **3** were, respectively, indicating that compound **3** possesses the highest radical scavenging potential

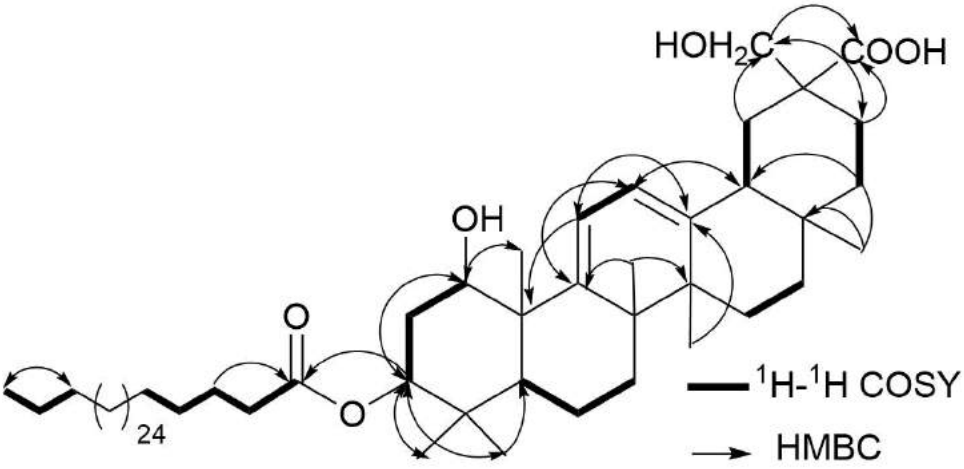


Figure 2 **Some Key ^{1}H-^{1}H COSY (–) and HMBC (H → C) correlations of 3.**
Full-size DOI: 10.7717/peerj.13040/fig-2

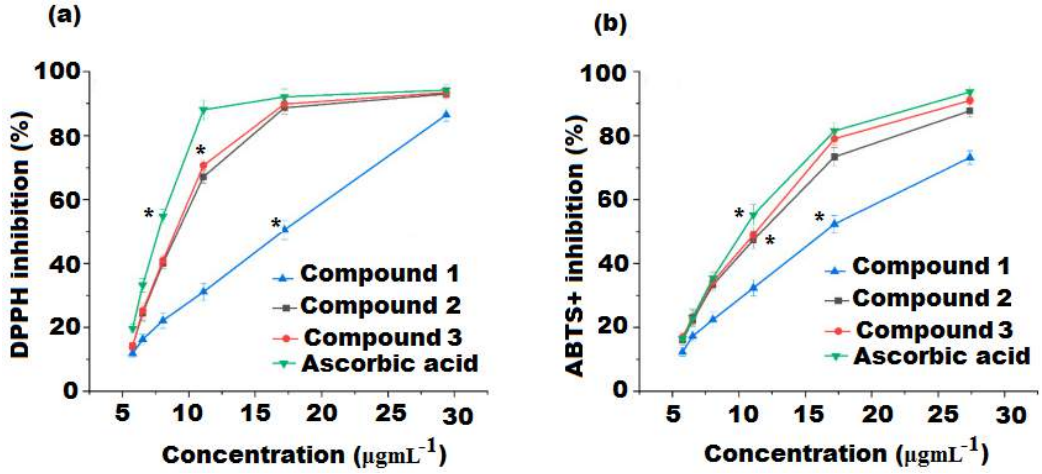


Figure 3 **(A) DPPH and (B) ABTS free radical scavenging activity of compound 1, 2, *3* and ascorbic acid.** Values were obtained as mean ± standard deviation and *mean significant difference compared to control ($p < 0.05$).
Full-size DOI: 10.7717/peerj.13040/fig-3

and was 1.55 fold lower than ascorbic acid (IC_{50} 3.12 ± 0.24 µgmL^{-1}), followed by compounds **2** and **1** ($p < 0.05$).

The ABTS^{+} assay is an additional important procedure for the quantification of radical scavenging potential that can provide parallel results to those obtained in the DPPH assay. The results showed that all the three isolated compounds exerted significant ABTS free radical scavenging activity and had an antioxidant potential proportional to that of ascorbic acid (Fig. 3B and Table 2). Compound **3** was found to be the most active radical scavenger with an IC_{50} value of 6.42 ± 0.25 µgmL^{-1} and 3.12 fold lower than that of ascorbic acid, followed by compound 2 and 1 with IC_{50} values of 8.18 ± 0.13 and 11.45

Table 2 Antioxidant activity and cholinesterase inhibitory of the extract and pure isolated compounds (1, 2 and 3) from of *A bidentatum*.

Sample	Antioxidant activity		Cholinesterase inhibitory	
	DPPH IC_{50} (μgmL^{-1})	$ABTS^{+}$ IC_{50} (μgmL^{-1})	AChE IC_{50} (μgmL^{-1})	BChE IC_{50} (μgmL^{-1})
Ethanol extract	16.34 ± 0.25e	18.65 ± 0.67e	132.56 ± 1.4e	142.35 ± 0.54e
Compound **1**	10.82 ± 0.24[d]	11.45 ± 0.37[d]	121.97 ± 1.61[d]	137.76 ± 0.67[d]
Compound **2**	7.60 ± 0.42[c]	8.18 ± 0.13[c]	68.65 ± 0.56[c]	49.52 ± 0.35[c]
Compound **3**	4.67 ± 0.28[b]	6.42 ± 0.25[b]	38.13 ± 0.07[b]	32.68 ± 0.37[b]
Ascorbic acid	3.12 ± 0.24[a]	4.45 ± 0.17[a]	–	–
Donepezil	–	–	9.32 ± 0.38[a]	–
Galantamine	–	–	–	10.27 ± 0.88[a]

Notes.
Means in each column with different subscript letters (a, b, c, d, e) differ significantly ($P < 0.05$).

± 0.37 μgmL^{-1}, respectively ($p < 0.05$). Compound **3** displayed both the highest DPPH and ABTS radical scavenging activities. A correlation between DPPH and ABTS methods applied to determine the antioxidant potential of the tested compounds was examined. The DPPH radical activity showed a strong correlation with ABTS radical activity (R_2 = 96.67%).

Cholinesterase inhibitory activity

The soluble fraction of the extract and compounds (**1–3**) were screened for AChE and BChE inhibition at different concentrations. The percent inhibition of AChE by the test compounds were presented in Fig. 4A. Donepezil used as a reference AChE inhibitor showed an IC_{50} value of 9.32 ± 0.38 μgmL^{-1}. The result revealed that all the investigated compounds (**1–3**) showed inhibition of the AChE enzyme in a dose-dependent manner. Among the compounds, high activity was displayed by compound **3** with IC_{50} values of 38.13 ± 0.07 μgmL^{-1}, followed by compound **2** (IC_{50} value = 68.65 ± 0.56 μgmL^{-1}). However, compound **1** showed low activity with IC_{50} of 121.97 ± 1.61 μgmL^{-1} (Table 2). Similarly, in BChE inhibitory assay, compounds **3** and **2** exerted high inhibitory potentials with IC_{50} of 32.68 ± 0.37 and 49.52 ± 0.35 μgmL^{-1}, respectively (Table 2, Fig. 4B). The IC_{50} value of compound **1** was 137.76 ± 0.67 μgmL^{-1} exhibited very negligible effects. Thus, compound **3** had appreciable activity towards both AChE and BChE enzymes.

Antimicrobial activity

The antimicrobial potential of isolated compounds (**1–3**) was assessed by calculating MIC values and their ability to inhibit the growth of the tested microbial strain. The assay was performed in 96-well microplates by applying resazurin as a developer. The well plates that exhibited the blue color after the addition of resazurin were considered as the MIC value for each microorganism (*Ostrosky et al., 2008*). The pure isolated compounds **1–3** were examined against *E. coli*, *P. aeruginosa,* and *S. aureus* bacterial strains for their antimicrobial potential. All the tested compounds exerted antimicrobial effect in the range of 125–1,000 μgmL^{-1} towards the pathogenic strains (Table 3). However, the prominent activity was shown by compound 2 with MIC ≤125 gmL^{-1}, ≤250 μgmL^{-1}, ≤150 μgmL^{-1}

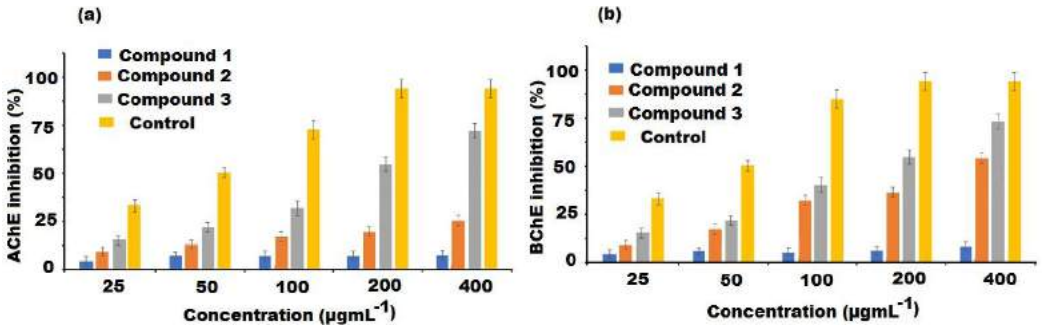


Figure 4 **Cholinesterase inhibitory activities of compound 1–3 isolated from *A. bidentatum* (a) Inhibition of acetylcholinesterase (AChE) by isolated compounds of *A. bidentatum* and standard donepezil, (B) Inhibition of butyrylcholinesterase (BChE) by compounds isolated from *A. bidentatum* and standard galantamine.**

Full-size DOI: 10.7717/peerj.13040/fig-4

Table 3 **Minimum inhibitory concentration (μgmL^{-1}) of extract and pure isolated compounds (1, 2 and 3) against selected pathogens.**

Sample	Microbial strain		
	E. coli	*P. aeruginosa*	*S. aureus*
Ethanol extract	≤600	≤1,000	≤1,000
Compound **1**	≤500	≤1,000	≤1,000
Compound **2**	≤125	≤250	≤150
Compound **3**	≤150	≤125	≤125
Chloramphenicol	≤40	≤40	≤40

and compound 3 with ≤150 μgmL^{-1}, ≤125 μgmL^{-1}, ≤125 μgmL^{-1}, against *E. coli, P. aeruginosa*, and *S.aureus*, respectively. There is no consistent classification with respect to MIC values (*Aligiannis et al., 2001*), but the values obtained ≤1,000 μgmL^{-1} were considered as satisfactory and sensitive (*Webster et al., 2008*). Thus, the MIC value of compounds **3** and **2** can be considered as promising.

CONCLUSION

A new oleanane-type triterpene ester, namely abubidentin A together with two known 2-hydroxydocosanoic acid and stigmasta-22-ene-3-β-ol were isolated from aerial parts of *A.bidentatum*. The extracts and compounds were investigated for antioxidant, cholinesterase inhibitory, and antimicrobial activities. The outcomes demonstrated that the newly isolated compound possesses a strong antioxidant effect towards DPPH and ABTS+ radical scavenging assays. This new triterpene exhibited high inhibition against acetylcholinesterase and butyrylcholinesterase. In addition, the new compound also showed promising antimicrobial effects against tested microorganisms. These results suggested that *A. bidentatum* is a promising source of useful natural products and the new compound offers opportunities to develop a novel drug.

ACKNOWLEDGEMENTS

This work was supported by Researchers Supporting Project No.(RSP-2021/294), King Saud University, Riyadh, Saudi Arabia.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by Researchers Supporting Project No. RSP-2021/294, King Saud University, Riyadh, Saudi Arabia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Researchers Supporting: RSP-2021/294.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Gadah A. Al-Hamoud conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Nawal M. Al-Musayeib analyzed the data, prepared figures and/or tables, and approved the final draft.
- Musarat Amina performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Sabrin R.M. Ibrahim conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data is available in the Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.13040#supplemental-information.

REFERENCES

Agrawal T. 2017. Abutilon a vulnerable genus: a review. *Research and Reviews: Journal of Pharmacognosy and Phytochemistry* **5(1)**:44–46.

Al Musayeib NM, Mothana RA, Ibrahim SR, El Gamal AA, Al-Massarani SM. 2014. Klodorone A and klodorol A: new triterpenes from *Kleinia odora*. *Natural Product Research* **28(15)**:1142–1146 DOI 10.1080/14786419.2014.915831.

Al-Musayeib NM, Mohamed GA, Ibrahim SR, Ross SA. 2013. Lupeol-3-O-decanoate, a new triterpene ester from *Cadaba farinosa* Forssk. growing in Saudi Arabia. *Medicinal Chemistry Research* **22(11)**:5297–5302 DOI 10.1007/s00044-013-0536-1.

Al-Shanwani M. 1996. *Plants used in the Saudi folk medicine*. Riyadh: King Abdul Aziz City for Science and Technology, 146.

Ali B, Ibrahim M, Hussain I, Hussain N, Imran M, Nawaz H, Jan S, Khalid M, Ghous T, Akash MSH. 2014. Pakistamide C, a new sphingolipid from *Abutilon pakistanicum*. *Revista Brasileira de Farmacognosia* **24**:277–281 DOI 10.1016/j.bjp.2014.07.002.

Aligiannis N, Kalpoutzakis E, Mitaku S, Chinou IB. 2001. Composition and antimicrobial activity of the essential oils of two *Origanum* species. *Journal of Agricultural and Food Chemistry* **49(9)**:4168–4170 DOI 10.1021/jf001494m.

Arbat AA. 2012. Pharmacognostic studies of stem of *Abutilon pannosum* (Forst f.). *Bioscience Discovery* **3(3)**:317–320.

Asaduzzaman M, Uddin MJ, Kader MA, Alam AH, Rahman AA, Rashid M, Kato K, Tanaka T, Takeda M, Sadik G. 2014. In vitro acetylcholinesterase inhibitory activity and the antioxidant properties of *Aegle marmelos* leaf extract: implications for the treatment of Alzheimer's disease. *Psychogeriatrics* **14(1)**:1–0 DOI 10.1111/psyg.12031.

Baquar SR. 1989. Medicinal and poisonous plants of Pakistan. In: *Medicinal and poisonous plants of Pakistan*. Karachi, Pakistan: Printas.

Caceres-Castillo D, Mena-Rejon GJ, Cedillo-Rivera R, Quijano L. 2008. 21 β-Hydroxy-oleanane-type triterpenes from *Hippocratea excelsa*. *Phytochemistry* **69(4)**:1057–1064 DOI 10.1016/j.phytochem.2007.10.016.

Connor WE, Wang Y, Green M, Lin DS. 2006. Effects of diet and metamorphosis upon the sterol composition of the butterfly *Morpho peleides*. *Journal of Lipid Research* **47(7)**:1444–1448 DOI 10.1194/jlr.M600056-JLR200.

El-Shanawany MA, Sayed HM, Ibrahim SR, Fayed MA. 2015. Stigmasterol tetracosanoate, a new stigmasterol ester from the Egyptian *Blepharis ciliaris*. *Drug Research* **65(07)**:347–353 DOI 10.1055/s-0034-1382064.

Ellman GL, Courtney KD, Andres Jr V, Featherstone RM. 1961. A new and rapid colorimetric determination of acetylcholinesterase activity. *Biochemical Pharmacology* **7(2)**:88–95 DOI 10.1016/0006-2952(61)90145-9.

Gomaa AA, Samy MN, Desoukey SY, Kamel MS. 2016. Pharmacognostical studies of leaf, stem, root and flower of *Abutilon hirtum* (Lam.) Sweet. *International Journal of Pharmacognosy Phytochemistry Research* **8**:199–216 *Available at http://www.ijppr.com*.

Gomaa AA, Samy MN, Desoukey SY, Kamel MS. 2018. Phytochemistry and pharmacological activities of genus Abutilon: a review (1972–2015). *Journal of Advanced Biomedical and Pharmaceutical Sciences* **1(2)**:56–74 DOI 10.21608/jabps.2018.3333.1000.

Ibrahim SRM, Khedr AIM, Mohamed GA, Zayed MF, El-Kholy AAS, Al Haidari RA. 2019. Cucumol B, a new triterpene benzoate from *Cucumis melo* seeds with cytotoxic effect toward ovarian and human breast adenocarcinoma. *Journal of Asian Natural Products Research* **21(11)**:1112–1118 DOI 10.1080/10286020.2018.1488832.

Ibrahim SR, Mohamed GA, Ross SA. 2016. Integracides F and G: New tetracyclic triterpenoids from the endophytic fungus *Fusarium* sp. *Phytochemistry Letters* **15**:125–130 DOI 10.1016/j.phytol.2015.12.010.

Ibrahim SR, Mohamed GA, Shaala LA, Banuls LM, Van Goietsenoven G, Kiss R, Youssef DT. 2012. New ursane-type triterpenes from the root bark of *Calotropis procera*. *Phytochemistry Letters* **5(3)**:490–495 DOI 10.1016/j.phytol.2012.04.012.

Inagaki M, Shibai M, Isobe R, Higuchi R. 2001. Constituents of ophiuroidea. 1. Isolation and structure of three ganglioside molecular species from the brittle star *Ophiocoma scolopendrina*. *Chemical and Pharmaceutical Bulletin* **49(12)**:1521–1525 DOI 10.1248/cpb.49.1521.

Jain R, Jain SC, Arora R. 1996. A new cholestane derivative of Abutilon bidentatum Hochst. and its bioactivity. *Pharmazie* **51(4)**:253 DOI 10.1002/chin.199637250.

Khadabadi SS, Bhajipale NS. 2010. A review on some important medicinal plants of *Abutilon* spp. *Research Journal of Pharmaceutical, Biological and Chemical Sciences* **1(4)**:718–729.

Litaudon M, Jolly C, Le Callonec C, Cuong DD, Retailleau P, Nosjean O, Nguyen VH, Pfeiffer B, Boutin JA, Guéritte F. 2009. Cytotoxic pentacyclic triterpenoids from *Combretum sundaicum* and *Lantana camaraas* inhibitors of Bcl-xL/BakBH3 domain peptide interaction. *Journal of Natural Products* **72(7)**:1314–1320 DOI 10.1021/np900192r.

Mahato SB, Kundu AP. 1994. 13C NMR spectra of pentacyclic triterpenoids—a compilation and some salient features. *Phytochemistry* **37(6)**:1517–1575 DOI 10.1016/S0031-9422(00)89569-2.

Migahid AM. 1978. *Flora of Saudi Arabia*. Second Edition. 1. Riyadh: Riyadh University Publication, 23.

Nascente PDS, Meinerz ARM, Faria ROD, Schuch LFD, Meireles MCA, Mello JRBD. 2009. CLSI broth microdilution method for testing susceptibility of Malassezia-pachydermatis to thiabendazole. *Brazilian Journal of Microbiology* **40**:222–226 DOI 10.1590/S1517-83822009000200002.

Ostrosky EA, Mizumoto MK, Lima ME, Kaneko TM, Nishikawa SO, Freitas BR. 2008. Metodos para avaliaçao da atividade antimicrobiana e determinaçao da concentraçao minima inibitoria (CMI) de plantas medicinais. *Revista Brasileira de Farmacognosia* **18**:301–307 DOI 10.1590/S0102-695X2008000200026.

Palomino JC, Martin A, Camacho M, Guerra H, Swings J, Portaels F. 2002. Resazurin microtiter assay plate: simple and inexpensive method for detection of drug resistance in *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy* **46(8)**:2720–2702 DOI 10.1128/AAC.46.8.2720-2722.2002.

Re R, Pellegrini N, Proteggente A, Pannala A, Yang M, Rice-Evans C. 1999. Antioxidant activity applying an improved ABTS radical cation decolorization assay. *Free Radical Biology and Medicine* **26(9–10)**:1231–1237 DOI 10.1016/S0891-5849(98)00315-3.

Rogers CB, Subramony G. 1988. The structure of imberbic acid, A 1 α-hydroxy pentacyclic triterpenoid from *Combretum imberbe*. *Phytochemistry* **27(2)**:531–533 DOI 10.1016/0031-9422(88)83135-2.

Shahwar D, Ahmad N, Ullah S, Raza MA. 2010. Antioxidant activities of the selected plants from the family Euphorbiaceae, Lauraceae, Malvaceae and Balsaminaceae. *African Journal of Biotechnology* **9(7)**:1086–1096 DOI 10.5897/AJB09.1622.

Survase SA, Jamdhade MS, Chavan S. 2012. Antibacterial activity of *Abutilon bidentatum* (Hochst.) leaves. *Science Research Reporter* **2(1)**:38–40.

Uddin MN, Afrin R, Uddin MJ, Uddin MJ, Alam AH, Rahman AA, Sadik G. 2015. *Vanda roxburghii* chloroform extract as a potential source of polyphenols with antioxidant and cholinesterase inhibitory activities: identification of a strong phenolic antioxidant. *BMC Complementary and Alternative Medicine* **15(1)**:1–9 DOI 10.1186/s12906-015-0728-y.

Wang CY, Chen YW, Hou CY. 2019. Antioxidant and antibacterial activity of seven predominant terpenoids. *International Journal of Food Properties* **22(1)**:230–238 DOI 10.1080/10942912.2019.1582541.

Webster NS, Xavier JR, Freckelton M, Motti CA, Cobb R. 2008. Shifts in microbial and chemical patterns within the marine sponge *Aplysina aerophoba* during a disease outbreak. *Environmental Microbiology* **10(12)**:3366–3376 DOI 10.1111/j.1462-2920.2008.01734.x.

Yasmin S, Akram Kashmiri M, Anwar K. 2011. Screening of aerial parts of *Abutilon bidentatum* for hepatoprotective activity in rabbits. *Journal of Medicinal Plants Research* **5(3)**:349–353.

Analysis of the role and mechanism of EGCG in septic cardiomyopathy based on network pharmacology

Ji Wu[1], Zhenhua Wang[1], Shanling Xu[2], Yang Fu[1], Yi Gao[1], Zuxiang Wu[1], Yun Yu[1], Yougen Yuan[3], Lin Zhou[3] and Ping Li[1]

[1] Department of Cardiovascular, The Second Affiliated Hospital of Nanchang University, Nan Chang, China
[2] Department of Cardiovascular, Medicine, Fuzhou First People's Hospital, Fu Zhou, China
[3] Department of Cardiovascular, The Three Affiliated Hospital of Nanchang University, Nan Chang, China

ABSTRACT

Background. Septic cardiomyopathy (SC) is a common complication of sepsis that leads to an increase in mortality. The pathogenesis of septic cardiomyopathy is unclear, and there is currently no effective treatment. EGCG (epigallocatechin gallate) is a polyphenol that has anti-inflammatory, antiapoptotic, and antioxidative stress effects. However, the role of EGCG in septic cardiomyopathy is unknown.

Methods. Network pharmacology was used to predict the potential targets and molecular mechanisms of EGCG in the treatment of septic cardiomyopathy, including the construction and analysis of protein-protein interaction (PPI) network, gene ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis and molecular docking. The mouse model of septic cardiomyopathy was established after intraperitoneal injection of LPS (lipopolysaccharide). The myocardial protective effect of EGCG on septic mice is observed by cardiac ultrasound and HE staining. RT-PCR is used to verify the expression level of the EGCG target in the septic cardiomyopathy mouse model.

Results. A total of 128 anti-SC potential targets of EGCGareselected for analysis. The GO enrichment analysis and KEGG pathway analysis results indicated that the anti-SC targets of EGCG mainly participate in inflammatory and apoptosis processes. Molecular docking results suggest that EGCG has a high affinity for the crystal structure of six targets (IL-6 (interleukin-6), TNF (tumor necrosis factor), Caspase3, MAPK3 (Mitogen-activated protein kinase 3), AKT1, and VEGFA (vascular endothelial growth factor)), and the experimental verification result showed levated expression of these 6 hub targets in the LPS group, but there is an obvious decrease in expression in the LPS + EGCG group. The functional and morphological changes found by echocardiography and HE staining show that EGCG can effectively improve the cardiac function that is reduced by LPS.

Conclusion. Our results reveal that EGCG may be a potentially effective drug to improve septic cardiomyopathy. The potential mechanism by which EGCG improves myocardial injury in septic cardiomyopathy is through anti-inflammatory and anti-apoptotic effects. The anti-inflammatory and anti-apoptotic effects of EGCG occur not only through direct binding to six target proteins (IL-6,TNF-α, Caspase3, MAPK3, AKT1, and VEGFA) but also by reducing their expression.

Submitted 4 August 2021
Accepted 2 February 2022
Published 9 March 2022

Corresponding author
Ping Li, lipingsydney@163.com

Academic editor
Gaurav Sharma

Additional Information and Declarations can be found on page 16

DOI **10.7717/peerj.12994**

How to cite this article Wu J, Wang Z, Xu S, Fu Y, Gao Y, Wu Z, Yu Y, Yuan Y, Zhou L, Li P. 2022. Analysis of the role and mechanism of EGCG in septic cardiomyopathy based on network pharmacology. *PeerJ* **10**:e12994 http://doi.org/10.7717/peerj.12994

Subjects Bioinformatics, Molecular Biology, Cardiology, Radiology and Medical Imaging, Histology
Keywords EGCG, Septic cardiomyopathy, Network pharmacology, Inflammation, Apoptosis

INTRODUCTION

Septic cardiomyopathy is a common complication of sepsis and an important factor in the high mortality of sepsis. The incidence of septic cardiomyopathy in sepsis is as high as 10–70% (*Sarah et al., 2018*). To date, the pathogenesis of septic cardiomyopathy is still not fully understood. Many mechanisms are involved in the occurrence and development of septic cardiomyopathies, such as inflammatory reactions, mitochondrial dysfunction, oxidative stress, apoptosis and abnormal calcium regulation (*Martin et al., 2019*). At present, the main methods for the treatment of septic cardiomyopathy are as follows: vasoconstrictor drugs, fluid resuscitation, cardiotonic drugs, heart rate control, mechanical support and other emerging treatments (*Heureux et al., 2020*). However, there is currently no drug fully effective at treating septic cardiomyopathy.

(-)-Epigallocatechin-3-gallate (EGCG) belongs to the family of catechins, is a secondary metabolite found in tea and is considered a potential substitute for synthetic food additives due to its antioxidant and antibacterial activities (*Nikoo, Regenstein & Ahmadi Gavlighi, 2018*). EGCG exhibits good performance in the study of some inflammation-related diseases (*Li et al., 2021b*; *Huang et al., 2021*; *Xie et al., 2020*; *Kar et al., 2019*). Recently, it has been reported that EGCG may participate in creating a protective effect against COVID-19 through anti-inflammatory and antioxidant effects (*Zhang et al., 2021b*). *Ma et al. (2021)* found that EGCG could protect against liver and intestinal tract injury induced by LPS by stabilizing intestinal flora. In addition, EGCG can also protect against LPS-induced neurological damage (*Cheng et al., 2021*) and lung injury (*Wang, Fan & Zhang, 2019*). However, the role of EGCG in septic cardiomyopathy is unknown.

A network pharmacology-based approach has previously proven successful in revealing treatable targets and mechanisms from bioinformatics assays (*Kibble et al., 2015a*). Therefore, in the present study, we tried to use the network pharmacology approach to explore predictive targets and therapeutic mechanisms underlying the action of EGCG against SC. We further verified the pharmacological effects of EGCG and the targets screened by network pharmacology by constructing a mouse model of septic cardiomyopathy.

METHODS

Screening of molecular targets of EGCG against septic cardiomyopathy

The plan of this study is shown in Fig. 1A. The pharmacological targets of EGCG were obtained by using the TCMID (*Yu et al., 2021*), TCMSP (*Guan et al., 2021*), STITCH (*Ding et al., 2021*), and SwissTargetPrediction (*Li et al., 2021a*) databases. Similarly, drugbank (*Li et al., 2020a*), Genecards (*Zhang et al., 2021a*), and OMIM (*Liu et al., 2021*) data were used to screen pathophysiological molecular targets for septic cardiomyopathy. A Venn diagram

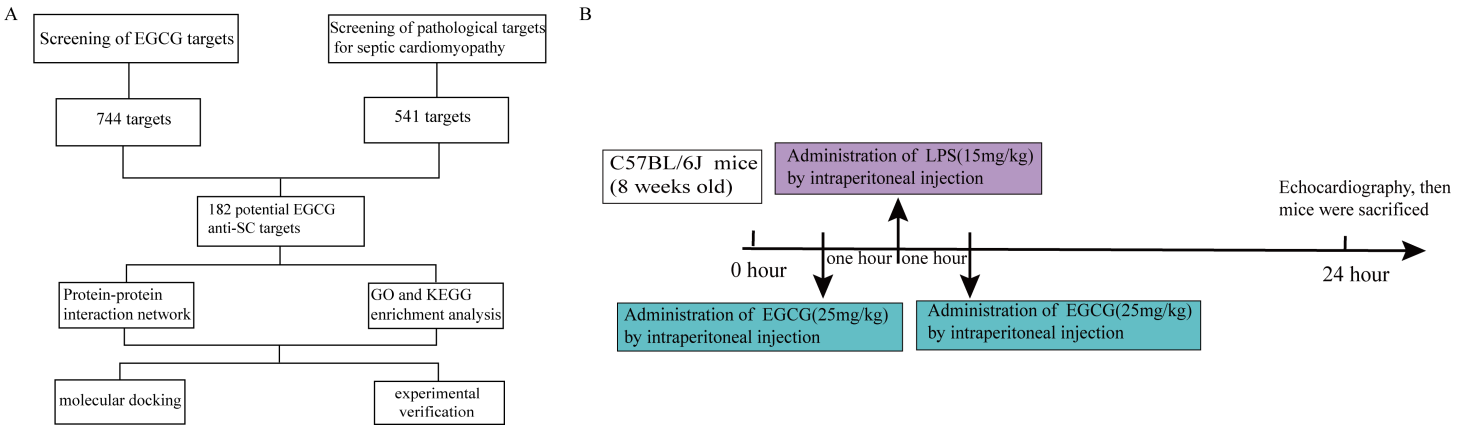


Figure 1 **Flow chart of the analysis process in this study.**

Full-size DOI: 10.7717/peerj.12994/fig-1

was used to evaluate the main targets of EGCG and septic cardiomyopathy to determine the potential targets of EGCG against septic cardiomyopathy.

Construction of PPI network and screening of key genes

The potential therapeutic targets of EGCG for septic cardiomyopathy were analyzed by a PPI network, and the species were set to humans using the STRING database (http://string-db.org/) with a confidence level greater than 0.4 (*Xiong et al., 2021*). Then, the Cytoscape software (version 3.7.2) was used to visualize the PPI network.

To search the highly connected subnetworks in the PPI, Molecular Complex Detection (MCODE) of Cytoscape was used. A vertex-weighting-based scheme is used to identify local high-density areas in the graph, which is depicted in the MCODE results. The subnetwork with a cutoff MCODE score $\geq$ 10 was used for further analysis. The score value of a node reflects the density of this node and its surrounding nodes.

CytoHubba plug-in Cytoscape software was used to acquire the hub genes for EGCG anti-SC, which included 6 topological analysis methods: MCC (maximal clique centrality), degree, closeness, radiality, and stress (*Chin et al., 2014*). The degree of protein correlation in this module was scored by the following criteria: the degree cutoff was 2, and the node score cutoff was 0.2. These criteria were used to score the degree of protein correlation in this module, following node score = 0.2, degree = 2, K-score = 2 and max depth = 100 (*Wang et al., 2020*).

GO and KEGG analysis of potential targets of EGCG Anti-SC

To comprehensively understand the potential target genes of EGCG in the treatment of septic cardiomyopathy, the DAVID database (https://david.ncifcrf.gov/) was used for GO and KEGG enrichment analysis (*Ashburner et al., 2000*; *Kanehisa et al., 2017*). $P < 0.05$ indicates that the entries and pathways are statistically significant. The ggplot2 package in R language was used to visualize the results.

Docking analysis of targets and EGCG

To obtain their specific relationships, the mode between EGCG and the six potential target proteins were evaluated through molecular docking which predicts the extent of interactions. The 3D structure of the target protein was downloaded from the PDB database (http://www.rcsb.org/), and the 2D structure of EGCG was downloaded from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/). Hetero and water molecules of the proteins were removed by PyMOL. AutoDock (version 4.2.6) was used to display the 3D grid box for molecular docking simulation. PyMOL and Discovery Studio 2020 were used to analyze the results.

Verify the role of EGCG and the level of hub genes

Materials and animal treatment

Twenty-four male C57BL/6J mice (aged 8–10 weeks, weighing 22 ± 5 g) were purchased from Hunan SJA Laboratory Animal Co., Ltd., Changsha, China. EGCG (E4143, purity ≥ 95%) and lipopolysaccharide (LPS, 2880) were purchased from Sigma (St. Louis, MO, USA).

All animal treatments were carried out in accordance with the ARRIVE guidelines and Use Committee of the Second Affiliated Hospital of Nanchang University, China (NO. SYXK 2015-0001). All animals were housed in plexiglass maintained at 24 ± 3 °C and a relative 60 ± 10% of humidity. The food and water were sufficiently provided during the 12-hour light/dark cycle. After one week of adaptive feeding, they were random divided into three groups: the control group ($n = 8$), the LPS group ($n = 8$), and the LPS+EGCG group ($n = 8$). The three groups of mice were treated in the following ways: (1) the control group was administered an intraperitoneal injection of normal saline; (2) the mice in the LPS group were administered an intraperitoneal injection of LPS (15 mg/kg); and (3) in the LPS + EGCG group, all mice were injected with EGCG (25 mg/kg) before and 1 h after receiving an intraperitoneal injection of LPS (Fig. 1B). All mice were anesthetized with pentobarbital sodium (50 mg/kg) and sacrificed by cervical dislocation. The experiment was suspended and the mice were euthanized.

Echocardiography

The changes in cardiac function in all mice were evaluated by echocardiography 24 h after intraperitoneal injection of LPS. We used Vevo770 (VisualSonics, Toronto, Canada) with a 30 Hz transducer for transthoracic echocardiography. A short-axis view under M-mode tracings was used to measure the systolic and diastolic sizes of the left ventricle. The formulas for calculating ejection fraction ((EF)) and minor axis shortening ((FS)) are as follows: EF (%) = [(LVIDd)3 − (LVIDs)3/(LVIDd)3] × 100%. LV fractional shortening (FS) was calculated as [(LVIDd − LVIDs)/LVIDd] × 100%.

Histological examination

Part of the mouse heart tissue was immersed in 4% paraformaldehyde solution for 48 h and dehydrated step by step in ethanol. The preparation process was routinely stained with hematoxylin and eosin and finally sealed with paraffin.

Table 1 Primers sequences of hub genes.

Gene	Primer nucleotide sequence
IL-6	Forward: 5′-CTGGTCTTCTGGAGTTCCGTTTCTAC-3′
	Reverse: 5′-GATGAGTTGGATGGTCTTGGTCCTTAG-3′
TNF	Forward: 5′-CCACGCTCTTCTGTCTACTGAACTTC-3′
	Reverse: 5′-GGTATGAAATGGCAAATCGGCTGAC-3′
MAPK3	Forward: 5′ATAGGCATCCGAGACATCCTCAGAG-3′
	Reverse: 5′-TTAAGGTCGCAGGTGGTGTTGATAAG-3′
VEGFA	Forward: 5′GGAGGAAGAGAAGGAAGAGGAGAGG-3′
	Reverse: 5′-CATGGTGGAGGTACAGCAGTAAAGC-3′
AKT1	Forward: 5′AGAGGCAGGAAGAAGAGACGATGG-3′
	Reverse: 5′-GCAGGACACGGTTCTCAGTAAGC-3′
Caspase3	Forward: 5′CACTGGAATGTCATCTCGCTCTGG-3′
	Reverse: 5′-GTCGCCTCTGAAGAAGCTAGTCAAC-3′

Quantitative reverse transcription-PCR analysis

Total RNA was extracted from heart tissue by TRIzol (Beyotime, R0016). The purity and concentration of total RNA were detected by an instrument (NanoDrop^{TM}), and the absorbance of 260/280 RNA samples between 1.8 and 2.2 was used for the next step of reverse transcription. cDNA was synthesized according to the instructions of the reverse transcription kit (Tiangen, KR116). Real-time fluorescence PCR was used to detect the expression of mRNA in heart tissue by the SYBR Green method (Tiangen, FP313). The primer sequences for the genes used for RT-PCR detection are as follows (Table 1).

Statistical analysis

All data are presented as means ± standard deviations (SD). GraphPad Prism 8.0.2 software (GraphPad Software Inc., San Diego, CA, U.S.) was utilized to conduct the statistical analyses. We used Student's *t*-test for comparison of two groups and one-way ANOVA test with post hoc contrasts by Student-Newman-Keuls test was used to evaluate differences between two or more groups. P values < 0.05 were considered statistically significant.

RESULT

Molecular formula and potential EGCG anti-SC molecular targets

The chemical formula of EGCG (Fig. 2A) was obtained from the PubChem database. The CAS number is 989-51-5. A total of 744 genes (after removing duplicates) related to septic cardiomyopathy were obtained from the DrugBank, GenBank and OMIM databases. Five hundred forty-one EGCG drug targets were obtained through the TCMID, TCMSP, SwissTargetPrediction and Stitch databases. The 182 potential targets of EGCG against septic cardiomyopathy were shown by using Venn diagrams (Fig. 2B).

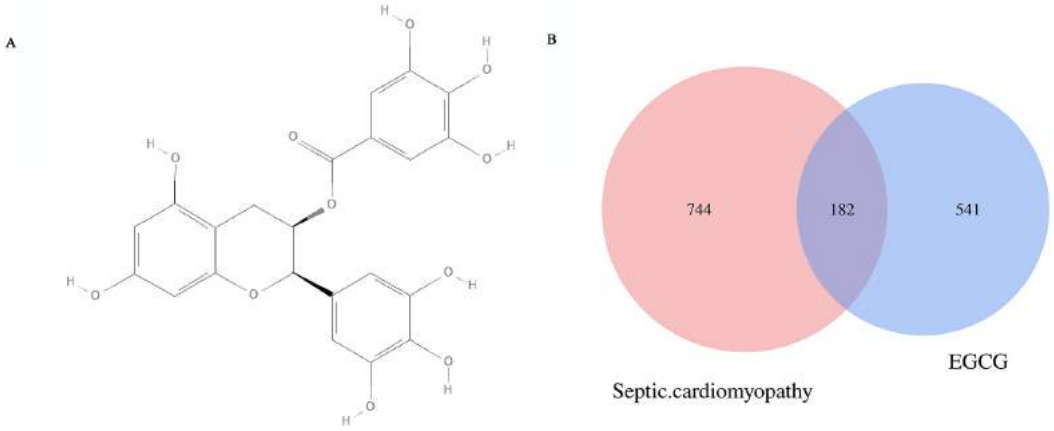


Figure 2 **(A) Molecular structure of EGCG. and potential targets. (B) The Venn diagram represen 541 EGCG targets and intersects 744 key genes of septic cardiomyopathy to obtain 182 potential targets. (C) The blue diamond represents the key gene of septic cardiomyopathy, the green octagon represents the target of EGCG, and the red square represents the potential target of EGCG in the treatment of septic cardiomyopathy.**

Full-size DOI: 10.7717/peerj.12994/fig-2

Acquisition results of hub genes for EGCG anti-SC and PPI network construction

The STRING database was applied to construct a function-related PPI network from the 182 targets by using a minimum required interaction score set to 0.4. First, we set the species to "Homo sapiens" and then entered 182 potential therapeutic targets for EGCG anti-SC to obtain the PPI network (Fig. S1). Next, we imported the TSV files into Cytoscape software (version 3.7.2) for further analysis and visualization. Cytoscape software was used to analyze the PPI network based on the MCODE topology, and similar functional clusters were selected to find closely connected areas. Four functional module clusters were detected according to the score, as shown in Fig. 3. Gene clusters with a score equal to 52.273 were used for further analysis.

The CytoHubba plug-in of Cytoscape software was used to obtain the hub genes for EGCG anti-SC based on the above cluster (score = 52.273), which currently contains 5 topological analysis methods (Fig. 4). The top 10 hub genes for EGCG anti-SC were obtained, and the results of 12 algorithms included 10 genes each. Finally, the hub genes included IL-6, TNF, Akt1, VEGFA, MAPK3, and Casp3, which were acquired by intersecting the genes obtained by the five algorithms (Fig. S2).

GO and KEGG pathway enrichment analysis

The results of GO analysis showed that there were 943 terms, including 750 terms (Table S1) for biological process (BP), 78 terms for cellular component (CC), and 115 terms for molecular function (MF). The results of each terms were sorted by the number of genes and significance, and the top 15 findings of each analysis were selected for display (Fig. 5A). Among them, 750 BP terms included positive and negative regulation of apoptosis, inflammatory response, response to lipopolysaccharide, *etc.*; 78 CC terms included

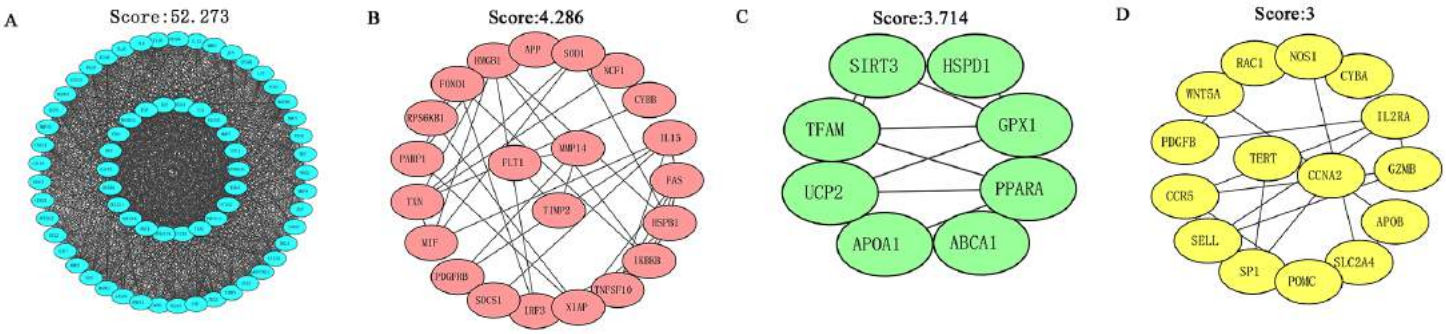


Figure 3 MCODE in Cytoscape software was used to analyze the PPI network and four functional module clusters were obtained according to different scores.

Full-size DOI: 10.7717/peerj.12994/fig-3

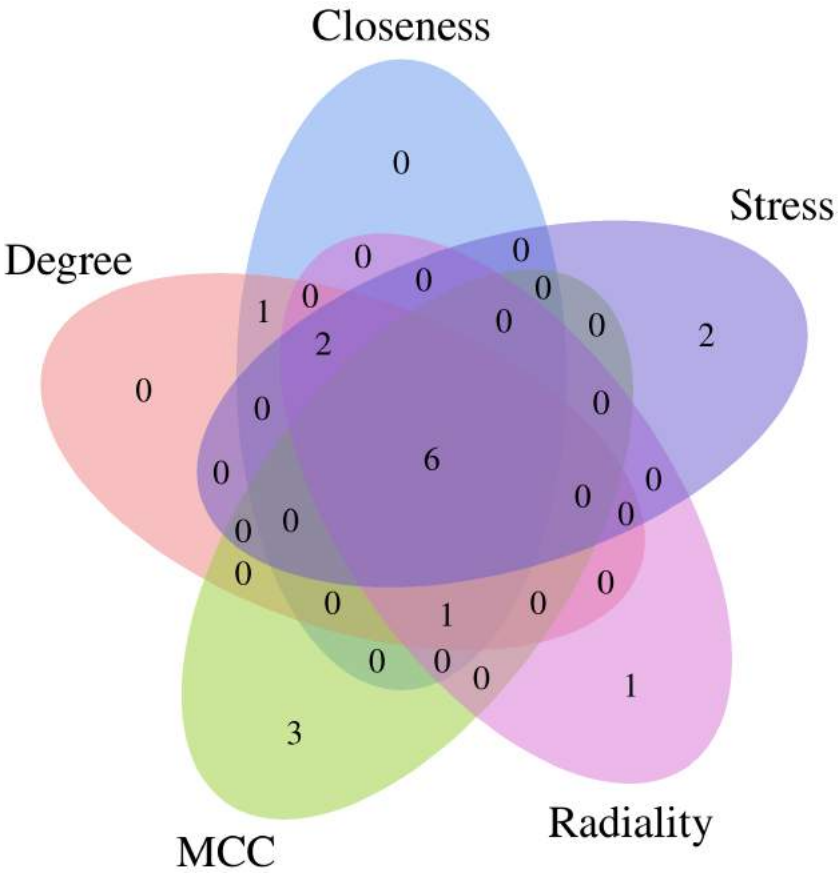


Figure 4 Five topological analysis methods of CytoHubba plug-in of Cytoscape software were used to analyze the gene clusters with a score of 52.273, and six key targets of EGCG in the treatment of septic cardiomyopathy were obtained.

Full-size DOI: 10.7717/peerj.12994/fig-4

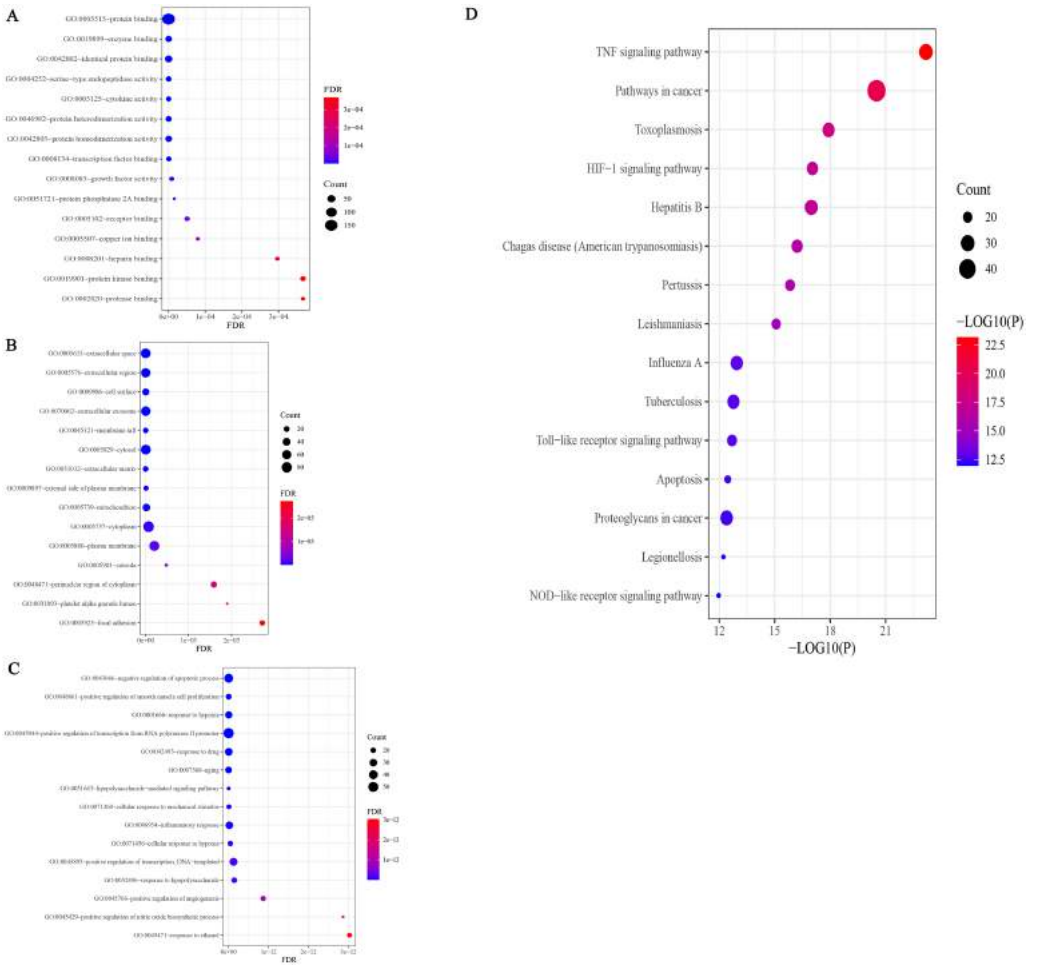


Figure 5 **GO enrichment analysis of potential targets.**

Full-size DOI: 10.7717/peerj.12994/fig-5

cytoplasm, plasma membrane, nucleus, and mitochondria, and 115 MF terms included protein binding, identical protein binding, enzyme binding, ATP binding, cytokine activity, and transcription factor binding, *etc*. The results of GO enrichment analysis of SC treated with EGCG showed that its biological process was mainly reflected in the regulation of apoptosis, signal transduction, immune response, and lipopolysaccharide response; the molecular function was mainly reflected in protein binding, cytokine activity, transcription factor activity, and ATP binding; and the composition was mainly reflected in the cytoplasm, plasma membrane, and nucleus. $P < 0.05$ was the filtering parameter of the cutoff point, and 119 KEGG pathway items were acquired (Table S2). The first 15 records filtered from small to large P values are shown in Fig. 5A. KEGG enrichment analysis involves infectious diseases caused by many pathogens, including the TNF signaling pathway, the HIF-1 signaling pathway, the Toll-like receptor signaling pathway, the apoptosis signaling pathway, the PI3K/AKT signaling pathway, and the NF-κB signaling pathway (Fig. 5B).

Table 2 The binding energy values of EGCG and core targets.

Compund	Targets	Binding affinity/ (kcal/mol)
EGCG	IL-6	−5.73
EGCG	TNF	−7.86
EGCG	MAPK3	−6.81
EGCG	VEGFA	−5.45
EGCG	AKT1	−6.26
EGCG	Caspase3	−7.07

Molecular docking results

Molecular docking is used to verify the binding affinity between EGCG and protein targets. The docking score indicates the binding affinity. Thus, the lowest score signifies the highest binding affinity between EGCG and the target protein. From this, the correct binding posture has been selected to analyze the interactions between EGCG and its target.The hub protein targets are filtered by the top 5 node degrees of the PPI network and seed nodes of clusters, including IL-6, TNF, MAPK3 (Erk1/2), VEGFA, AKT1, and CASP3. The results of the molecular docking of drugs and target proteins are shown in Table 2. EGCG is compared with each target under the condition that all target proteins met physiological pH (pH = 7.35). The grid box is located in the center, covering the active binding site and all necessary residues. For IL6, the grid box (100 Å × 100 Å × 100 Å) is centered at (30.183, 58.501, 20.278) Å; for TNF, the grid box (92 Å × 92 Å × 92 Å) is centered at (−33.22, 41.788, 39.681) Å; for MAPK3, the grid box (126 Å × 126 Å × 126 Å) is centered at (57.761, 4.848, −3.43) Å; for VEGFA, the grid box (40 Å × 40 Å × 40 Å) is centered at (10.383, −18.002, 25.541) Å; for AKT1, the grid box (118 Å ×106 Å × 86 Å) center at (21.763, 14.444, 9) Å; for the CASP3, grid box (60 Å × 60 Å × 60 Å) center at (24.892, 53.581, 12.294) Å. At present, there is no unified standard to screen active molecules. According to the literature (*Guan et al., 2021*), we select the active components with binding energy ≤ −5.0 kJ/mol as the basis for screening.It can be seen from Table 2 that the binding affinity of EGCG to the crystal structure of 6 core proteins is lower than -5 kcal/mol, indicating that EGCG exhibits notable binding affinity for the native structure of the target proteins. EGCG show the high binding affinity of −5.73 kcal/mol in IL6, −7.86 kcal/mol in TNF, −6.81 kcal/mol in MAPK3, −5.45 kcal/mol in VEGFA, −6.26 kcal/mol in AKT1, and −7.07 kcal/mol in CASP3, and the exhaustive value of all molecular docking results is 6. The results show that CASP3 is the most effective target for sepsis cardiomyopathy, both in terms of binding energy and the number of hydrogen bonds.

After docking we have obtained six docked poses in each case or for each protein target. According to the number and distance of hydrogen bond forces (distance ≤ 3.0 Å), it is concluded that the tertiary structure 3h0e of CASP3 is the best structure for molecular docking. Small molecular compounds bind closely to protein residues through various interactions (Fig. 6).

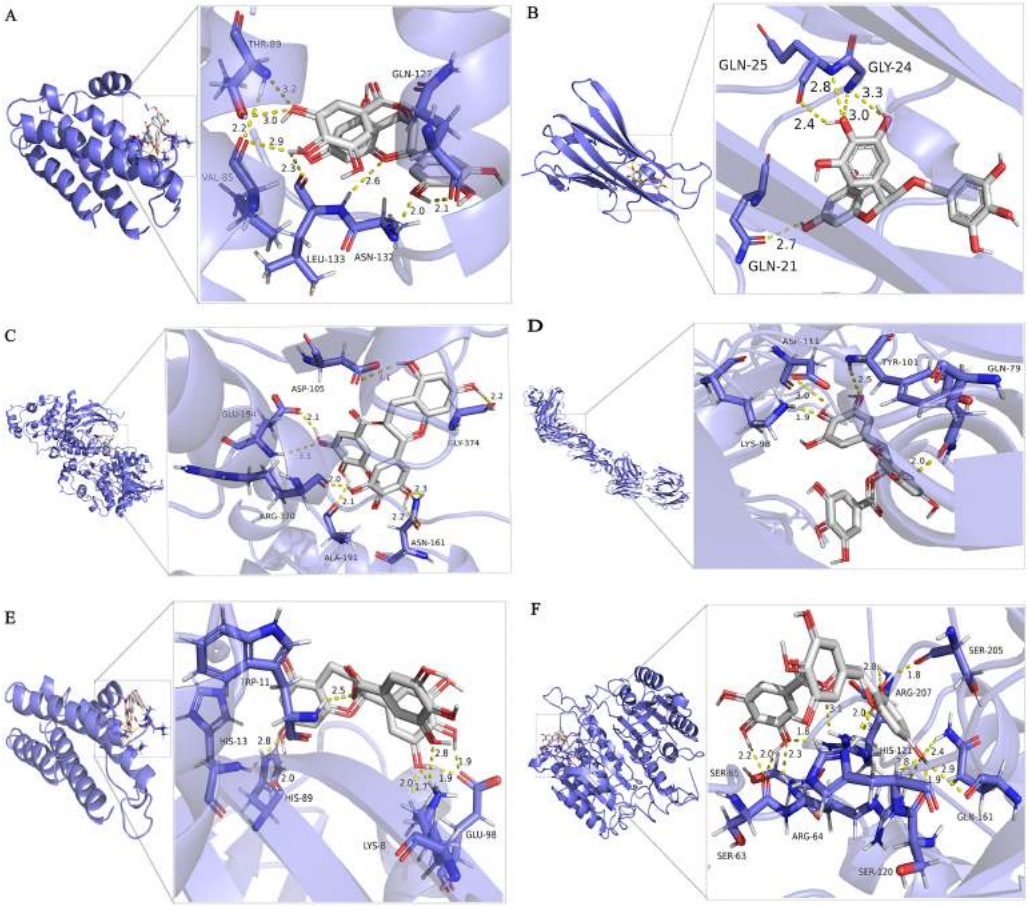


Figure 6 **Molecular docking simulation of EGCG-target binding.**
Full-size DOI: 10.7717/peerj.12994/fig-6

Effect of EGCG on functional and morphological changes of echocardiography induced by LPS

The echocardiography results in mice showed that LPS significantly increased the left ventricular end-systolic diameter (LVESD), decreased the ejection fraction and shortening fraction and had little effect on the left ventricular end-diastolic diameter (LVEDD) (Fig. 7B). In addition, EF and FS in LPS + EGCG group were significantly higher than those in LPS group. It is suggested that EGCG can improve the cardiac function of mice induced by LPS. The morphological changes of the myocardium were evaluated by HE staining. As shown in the figure (Fig. 7C), compared with the control group, the arrangement of cardiomyocytes in the LPS group was more disordered, with interstitial edema, some infiltration of inflammatory cells and a small amount of exudation of red blood cells that could be seen in the interstitium. Compared with the LPS group, the cardiomyocytes in the LPS+EGCG group were neatly arranged, interstitial edema was alleviated, and there was no obvious inflammatory cell infiltration or erythrocyte exudation.

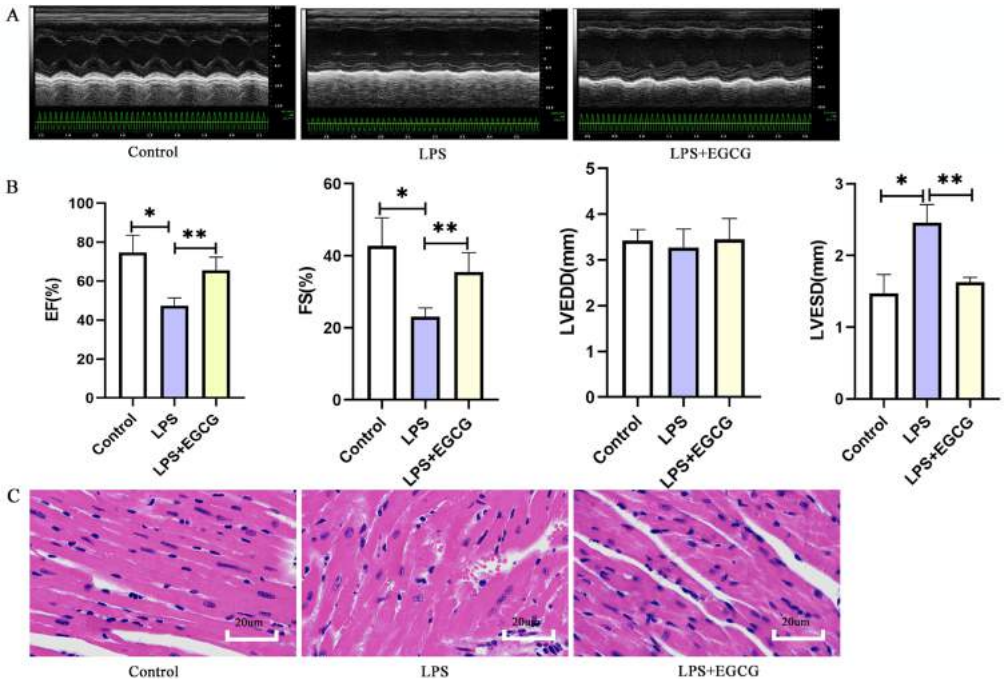


Figure 7 **Treatment of EGCG ameliorated cardiac dysfunction and histopathological injury in heart induced by LPS.**

Full-size DOI: 10.7717/peerj.12994/fig-7

Experimental verification of target genes expression in LPS-induced mouse model

As shown in Fig. 8, the RT-PCR results showed that the expression of key genes (IL-6, TNF-a, Caspase-3, VEGFA, Akt1, MAPK3) in the LPS group was higher than in the control group. Compared with that in the LPS group, the expression of key genes (IL-6, TNF-a, caspase-3, VEGFA, Akt1, MAPK3) in the LPS+EGCG treatment group was decreased. From the Genecards database, IL-6, TNF-α, MAPK3 and VEGFA are related to inflammation, while Caspase-3 and AKT1 are related to apoptosis. Therefore, EGCG can inhibit inflammation and apoptosis.

DISCUSSION

Septic cardiomyopathy is a common complication of sepsis, with an incidence of 10–70% in patients with sepsis (*Zhang et al., 2021b*). Septic cardiomyopathy is often characterized by systolic or diastolic dysfunction, which causes further ischemia and hypoxia in peripheral tissue, aggravates the dysfunction of other organs, and leads to increased mortality (*Martin et al., 2019*). Therefore, it is necessary to identify an effective drug for the treatment of septic cardiomyopathy. EGCG is a polyphenol component of green tea that has anti-inflammatory, antioxidant, antiviral and antitumor effects (*Eng, Thanikachalam & Ramamurthy, 2018*). Some studies have shown that EGCG plays an important role in preventing cardiovascular disease, participating in antiplatelet aggregation (*Joo et al., 2018*; *Lee et al., 2013*), decreasing atherosclerosis (*Duan et al., 2020*;

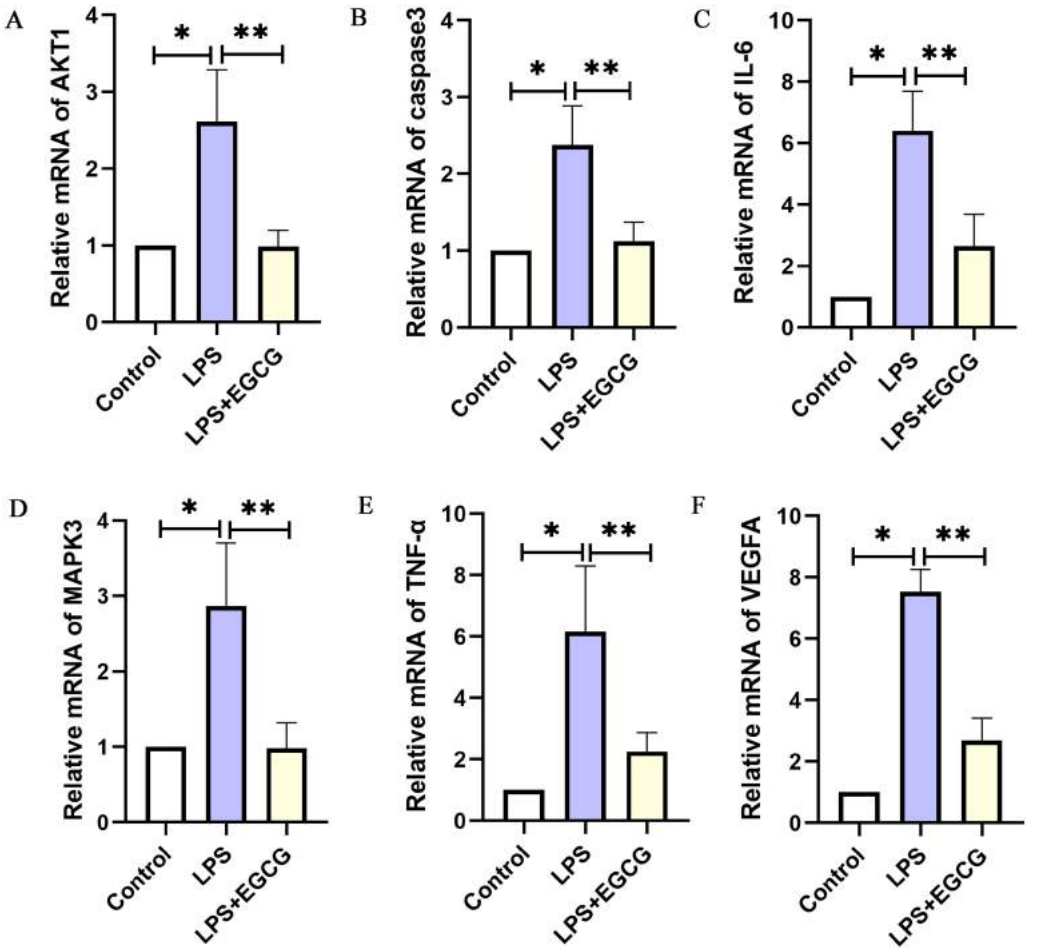


Figure 8 **The mRNA expression of the key genes of EGCG in the treatment of septic cardiomyopathy in each group of mice.**

Full-size DOI: 10.7717/peerj.12994/fig-8

Yamagata, 2020), counteracting antiventricular remodeling (*Cai et al., 2021*; *Ma et al., 2019*) and regulating blood lipids (*Ma et al., 2019*; *Liu et al., 2013*). Network pharmacology is used to study the intervention mechanism of drugs on diseases through the construction of biological interaction networks, which systematically reveal the therapeutic effects of drugs on diseases from the interaction of drugs, targets and diseases (*Kibble et al., 2015b*). In previous studies, EGCG improved intestinal injury (*Ma et al., 2021*), cerebral nervous system injury (*Cheng et al., 2021*), lung injury (*Wang, Fan & Zhang, 2019*) and endothelial injury (*Baek et al., 2019*) in LPS-induced sepsis in mice. *Li et al. (2020b)* found that EGCG can inhibit AKT phosphorylation and the ERK signaling pathway to improve LPS-induced H9C2 cell injury. As shown in Fig. 7, EGCG improved cardiac function and myocardial injury in LPS-induced mice. Therefore, we conclude that EGCG may be involved in exerting a protective effect in septic cardiomyopathy.

While the pathogenesis of sepsis is not completely clear, it is mainly related to the release of circulating myocardial inhibitory factors, downregulation of sarcomere- and

mitochondrial-related genes, downregulation of the adrenergic pathway, changes in coronary microcirculation, activation of the inflammatory response, mitochondrial dysfunction, increased oxidative stress and abnormal calcium regulation (*Hollenberg & Singer, 2021*). Inflammation, as the initiator of septic cardiomyopathy, promotes an increase in intracellular oxidative stress, which can cause cardiomyocyte apoptosis and lead to cardiac dysfunction. The anti-inflammatory effect of EGCG has been verified in many studies. EGCG ameliorates cigarette smoke-induced myocardial inflammation through MAPK and NF-κB (*Liang, Ip & Mak, 2019*). After administration of EGCG for one month, the level of serum proinflammatory factors (IL-1β, TNF-α, IL-6) in streptozotocin-induced diabetic mice decreased (*Othman et al., 2017a*).

The results of GO enrichment analysis further confirmed that the candidate target protein of EGCG is mainly involved in the process of inflammation and apoptosis. In our results, BP enrichment of anti-inflammatory factors included an inflammatory response, signal transduction, response to lipopolysaccharide, and cell adhesion, suggesting that the anti-inflammatory effects of EGCG range from inflammatory cell adhesion and inhibition of the cell response to lipopolysaccharide and intracellular signal transduction. In addition to the anti-inflammatory process, the enrichment of BP also includes apoptosis and oxidation–reduction processes, which is consistent with the pathogenesis of septic cardiomyopathy, indicating that the protective effect of EGCG on septic cardiomyopathy is multifaceted.

Six targets with a high degree of PPI network (IL-6, TNF, MAPK3, VEGFA, AKT1 and Caspase3) are identified according to the analysis of compound-target interactions. In the molecular docking analysis of our study, the binding mode with the highest docking fraction is chosen to analyze the interaction between EGCG and protein receptors. These six targets of EGCG in the treatment of septic cardiomyopathy are screened out by network pharmacology, and they are mainly related to inflammation and apoptosis. This finding is consistent with the results of GO enrichment. IL-6 and TNF are common markers of inflammation. EGCG protects against LPS-induced lung injury by promoting PRCKA and inhibiting the expression of IL-6 and TNF (*Wang et al., 2021*). EGCG protects the myocardium of streptozotocin-nicotinamide-induced diabetic rats through anti-inflammatory and anti-apoptotic pathways, mainly through diminishing the levels of IL-6, TNF-α and Caspase3 (*Othman et al., 2017a*). In addition, EGCG can inhibit cardiomyocyte apoptosis induced by isoproterenol by reducing the level of Caspase3 (*Othman et al., 2017b*). MAPK3, also known as ERK1, is involved in inflammation (*Yang et al., 2019*), proliferation (*Mebratu & Tesfaigzi, 2009*) and apoptosis (*Cook et al., 2017*). AKT1, also known as protein kinase B, is one of the three subunits of the AKT family and is involved in cell proliferation, inflammation, survival, metabolism and angiogenesis. Many studies have suggested that Akt1 has a protective effect in cardiovascular disease, but some studies still disagree. AKT1 deficiency can prolong the life span in the DKO mouse model, improve cardiac function and reduce myocardial hypertrophy (*Kerr et al., 2013*). Akt1 increases oxidative stress in DKO cells and promotes cell senescence and apoptosis by inhibiting the expression of ROS scavengers downstream of FoxO (*Nogueira et al., 2008*). The Akt1/NF-κB axis is involved in the inflammatory injury process of collagen-induced

arthritis in mice (*Yang et al., 2020*). The docking scores of EGCG targets are all less than -5 kcal/mol, indicating that EGCG has a good binding affinity for all six targets. Table 2 shows that the binding energy of EGCG and AKT1 is 6.26 kcal/mol, which shows that EGCG can combine well with Akt1. These results suggest that EGCG may be involved in inhibiting the AKT1/NF-κB axis and improving myocardial injury in septic cardiomyopathy. VEGFA is a member of the PDGF/VEGF growth factor family. It is essential for physiological and pathological angiogenesis to induce the proliferation and migration of vascular endothelial cells. Serum VEGFA in peritoneal dialysis patients is proportional to inflammatory factors such as IL-6. However, miR-15a-5p suppresses the inflammatory response of human peritoneal interstitial cells by targeting VEGFA (*Shang et al., 2018*). VEGFA promotes cardiomyocyte apoptosis induced by hypoxia, and miR-448-5p targeting VEGFA can inhibit the expression of Caspase3 (*Tang et al., 2021*). The binding energy of EGCG and VEGFA is 5.45 kcal/mol. EGCG combined with VEGFA may be involved in the inhibition of inflammation and apoptosis. Caspase3 is a downstream protein of the apoptotic pathway and participates in the progression of septic cardiomyopathy (*Cao et al., 2020*). Sulfur dioxide reduces Caspase3, inhibits cardiomyocyte apoptosis and improves myocardial injury in mice with CLP (*Yang, Zhang & Chen, 2018*). Therefore, we believe that the myocardial protective effects of EGCG may be associated with the regulation of caspase-3.

CONCLUSION AND LIMITATION

As shown in Fig. 9, EGCG probably improves myocardial injury in sepsis through anti-inflammatory and anti-apoptotic effects. Through the study of network pharmacology, it is concluded that EGCG can not only bind directly to inflammation-related proteins to inhibit inflammation, but it also binds to apoptosis-related proteins to inhibit apoptosis. Through the construction of the LPS-induced sepsis model to verify the key targets, it can be concluded that EGCG can reduce the expression of inflammation- and apoptosis-related genes. However, there are still some limitations to this study. First, we screened the molecular targets of human septic cardiomyopathy in the database, but we verified it using a mouse model. Second, through the screening of network pharmacology, we are bound to miss the important target of EGCG. Finally, we only verified the effect of EGCG on the target at the mRNA level, not the protein level of the target. In summary, our findings may provide a solution for the treatment of septic cardiomyopathy.

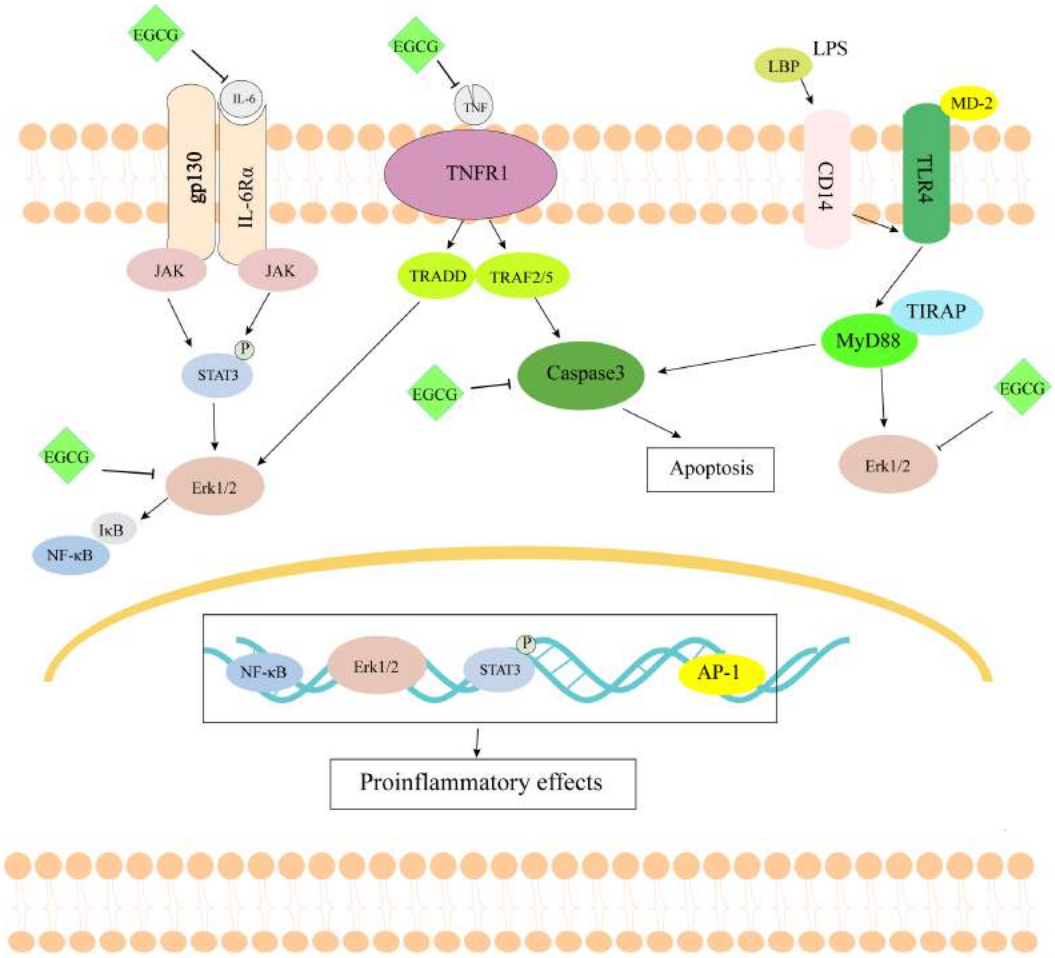


Figure 9 **The schematic diagram of EGCG in the treatment of septic cardiomyopathy.**
Full-size DOI: 10.7717/peerj.12994/fig-9

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Natural Science Foundation of China (NO. 81860058, 81960088), and special funds for guiding local scientific and technological development by the central government of China (NO. S2019CXSFG0016). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 81860058, 81960088.
Central government of China: S2019CXSFG0016.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Ji Wu and Zhenhua Wang conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Shanling Xu performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Yang Fu performed the experiments, prepared figures and/or tables, and approved the final draft.
- Yi Gao performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Zuxiang Wu, Yougen Yuan analyzed the data, prepared figures and/or tables, and approved the final draft.
- Yun Yu analyzed the data, authored or reviewed drafts of the paper, analysis tools, and approved the final draft.
- Lin Zhou conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Ping Li conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Animal Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

Committee of the Second Affiliated Hospital of Nanchang University(SYXK(èţč)2015-0001)

Data Availability

The following information was supplied regarding data availability:

The raw measurements are available in the Supplementary Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.12994#supplemental-information.

REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genetics* **25(1)**:25–29 DOI 10.1038/75556.

Baek CH, Kim H, Moon SY, Park S-K, Yang WS. 2019. Epigallocatechin-3-gallate downregulates lipopolysaccharide signaling in human aortic endothelial cells by inducing ectodomain shedding of TLR4. *European Journal of Pharmacology* **863**:172692 DOI 10.1016/j.ejphar.2019.172692.

Cai Y, Yu SS, He Y, Bi XY, Gao S, Yan TD, Zheng GD, Chen TT, Ye JT, Liu PQ. 2021. EGCG inhibits pressure overload-induced cardiac hypertrophy via the PSMB5/Nmnat2/SIRT6-dependent signalling pathways. *Acta Physiologica* **231(4)**:e13602.

Cao Y, Han X, Pan H, Jiang Y, Peng X, Xiao W, Rong J, Chen F, He J, Zou L, Tang Y, Pei Y, Zheng J, Wang J, Zhong J, Hong X, Liu Z, Zheng Z. 2020. Emerging protective roles of shengmai injection in septic cardiomyopathy in mice by inducing myocardial mitochondrial autophagy via caspase-3/Beclin-1 axis. *Inflammation Research* **69(1)**:41–50 DOI 10.1007/s00011-019-01292-2.

Cheng C-Y, Barro L, Tsai S-T, Feng T-W, Wu X-Y, Chao C-W, Yu R-S, Chin T-Y, Hsieh MF. 2021. Epigallocatechin-3-gallate-loaded liposomes favor anti-inflammation of microglia cells and promote neuroprotection. *International Journal of Molecular Sciences* **22(6)**:3037 DOI 10.3390/ijms22063037.

Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. 2014. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Systems Biology* **8**:S11 DOI 10.1186/1752-0509-8-S4-S11.

Cook SJ, Stuart K, Gilley R, Sale MJ. 2017. Control of cell death and mitochondrial fission by ERK1/2 MAP kinase signalling. *The FEBS Journal* **284(24)**:4177–4195 DOI 10.1111/febs.14122.

Ding H, Chen L, Hong Z, Yu X, Wang Z, Feng J. 2021. Network pharmacology-based identification of the key mechanism of quercetin acting on hemochromatosis. *Metallomics* **13(6)**:mfab025 DOI 10.1093/mtomcs/mfab025.

Duan J, Chen Z, Liang X, Chen Y, Li H, Tian X, Zhang M, Wang X, Sun H, Kong D, Li Y, Yang J. 2020. Construction and application of therapeutic metal-polyphenol capsule for peripheral artery disease. *Biomaterials* **255**:120199 DOI 10.1016/j.biomaterials.2020.120199.

Eng QY, Thanikachalam PV, Ramamurthy S. 2018. Molecular understanding of Epigallocatechin gallate (EGCG) in cardiovascular and metabolic diseases. *Journal of Ethnopharmacology* **210**:296–310 DOI 10.1016/j.jep.2017.08.035.

Guan M, Guo L, Ma H, Wu H, Fan X. 2021. Network pharmacology and molecular docking suggest the mechanism for biological activity of rosmarinic acid. *Evidence-Based Complementary and Alternative Medicine* **2021**:1–10.

Heureux ML, Sternberg M, Brath L, Turlington J, Kashiouris MG. 2020. Sepsis-induced cardiomyopathy: a comprehensive review. *Current Cardiology Reports* **22(5)**:35 DOI 10.1007/s11886-020-01277-2.

Hollenberg SM, Singer M. 2021. Pathophysiology of sepsis-induced cardiomyopathy. *Nature Reviews Cardiology* **18(6)**:424–434 DOI 10.1038/s41569-020-00492-2.

Huang HT, Cheng TL, Yang CD, Chang CF, Ho CJ, Chuang SC, Li J-Y, Huang S-H, Lin Y-S, Shen H-Y, Yu T-H, Kang L, Lin S-Y, Chen C-H. 2021. Intra-articular

injection of (-)-epigallocatechin 3-gallate (EGCG) ameliorates cartilage degeneration in guinea pigs with spontaneous osteoarthritis. *Antioxidants* **10(2)**:178 DOI 10.3390/antiox10020178.

Joo HJ, Park JY, Hong SJ, Kim KA, Lee SH, Cho JY, Park JH, Yu CW, Lim D-S. 2018. Anti-platelet effects of epigallocatechin-3-gallate in addition to the concomitant aspirin, clopidogrel or ticagrelor treatment. *The Korean Journal of Internal Medicine* **33(3)**:522–531 DOI 10.3904/kjim.2016.228.

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45**:D353–D361 DOI 10.1093/nar/gkw1092.

Kar AK, Singh A, Dhiman N, Purohit MP, Jagdale P, Kamthan M, Singh D, Kumar M, Ghosh D, Patnaik S. 2019. Polymer-assisted in situ synthesis of silver nanoparticles with epigallocatechin gallate (EGCG) impregnated wound patch potentiate controlled inflammatory responses for brisk wound healing. *International Journal of Nanomedicine* **14**:9837–9854 DOI 10.2147/IJN.S228462.

Kerr BA, Ma L, West XZ, Ding L, Malinin NL, Weber ME, Tischenko M, Goc A, Somanath PR, Penn MS, Podrez EA, Byzova TV. 2013. Interference with Akt signaling protects against myocardial infarction and death by limiting the consequences of oxidative stress. *Science Signaling* **6(287)**:a67.

Kibble M, Saarinen N, Tang J, Wennerberg K, Mäkelä S, Aittokallio T. 2015a. Network pharmacology applications to map the unexplored target space and therapeutic potential of natural products. *Natural Product Reports* **32(8)**:1249–1266 DOI 10.1039/C5NP00005J.

Kibble M, Saarinen N, Tang J, Wennerberg K, Mäkelä S, Aittokallio T. 2015b. Network pharmacology applications to map the unexplored target space and therapeutic potential of natural products. *Natural Product Reports* **32(8)**:1249–1266 DOI 10.1039/C5NP00005J.

Lee DH, Kim YJ, Kim HH, Cho HJ, Ryu JH, Rhee MH, Park H-J. 2013. Inhibitory effects of epigallocatechin-3-gallate on microsomal cyclooxygenase-1 activity in platelets. *Biomolecules & Therapeutics* **21(1)**:54–59 DOI 10.4062/biomolther.2012.075.

Li R, Guo C, Li Y, Qin Z, Huang W. 2020a. Therapeutic targets and signaling mechanisms of vitamin C activity against sepsis: a bioinformatics study. *Briefings in Bioinformatics* **22(3)**:bbaa079.

Li ZH, Shi Z, Tang S, Ping Yao H, Lin X, Wu F. 2020b. Epigallocatechin-3-gallate ameliorates LPS-induced inflammation by inhibiting the phosphorylation of Akt and ERK signaling molecules in rat H9c2 cells. *Experimental and Therapeutic Medicine* **20(2)**:1621–1629 DOI 10.3892/etm.2020.8827.

Li NN, Xiang SY, Huang XX, Li YT, Luo C, Ju PJ, Xu Y-F, Chen J-H. 2021a. Network pharmacology-based exploration of therapeutic mechanism of Liu-Yu-Tang in atypical antipsychotic drug-induced metabolic syndrome. *Computers in Biology and Medicine* **134**:104452 DOI 10.1016/j.compbiomed.2021.104452.

Li Y, Zhao Y, Han J, Wang Y, Lei S. 2021b. Effects of epigallocatechin gallate (EGCG) on the biological properties of human dental pulp stem cells and inflammatory pulp tissue. *Archives of Oral Biology* **123**:105034 DOI 10.1016/j.archoralbio.2020.105034.

Liang Y, Ip MSM, Mak JCW. 2019. (-)-Epigallocatechin-3-gallate suppresses cigarette smoke-induced inflammation in human cardiomyocytes via ROS-mediated MAPK and NF-κB pathways. *Phytomedicine: International Journal of Phytotherapy and Phytopharmacology* **58**:152768 DOI 10.1016/j.phymed.2018.11.028.

Liu F, Li L, Chen J, Wu Y, Cao Y, Zhong P. 2021. Calculus bovisa network pharmacology to explore the mechanism of in the treatment of ischemic stroke. *BioMed Research International* **2021**:6611018.

Liu Z, Li Q, Huang J, Liang Q, Yan Y, Lin H, Xiao W, Lin Y, Zhang S, Tan B, Luo G. 2013. Proteomic analysis of the inhibitory effect of epigallocatechin gallate on lipid accumulation in human HepG2 cells. *Proteome Science* **11(1)**:32 DOI 10.1186/1477-5956-11-32.

Ma Y, Hu Y, Wu J, Wen J, Li S, Zhang L, Zhang J, Li Y, Li J. 2019. Epigallocatechin-3-gallate inhibits angiotensin II-induced cardiomyocyte hypertrophy via regulating Hippo signaling pathway in H9c2 rat cardiomyocytes. *Acta Biochimica et Biophysica Sinica* **51(4)**:422–430 DOI 10.1093/abbs/gmz018.

Ma Y, Liu G, Tang M, Fang J, Jiang H. 2021. Epigallocatechin gallate can protect mice from acute stress induced by LPS while stabilizing gut microbes and serum metabolites levels. *Frontiers in Immunology* **12**:640305 DOI 10.3389/fimmu.2021.640305.

Martin L, Derwall M, Al Zoubi S, Zechendorf E, Reuter DA, Thiemermann C, Schuerholz T. 2019. The septic heart: current understanding of molecular mechanisms and clinical implications. *Chest* **155(2)**:427–437 DOI 10.1016/j.chest.2018.08.1037.

Mebratu Y, Tesfaigzi Y. 2009. How ERK1/2 activation controls cell proliferation and cell death: Is subcellular localization the answer? *Cell Cycle* **8(8)**:1168–1175 DOI 10.4161/cc.8.8.8147.

Nikoo M, Regenstein JM, Ahmadi Gavlighi H. 2018. Antioxidant and antimicrobial activities of (-)-epigallocatechin-3-gallate (EGCG) and its potential to preserve the quality and safety of foods. *Comprehensive Reviews in Food Science and Food Safety* **17(3)**:732–753 DOI 10.1111/1541-4337.12346.

Nogueira V, Park Y, Chen C-C, Xu P-Z, Chen M-L, Tonic I, Unterman T, Hay N. 2008. Akt determines replicative senescence and oxidative or oncogenic premature senescence and sensitizes cells to oxidative apoptosis. *Cancer Cell* **14(6)**:458–470 DOI 10.1016/j.ccr.2008.11.003.

Othman AI, El-Sawi MR, El-Missiry MA, Abukhalil MH. 2017. Epigallocatechin-3-gallate protects against diabetic cardiomyopathy through modulating the cardiometabolic risk factors, oxidative stress, inflammation, cell death and fibrosis in streptozotocin-nicotinamide-induced diabetic rats. *Biomedicine & Pharmacotherapy* **94**:362–373 DOI 10.1016/j.biopha.2017.07.129.

Othman AI, Elkomy MM, El-Missiry MA, Dardor M. 2017b. Epigallocatechin-3-gallate prevents cardiac apoptosis by modulating the intrinsic apoptotic pathway

in isoproterenol-induced myocardial infarction. *European Journal of Pharmacology* **794**:27–36 DOI 10.1016/j.ejphar.2016.11.014.

Sarah JB, Weber G, Sarge T, Nikravan S, Grissom CK, Lanspa MJ, Shahul S, Brown S. 2018. Septic cardiomyopathy. *Critical Care Medicine* **46**(**4**):625–634.

Shang J, He Q, Chen Y, Yu D, Sun L, Cheng G, Liu D, Xiao J, Zhao Z. 2018. miR-15a-5p suppresses inflammation and fibrosis of peritoneal mesothelial cells induced by peritoneal dialysis via targeting VEGFA. *Journal of Cellular Physiology* **234**(**6**):9746–9755.

Tang H, Zhang S, Huang C, Li K, Zhao Q, Li X. 2021. MiR-448-5p/VEGFA axis protects cardiomyocytes from hypoxia through regulating the FAS/FAS-L signaling pathway. *International Heart Journal* **62**(**3**):647–657 DOI 10.1536/ihj.20-600.

Wang J, Fan SM, Zhang J. 2019. Epigallocatechin-3-gallate ameliorates lipopolysaccharide-induced acute lung injury by suppression of TLR4/NF-κB signaling activation. *The Brazilian Journal of Medical and Biological Research* **52**(**7**):e8092 DOI 10.1590/1414-431x20198092.

Wang Y, Liu T, Ma F, Lu X, Mao H, Zhou W, Yang L, Li P, Zhan Y. 2020. A network pharmacology-based strategy for unveiling the mechanisms of tripterygium wilfordii hook F against diabetic kidney disease. *Journal of Diabetes Research* **2020**:2421631.

Wang M, Zhong H, Zhang X, Huang X, Wang J, Li Z, Chen M, Xiao Z. 2021. EGCG promotes PRKCA expression to alleviate LPS-induced acute lung injury and inflammatory response. *Scientific Reports* **11**(**1**):11014 DOI 10.1038/s41598-021-90398-x.

Xie LW, Cai S, Zhao TS, Li M, Tian Y. 2020. Green tea derivative (-)-epigallocatechin-3-gallate (EGCG) confers protection against ionizing radiation-induced intestinal epithelial cell death both in vitro and in vivo. *Free Radical Biology & Medicine* **161**:175–186 DOI 10.1016/j.freeradbiomed.2020.10.012.

Xiong Z, Yang F, Li W, Tang X, Ma H, Yi P. 2021. Deciphering pharmacological mechanism of Buyang Huanwu decoction for spinal cord injury by network pharmacology approach. *Evidence-Based Complementary and Alternative Medicine* **2021**:1–20.

Yamagata K. 2020. Protective effect of epigallocatechin gallate on endothelial disorders in atherosclerosis. *Journal of Cardiovascular Pharmacology* **75**(**4**):292–298 DOI 10.1097/FJC.0000000000000792.

Yang J, Cheng M, Gu B, Wang J, Yan S, Xu D. 2020. CircRNA_09505 aggravates inflammation and joint damage in collagen-induced arthritis mice via miR-6089/AKT1/NF-κB axis. *Cell Death & Disease* **11**(**10**):833 DOI 10.1038/s41419-020-03038-z.

Yang L, Zhang H, Chen P. 2018. Sulfur dioxide attenuates sepsis-induced cardiac dysfunction via inhibition of NLRP3 inflammasome activation in rats. *Nitric Oxide* **81**:11–20 DOI 10.1016/j.niox.2018.09.005.

Yang L, Zheng L, Chng WJ, Ding JL. 2019. Comprehensive analysis of ERK1/2 substrates for potential combination immunotherapies. *Trends in Pharmacological Sciences* **40**(**11**):897–910 DOI 10.1016/j.tips.2019.09.005.

Yu S, Gao W, Zeng P, Chen C, Zhang Z, Liu Z, Liu J. 2021. Exploring the effect of Gupi Xiaoji prescription on hepatitis B virus-related liver cancer through network pharmacology and in vitro experiments. *Biomedicine & Pharmacotherapy* **139**:111612 DOI 10.1016/j.biopha.2021.111612.

Zhang P, Chen H, Shen G, Zhang Z, Yu X, Shang Q, Zhao W, Li D, Li P, Chen G, Liang D, Jiang X, Ren H. 2021a. Network pharmacology integrated with experimental validation reveals the regulatory mechanism of plastrum testudinis in treating senile osteoporosis. *Journal of Ethnopharmacology* **276**:114198 DOI 10.1016/j.jep.2021.114198.

Zhang Z, Zhang X, Bi K, He Y, Yan W, Yang CS, Zhang J. 2021b. Potential protective mechanisms of green tea polyphenol EGCG against COVID-19. *Trends in Food Science & Technology* **114**:11–24 DOI 10.1016/j.tifs.2021.05.023.

Investigating effect of mutation on structure and function of G6PD enzyme: a comparative molecular dynamics simulation study

Sadaf Rani[1], Fouzia Perveen Malik[1], Jamshed Anwar[2] and Rehan Zafar Paracha[1]

[1] School of Interdisciplinary Engineering and Sciences (SINES), National University of Sciences and Technology, Islamabad, Federal, Pakistan

[2] Department of Chemistry, Lancaster University, UK, Lancaster, United Kingdom, UK

ABSTRACT

Several natural mutants of the human G6PD enzyme exist and have been reported. Because the enzymatic activities of many mutants are different from that of the wildtype, the genetic polymorphism of G6PD plays an important role in the synthesis of nucleic acids via ribulose-5-phosphate and formation of reduced NADP in response to oxidative stress. G6PD mutations leading to its deficiency result in the neonatal jaundice and acute hemolytic anemia in human. Herein, we demonstrate the molecular dynamics simulations of the wildtype G6PD and its three mutants to monitor the effect of mutations on dynamics and stability of the protein. These mutants are Chatham (A335T), Nashville (R393H), Alhambra (V394L), among which R393H and V394L lie closer to binding site of structural $NADP^+$. MD analysis including RMSD, RMSF and protein secondary structure revealed that decrease in the stability of mutants is key factor for loss of their activity. The results demonstrated that mutations in the G6PD sequence resulted in altered structural stability and hence functional changes in enzymes. Also, the binding site, of structural $NADP^+$, which is far away from the catalytic site plays an important role in protein stability and folding. Mutation at this site causes changes in structural stability and hence functional deviations in enzyme structure reflecting the importance of structural $NADP^+$ binding site. The calculation of binding free energy by post processing end state method of Molecular Mechanics Poisson Boltzmann SurfaceArea (MM-PBSA) has inferred that ligand binding in wildtype is favorable as compared to mutants which represent destabilised protein structure due to mutation that in turn may hinder the normal physiological function. Exploring individual components of free energy revealed that the van der Waals energy component representing non-polar/hydrophobic energy contribution act as a dominant factor in case of ligand binding. Our study also provides an insight in identifying the key inhibitory site in G6PD and its mutants which can be exploited to use them as a target for developing new inhibitors in rational drug design.

Subjects Biochemistry, Bioinformatics, Molecular Biology, Computational Science

Keywords G6PD mutants, MD simulations, Binding site, Structural stability, Binding free energy, Catalysis

Submitted 1 October 2021
Accepted 1 February 2022
Published 29 March 2022

Corresponding authors
Fouzia Perveen Malik,
fouzia@rcms.nust.edu.pk
Jamshed Anwar,
j.anwar@lancaster.ac.uk

Academic editor
Pedro Silva

Additional Information and Declarations can be found on page 28

DOI **10.7717/peerj.12984**

How to cite this article Rani S, Malik FP, Anwar J, Zafar Paracha R. 2022. Investigating effect of mutation on structure and function of G6PD enzyme: a comparative molecular dynamics simulation study. *PeerJ* **10**:e12984 http://doi.org/10.7717/peerj.12984

INTRODUCTION

Human Glucose-6-Phosphate Dehydrogenase (G6PD) is a rate-limiting enzyme of the pentose phosphate pathway (PPP) involved in the formation of 6-phosphogluconolactonere leasing NADPH which directs the formation of ribulose 5- phosphate producing nucleotides at the end of the pathway. The PPP is the only source of NADPH in red blood cells (RBC's) required to protect the cells from oxidative damage caused by reactive oxygen species (ROS) (*WHO Working Group, 1989*). NADPH maintains the level of reduced glutathione which is vital for the reduction of H_2O_2 and oxygen free radicals, thus controlling the concentration of RBC's proteins including hemoglobin (*Desnick et al., 2001*). An elevated level of reactive oxygen species causes loss of membrane integrity leading to hemolysis in RBCs (*Dessi et al., 1984*; *Rao, Kottapally & Shinozuka, 1984*; *Townsend & Tew, 2003*).

The *G6PD* gene is located on the x- chromosome consisting of 13 exons and 12 introns (*Persico et al., 1986*), and the product of this gene is a protein consisting of 515 amino acids with a molecular weight of 59 kDa located in the cytosol (*Rattazzi, 1968*). Mutations in the *G6PD* gene result in an x-linked hereditary disease known as G6PD deficiency, which is associated with the protein variants having different levels of enzyme activity leading to a wide variety of biochemical and clinical phenotypes (*Desnick et al., 2001*). G6PD deficiency is the most common RBC enzyme deficiency which affects more than 400 million individuals worldwide (*Cappellini & Fiorelli, 2008*). More than 300 G6PD mutations have been reported to date, and most amongst them are the result of a single base pair change in the amino acid sequence of G6PD protein. G6PD deficiency is characterized in five classes based on clinical manifestations and residual enzyme activity, which are class I to V (*WHO Working Group, 1989*). Generally, G6PD deficient individuals remain asymptotic throughout their life except when exposed to certain drugs or with intake of certain food such as fava beans, resulting in hemolysis (*WHO Working Group, 1989*). The most common clinical symptoms of G6PD deficiency are hemolytic anemia, neonatal jaundice and chronic non-spherocytic hemolytic anemia (CNSHA) (*Desnick et al., 2001*).

The structure of G6PD exists in dimer or tetramer equilibrium based on electrostatic interactions, pH and ionic strength. The dimer has two subunits located along symmetrical β-sheets. Each monomer of G6PD has a catalytic binding site which contains substrate G6P and coenzyme $NADP^+$ along with a structural $NADP^+$ binding site located 15 Å away from catalytic site. This structural $NADP^+$ site is present only in higher organisms (*Kotaka et al., 2005*) which play an important role in long term structural stability in human G6PD (*Wang et al., 2008*).

Point mutation at structural $NADP^+$ site causes decreased enzyme stability leading to Class-I deficiency according to WHO classification exhibiting a clinical condition known as chronic non spherocytic hemolytic anemia (CNSHA). Mutations which cause severe G6PD deficiency other than CNSHA are categorized as class-II deficiency. The nature of the binding site of structural $NADP^+$ indicates its structural importance in G6PD structure. Mutations other than class-I and class-II causing G6PD deficiency are away from the structural domain of G6PD protein. Studies show that point mutations in class I,

class II and III have a low catalytic ability, decreased efficiency and less compact structure compared to wildtype G6PD (*Gómez-Manzo et al., 2014*; *Au et al., 2000*).

Mutations close to the binding site alter initial properties and interaction of G6PD with substrate and coenzyme $NADP^+$, resulting in a change of thermodynamic properties and interaction of enzyme (*Nguyen, Nguyen & Le, 2016*).

More than 300 mutants of G6PD exist in nature; therefore, experimental studies for all mutants are deemed to be impractical. On the other hand, computational methods, in particular molecular dynamics simulation, are a robust and efficient tool for providing detailed insight into the dynamical properties of proteins and structural changes associated with them. The polymorphic nature of G6PD makes the enzyme a good candidate to explore the dynamical behavior and structural stability of its mutants with varying thermodynamics using molecular dynamics simulation. Only a few G6PD mutants have been previously studied via computational methods (*Nguyen, Nguyen & Le, 2016*; *Doss et al., 2016*); however, mutations in the dimer interface and their effect on the structural and functional properties of the G6PD enzyme still need attention to predict and design potent inhibitors against these mutants.

Here, we have focused on G6PD mutants with known pathological conditions which have mutations close to the dimer interface; to understand the effect of structural changes and their effects on substrate binding site associated with these mutations. Among the selected mutants A335T known as G6PD Chatham with 1003G>A cDNA substitution at Exon 9 lying at the βH- βI loop of the protein. This variant has been reported in Italy, Asia and Africa. According to the pathological condition it is categorized as class-II variant.

The second mutant R393H is known as Nashville, with 1178G->A cDNA substitution at exon 10 lying at dimer interface at β sheet βL. The variant is categorized in class-I based on severity of pathogenicity. This variant is reported in the USA, Italy and Portugal.

Mutant V394L, known as Alhambra, with 1180G->A cDNA substitution also lies at the dimer interface at β sheet βL adjacent to R393H. The mutation in the dimer interface causes severe clinical manifestation and categorized in class-I similar to R393H. This variant has been reported in Sweden and Finland (*Minucci et al., 2012*).

For these selected mutants, we have investigated the effect of amino acid mutations on 3D structures and structural flexibility of proteins is investigated using MD simulations of the wildtype and three mutants of G6PD. The objective of the work is to explore structural and dynamic effect of mutations that may act as guide for potential design of inhibitors to treat mutants for their normal functioning leading to cure of diseases associated with these mutations. Moreover the structural changes in enzymatic activities of the mutant are evaluated from the simulation results.

MATERIALS & METHODS

Model preparation

X-ray crystallographic structures of wildtype G6PD enzyme complexed with G6P substrate, coenzyme $NADP^+$, and structural $NADP^+$ were retrieved from the protein data bank with PDB ID 2BHL and 2BH9 (*Kotaka et al., 2005*). The substrate, the coenzyme, and

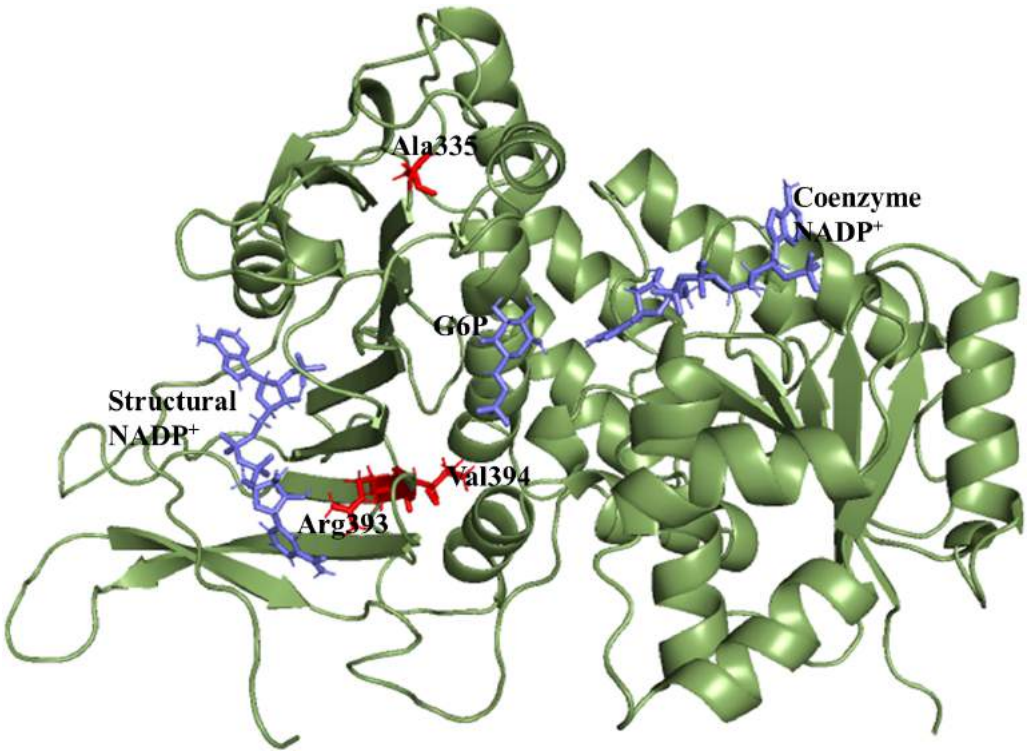


Figure 1 Modelled structure of G6PD enzyme. 3D model of G6PD. Ligand G6P, Coenzyme $NADP^+$ and Structural $NADP^+$ are represented in blue. Mutants are shown in red color.
Full-size DOI: 10.7717/peerj.12984/fig-1

the structural $NADP^+$ are present as ligands in the wildtype human G6PD for normal functioning. However; no crystal structure of the human G6PD enzyme is available so far, with the three ligands altogether in the form of a single structure in the protein data bank. To build the wildtype G6PD enzyme structure, we developed a model structure by superimposing both the 3D structures (PDB IDs: 2BHL and 2BH9) with the help of PyMol software (*De Lano, 2002*) (Fig. 1).

Sequence alignment

UNIPROT and dBSNP databases were (accessed on Nov 2020) retrieved to obtain the sequence and position of amino acid mutations in wildtype G6PD. FASTA sequence of each mutant structure was retrived from dbSNP database. In order to identify the mutation position sequence alignment of above mentioned modeled structure of wildtype was carried out with mutant FASTA sequence (obtained from dbSNP database) using online T-coffee multiple sequence alignment tool (*Notredame, Higgins & Heringa, 2000*) via ClustalW (*Sievers et al., 2011*) with default parameter. To create mutant structure; In-silico site-directed mutagenesis was carried out at the corresponding mutation sites in the modeled wildtype G6PD structure using PyMol software (*De Lano, 2002*) choosing rotamer with minimum energy and clashes. Few of pathogenic mutations lying within binding site of structural $NADP^+$ and dimer interface belonging to classes I and class- III) were chosen for this study Table 1.

Table 1 Selected variants dbSNP accession number, classes, amino acid position and change. Data represents dSNP accession number, classes, amino acid position and change in each amino acid. Column V, depicts each amino acid of enzyme replaced in mutants, R by H, V by L and A by T.

Ser	Accession no	Amino acid/mutation	Class	Mutation
1	rs137852316	393 (Nashville/Anaheim)	Class-I	R>H
2	rs137852335	394 (Alhambra)	Class-I	V>L
3	rs5030869	335 (Chatham)	Class-III	A>T

Ligand modeling

Ligands were geometry optimized at Density functional (DFT) level of theory using B3LYP functional and 6-31G(d,p) basis set in Gaussian 09 (*Frisch et al., 2016*). From the optimized geometries ESP charge were calculated at HF/6-31G* level of theory followed by RESP charge in antechamber module of Amber 20 followed by parameterization of ligands using Generalized AMBER Force Field (GAFF) (*Wang et al., 2004*).

Molecular dynamics simulation

All the protein structures were subjected to molecular dynamics simulation with the help of GPU based pmemd module implemented in the AMBER 20 simulation package. The protein structure was treated with the Amber 99SB-ILDN force field (*Lindorff-Larsen et al., 2010*), whereas the ligands (G6P, coenzyme $NADP^+$, and structural $NADP^+$) were parameterized with the generalized AMBER force field (GAFF). Initial configuration of the simulation systems was obtained with the help of the xLEaP module of the simulation package. The systems were protonated, neutralized with counter ions, and solvated with TIP3P water models thus resulting in a dodecahedron box under periodic boundary conditions.

Simulation protocol

The solvated systems were minimized in three steps to remove bad contacts. In the first step, the steepest-descent minimization was performed while applying a harmonic restraint of 25 kcal mol^{-1} $Å^2$ on the solute for 100,000 steps following 10,000 steps of conjugate-gradient minimization. In the subsequent steps, restraints were gradually removed with a difference of 5 kcal mol^{-1} $Å^2$ on the solute using a steepest-decent algorithm. In the final step, an unrestrained minimization of 100,000 steps was performed for the whole system. The minimizations were followed by a five-stage heating step for 40 ps each in which the temperature was raised from 0 to 300 K with a difference of 50 K while applying the harmonic restraint of 25 kcal mol^{-1} $Å^2$ and removing 5 kcal mol^{-1} $Å^2$ in each step with a time step of 0.5 fs in canonical (NVT) ensemble. This was then followed by 2 ns equilibration in isobaric-isothermal (NPT) ensemble at a constant pressure of 1 bar using Berendsen barostat (*Berendsen et al., 1984*) while keeping a harmonic contact of 25 kcal mol^{-1} $Å^2$ on solute atoms that was subsequently removed with a difference of 25 kcal mol^{-1} $Å^2$ (*Ryckaert, Ciccotti & Berendsen, 1977*). On getting well-equilibrated, production MD was performed for 100 ns for all systems in the NPT ensemble. During simulation, the Particle-mesh Ewald summation method was used to calculate electrostatic interaction (*Darden, York & Pedersen, 1993*) using the the cut-off of 1.4nm for both the

electrostatic and van der Waals (vdW) interactions SHAKE algorithm was used to constrain bonds involving hydrogen atoms while setting the time-step of 2 fs for all the simulation. The sampling of the trajectories was carried out every 2 ps which were then processed for the analysis using CPPTRAJ package (*Roe & Cheatham III, 2013*) of AMBERTOOLS v.20. XMGRACE software program and R-studio were used to generate plots (*Turner, 2005*).

Binding free energy calculation

The binding free energy was calculated using the Molecular Mechanics-Poisson Boltzman surface Area (MMPBSA) method (*Homeyer & Gohlke, 2012*; *Miller III et al., 2012*). The binding free energy of the complex is given as below (*Miller III et al., 2012*; *Rastelli et al., 2010*):

$$\Delta G_{bind} = \Delta G_{complex} - \left(G_{protein} + G_{ligand}\right). \tag{1}$$

Total binding energy is the contribution of two thermodynamic quantities *i.e.*, ΔH and entropy ΔS (*Salmas et al., 2015*; *Hou et al., 2011*) as mentioned below:-

$$\Delta G_{bind} = \Delta H - T(\Delta S) \tag{2}$$

whereas enthalpy is the contribution of molecular mechanics energy and free energy of solvation as mentioned below (*Rastelli et al., 2010*)

$$\Delta H = \Delta \mathrm{E}_{MM} + \Delta G_{sol} \tag{3}$$

ΔE_{MM} signifies the molecular mechanics energy components of bonded and non-bonded forces of interactions (*Verma et al., 2016*)

$$\Delta E_{MM} = \Delta E_{\mathrm{bonds}} + \Delta E_{\mathrm{angles}} + \Delta E_{\mathrm{torsions}} + \Delta E_{vdW} + \Delta E_{\mathrm{electrostatics}}. \tag{4}$$

Molecular mechanics energy is the contribution of energy of bonded terms *e.g.*, bonds, angles, dihedrals and non-bonded energy van der Waals and electrostatic energy. For protein ligand complex the contribution of free energy due to bonded terms are excluded by taking it zero so Eq. (4) is modified as:-

$$\Delta E_{MM} = \Delta \mathrm{E}_{vdw} + \Delta \mathrm{E}_{elec} \tag{5}$$

The $\Delta \mathrm{E}_{elec}$ contributes to the polar contribution of solvation free energy and non-polar energy (*Rastelli et al., 2010*; *Salmas et al., 2015*; *Hou et al., 2011*)

$$\Delta G_{sol} = \Delta G_{polar} + \Delta G_{non-polar}. \tag{6}$$

Contribution of polar energy can be calculated using Poison Boltzmann model whereas nonpolar contribution can be calculated using the following equation:

$$\Delta \mathrm{G}_{non-polar} = \gamma SASA + \beta \tag{7}$$

where γ and β (*Homeyer & Gohlke, 2012*; *Salmas et al., 2015*; *Hou et al., 2011*; *Verma et al., 2016*) are effective surface tension parameter and offset value respectively.

MM-PBSA calculations

To find out binding free energy, the single MD trajectory approach (*Homeyer & Gohlke, 2012*; *Srinivasan et al., 1998*) was applied using holo trajectory to extract snapshots of complex, ligand and protein. Five hundred frames were extracted throughout the simulation. PBSA program in amber was used employing Parse atomic radii (*Sitkoff, Sharp & Honig, 1994*; *Syeda Rehana & Zaheer, 2018*). Salt concentration of 0.150 and GB model with igb = 2 with mbondi2 was used. The surface tension (γ) value was fixed to 0.005 kcal/(mol Å^2). The solvent probe radius of 1.4 Å was used to calculate the SASA.

RESULTS

Molecular dynamics simulations

Before performing molecular dynamics simulation we calculated protein-ligand interaction profiles to know the key residues involved in the binding of ligand with protein and to get an insight whether these interactions would be maintained after MD simulation as a result of the effect of mutation on these interaction. Ligand interaction profile of the ligands, G6P and $NADP^+$ with the neighboring residues of wildtype G6PD enzyme is depicted in Fig. 2. G6PD enzyme provides three binding sites such as the site in the catalytic domain for the substrate (G6P) binding, the site for the coenzyme $NADP^+$ binding in close proximity to the G6P site and the structural $NADP^+$ binding, which is away from the catalytic domain. The binding site for the structural $NADP^+$ provides enhanced stability to the enzyme structure (*Wang et al., 2008*). Figure 2A illustrated G6P binding site in which Gln395, Lys205 interacted with the phosphate group of G6P via donor–acceptor interactions, whereas Lys171 belonging to a conserved peptide and His263 interacted with the hydroxyl group of G6P via hydrogen bonding. Figure 2B displayed interaction plot of the coenzyme $NADP^+$ with G6PD residues in which Arg72 interacted with the 2′-phosphate via hydrogen bonding, and Pro143 formed arene-arene interaction with the adenine ring of NADP. The main chain residues *i.e.*, Gly41 and Asp42 form hydrogen bond with adenine ribose 3′-hydroxyl and the bisphosphate O atoms of the coenzyme $NADP^+$ (Fig. 2B). Structural $NADP^+$ binding site is represented in the Fig. 2C. As seen in the Fig. 2C, at the structural $NADP^+$ binding site; Arg393 interacted with the nicotinamide group of the structural ligand, and Tyr401 and Tyr503 formed π–π stacking interaction with the nicotinamide and the adenine ring, respectively. Lys238, Lys366 and Arg487 interacted with the 2′ phosphate group through donor–acceptor interactions. Arg393, Glu239 and Pro396 are also found to be involved in the binding and defining shape of the binding site. The interaction profiles of these three ligands provided a static picture of the enzyme-ligand interaction to depict how the ligands attained stability in the binding site; however the dynamics of these interactions needed to be evaluated in detail through the simulation studies.

Effect of mutations on residual interactions

Mutations have pronounced effect on residual interactions of enzyme. Comparison of residual contact pattern for wildtypes and mutants have been depicted in Fig. 3. A335T lies at the end of βi of an extensive nine stranded β-sheet which forms the part of dimer

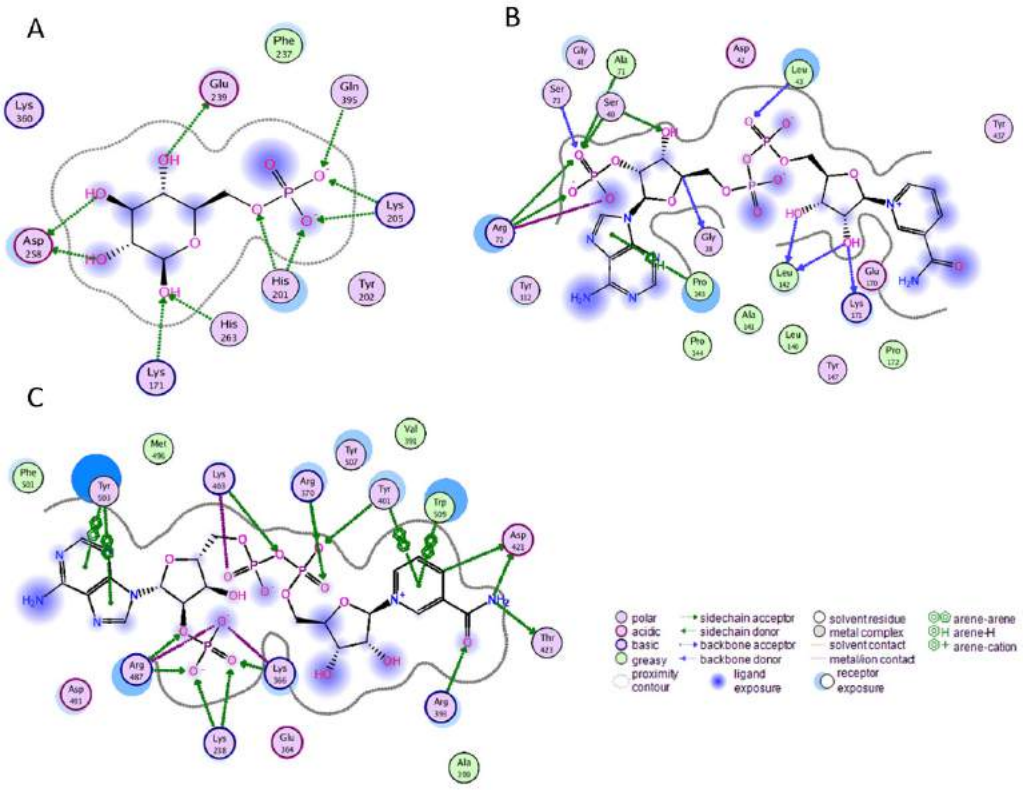


Figure 2 Ligand interactions of G6PD enzyme. 2D ligplot showing interactions of G6PD in binding sites with (A) G6P, (B) coenzyme NADP+, and (C) structural NADP+ interacting with residues of G6PD.
Full-size DOI: 10.7717/peerj.12984/fig-2

interface accommodating structural $NADP^+$. Alanine in wildtype is the part of loop region at the end of extensive β-sheet network close to dimer interface (Fig. 3A). Introduction of large sized threonine instead of alanine resulted in the formation of polar contacts with Ile 255 (Fig. 3B) in adjacent αi- helix.

Based on the severity of pathogenicity; the mutants in class-I of G6PD are characterized into two categories *i.e.*, lying in structural $NADP^+$ binding site and at dimer interface. R393H and V394L fall in both of these categories. Figure 3C depicted that guanidino group of Arg393 in wildtype interacted with amide oxygen of nicotinamide ring of structural $NADP^+$, Asn397 and Ala399, whereas in mutant R393H; replacement of Arg393 with Histidine resulted in the loss of these interactions. Histidine developed polar interaction with Gln395 and Ala399 (Fig. 3D). Loss of charge interaction may have resulted in the movement of nicotinamide ring of $NADP^+$ away leading to loss of key residue interactions.

Similarly; V394L lies in the binding site of structural $NADP^+$ being the part of large extended β-sheet region. In wildtype backbone contact of Valine 394 with Ala 367 is clearly observable according to Fig. 3E, however this contact of Ala 267 was maintained with Leucine as well in mutant V394L (Fig. 3F). Part of the β-sheet and αf helix are enriched with the hydrophobic side chain interactions between Ile392, V394, and V400 with M207, L211 and L214. In V394L; Incorporation of larger Leucine in V394L as compared to

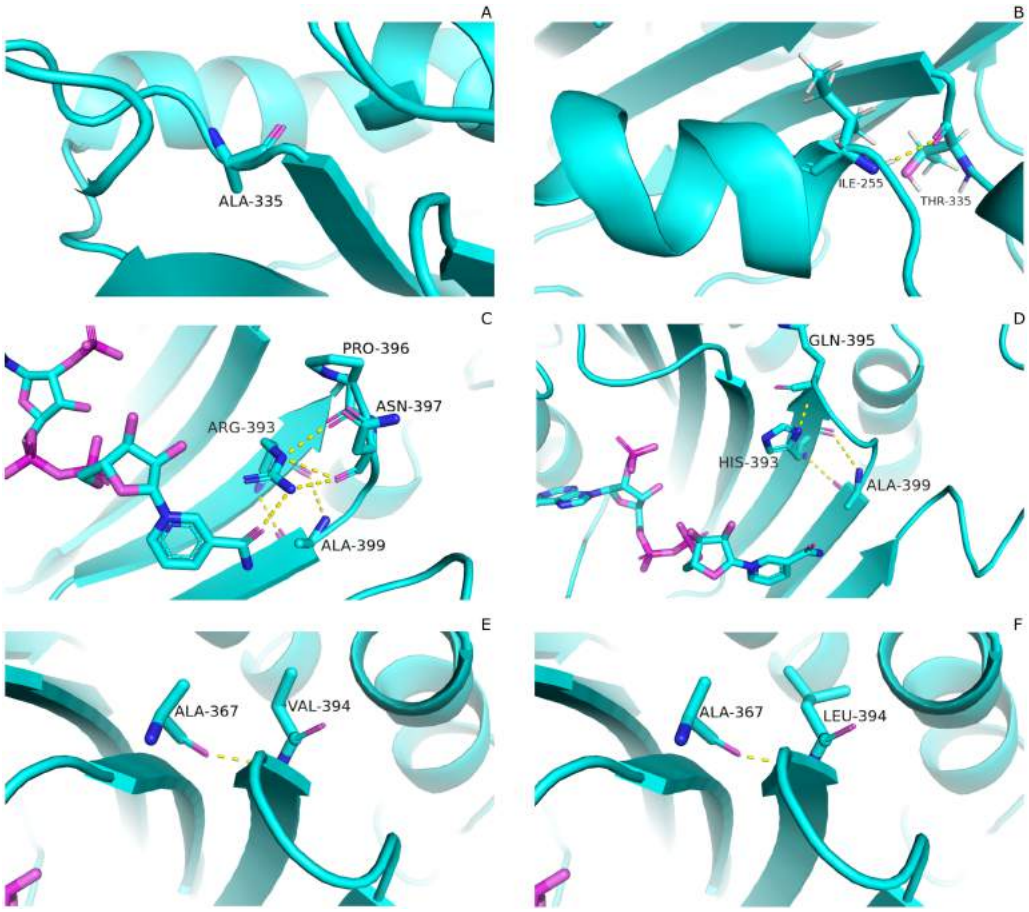


Figure 3 **Comparartive analysis of interaction G6PD and mutant amino acids.** (A) Wild type 335, (B) Mutant 335T, (C) Wild type 393R, (D) Mutant 393H, (E) Wild type 394V, (F) Mutant 394L. Mutants show change in the interaction pattern of neighboring residues.

Full-size DOI: 10.7717/peerj.12984/fig-3

smaller sized Valine may have displaced the adjoining residues causing week hydrophobic interactions and hence higher flexibility of this region. As a consequence, perturbation in αf helix may occur which is involved in the substrate binding and hence protein stability. Valine is a Cβ branched hydrophobic residue which may not be involved in catalysis however; its neighboring residue R393 is involved in forming hydrogen bond interaction with nicotinamide ring of structural $NADP^+$. There may be a change in the interactions at structural $NADP^+$ binding site with subsequent movement of $NADP^+$ resulting in overall change in dynamics.

Structural stability of wildtype G6PD and its mutants

Overall structural stability of wildtype complex was accessed by calculating root mean square deviation (RMSD) of C α atoms during the course of 100 ns simulation. Duplicate runs were performed and the average RMSD was found to be 2.5 Å as shown in Figs. 4A–4D. The system was equilibrated between 30 ns to 50 ns for first run. There was an increase in RMSD between 50 ns to 80 ns up to 3.2 Å after which it was equilibrated from 80 ns to 100

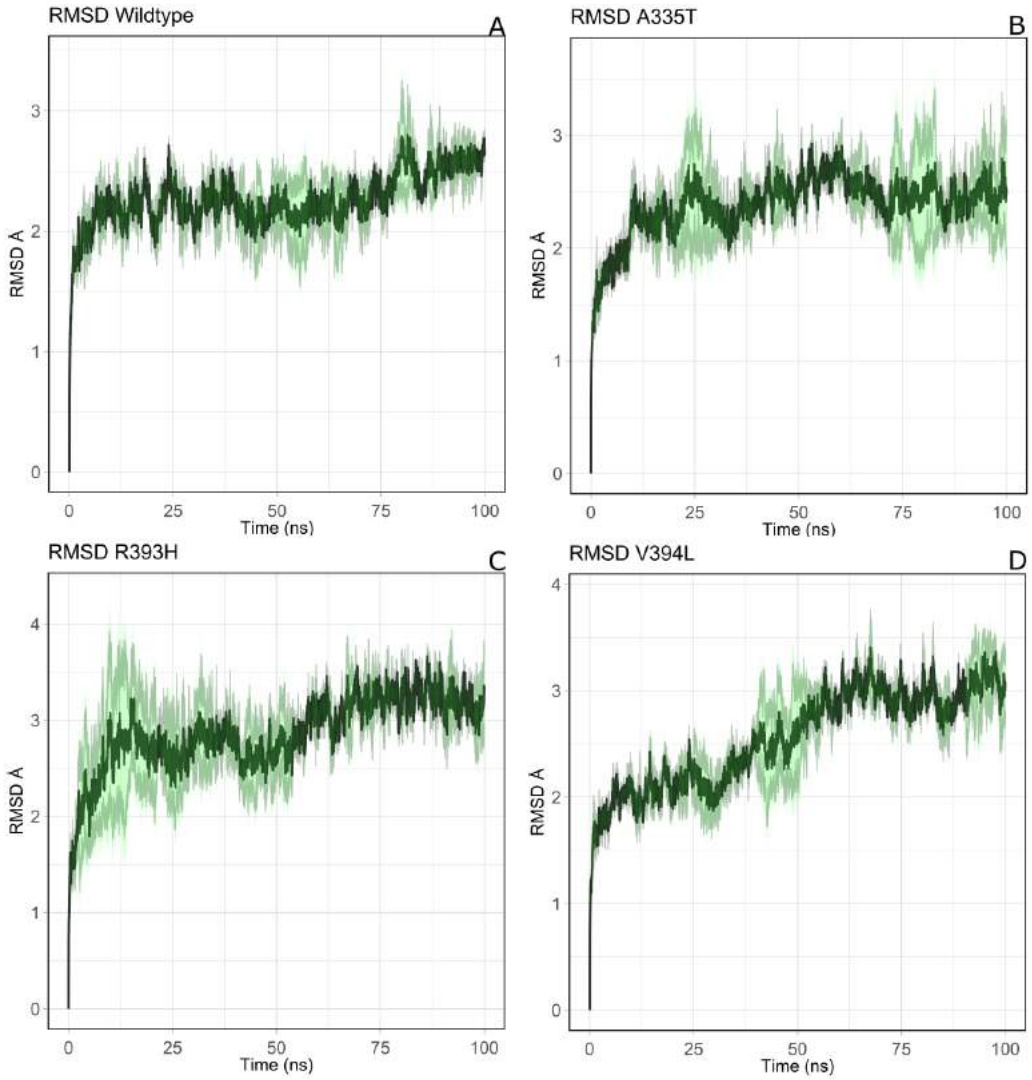


Figure 4 The comparative backbone RMSD of wild type and mutants in duplicate simulation. Run1 is shown in black and run2 shown in red color, (A) wildtype, (B) A365T, (C) R393H, (D) V394L. The reference structures of RMSD calculation were the starting structure of MD simulations.

Full-size DOI: 10.7717/peerj.12984/fig-4

ns. In duplicate simulation the system was equilibrated with an average RMSD of 2.11 Å. Overall average RMSD of the two structures was 2.5Å (Fig. 4A). No significant difference in stability of the two structures was observed which clearly indicated the reliability of results and enough stability of structure. The stable structures were further selected for comparison of stability between wildtype and its three mutants *i.e*; A335T, R393H and V394L via RMSD of their backbone atoms. For the clarity of differences between the structure of wildtype and mutants, residual contacts are presented in Fig. 3.

For the mutant A335T both runs seem to get equilibrated with an average RMSD of 2.6 Å Both systems tend to get stabilized between 80 ns to 100 ns with an RMSD of 2.1 Å. Overall average RMSD of the two structures was 2.3 Å (Fig. 4B). In mutant R393H; a

similar trend was observed in which the system shows stability between 40 ns to 60 ns with an average RMSD of 2.52 Å for both of the runs. A gradual increase in RMSD was observed at 80 ns for both structures which tend to stabilize with RMSD value of 2.8 Å at 100 ns. Overall average RMSD of the duplicate structures was 2.4 Å as depicted in the Fig. 4C. In V394L RMSD was found to be 3.2 Å on average. Both runs exhibited stability between 60 ns to 80 ns after which a fluctuation was seen for both of the structures with an average RMSD value of 3.1 Å (Fig. 4D).

The variation in RMSD signifies the conformational changes in the protein structure happening throughout the simulation. As it is evident from the Fig. 4A, backbone RMSD of wildtype G6PD fluctuated in the range between 2.0 to 2.5 Å, however; it was converged at an average value of 2.5 Å. On the other hand, the RMSD of A335T ranging from 2–2.8 Å converged at an average of 2.5 Å depicted in the Fig. 4B, however, R393H ranging from 2.5–3.3 Å converged at an average of 2.8 Å (Fig. 4C). RMSD of V394L ranged between 2.0 to 3.5 Å and converged at an average of 3.0 Å as evident from the Fig. 4D. Overall there was a slight difference in RMSD of wild type and mutant A335T. Also, RMSD plot of R393H AND V394L showed variation in RMSD as compared to wildtype. A slight difference in global RMSD of A335T may be attributed to the position of mutation away from the binding site of substrate and coenzyme indicating that the mutation did not alter significant change in protein stability. RMSD of R393H and V394L was higher as compared to wildtype providing an indication that mutation of essential residues in the binding site of structural $NADP^+$ could be ascribed to the low stability of the enzyme structure. In R393H; Arginine with electrically charged side chain has been replaced by Histidine both serving as basic residues. The replacement of guanidine group of Arginine with imidazole ring of Histidine resulted in the change of interactions with neighboring residues as shown in Figs. 3C–3D however; the effect would not be so pronounced to show a prominent change in the global RMSD.

In the same way, in V394L Valine having small side chain has been replaced with Leucine. Both residues belong to the same class of amino acids where the side chain is hydrophobic. The larger size of Leucine as compared to Valine may have changed the local distances of binding site resulting in a change in overall RMSD of the protein as shown in Fig. 4D.

The validation of changes in structural stability was further evaluated by the computation of averaged root mean square fluctuation (RMSF).

Residual conformational changes and associated flexibility

It was assumed after the preliminary RMSD calculations that mutations of the essential residues may lead to changes in the residual dynamics of the enzyme which were assessed as averaged RMSF for the backbone atoms of wild type and its mutants. Proteins in loop region are more flexible than rest of secondary structure elements. In order to evaluate the flexibility of G6PD enzyme; RMSF of the two duplicate simulations were compared and displayed in Figs. 5A–5D. RMSF plot disclose that loop regions in G6PD exhibited greater flexibility and randomness as compared to the other secondary structural elements.

Residue region 298–305, 350–358 and 384–386 representing loop region exhibited higher RMSF values for both of simulations in wildtype (Fig. 5A). Similarly the loop region

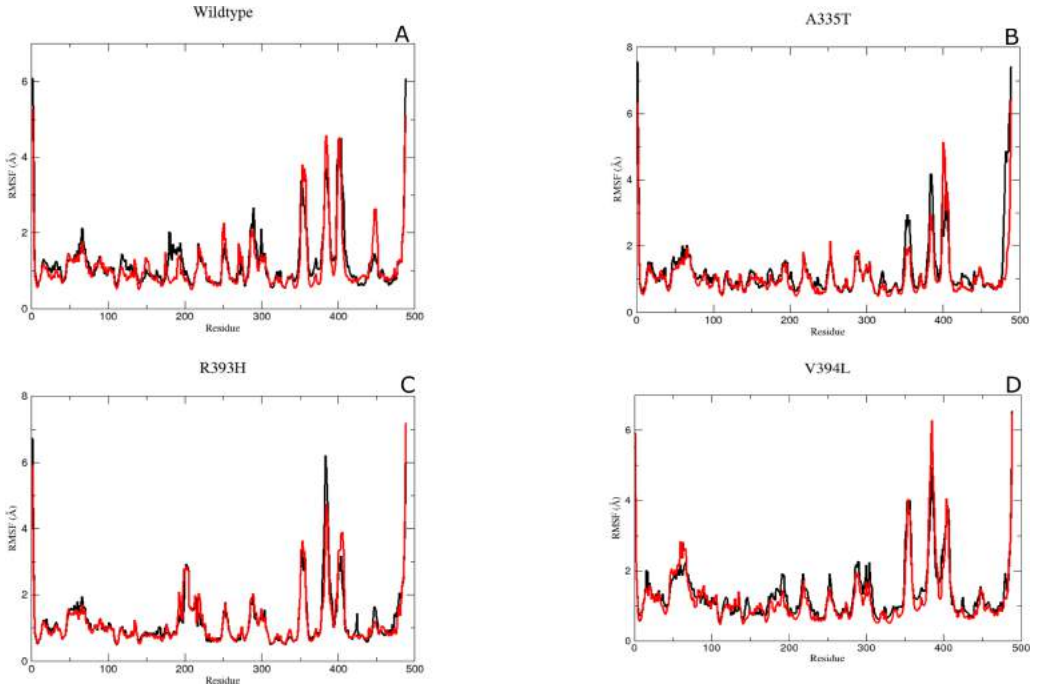


Figure 5 The comparative Backbone RMSF of wild type and mutants in duplicate simulation (run1 shown in black and run2 shown in red color) for flexibility analysis. (A) Wildtype, (B) A365T, (C) R393H, (D) V394L. The comparison of both the runs show the same trend in RMSF depicting reliability of simulation.

Full-size DOI: 10.7717/peerj.12984/fig-5

corresponding to residues 190–197 exhibited higher flexibility. This region is adjacent to conserved residue peptide (198–205) which is involved in substrate binding. Residue region 426–431 representing loop region and connecting the large βM-sheet with αI helix and residues 488–489 representing loop region of-c-terminus tail show higher fluctuation with RMSF value of 5.5 Å. Residue region 350–358 signified the loop region connecting two βL and βK.380-405 region exhibiting βM sheet showed RMSF values of 2.8 Å and 5.8 Å respectively revealing flexible nature of residues in this region of protein. This region belongs to the β + α domain of G6PD accommodating structural $NADP^+$ binding site. Higher RMSF may indicate the mobile nature of the region.

RMSF peaks for mutant A335T were comparable to the wildtype indicating the perseverance of flexibility of G6PD enzyme. *i.e.*, RMSF of the region harboring residues 350–358 exhibited value of 2.5 Å and 4 Å for residues 385–401 revealing the flexibility of these regions was not affected by mutation (Fig. 5B).

The RMSF plot for R393H and V394L (Figs. 5C & 5D) showed higher flexibility for the loop region connecting the βM-sheet with αI helix *i.e.*, 6Å as compared to wildtype. Interestingly the residue region involving the substrate binding *i.e.*, 198–201 exhibited higher flexibility in V394L as compared to wildtype.

Overall an increase in flexibility showed the instability of substrate binding site residues.

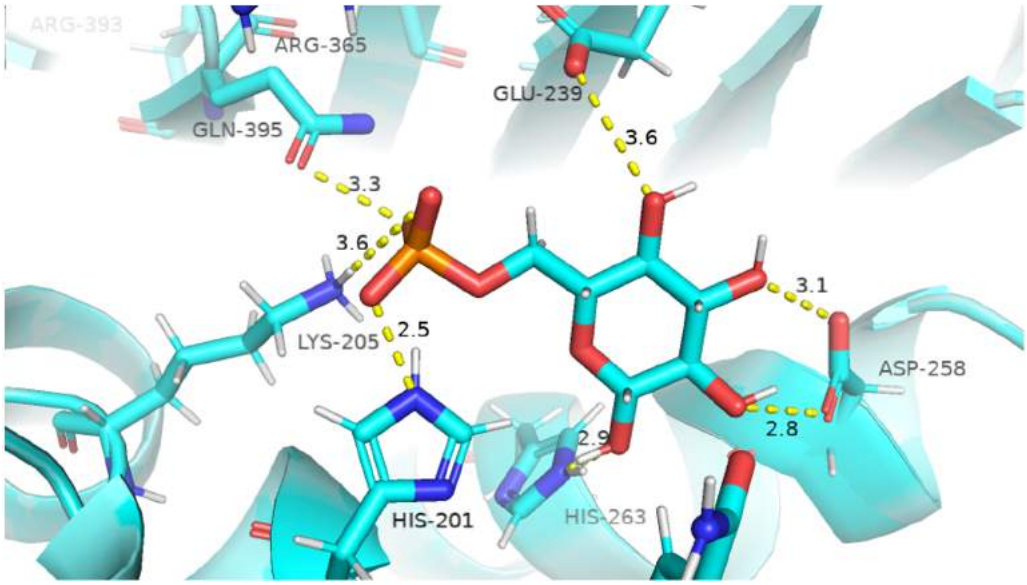


Figure 6 Native contacts between G6P and residues in the binding site. Native contacts between G6P and residues in the binding site. The native contacts were calculated based from crystal structure.
Full-size DOI: 10.7717/peerj.12984/fig-6

Flexibility of Residues in the binding site

Mutants R393H and V394L showed a decrease in flexibility for dinucleotide binding finger print region containing residues 38–44 demonstrating that these mutants did not affect the binding of coenzyme $NADP^+$ site.

The residue region belonging to the conserved peptide *i.e.*, 198–206 plays an important role in the substrate binding *e.g.,* His 201 is involved in stabilizing the G6P ring in proper orientation for catalysis. RMSF value for this region in wildtype and A335T was 1.8–1.9 Å as compared to the RMSF of R393H and V394L which was 2.5 Å. This difference manifested that mutation resulted in the increase in the flexibility of binding site residues in mutant R393H and V394L (Figs. 5A–5D)

Substrate interaction with protein and variation in residual contacts. The visual description of interactions of substrate G6P with the neighboring protein residues in crystal structure are depicted in Fig. 6. To describe the changes in dynamics of these interacting residues during the simulation; we compared the time evolution of average distance of residues in binding site of G6P in wildtype and mutants. (Fig. 7).

The interactions of G6P with neighboring residues were preserved in wildtype during the course of simulation. As it is apparent from the Fig. 6 that Arg 365 interacts with O3 atom of G6P in crystal structure. The time evolution graph of interaction of G6P with Arg365 for wildtype and mutants shown in Figs. 7A–7D. It is apparent from the Fig. 7A that in wildtype, at the start of simulation the average distance was 4 Å till 10 ns which gradually reduced to 3 Å until 60 ns. Upon reaching 100 ns the average distance maintained itself to 4 Å. In A335T this distance gradually increased from 5 Å to 7.5 Å till 50 ns after which it reduced to 6 Å on average as evident from the Fig. 7B. Likewise the average distance in

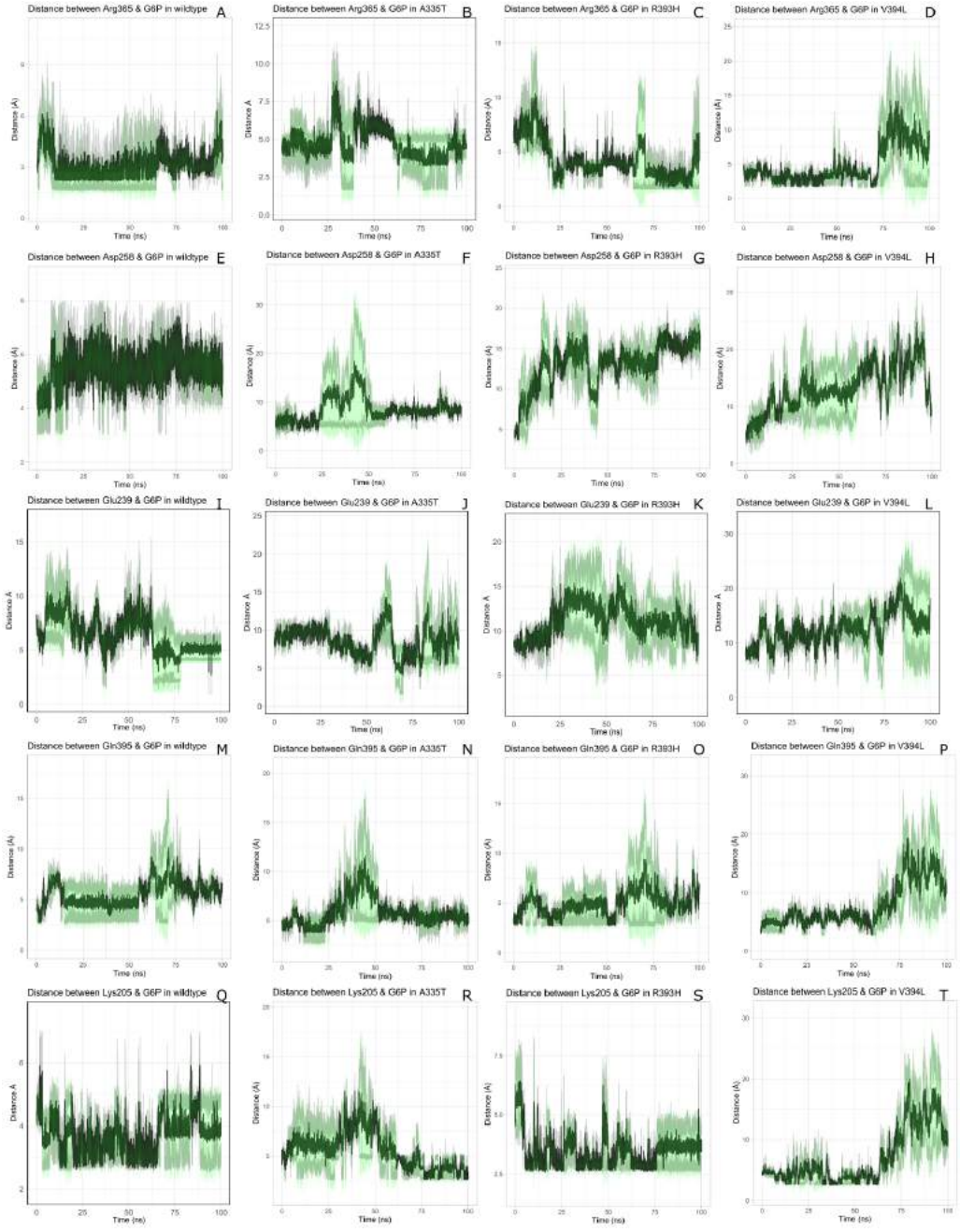


Figure 7 **Time evolution of the distance of G6P substrate.** Time evolution of the distance of G6P substrate with the key interacting residues in the binding site of wild type and mutant. The shaded lines represent individual run while dark green line show the average of the two runs.

Full-size DOI: 10.7717/peerj.12984/fig-7

distance R393H was found to be 4.5 Å between 20 ns to 60 ns and 70 ns to 90 ns however; fluctuations in the average distance up to 7.5 Å was observed from 0 ns to 20 ns. Also these fluctuation went up to 6 Å between 60 ns–65 ns and 80 ns–100 ns showing destabilization of hydrogen bond distance (Fig. 7C). In V394L the average distance remained up to 4.5 Å till 70 ns and increased up to 10 Å till 100 ns (Fig. 7D).

Average distances of Asp 258 and G6P is manifested in (Figs. 7E–7H) for wildtype and mutants respectively. Asp 258 in the binding site of G6P interacts with O5 and O4 atom of the G6P ring in crystal structure as depicted in Fig. 7E. In the wildtype the average distance remained 5 Å during 100 ns of simulation. A fluctuation in distance was observed for all of the three mutants *i.e.*, in A335T the average distance of 7 Å was maintained, In R393H and V394L the distance increased up to 12 Å during the simulation according to Figs. 7F–7H respectively. The fluctuations in the distances between residues in the binding sites with substrate in mutant structures showed that mutations resulted in perturbation of native residual contacts resulting in the overall less stable structure.

Figures 7I–7L demonstrated average distances of Glu 239 with G6P. As Fig. 6 indicated that Glu 239 interacts with O7 atom of the G6P in the crystal structure, a deeper analysis of Fig. validated that this contact was preserved in the wildtype during course of simulation with an average distance of 3.5 Å (Fig. 7I). In mutant A335T; a large fluctuation in distance was observed during simulation between 20 ns to 40 ns with an average of 15 Å which was reduced to an average of 4.5 Å from 50 ns to 100 ns indicating that the hydrogen bond was not preserved during some parts of simulation. Moreover A335T showed some conformational change in the larger loop movement as well (Fig. 7J). For R393H, Fig. 7K revealed that large fluctuation in distance were found with an average value of 7.5 Å. For mutant V394L initially this distance was 10 Å which gradually increased up to 15 Å from 75 ns to 100 ns indicating the movement of β-sheet away from the substrate (Fig. 7L).

Average distances between Gln395 and G6P were also calculated for wildtype and mutants and represented in Figs. 7M–7P). In the crystal structure of wildtype, the Gln395 interacts with O atom of G6P. This interaction was sustained till 55 ns after which a change in distance was observed between 60 ns to 70 ns which gradually reduced to 5 Å for wild type (Fig. 7M). This distance was increased up to 6.5 Å between 0 ns to 50 ns for mutant A335T t after which it was stabilized (Fig. 7N).

For mutant R393H, Fig. 7O signifies an increase in average distance to 7.5 Å manifesting that change in structural $NADP^+$ site resulted in the displacement of key G6P binding site residues as well. For V394L the distance between G6P and Gln 395 was 5 Å till 60 ns after which a gradual increase was observed with an average distance of 15 Å from 60 ns to 100 ns (Fig. 7P) Residue Lys205 from the conserved peptide region interact with the O2 atom of the phosphate group of G6P. Average distance of Lys205 with G6P are demonstrated in the Figs. 7Q–7T) for wildtype and mutants. Though this contact has been preserved in the wildtype with an average distance of 3.5 Å as reflected through Fig. 7Q. However; in A335T the distance increases from 5 Å to 7 Å till 50 ns after which a gradual decrease was observed maintaining a distance of 4 Å (Fig. 7R). Correspondingly, a fluctuation in the distance was observed in R393H from 0 ns to 70 ns after which it was maintained up to

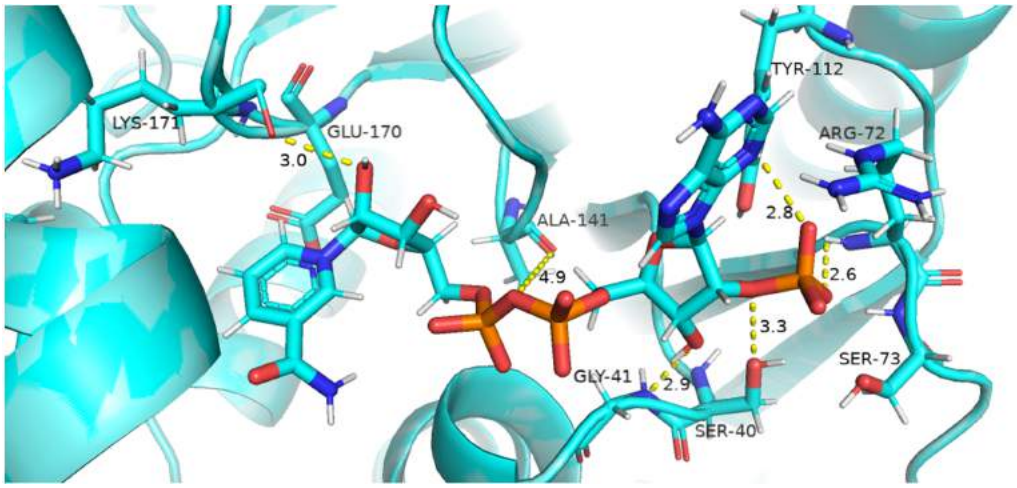


Figure 8 Native contacts between Coenzyme NADP$^+$ and residues in the binding site. The native contacts of *Coenzyme NADP*$^+$ were calculated based from crystal structure.
Full-size DOI: 10.7717/peerj.12984/fig-8

3.5 Å (Fig. 7S). The distance was not preserved for the mutant V394L as the simulation proceeded with a gradual increase to 10 Å between 75 ns–100 ns according to the Fig. 7T.

Coenzyme NADP$^+$ interaction with protein and variation in residual contacts. Residual contacts between key interacting residues and coenzymes NADP+ are described in the Fig. 8 and time evolution of key residues interaction with the cofactor NADP$^+$ were calculated as depicted in the Fig. 9.

Figure 8 presented that in crystal structure Gly 41 of the nucleotide binding finger print region (GASGDLA) form hydrogen bond with 3′ hydroxyl group of NADP$^+$. Average distances of interacting residues with coenzyme NADP$^+$ are depicted in Fig. 9.

Average distances of reside Gly41 with NADP+ for wild type, mutants A335, R393H and V394L are depicted in the Figs. 9A–9D) respectively. For wildtype the distance was reduced from 4 Å to 3.5 Å gradually during the course of simulation. Importantly; this distance of coenzyme NADP$^+$ remained conserved with an average value of 3.8 Å in A335T. For mutant R393H the average distance was 4 Å initially at 0 ns which gradually reduced to 3.5 Å. An average distance of 3.5 Å was observed in V394L.

Interactions of Arg72 with NADP$^+$ for wildtype and mutants are evident from Figs. 9E–9H. Guanidino group of Arg72 and 2′phosphate of NADP$^+$ form hydrogen bond. This interaction was maintained for wildtype with at an average distance of 2.5 Å. In A335T the distance was 3 Å till 70 ns after which a peak was observed from 70 ns to 95 ns with an average value of 4 Å. In R393H, this distance remained 2.9 Å with a peak between 40 ns to 45 ns. In V394L the average distance was 2.9 Å which increased gradually from 25 ns to 100 ns with an average value of 3.5 Å.

2′ hydroxyl group of nicotinamide ribose form hydrogen bond with the Lys 171 of the βE- αe turn in crystal structure (Fig. 8). The average distance for the wildtype was calculated as 7 Å. In A335T average distance of 6.5 Å was maintained till 50 ns which

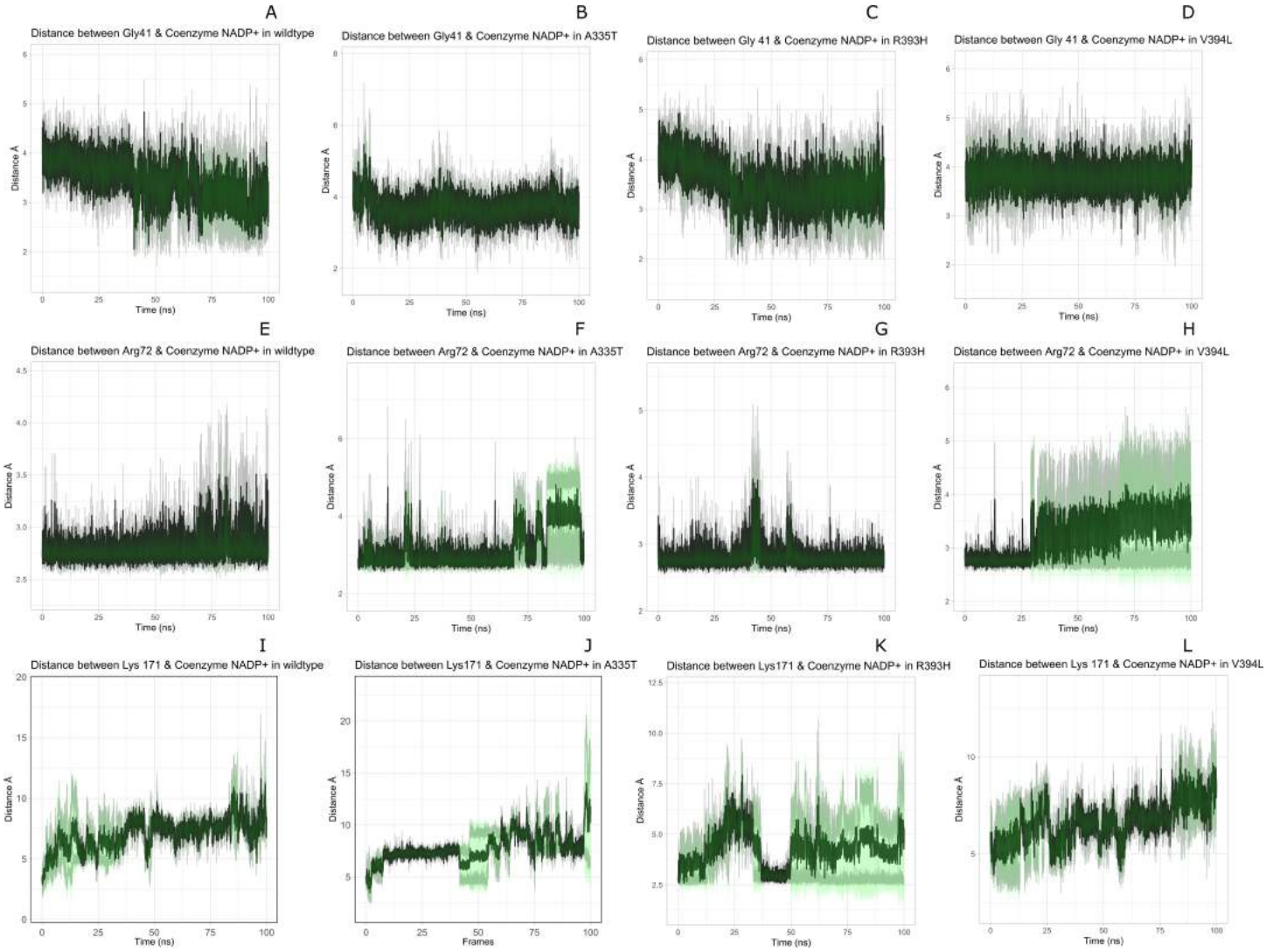


Figure 9 **Time evolution of distance of Coenzyme NADP$^+$ with the key interacting residues in the binding site for wild type and mutant (A–L).** Insignificant change in the distances indicate that mutations did not have significant alterations co-enzyme NADP$^+$ site.

Full-size DOI: 10.7717/peerj.12984/fig-9

gradually increased up to 10 Å with large fluctuations. In R393H a fluctuation in the distance was observed between 0 ns to 30 ns from 3 Å to 6.5 Å which gradually reduced to 3 Å. In V394L a gradual increase in distance ranging from 5 Åto 10 Å was observed during the course of simulation indicating the loop movement away from the cofactor. Lys 171 is the second residue of conserved peptide which keeps the integrity of binding site by interacting with ring O atom of G6P. A change in distance in mutant V394L indicate that conformational change at this region may occurred in the binding site resulting in the increase in distance.

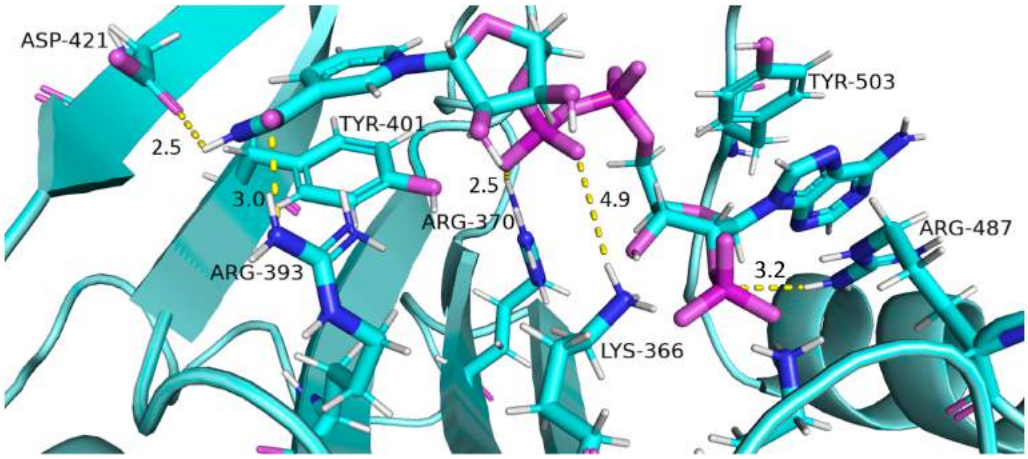


Figure 10 Residue interactions within binding site of Structural NADP$^+$. The native contacts for interactions within binding site of Structural NADP$^+$ were calculated based from crystal structure.
Full-size DOI: 10.7717/peerj.12984/fig-10

Structural NADP$^+$ interaction with protein and variation in residual contacts. Interaction of protein residues Tyr 401, Asp 421, Lys 366 and Arg370 with structural NADP$^+$has been represented in Fig. 10 and their distances indicated in the Fig. 11.

Lys 366 interconnects with 2′ phosphate group of structural NADP$^+$ in crystal structure. The average distance of 5.5 Å was observed in wildtype from 0 ns to 50 ns which gradually reduced to 5 Å. Figure 9A. On average distance of 5.5 Å was observed for both mutants *i.e.*, A335T and R393H however R393H showed higher fluctuation from 75 ns to 100 ns. A similar trend in distance was observed for the mutant V394L where the average distance was 6 Å from 0 ns to 25 ns and gradually reduced to 5.5 Å. Displacement of structural NADP$^+$ due to loss of hydrogen bond with Asp 421 and Try 401 oriented 2′ phosphate closer to the Lys 366 in R393H with a decrease in average distance of 3.5 Å.

Arg370 interact with the bisphosphate of via hydrogen bond in crystal structure. In wildtype the average distance of 3.0 Åwas maintained as compared to mutant A335T 3.3 Å. In R393H the average distance was 6 Å which increased from 30 ns to 80 ns to 7.5 Å. In V394L the distance remained 3.5 Å till 60 ns which increased up to 3.7 Å till 100 ns. R393H showed loss of hydrogen bond as a result of which increase in distance was observed.

Tyr 401 form π–π stacking interaction with nicotinamide ring of NADP$^+$ in crystal structure. This interaction was conserved in wildtype and mutant A335T, V394L with an average distance of 2.5 Å between Oxygen of bisphosphate and hydroxyl group of Tyr401. However this π-stacking interaction was lost with the adenine ring of NADP $^+$ in R393H due to the increase in the distance.

Asp 421 forms the polar interaction with amide hydrogen of nicotinamide in crystal structure. Average distance for the wildtype was 4.5 Å. In A335T the average distance was higher *i.e.*, 6 Å till 40 ns which gradually reduced to 5 Å till 100 ns. In R393H the average distance increases from 5 Å to 7 Å during the course of simulation due to movement of adenine ring of NADP$^+$ away resulting in loss of this bond. In V394L the distance was 4.5 Å till 70 ns after which it increased up to 5 Å. The average bond length of this interaction

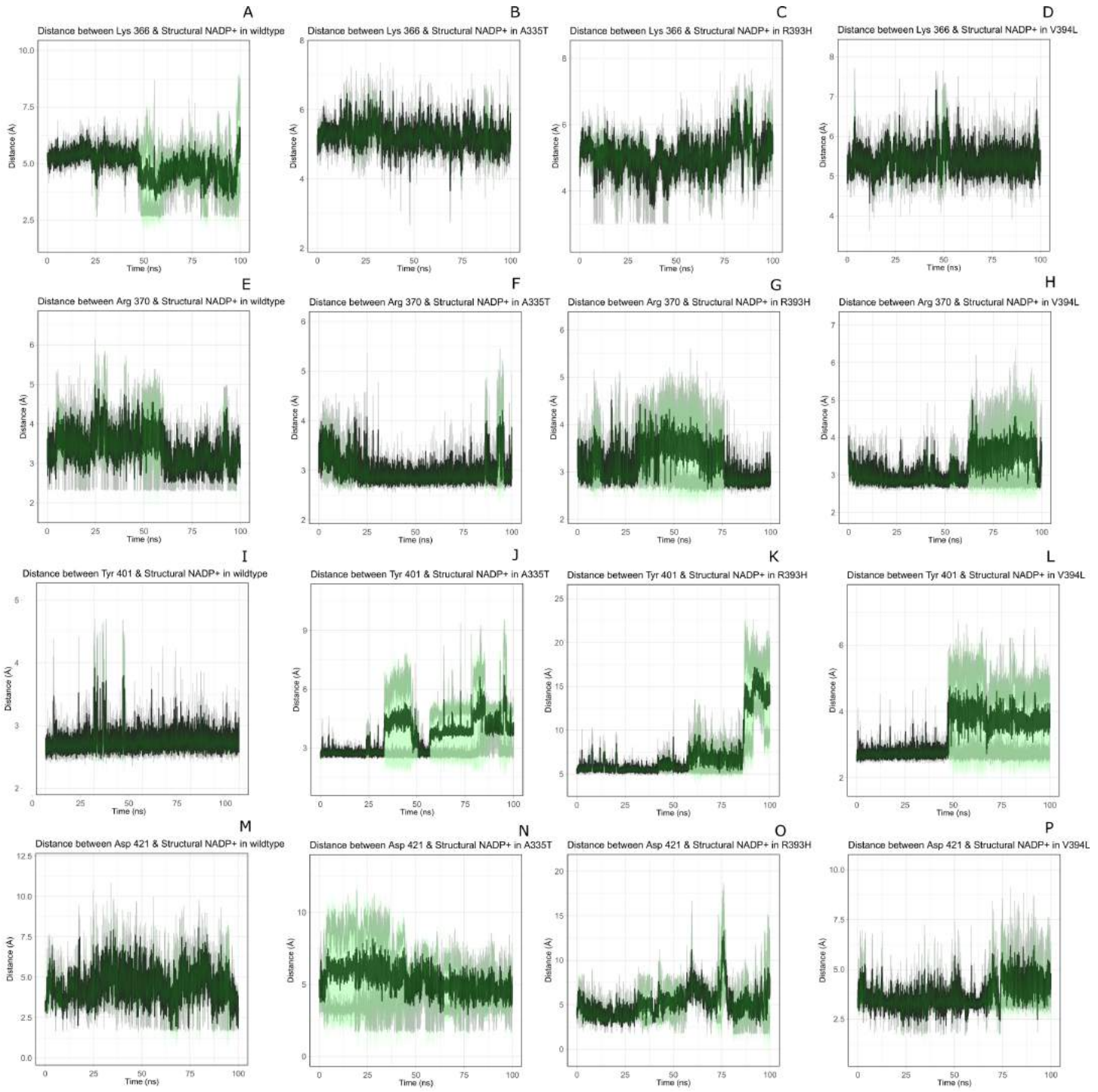


Figure 11 **Time evolution of distance of structural NADP$^+$ with the key interacting residues in the binding site for wild type and mutant (A–L).** Changes in the distances depict the change in interaction of mutants with protein in the structural NADP$^+$ binding site.

Full-size DOI: 10.7717/peerj.12984/fig-11

was 8 Å in mutant R393H. Asp 421 lies at the center of dimer interface. Loss of interaction in the R393H results in the displacement of nicotinamide ring of $NADP^+$ away from the histidine.

In wildtype guanidine group of Arginine 393 form hydrogen bond with 2′ hydroxyl ribose ring of structural $NADP^+$ as depicted in Fig. 3C. Mutation from Arginine to Histidine resulted in the loss of this hydrogen bond Fig. 3D.

Effect of mutations on conformation and catalysis

As described earlier that A335T lies at the end of βi of extensive 9 stranded β-sheet which forms the part of dimer interface accommodating structural $NADP^+$. Alanine in wildtype is the part of loop region at the end of extensive β-sheet network close to dimer interface. Introduction of large sized threonine instead of alanine resulted in the formation of polar contacts with Ile 255 (Fig. 3B) in adjacent αi- helix. Resultantly; loss of α helix structure in to loop at residue region containing amino acids 252–254 EFG took place making this region more flexible. The adjacent region of αi- helix showed disordered helix containing residues 258–263 which play an important role in substrate binding *i.e.*, Asp 258 is involved in the formation of polar contacts with ring oxygen atoms of G6P. Due to increase in flexibility the distance between G6P and Asp 258 increases resulting in the movement of Asp 258 away from the binding site. Similarly; Orientation of G6P has also been changed resulting in the movement of ring O atom away from the His 263 considered to be involved in the catalysis.(Fig. 12B)

Guanidino group of Arg393 in wildtype interacted with amide oxygen of nicotinamide ring in structural $NADP^+$, Asn397 and Ala399 Fig. 3C. In mutant R393H; replacement of Arg393 with Histidine resulted in the loss of these interactions (Fig. 3D). Loss of charge interaction resulted in the movement of nicotinamide ring of $NADP^+$ away leading to loss of π - π stacking interaction of Tyr 401 and Tyr 509 with nicotinamide ring of $NADP^+$ as depicted in Fig. 12C. Similarly interaction between nicotinamide ring of $NADP^+$ with Asp 421 and Arg 370 with oxygen atom of bisphosphate (Figs. 8B–8C) was lost resulting in the movement of nicotinamide ring away from the key interacting residues between substrate and $NADP^+$ making the substrate binding site more flexible due to which substrate may have lesser stability in the binding site.

Based on the severity of pathogenicity the mutants of G6PD are characterized into two categories *i.e.*, lying in structural $NADP^+$ binding site and at dimer interface. R393H and V394L fall in both of categories. V394L lies in the binding site of structural $NADP^+$ being the part of large extended β-sheet region. Incorporation of larger Leucine in V394L as compared to smaller sized Valine in wildtype have changed the local contacts. Valine is a C β branched hydrophobic residue which is not involved in catalysis however it is directly attached to the neighboring residue R393. R393 is involved in forming hydrogen bond interaction with nicotinamide ring of structural $NADP^+$. A change in the structure have resulted in the change of interactions in structural $NADP^+$ binding site with subsequent movement of residues in the binding site resulting in overall change in dynamics.

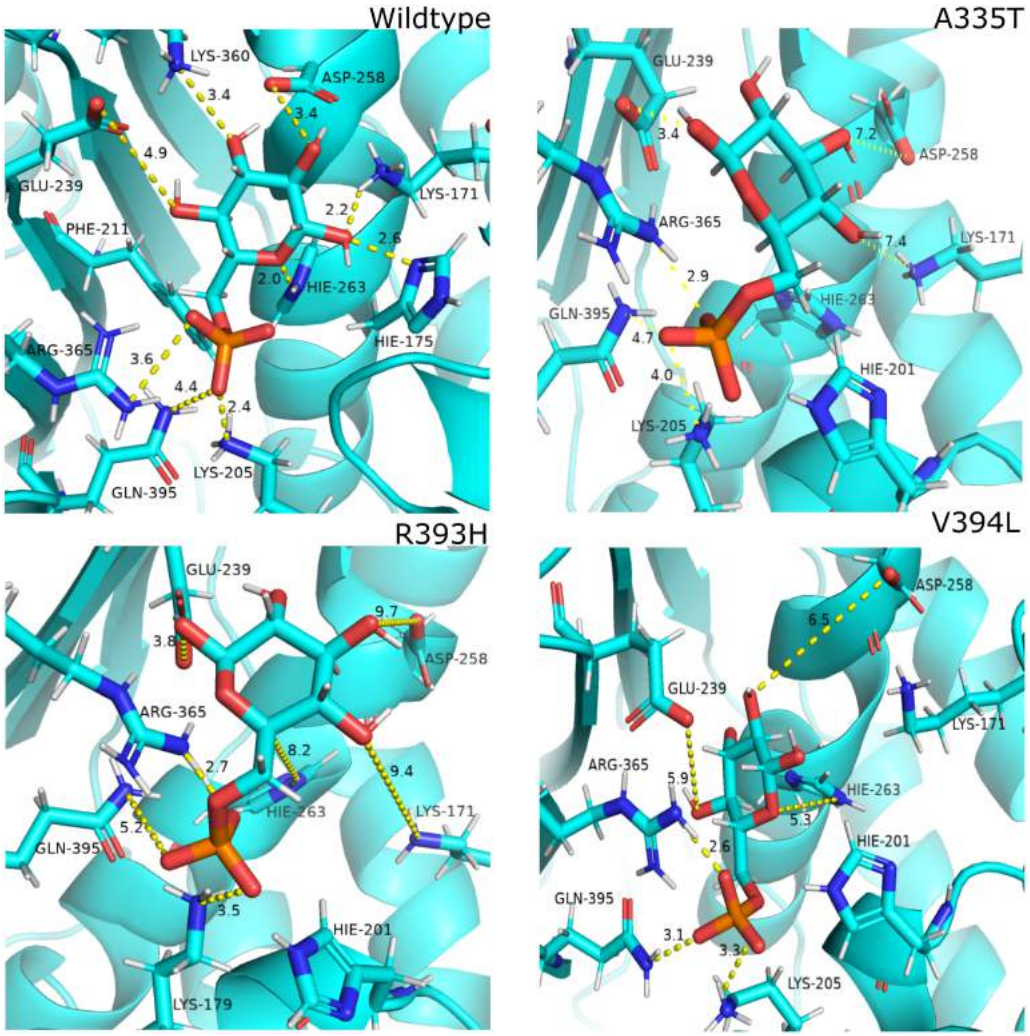


Figure 12 **Most common centroid structure showing mean change in distances in G6P binding site.** (A) Wildtype (B) A335T, (C) R393H, (D) V394L. Variation in interactions of G6P with neighboring residues were observed in mutants as compared to wildtype.

Full-size DOI: 10.7717/peerj.12984/fig-12

Most populated cluster conformation

In order to measure the interactions of substrate G6P, Coenzyme $NADP^+$ and structural $NADP^+$; a map of interactions of most populated cluster was produced after cluster analysis for wildtype as well as mutants. There was a variation in the distance of the binding site residues with substrate and cofactors for wildtype and mutants. Figure 12 describes the interaction established in the most populated cluster in the presence of G6P in wildtype and mutants (Fig. 12A). G6P develops hydrogen bond with Lys360, Asp258, Lys171, Arg365, Lys205 and Hie 263. A variation in the distance between these residues was observed upon mutation, *i.e.* loss of hydrogen bond with Asp 258, Lys 171 was observed for A335T as evident from the Fig. 12B. In R393H; loss of hydrogen bond took place for Asp25, Lys

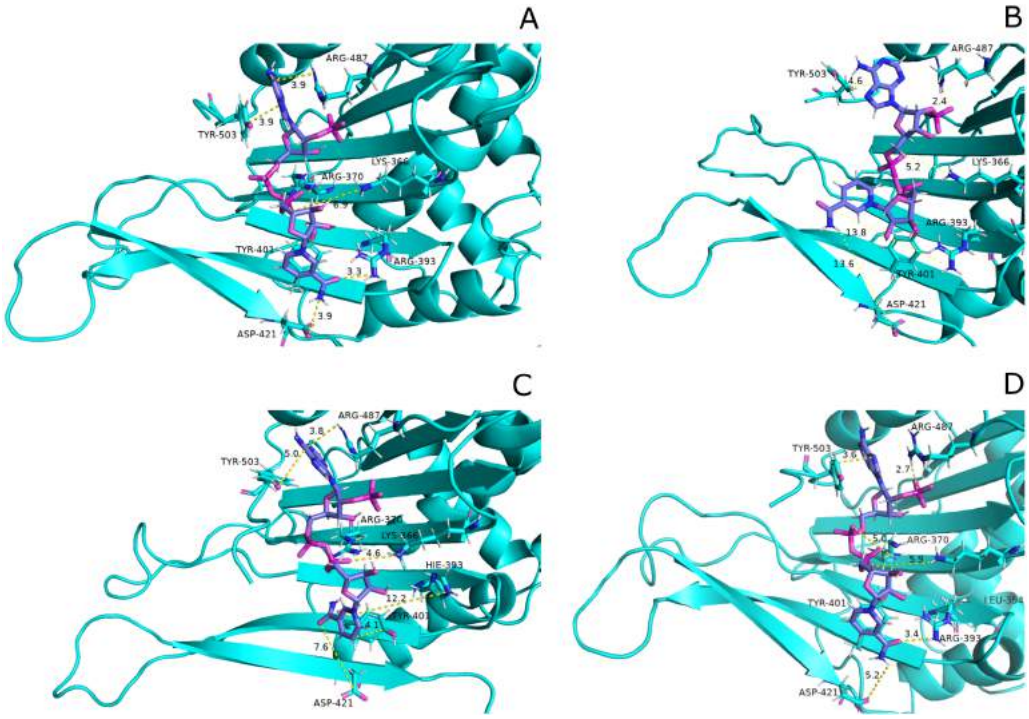


Figure 13 Most common centroid structure showing mean change in distances in structural NADP$^+$ binding site. (A) Wildtype (B) A335T, (C) R393H, (D) V394L. Significant changes in interactions of nicotinamide ring with neighboring residues were observed in mutants as compared to wildtype.
Full-size DOI: 10.7717/peerj.12984/fig-13

171, Gln395, Hie 263 (Fig. 12C) and in V394L hydrogen bond was lost between G6P and Asp258, Glu 238, Gie263, Arg365, Lys171 as apparent in the Fig. 12D.

These variations were also observed in the form of change in interactions in structural NADP$^+$ binding site specifically in the nicotinamide ring region indicate the formation of hydrogen bond between nicotinamide ring of structural NADP$^+$ with Arg393 and Asp421(Fig. 13A). Similarly; π-stacking interaction was developed between Tyr 401 and nicotinamide ring. Π-staking interaction was also maintained between Tyr503 and adenine ring of structural NADP $^+$. The π-stacking interactions were lost in mutants A335T, R393H and V394L in addition to the loss of hydrogen bond with Arg393 and Asp421 (Figs. 13B–13D) indicative of the fact that the change in binding mode of key residues in structural NADP$^+$ binding site had influenced G6P binding site as well upon mutation.

Secondary structure analysis

Maintenance of secondary structure is crucial for studying the dynamic behavior of protein. Secondary structure analysis of the G6PD enzyme and its mutants was performed to monitor the structural changes associated with mutation as compared to wildtype in structural elements such as alpha helices, beta sheets, and coils during the simulation period. Trajectories at 50 ns and 100 ns were collected and compared to the structure at the first frame to investigate the structural changes during the simulation. Overall structural

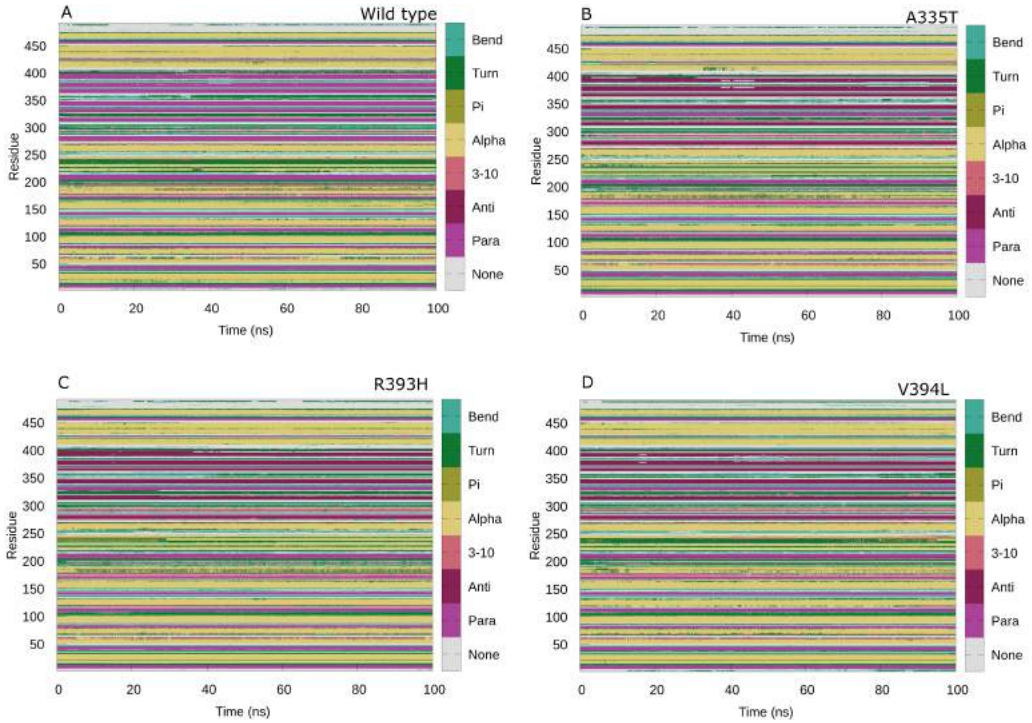


Figure 14 **Protein secondary structure analysis.** Secondary structure analysis of wild type G6PD and its mutant as a function of time from 0 to 100 ns. (A) Wildtype, (B) A335T, (C) R393H, (D) V394L.
Full-size DOI: 10.7717/peerj.12984/fig-14

changes throughout the trajectory are represented in the Fig. 14. Small secondary structural changes associated with amino acid and their position is depicted in Table 2. In A335T change of secondary structure elements from alpha helix to turn for residues 410–415 of β + α domain was observed between 30 ns to 50 ns during the course of trajectory (Fig. 14B). In mutant R393H loss of loop took place for residues 230–240 starting from 40 ns to 100 ns and 30 ns to 100 ns as evident from Fig. 14C. Alike change in the structure was observed at 90 ns in V394L according to Fig. 14D. These residues belong to the β + α domain of G6PD; the structural $NADP^+$ lies in the β + α domain. The perturbation in this region may have an effect on $NADP^+$ binding.

The change in secondary structure elements in this region represented that mutation had effected dynamical behavior of structural $NADP^+$binding site thus affecting the structural stability of mutants. No significant difference in secondary structure was observed as evident from previous experimental studies (*Gómez-Manzo et al., 2014*) however; few local changes in secondary structures were observed throughout the simulation indicating the loss of stability as shown in Table 2 and Figs. 14A-14D.

Binding free energy calculation by MM-PBSA

Binding free energy calculations for interaction of substrate G6P with wildtype and its mutants was done by MM-PBSA. The decomposition of binding free energy into its components has been shown in the Table 3 depicting the electrostatic energy component

Table 2 Changes in secondary structure and associated amino acids at 50 and 100 ns. Secondary structural analysis represents amino acid position, their sequence and structural change as a result of mutations at 50 ns and 100 ns.

Variants	Time (ns)	Position	Amino acid Sequence	Structural change
Wild type	50	ii. 300-301	ANN	Loop to turn
	100	i. 193-195	EDQ	Loop to turn
		ii. 201-203	RDN	Loop to turn
		iii. 300-301	ANN	Loop to turn
A335T	50	i.219-221	LMVLRF	α-helix to loop
		ii. 219-221	RIF	Turn to loop
	100	i. 144-146	TVY	Loop to turn
		ii. 316-319	EAT	Loop to turn
		iii.321-323	YLD	Loop to turn
		iv. 252-253	EF	α-helix to loop
R393H	50	i.85-87	EPF	α-helix to loop
		ii.144-146	PTV	Turn to loop
		iii. 219-221	RIF	loop to turn
		iv. 427-429	RYKN	Loop to turn
		v. 379-381	DIF	Loop to turn
	100	i. 300-302	ANN	Loop to turn
		ii. 379-381	DIF	Loop to turn
		iii. 448-451	SQMH	Loop to turn
		iv. 316-318	EAT	Loop to turn
V394L	50	i.221-225	FGPI	α-helix to loop
	100	i.221-225	FGPI	Loop to turn
		ii.156-158	ESC	Loop to turn

being the major contributor of binding fee energy. The Molecular mechanics energy (E_{MM}) constitutes electrostatic and van der Waals energy components. The calculated E_{MM} for wildtype was −188.5 kcal/mol and; −147.5 kcal/mol, 138.8 kcal/mol, −150.7 kcal/mol for A335T, R393H and -V394L respectively. The molecular mechanics energy of wildtype indicated its structure stabilization as compared to mutants.

The total solvation free energy which is the blend of polar and non-polar energy terms was unfavorable for all structure *i.e.*, wildtype, A335T, R393H and V394L.

The total electrostatic contribution ($\Delta G_{polar} + \Delta E_{electrostatics}$) was least for the wildtype *i.e.*, −25.6 kcal/mol and highest for V394L having value −9.2 kcal/mol displaying lesser affinity of V394L for the substrate binding as compared to wildtype and other mutants. The impact of vander Waals energy together with the non-polar solvation free energy ($\Delta E_{vdW} + \Delta G_{non\text{-}polar}$) constituted the non-polar energy term, which considerably contributed to the overall binding free energy by a value of −28 kcal/mol for the wildtype and for mutants being A335T; −27.1 kcal/mol, R393H; −16.3 kcal/mol and V394L; −28.8 kcal/mol. The net binding enthalpy ($\Delta H = \Delta E_{MM} + \Delta G_{sol}$) for wildtype was found to be −83.5 kcal/mol, as compared to A335T; −47.5 kcal/mol, R393H; −23.3 kcal/mol and V394L; −38 kcal/mol in which the molecular mechanics van der Waals (ΔE_{vdW}) energetic contribution is the

Table 3 Binding free energy values calculated by using MMPBSA method. Binding free energy components included electrostatic and vdW energy in kcal/mol. The total solvation free energy which is the blend of polar and non-polar energy terms.

Free Energy Component	Amber output term	Wildtype	A335T	R393H	V394L
		Energy (kcal/mol) Mean ± Std.dev	Energy (kcal/mol) Mean ± Std.dev	Energy (kcal/mol) Mean ± Std.dev	Energy (kcal/mol) Mean ± Std.dev
$\Delta E_{electrostatic}$	EEL	−210.2695 ± 36.0636	−200.8102 ± 45.7379	−116.6845 ± 48.1122	−86.5013 ± 47.0398
ΔE_{vdw}	VDWAALS	−8.2111 ± 4.9705	−3.2591 ± 4.2250	−4.1440 ± 3.9698	−7.5739 ± 3.5425
$\Delta E_{MM} = \Delta E_{electrostatic} + \Delta E_{vdw}$		−218.4806	−204.0693	−120.8285	−94.0752
MM-PBSA					
ΔG_{polar}	EPB	180.2649 ± 27.5577	180.3975 ± 40.7232	100.5207 ± 41.2799	74.9024 ± 43.4364
$\Delta G_{non-polar}$	ENPOLAR	−14.4928 ± 2.8196	−11.0225 ± 2.2161	−10.4887 ± 2.8433	−15.786 ± 1.6303
$\Delta G_{sol} = \Delta G_{polar} + \Delta G_{non-polar}$		165.7721	169.375	90.032	59.1164
$\Delta G_{polar} + \Delta E_{electrostatic}$		−30.0046	−20.4127	−16.1638	−11.5989
ΔGnonpolar solvation= ΔEvdw+ ΔGnon-polar		−22.7039	−14.2816	−14.6327	−23.3599
$\Delta H = \Delta E_{MM} + \Delta G_{sol}$		−52.7085	−34.6943	−30.7965	−34.9588
Total Binding Energy MM(PBSA)		−25.8260 ± 13.3498	−16.4031 ± 8.1141	−12.7896 ± 8.5373	−8.4411 ± 7.3496

most dominating enthalpic factor driving protein-ligand binding. It is evident from the results of binding free energy calculations that substrate complexes with mutants have structural stability much lower than that of wildtype protein (Table 3). Also mutations in the wildtype have induced a change in substrate binding pattern leading to structural distortion of mutants.

DISCUSSION

The structural stability of protein is directly correlated to its function. Mutation can induce a stabilizing or destabilizing effect resulting in the change of physicochemical properties of mutants. G6PD deficient red blood cells face the challenge of low stability of already synthesized G6PD as well as unavailability of a nucleus to synthesize new enzymes to replace altered enzymes with low activity. As a result, cells may burst due to the high level stress of ROS. Here molecular dynamics simulation of selected G6PD mutants in the dimer interface was carried out to understand the effect of mutation on protein stability and flexibility. RMSD was selected as primary criteria to determine structural stability. A higher value of RMSD indicated the compromised stability of protein. To find out the possible reason behind the change in stability, flexibility of protein was accessed by calculating RMSF from last frame of stable trajectories which showed variations as compared to wildtype. Secondary structure analysis was carried out to understand the changes in secondary structure elements.

Mutant A335T, known as G6PD Chatham, is the second most abundant G6PD mutation after G6PD Mediterranean which is associated with neonatal jaundice. RMSD of A335T was comparable to the wildtype. This could be attributed due to the nature of the wildtype amino acid versus that in mutated residue. *e.g.*, in A335T, with alanine having the hydrophobic side chain replaced by threonine in A335T with the uncharged polar side chain which does not alter any side chain interactions with the protein, as shown in Figs. 3A–3B. The fluctuation in RMSD as compared to wildtype indicated structural perturbation which is also evident from the experimental studies showing that mutation of Alanine at position 335 with Threonine exhibits decreased affinity for the substrate *i.e.*, G6P (*Vulliamy et al., 1988*).The Threonine is polar amino acid with bigger uncharged side chain as compared to Alanine which has different affinity for substrate in wildtype. Also, Alanine in wildtype is the part of loop region at the end of extensive β-sheet network close to dimer interface (Fig. 3A). In order to accommodate large threonine there may be a change in the native contacts of protein resulting in the decrease of overall stability of protein.

Mutant R393H, also known as G6PD Nashville, lies in the region closer to the structural $NADP^+$ binding site between the dimer interfaces. Comparisons of kinetics between wildtype and R393H indicated that k_m values for $NADP^+$ and G6P were higher for R393H mutant as compared to wildtype G6PD (*Syeda Rehana & Zaheer, 2018*). The deleterious effect of the mutation could be plausible leading to lower binding capacity with coenzyme and lesser stability of protein as indicated by higher RMSD of mutant R393H when compared with wildtype. A decrease in native contacts has reduced the flexibility of binding site residues indicating that there exist three groups of interacting residues in

dimer interface on βG, βK, βL lying closer in sequence but stretch from G6P binding site to structural NADP $^{+}$ binding site. It implied that residues Asn363 and Glu364 lying in the loop between βJ and βK interacted indirectly to 2′ phosphate group of structural $NADP^{+}$ via bound water molecule. Residue next in sequence *i.e.*, Arg365 interacted to the O7 and O9 atoms of phosphate in G6P while neighboring Lys 366 developed direct contact to 2′ phosphate group of structural $NADP^{+}$. In beta sheet; βG Lys 238 interacted indirectly with structural $NADP^{+}$ whereas adjacent residue; Glu 239 got involved in hydrogen bond formation with O4 atom of G6P. Similarly Gln 395 of βL strand interacted with O atom of phosphate in G6P whereas Arg393 built interaction with nicotinamaide ring of structural $NADP^{+}$ (*Gómez-Manzo et al., 2014*). The presence of substrate in the binding site is connected to structural $NADP^{+}$ via these interacting residues. Mutation in these sites may have a disruptive effect on the G6P binding site as well resulting in overall decrease in structural flexibility (*Gómez-Manzo et al., 2014*). Experimental studies show that mutant R393H has more exposed hydrophobic areas on treatment with different concentrations of urea as compared to wildtype which indicated that mutation greatly reduces the activity of enzymes when compared to wildtype (*Wang, Lam & Engel, 2006*). Our RMSD and RMSF results showing decrease in structural stability and flexibility complemented the experimental findings of researcher however exact mechanism of dynamics was unknown previously which has been explored here providing information even at residual contact level.

Mutant V394L, known as G6PD Alhambra, has been classified as class I deficiency. The amino acid residue lies in the extensive β-sheet network in dimer interface in vicinity of structural $NADP^{+}$. Residues in the β-sheets are known to form hydrogen bonds between the strands. The mutations in the β-sheet may be expected to obstruct normal hydrogen bond formation which was observed in V394L. Our study revealed that there is no direct contact of the residue with structural $NADP^{+}$ however residues in the vicinity are in direct contact with the cofactor *i.e.*, Arginine 393 (Fig. 12) . Incorporation of larger Leucine in V394L as compared to smaller sized Valine in wildtype may have changed the local contacts and hence stability which is evident from higher RMSD and RMSF values. This local conformational change may affect the binding of structural NADP $^{+}$ disrupting the normal protein stability and function.

Our study is in line with the previously reported (*Nguyen, Nguyen & Le, 2016*) computational studies indicating that the mutation in the binding site of substrate and structural $NADP^{+}$ result in the altered functioning of the enzyme by changing the alteration in the interaction of neighboring residues.

CONCLUSIONS

The current study focuses on the dynamic properties of the G6PD wildtype and its three mutants to exploit their effect on the stability of the G6PD structure by MD simulations. Among these, A335T exhibited fewer fluctuations in these properties. Mutants R393H and V394L lying in the dimer interface and closer to structural $NADP^{+}$, caused instability in protein structure. A decrease in stability may lead to a decrease in the catalytic activity of

enzymes as a result of conformational modifications and reorganization which indicated that the changes in the region around the structural $NADP^+$ had an effect on protein stability and hence activity. Alteration in this region severely affects the protein function leading to class-I deficiency.

Mutations in the dimer interface resulted in the change of interaction pattern of residues in the vicinity indicating the predominant role of dimer interface residue in maintaining the protein stability and flexibility. The protein stability is a key factor in determining the functionality of protein. The stability of protein can be affected by mutations in two ways: either by destabilizing or overstabilizing; consequently, deterioration of protein function by change of physico-chemical properties as a result of mutation. The binding free energy calculations performed by Molecular Mechanics Poisson Boltzmann Surface Area (MM-PBSA) quantified greater stability and spontaneity of wildtype complex with ligands as compared to A335T, R393H and V394L complexes with ligand. Free energy decomposition analysis revealed vdW energy as the major important factor facilitating the protein-ligand binding.

Naturally, a large number of G6PD mutants are found; therefore, the experimental observations of all mutants are not practical. However, the computational procedures can efficiently predict and support the experiments in evaluating a significant number of mutants. As no drug for G6PD is available to date, our study will provide insight into the structure-based drug designing to inhibit the effect of mutation and maintain the normal enzyme function. Also, exploiting a particular mutation can provide an insight to develop structure-based inhibitors against G6PD for cancer cells.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Department of Chemistry, Lancaster University for providing computational resources and facility to support this work.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Sadaf Rani performed the experiments, prepared figures and/or tables, and approved the final draft.
- Fouzia Perveen Malik conceived and designed the experiments, authored or reviewed drafts of the paper, provided high end workstations, and approved the final draft.
- Jamshed Anwar and Rehan Zafar Paracha analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data is available in the Supplemental Figures, Table 1, and UniProt: rs137852316, rs137852335 , and rs5030869.

REFERENCES

Au SW, Gover S, Lam VM, Adams MJ. 2000. Human glucose-6-phosphate dehydrogenase: the crystal structure reveals a structural $NADP^+$ molecule and provides insights into enzyme deficiency. *Structure* **8(3)**:293–303 DOI 10.1016/S0969-2126(00)00104-0.

Berendsen HJ, Postma Jv, van Gunsteren WF, Di Nola A, Haak JR. 1984. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81(8)**:3684–3690 DOI 10.1063/1.448118.

Cappellini MD, Fiorelli G. 2008. Glucose-6-phosphate dehydrogenase deficiency. *The Lancet* **371(9606)**:64–74 DOI 10.1016/S0140-6736(08)60073-2.

Darden T, York D, Pedersen L. 1993. Particle mesh Ewald: An N · log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **98(12)**:10089–10092 DOI 10.1063/1.464397.

De Lano WL. 2002. Pymol: an open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography* **40(1)**:82–92.

Desnick R, Ioannou Y, Eng C, Scriver C, Beaudet A, Sly W, Valle D. 2001. The metabolic and molecular bases of inherited disease. Seventh edition. New York: McGraw-Hill DOI 10.1016/S0307-4412(96)80019-7.

Dessi S, Batetta B, Laconi E, Ennas C, Pani P. 1984. Hepatic cholesterol in lead nitrate induced liver hyperplasia. *Chemico-Biological Interactions* **48(3)**:271–279 DOI 10.1016/0009-2797(84)90140-6.

Doss CGP, Alasmar DR, Bux RI, Sneha P, Bakhsh FD, Al-Azwani I, El Bekay R, Zayed H. 2016. Genetic epidemiology of Glucose-6-phosphate dehydrogenase deficiency in the Arab world. *Scientific Reports* **6(1)**:1–11.

Frisch M, Clemente F, Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zhe G. 2016. Gaussian 09, Revision A. 01. Wallingford: Gaussian, Inc, 20–44.

Gómez-Manzo S, Terrón-Hernández J, Mora DlaMora-Dela, González-Valdez A, Marcial-Quino J, García-Torres I, Vanoye-Carlo A, López-Velázquez G, Hernández-Alcántara G, Oria-Hernández J, Reyes-Vivas H, Enríquez-Flores S. 2014. The stability of G6PD is affected by mutations with different clinical phenotypes. *International Journal of Molecular Sciences* **15(11)**:21179–21201 DOI 10.3390/ijms151121179.

Homeyer N, Gohlke H. 2012. Free energy calculations by the molecular mechanics Poisson − Boltzmann surface area method. *Molecular Informatics* **31(2)**:114–122 DOI 10.1002/minf.201100135.

Hou T, Wang J, Li Y, Wang W. 2011. Assessing the performance of the MM/PBSA and MM/GBSA methods 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling* **51(1)**:69–82 DOI 10.1021/ci100275a.

Kotaka M, Gover S, Vandeputte-Rutten L, Au SW, Lam VM, Adams MJ. 2005. Structural studies of glucose-6-phosphate and $NADP^+$ binding to human glucose-6-phosphate dehydrogenase. *Acta Crystallographica Section D: Biological Crystallography* **61(5)**:495–504 DOI 10.1107/S0907444905002350.

Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* **78(8)**:1950–1958 DOI 10.1002/prot.22711.

Miller III BR, Mc Gee Jr TD, Swails JM, Homeyer N, Gohlke H, Roitberg AE. 2012. MMPBSA py: an efficient program for end-state free energy calculations. *Journal of Chemical Theory and Computation* **8(9)**:3314–3321 DOI 10.1021/ct300418h.

Minucci A, Moradkhani K, Hwang MJ, Zuppi C, Giardina B, Capoluongo E. 2012. Glucose-6-phosphate dehydrogenase (G6PD) mutations database: review of the old and update of the new mutations. *Blood Cells, Molecules and Diseases* **48(3)**:154–165 DOI 10.1016/j.bcmd.2012.01.001.

Nguyen H, Nguyen T, Le L. 2016. Computational study of glucose-6-phophate-dehydrogenase deficiencies using molecular dynamics simulation. *South Asian Journal of Life Sciences* **4(1)**:32–39 DOI 10.14737/journal.sajls/2016/4.1.32.39.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302(1)**:205–217 DOI 10.1006/jmbi.2000.4042.

Persico MG, Viglietto G, Martini G, Toniolo D, Paonessa G, Moscatelli C, Dono R, Vulliamy T, Luzzatto L, D'Urso M. 1986. Isolation of human glucose-6-pbosphate debydrogenase (G6PD) cDNA clones: primary structure of the protein and unusual 5′ non-coding region. *Nucleic Acids Research* **14(6)**:2511–2522 DOI 10.1093/nar/14.6.2511.

Rao KN, Kottapally S, Shinozuka H. 1984. Acinar cell carcinoma of rat pancreas: Mechanism of deregulation of cholesterol metabolism. *Toxicologic Pathology* **12(1)**:62–68 DOI 10.1177/019262338401200110.

Rastelli G, Rio AD, Degliesposti G, Sgobba M. 2010. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *Journal of Computational Chemistry* **31(4)**:797–810.

Rattazzi MC. 1968. Glucose 6-phosphate dehydrogenase from human erythrocytes: molecular weight determination by gel filtration. *Biochemical and Biophysical Research Communications* **31(1)**:16–24 DOI 10.1016/0006-291X(68)90024-7.

Roe DR, Cheatham III TE. 2013. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation* **9(7)**:3084–3095 DOI 10.1021/ct400341p.

Ryckaert J-P, Ciccotti G, Berendsen HJ. 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23(3)**:327–341 DOI 10.1016/0021-9991(77)90098-5.

Salmas RE, Mestanoglu M, Yurtsever M, Noskov SY, Durdagi S. 2015. Molecular simulations of solved co-crystallized X-ray structures identify action mechanisms of PDE δ inhibitors. *Biophysical Journal* **109(6)**:1163–1168 DOI 10.1016/j.bpj.2015.08.001.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7(1)**:539 DOI 10.1038/msb.2011.75.

Sitkoff D, Sharp KA, Honig B. 1994. Accurate calculation of hydration free energies using macroscopic solvent models. *The Journal of Physical Chemistry* **98(7)**:1978–1988 DOI 10.1021/j100058a043.

Srinivasan J, Cheatham TE, Cieplak P, Kollman PA, Case DA. 1998. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate − DNA helices. *Journal of the American Chemical Society.* **120(37)**:9401–9409 DOI 10.1021/ja981844+.

Syeda Rehana Z, Zaheer U-H. 2018. Molecular dynamics simulation of interleukin-2 and its complex and determination of the binding free energy. *Molecular Simulation* **44(17)**:1411–1425 DOI 10.1080/08927022.2018.1513651.

Townsend DM, Tew KD. 2003. The role of glutathione-S-transferase in anti-cancer drug resistance. *Oncogene* **22(47)**:7369–7375 DOI 10.1038/sj.onc.1206940.

Turner P. 2005. *XMGRACE, Version 5.1.19.* Beaverton: Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology.

Verma S, Grover S, Tyagi C, Goyal S, Jamal S, Singh A, Grover A. 2016. Hydrophobic interactions are a key to MDM2 inhibition by polyphenols as revealed by molecular dynamics simulations and MM/PBSA free energy calculations. *PLOS ONE* **11(2)**:e0149014 DOI 10.1371/journal.pone.0149014.

Vulliamy T, D'urso M, Battistuzzi G, Estrada M, Foulkes N, Martini G, Calabro V, Poggi V, Giordano R, Town M. 1988. Diverse point mutations in the human glucose-6-phosphate dehydrogenase gene cause enzyme deficiency and mild or severe hemolytic anemia. *Proceedings of the National Academy of Sciences of the United States of America* **85(14)**:5171–5175 DOI 10.1073/pnas.85.14.5171.

Wang XT, Chan TF, Lam VM, Engel PC. 2008. What is the role of the second structural $NADP^+$-binding site in human glucose 6-phosphate dehydrogenase? *Protein Science* **17(8)**:1403–1411 DOI 10.1110/ps.035352.108.

Wang X-T, Lam VM, Engel PC. 2006. Functional properties of two mutants of human glucose 6-phosphate dehydrogenase, R393G and R393H, corresponding to the clinical variants G6PD Wisconsin and Nashville. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1762(8)**:767–774 DOI 10.1016/j.bbadis.2006.06.014.

Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. 2004. Development and testing of a general amber force field. *Journal of Computational Chemistry* **25(9)**:1157–1174 DOI 10.1002/jcc.20035.

WHO Working Group. 1989. Glucose-6-phosphate dehydrogenase deficiency. *Bulletin of the World Health Organization* **67(6)**:601–611.

Combining biomedical knowledge graphs and text to improve predictions for drug-target interactions and drug-indications

Mona Alshahrani[1], Abdullah Almansour[1], Asma Alkhaldi[1], Maha A. Thafar[2,3], Mahmut Uludag[3], Magbubah Essack[3] and Robert Hoehndorf[3]

[1] National Center for Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA), Riyadh, Saudi Arabia

[2] College of Computers and Information Technology, Taif University, Taif, Saudi Arabia

[3] Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

ABSTRACT

Biomedical knowledge is represented in structured databases and published in biomedical literature, and different computational approaches have been developed to exploit each type of information in predictive models. However, the information in structured databases and literature is often complementary. We developed a machine learning method that combines information from literature and databases to predict drug targets and indications. To effectively utilize information in published literature, we integrate knowledge graphs and published literature using named entity recognition and normalization before applying a machine learning model that utilizes the combination of graph and literature. We then use supervised machine learning to show the effects of combining features from biomedical knowledge and published literature on the prediction of drug targets and drug indications. We demonstrate that our approach using datasets for drug-target interactions and drug indications is scalable to large graphs and can be used to improve the ranking of targets and indications by exploiting features from either structure or unstructured information alone.

Submitted 28 August 2020
Accepted 13 February 2022
Published 4 April 2022

Corresponding author
Mona Alshahrani,
mona.alshahrani@kaust.edu.sa

Academic editor
Shuihua Wang

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.13061**

Subjects Computational Biology, Computational Science, Data Mining and Machine Learning, Data Science

Keywords Biomedical literature , Biomedical knowledge graphs, Drug-target interactions, Drug-indications, Multi-modal learning, Bio-ontologies, Linked Data

INTRODUCTION

Over the recent years, knowledge graphs have become an effective data model to store, retrieve, share and link domain-specific knowledge in healthcare and biomedicine (*Bizer, Heath & Berners-Lee, 2011*; *Berners-Lee, Hendler & Lassila, 2001*). Knowledge graphs refer to a form of knowledge representation that describes entities and the binary relations in which they stand (*Paulheim, 2017*; *Ehrlinger & Wöß, 2016*). Biomedical data from structured databases is often represented in the form of knowledge graphs, for example using the Resource Description Framework (RDF) (*Brickley & Guha, 2004*) as a way to link and cross-reference different databases (*Jupp et al., 2014b*; *The UniProt Consortium, 2016*). However, voluminous biological and biomedical scientific findings are recorded in

How to cite this article Alshahrani M, Almansour A, Alkhaldi A, Thafar MA, Uludag M, Essack M, Hoehndorf R. 2022. Combining biomedical knowledge graphs and text to improve predictions for drug-target interactions and drug-indications. *PeerJ* **10**:e13061
http://doi.org/10.7717/peerj.13061

the form of disparate unstructured knowledge available as free text in journals, papers, book chapters, *etc.*, with only a limited amount of curated information available in public databases. PubMed database alone stores more than 32 million research abstracts from biomedical and life sciences, while PubMed Central (PMC) provides free full-text access for about 7.3 million articles. Knowledge graphs embedding methods have emerged as a novel paradigm for analyzing and learning from knowledge graphs within and across different subject domains (*Nelson et al., 2019*; *Ali et al., 2018*; *Alshahrani, Thafar & Essack, 2021*). Several methods have been developed for information represented as graphs (*Perozzi, Al-Rfou & Skiena, 2014*), knowledge graphs (*Ristoski & Paulheim, 2016*; *Nickel et al., 2016*), or formal knowledge bases (*Gutiérrez-Basulto & Schockaert, 2018*). The key idea is to map knowledge graph entities and their relations into a vector representation which preserves some local structure of individual nodes, and possibly some global structure of the graph, and use the resulting representations in machine learning tasks such as link prediction, entity classification, relation extraction, and entity resolution (*Nickel et al., 2016*). Machine learning models developed using these methods can perform comparatively to traditional predictive methods that rely on manual feature engineering (*Alshahrani et al., 2017*).

Learning representations of entities is not restricted to entities retrieved from structured databases; representation learning has been applied to many other types of data such as text, images, or videos (*LeCun, Bengio & Hinton, 2015*). Word2Vec (*Mikolov et al., 2013*) or GLOVE (*Pennington, Socher & Manning, 2014*) can learn representations of words that preserve some word semantics under certain vector operations and can, therefore, be utilized for downstream analysis. Knowledge graphs are also used in the development of many natural language processing (NLP) systems (*Xie et al., 2016*; *Hoffmann et al., 2011*), where they provide background knowledge for purposes such as disambiguating word mentions (*Dietz, Kotov & Meij, 2018*). Computationally predicting new drug-target interactions (DTI) and drug indications is a challenge in drug repurposing that relies on information in several knowledge bases, such as Bio2RDF (*Belleau et al., 2008*), UniProt (*Jupp et al., 2014a*), and others (*Williams et al., 2012*). It has become more common to predict new uses for known drugs (*i.e.,* drug repurposing) using the information in such databases combined with information derived from in *silico* cheminformatics and structural bioinformatics methods (*Chen et al., 2015*; *Pryor & Cabreiro, 2015*). A recent example of computational drug repurposing for COVID-19 used graph techniques to identify six drugs (*Gysi et al., 2021*). All six drugs exhibit the ability to reduce viral infections experimentally. Moreover, four of the drugs show very strong anti–SARS-CoV-2 response, which suggests they can be repurposed to treat COVID-19 (*Gysi et al., 2021*). Overall, the computational approaches developed to predict DTI and drug indications (*Ezzat et al., 2018*; *Thafar et al., 2019*; *Muñoz, Nováček & Vandenbussche, 2017*; *Mohamed, Nováček & Nounu, 2019*) differ in the algorithms they employed and the data sources utilized. That is, the network-based approaches (*i.e.,* graph-based methods) developed for drug repurposing utilize different data sources, including genomic and chemical similarities and various other drugs and target interactions profiles or descriptors (*Yamanishi et al., 2008*; *Wang et al., 2014a*), integrate information related to drug mechanisms, and use machine learning techniques or graph inference methods to predict novel DTIs (*Seal, Ahn & Wild,*

2015; *Thafar et al., 2020b*; *Fu et al., 2016*; *Chen et al., 2012*; *Thafar et al., 2020a*; *Thafar et al., 2021*).

Graph embeddings applied on the knowledge graphs improves the DTI prediction performance through the learning of low-dimensional feature representation of drugs or targets, used with the machine learning models. For example, the recently developed DTINet (*Luo et al., 2017*) used graph embedding approaches and matrix factorization, to predict novel DTIs from a heterogeneous graph. DTINet combines different types of drug and target (*i.e.,* protein) information such as drug–disease associations, drug–side effect associations, drug–drug similarity, drug–drug interactions, protein–protein interaction, protein–disease associations, and protein–protein similarities to construct a full heterogeneous graph. Another recent example of a knowledge graph-based method, TriModel (*Mohamed, Nováček & Nounu, 2019*), formulates DTI prediction as a link prediction problem associated within a knowledge graph. It learns feature representations (*i.e.,* knowledge graph embeddings) for entities and relations from a knowledge graph that integrated information from multiple structured databases similar to DTINet, and then predicts novel DTIs based on their interaction scores calculated using trained tensor factorization applied on the knowledge graph embeddings.

Some other approaches to drug repurposing rely on integrating entities text-mined from the biomedical literature (unstructured text) into knowledge graphs to predict novel associations between drugs and targets or drugs and diseases (*Swanson, 1990*; *Andronis et al., 2011*; *Frijters et al., 2010*; *Agarwal & Searls, 2008*). One such example is the biomedical knowledge graph-based method, SemaTyP (Semantic Type Path) (*Sang et al., 2018*). SemaTyP predicts candidate drugs for diseases by text-mining entities in published biomedical literature. This method first constructed a semantic biomedical knowledge graph, SemKG, with extracted relations from PubMed abstracts, then a logistic regression model is trained by learning the semantic types of paths of known drug therapies existing in the biomedical knowledge graph. Finally, the learned model, SemaTyP, is applied to exploit the semantic types of paths to discover drug therapies for new diseases. SemaTyP is the first method focused on drug repurposing that uses entities text-mined from biomedical literature and knowledge graph to predict candidate drugs. Another such recent method focused on drug repurposing, GNBR (Global Network of Biomedical Relationships) (*Percha & Altman, 2018*), also uses a large, heterogeneous knowledge graph to leverage integrated biomedical information across the literature of pharmacology, genetics, and pathology. The GNBR knowledge graph is generated based on three types of entities (drugs, diseases, and target proteins) that are connected by semantic relationship derived from the biomedical literature abstracts. The embedding method applied to this knowledge graph explicitly models the uncertainty associated with literature-derived relationships. Thus, GNBR is the first method that incorporates uncertainty (*i.e.,* noise) into a literature-based graph embedding method, allowing for a more precise and nuanced drug repurposing model. The GNBR method for drug repurposing produced treatment hypotheses with strong evidence from published literature, evaluated using gold-standard drug indications. Furthermore, they applied their model to generate novel drug repurposing hypotheses and assess their scientific validity using a variety of sources.

Despite several methods extracting biological relations from text, data integration issues remain between knowledge graphs and biomedical literature. First, biological entities are mentioned in knowledge graphs and biomedical literature using different vocabularies and thesaurus, which leads to low coverage when integrating structured knowledge graphs and unstructured biomedical literature. We address this issue by utilizing bio-ontologies for normalizing and unifying mentions of biological entities at the token level. Another problem is that knowledge graph learning or text-based methods, when used alone, fail in the "zero-shot scenario" when an entity is absent from either the knowledge graph or the text corpus and therefore can not be seen during training. Also, both modes of representation lack automatic feature generation. This work presents a method that combines knowledge graphs with rich textual content in the scientific literature in a unified representation learning framework. Additionally, our approach addresses the issues mentioned above of low coverage and different mentions of biological entities by utilizing bio-ontologies for normalization at the token level. We also tackle the "zero-shot scenario" through joint representation learning between knowledge graphs and literature. The primary goal is to complement the knowledge graph representation model presented previously with the model that utilizes background knowledge of biological entities available in the biomedical literature. We demonstrate that this multimodal view of feature representation enhances the prediction results of biological relations such as drugs targets and indications.

MATERIALS AND METHOD

Data sources and benchmark datasets

To construct the knowledge graph we used three ontologies, Gene Ontology(GO) (*Ashburner et al., 2000*), Disease Ontology (DO) (*Schriml et al., 2011*), and the Human Phenotype Ontology (HPO) (*Köhler et al., 2014*) (see Fig. 1). It also includes several biological entities such as diseases, genes (we do not distinguish between genes and proteins in the graph), and chemicals/drugs. The graph further includes relations between entities such as the protein-protein interactions obtained from STRING (*Szklarczyk et al., 2010*) (file: `protein.actions.v10.txt.gz`), chemical–protein interactions from STITCH (*Kuhn et al., 2012*) (file: `9606.actions.v4.0.tsv`), and drugs and their side-effects and indications from SIDER (*Kuhn et al., 2015*) (file: `meddra_all_indications.tsv`). We downloaded all the above-mentioned data on 11 March 2018 and used it to build the knowledge graph using RDF.

For the text-derived corpus 2, we used the pre-annotated Medline corpus provided by the PubTator project (*Wei, Kao & Lu, 2013*), downloaded on 18 December 2017. PubTator is a web-based tool designed to assist manual biocuration (*e.g.*, annotating biological entities and their relationships) through the use of advanced text-mining techniques. This corpus contains 27,599,238 abstracts together with annotations for chemicals, genes/proteins, and diseases. PubTator has annotations for 17,505,118 chemicals that represent 129,085 distinct drugs using either CHEBI or MESH identifiers. PubTator also contains 17,260,141 gene mentions covering 137,353 distinct genes in different species, of which 35,466 refers to human genes. We used 9,545 of the STITCH identifier (using the

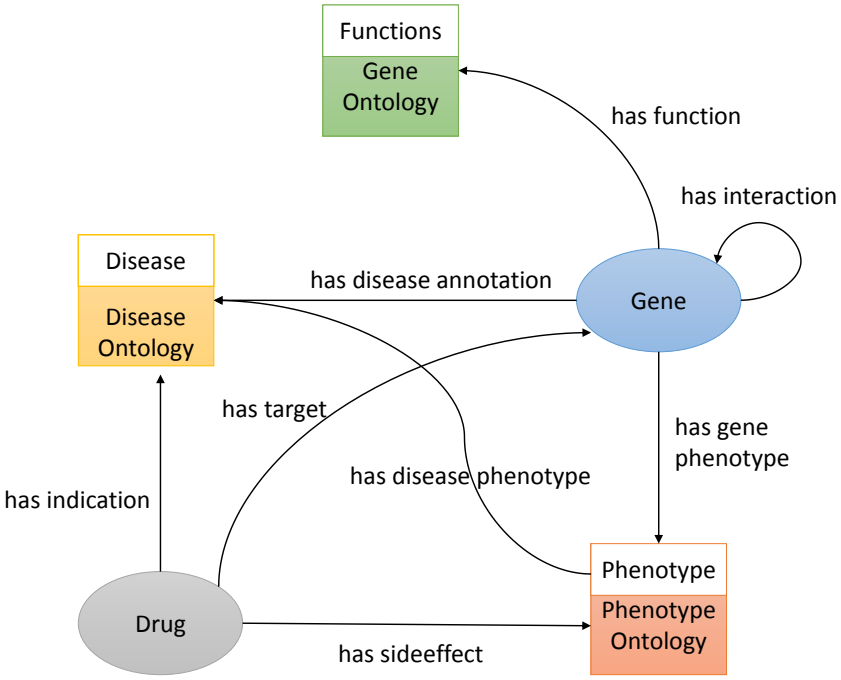


Figure 1 **An illustration of the knowledge graph used to repurpose the drugs.** To predict drug targets, we removed all the *has-target* links in the graph before applying our random walk algorithm. Similarly, for predicting drug indications, we removed all the *has-indication* links in the graph before applying our random walk algorithm.

Full-size DOI: 10.7717/peerj.13061/fig-1

file `9606.protein.aliases.v10.txt` provided by STITCH). PubTator further contains 81,655,248 diseases that represent 8,143 distinct diseases in MESH. We used DO to map diseases to 2,581 distinct DO classes. Table 1 provides the statistics for the DTI and drug indication data used to evaluate the models.

Additionally, we added gold standard datasets Table 2 commonly used in the literature to evaluate DTI prediction methods, *i.e.,* the Yamanishi (*Yamanishi et al., 2008*) and DrugBank datasets (*Wishart et al., 2008*). The Yamanishi dataset consists of interactions of drugs with four types of proteins, namely: Enzyme (E), Ion Channel (IC), G-protein-coupled receptor (GPCR), and Nuclear receptor (NR). We utilized the Enzyme and Ion Channel groups as they contain the largest number of interactions and can be found in our graphs after mapping of drugs ID (KEGG IDs) to our graph IDs (PubChem IDs).

Knowledge graph construction

We build the RDF graph by linking biological entities with relations from each database. For example, we link drug and protein targets from STITCH by the *has target* relations. The relations between the different biological entities are shown in Fig. 1. We also added classes from GO, HPO and DO onologies. For example, we link the disease *primary pulmonary hypertension* (`DOID:14557`) to the phenotype *arrhythmia* (`HP:0011675`) (using a `has phenotype` relation), we link the gene *CAV1* to disease *primary pulmonary hypertension* (`DOID:14557`) (using a `has disease association` relation), and we link the drug

Table 1 Statistics of the datasets used in model training and evaluation.

Dataset	Overlap in KG and literature		Training data (80%)		Testing data (%)	
	No. of drugs	No. of targets or (of diseases)	No. of positive samples	No. of negative samples	No. of positive samples	No. of negative samples
Drug–target interactions	820	17,380	65,379	65,379	16,345	16,345
Drug–indication associations	754	2,552	6,363	6,363	1,591	1,591

Table 2 Statistics of the datasets.

Dataset	No. drugs	No. of targets	No. of positive assoc.	No. of unknown assoc.
Enzyme (E)	445	664	2926	292,554
Ion Channel (IC)	210	204	1476	41,364
Drugbank dataset	1,482	1,408	9881	2,076,775

Tadalafil (`CID00110635`) to phenotype *abdominal pain* (`HP:0002027`) (using a `has side effect` relation), as well as disease *connective tissue disease* (`DOID:65`) (using a `has indication` relation):

```
@prefix doid: <http://purl.obolibrary.org/obo/DOID_> .
@prefix hp:  <http://purl.obolibrary.org/obo/HP_> .
@prefix b2v: <http://bio2vec.net/relation/> .
@prefix entrez: <http://www.ncbi.nlm.nih.gov/gene/> .
@prefix stitch: <http://bio2vec.net/CID> .

doid:14557 b2v:has_disease_phenotype hp:0011675 .
entrez:857 b2v:has_disease_association doid:14557.
stitch:00110635 b2v:has_sideeffect hp:0002027 .
stitch:00110635 b2v:has_indication doid:65 .
```

Integrating structured biomedical knowledge and literature

We use RDF (*Beckett, 2004*) to express and integrate structured information considered to be useful for predicting DTI and drug indication associations. In RDF, knowledge is expressed in a graph-based format in which entities (*i.e.,* nodes) are represented by an Internationalized Resource Identifier (IRI), and the relations between entities are represented as edges (*i.e.,* an edge connects two entities). Specifically, to integrate several datasets related to drug actions and diseases in a knowledge graph using RDF as representation language, we combine information about drugs and their targets (*Kuhn et al., 2012*) and indications (*Kuhn et al., 2015*), gene–disease associations (*Piñero et al., 2016*), and disease phenotypes (*Hoehndorf, Schofield & Gkoutos, 2015*), as well as gene functions and interactions between gene products (*Szklarczyk et al., 2010*). We further added biological background knowledge expressed in the HPO, GO, and DO ontologies, directly to this RDF graph so that the superclasses of phenotypes can be accessed and used by the machine learning model.

We generate a corpus from the RDF graph by applying iterated random walks (*Alshahrani et al., 2017*). We considered each random walk as a sequence that expresses a chain of statements following a random path through the knowledge graph. Subsequently, we align the entities that occur in our knowledge graph with the information contained in the biomedical literature. For this purpose, we normalized the entities in the abstracts of the biomedical literature to the entities in the knowledge graph using named entity recognition and entity normalization approaches (*Rebholz-Schuhmann, Oellrich & Hoehndorf, 2012*). Specifically, we normalized the drug, gene, and disease names/symbols to the knowledge graph using the annotated literature in PubMed abstracts provided by the PubTator (*Wei, Kao & Lu, 2013*) database, and the mappings provided between different vocabularies of drugs and diseases. PubTator aggregates different entity normalization approaches such as GNorm (*Wei, Kao & Lu, 2015*) or DNorm (*Leaman, Islamaj Doğan & Lu, 2013*), which can also be used directly with new text. We then processed the annotated PubMed abstract corpus by replacing each entity (*i.e.,* gene, drug/chemical compound, or disease) with the IRI used to represent the synonymous entities in the knowledge graph. This replacement ensures that our literature entities and knowledge graph entities overlap on the token level. Figure 2 provides an illustration of the normalization step between literature entities and knowledge graph entities overlap. We then used the knowledge graph to generate corpus 1 using an edge-labeled iterated random walk of fixed length without restart (*Alshahrani et al., 2017*). For each node in the graph, we generated a sequence based on a short random walk, where each walk is a sequence of nodes and edges (refer to Table S1 for more information). We used two hyperparameters to generate the corpus: walk-length (the size of each walk sequence) and the number of walks (the total number of walks generated for each node).

These processing steps led to the generation of two corpora: Corpus 1 generated from random walks starting from nodes in our knowledge graph, and Corpus 2 generated from annotated literature abstracts in which entities in the literature that also appear in our graph have been replaced by the IRI of the entities in the knowledge graph. These two corpora form the foundation of our feature learning step. Figure 3 provides an overview of the workflow.

Generating embeddings

Word2Vec is a vector space model that maps words to vectors based on the co-occurrence of words within a context window across the text corpus. Thus, in our graph, these semantics are captured by the random walks representing the co-occurrence of different entities and relations. We used the Word2Vec skip-gram model (*Mikolov et al., 2013*) to generate embeddings for the corpus generated by random walks on the knowledge graph (corpus 1) and for the Medline corpus (corpus 2). For both corpora, we used negative sampling with 5 words drawn from the noise distribution, a window size of 10, and an embedding dimension of 128, based on the parameter optimization results. Additionally, we generate embeddings by using the TransE (*Bordes et al., 2013*) knowledge graph embedding method. TransE is an embedding model specifically designed for knowledge graphs; it leverages the translation in the vector space. That is, given Given a triple *(subject, predicate, object)*

(a) Sample of original PubMed title and abstract corpus

24615250|t|Influence of SREBP-2 and SCAP gene polymorphismson lipid-lowering response to atorvastatin in a cohort of Chilean subjects with Amerindian background.
24615250|a|BACKGROUND AND OBJECTIVES: This study evaluated the influence of SREBP-2 and SCAP genes, respectively, on the response to atorvastatin treatment in a cohort of Chilean subjects with Amerindian background. METHODS: A total of 142 hypercholesterolemic individuals underwent atorvastatin therapy (10 mg/day/1 month).

(b) PubMed title and abstract normalized with KG

24615250|t|Influence of <http:www.ncbi.nlm.nih.gov/gene/6721> and <http:www.ncbi.nlm.nih.gov/gene/22937> gene polymorphismson lipid-lowering response to <http://bio2vec.net/chem/CID0002250> in a cohort of Chilean subjects with Amerindian background.
24615250|a|BACKGROUND AND OBJECTIVES: This study evaluated the influence of <http:www.ncbi.nlm.nih.gov/gene/6721> and <http:www.ncbi.nlm.nih.gov/gene/22937> genes, respectively, on the response to <http://bio2vec.net/chem/CID0002250> treatment in a cohort of Chilean subjects with Amerindian background. METHODS: A total of 142 <http://purl.obolibrary.org/obo/DOID_13810> individuals underwent <http://bio2vec.net/chem/CID0002250> therapy (10 mg/day/1 month).

(c) Sample of knowledge graph corpus

<http:www.ncbi.nlm.nih.gov/gene/6721><http://bio2vec.net/relation/has_gene_phenotype><http://purl.obolibrary.org/obo/HP_0001114><http://www.w3.or/2000/02/rdf-schema#subClassOf><http://purl.obolibrary.org/obo/HP_0000991>
…….. …….. ……..
…….. …….. ……..
<http:www.ncbi.nlm.nih.gov/gene/6723><http://bio2vec.net/relation/has_gene_phenotype><http://purl.obolibrary.org/obo/HP_0001114><http://www.w3.or/2000/02/rdf-schema#subClassOf><http://purl.obolibrary.org/obo/HP_0000991>
<http:www.ncbi.nlm.nih.gov/gene/6721><http://bio2vec.net/relation/has_gene_interaction><http:www.ncbi.nlm.nih.gov/gene/22937>

(d) Sample of PubMed abstract concatenated with KG corpus

24615250|t|Influence of <http:www.ncbi.nlm.nih.gov/gene/6721> and <http:www.ncbi.nlm.nih.gov/gene/22937> gene polymorphismson lipid-lowering response to <http://bio2vec.net/chem/CID0002250> in a cohort of Chilean subjects with Amerindian background.
24615250|a|BACKGROUND AND OBJECTIVES: This study evaluated the influence of <http:www.ncbi.nlm.nih.gov/gene/6721> and <http:www.ncbi.nlm.nih.gov/gene/22937> genes, respectively, on the response to <http://bio2vec.net/chem/CID0002250> treatment in a cohort of Chilean subjects with Amerindian background. METHODS: A total of 142 <http://purl.obolibrary.org/obo/DOID_13810> individuals underwent <http://bio2vec.net/chem/CID0002250> therapy (10 mg/day/1 month).
<http:www.ncbi.nlm.nih.gov/gene/6721><http://bio2vec.net/relation/has_gene_phenotype><http://purl.obolibrary.org/obo/HP_0001114><http://www.w3.or/2000/02/rdf-schema#subClassOf><http://purl.obolibrary.org/obo/HP_0000991>
<http:www.ncbi.nlm.nih.gov/gene/6723><http://bio2vec.net/relation/has_gene_phenotype><http://purl.obolibrary.org/obo/HP_0001114><http://www.w3.or/2000/02/rdf-schema#subClassOf><http://purl.obolibrary.org/obo/HP_0000991>

Figure 2 **(A) Sample of the original Pubmed title and abstract; (B) Illustration of how we normalize literature abstracts to our knowledge graph to ensure that both overlap on the level of tokens. It shows the use of ontologies to normalize synonymous or similar terms to their respective ontology identifiers as in *hypercholesterolemic*. We refer to NCBI semantic web links for genes. For other entities with no standard semantic web links, we assign them to links that start with http://bio2vec.net/. (C) Sample of the knowledge graph corpus. (D) Sample of the knowledge graph corpus concatenated with the PubMed abstract corpus.**

Full-size DOI: 10.7717/peerj.13061/fig-2

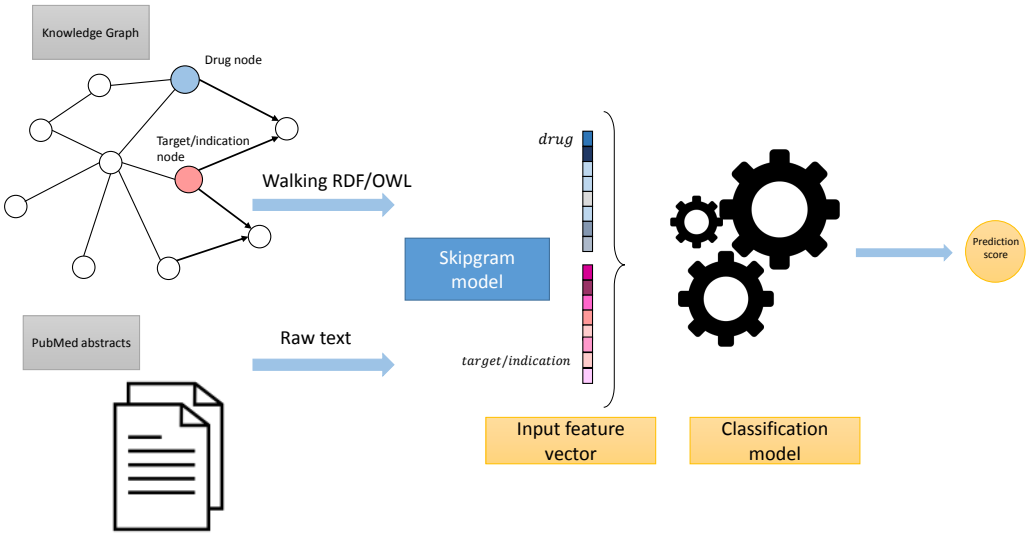


Figure 3 **High-level overview of the workflow.**

Full-size DOI: 10.7717/peerj.13061/fig-3

or simply *(s,p,o)*, it aims to make the sum of the subject and predicate vectors as close as possible to the object vector (*i.e.*, $\vec{s}+\vec{p}\approx\vec{o}$) when *(s,p,o)* holds, and the sum is far away otherwise. This is done based on some distance measure $d(\vec{s}+\vec{p},\vec{o})$, which is chosen to be L_1 or L_2 norms. The loss function is the pairwise ranking loss as follows:

$$\mathcal{L}=\sum_{(s,p,o)\in\mathcal{S}}\sum_{(s',p,o')\in\mathcal{S}'}[\gamma+d(\vec{s}+\vec{p},\vec{o})-d(\vec{s}'+\vec{p},\vec{o}')] \quad (1)$$

The TransE model deals with only one-to-one relations, but it fails to account for other types of relations and mapping properties such as one-to-many, many-to-one, and many-to-many which are mitigated by other knowledge graphs embeddings variants such as TransH (*Wang et al., 2014b*), TransR (*Lin et al., 2015*) and others (*Ji et al., 2021*).

Training the prediction models

We evaluated the performance of each method by using the embedding vectors to predict DTI and drug indication associations in a supervised manner. For prediction models, we used neural networks-based models such as: Artificial neural networks (ANN) and Siamese Networks (*Bertinetto et al., 2016*). The Siamese network uses a unique structure to learn similarity between inputs even with the presence of one or training example and able to generalize to data from complete different distributions with new classes. Although they have been widely used for images, they could be also applied to learning similarity between any two different entities encoded as feature vectors. Moreover, we have used Random forests (RF), and logistic regression (LR) classifiers as basic and self-explained machine learning models. For each model, the dataset was randomly split into 80% and 20%. proportions for the training set and testing set, respectively. The models were trained as binary classification models to predict whether there is an interaction between drug and target or not (based on the drug-target dataset), or if there is an association between drug and disease or not (based on the drug indications dataset). Table 1 provides all the statistics for the DTI and drug indication data used to evaluate the models.

For ANN model training, we implemented an architecture with a single hidden layer that is twice the size of the input vector. We used the Rectified Linear Unit(ReLU) (*Nair & Hinton, 2010*) as an activation function for the hidden layer and a sigmoid function as the activation function for the output layer. We also used cross-entropy as the loss function, RMSprop optimizer (*Hinton, Srivastava & Swersky, 2012*) to optimize the ANN parameters, and we implemented all these steps using Keras library in Python (*Gulli & Pal, 2017*). We optimized the ANN architecture and the size of the embeddings using a narrow search (see Tables S2 and S3), we have also optimized the learning rate and the number of dense layers of the Siamese networks. To train the RF classifier, we specified the number of trees to be 50, with the minimum number of one for the training samples in leaf nodes, and used the Gini impurity index to measure the quality of the split. For the LR, we optimized the LR concerning two of its most effective hyperparameters: the penalty term [L1,L2] and the C = [100,10,1.0,0.1,0.01] (the inverse of regularization), which controls the strength of the penalty. Small values of this hyperparameter cause stronger regularization. We found that L2 and $C = 10$ are the optimal values. We trained the LR classifier using scikit-learn (version 0.17.1) in Python (*Pedregosa et al., 2011*).

RESULTS

Learning and combining features

We integrated both data sources intending to leverage the information in a single predictive model. To achieve this goal, we obtained embeddings for all entities. We used two embedding approaches for the knowledge graph including the Word2Vec skip-gram model (*Mikolov et al., 2013*), and TransE (*Bordes et al., 2013*), and for biomedical literature only the Word2Vec skip-gram model. We used two different approaches to combine the embeddings from corpus 1 and 2. First, we generated the embeddings for each corpus, then concatenated the embedding vectors from both corpus. Second, we concatenated the two corpora, then generated jointly-learned embeddings from the combined corpus. Here, it should be noted that not all entities in the knowledge graph have a representation in literature, and not all entities (drugs, diseases, and genes) mentioned in literature are included in the knowledge graph. Nonetheless, we obtained embeddings for all entities in corpus 2, in particular, for the entities which we normalized to our knowledge graph. Figure S1 shows the overlap between the two datasets. Figures 4 and 5 show a visualization of the embeddings (from the knowledge graph, literature, and combined) using t-SNE (*Van der Maaten & Hinton, 2008*). Disease embeddings are coloured based on their top-level DO class, and drug embeddings based on their top-level class in the Anatomical Therapeutic Chemical (ATC) Classification System. The clustering by both top-level DO classes and the top-level ATC categories shown in both figures indicate that the embeddings cluster into biologically meaningful groups.

Evaluating the prediction performance

We evaluated the performance of our method by predicting drug-target interactions and drug indications. For this purpose, we used five different evaluation methods: (1) embeddings generated from the knowledge graph *via* the TransE model, (2) embeddings generated *via* the Word2Vec skip-gram model from the knowledge graph alone after the corpus generation through random walks (Walking RDF/OWL), (3) embeddings generated from the literature corpus alone *via* the Word2Vec skip-gram model, (4) concatenated embeddings from (2) and (3), and (5) jointly-learned embeddings generated from combining corpus 1 and corpus 2 on which we applied the Word2Vec skip-gram model. We used drug-target interactions from the STITCH database and drug indications from the SIDER database as evaluation datasets. Furthermore, to clearly distinguish and evaluate the contributions of the different data sources (*via* the five evaluation methods), we only used the entities in the evaluation dataset that have a representation in both the knowledge graph and the literature corpus. Before training the model, we removed all `has-target` edges (when predicting the drug-target interactions) or `has-indication` edges (when predicting drug indications) in the graph before generating the predictions for drug-target interactions and drug indications, respectively.

Consequently, from the drug-target interactions dataset, we obtained 81,724 positive samples, and from the drug indications dataset, 7,954 positive samples. For the negative samples, we randomly selected the same number of negative samples as positive samples

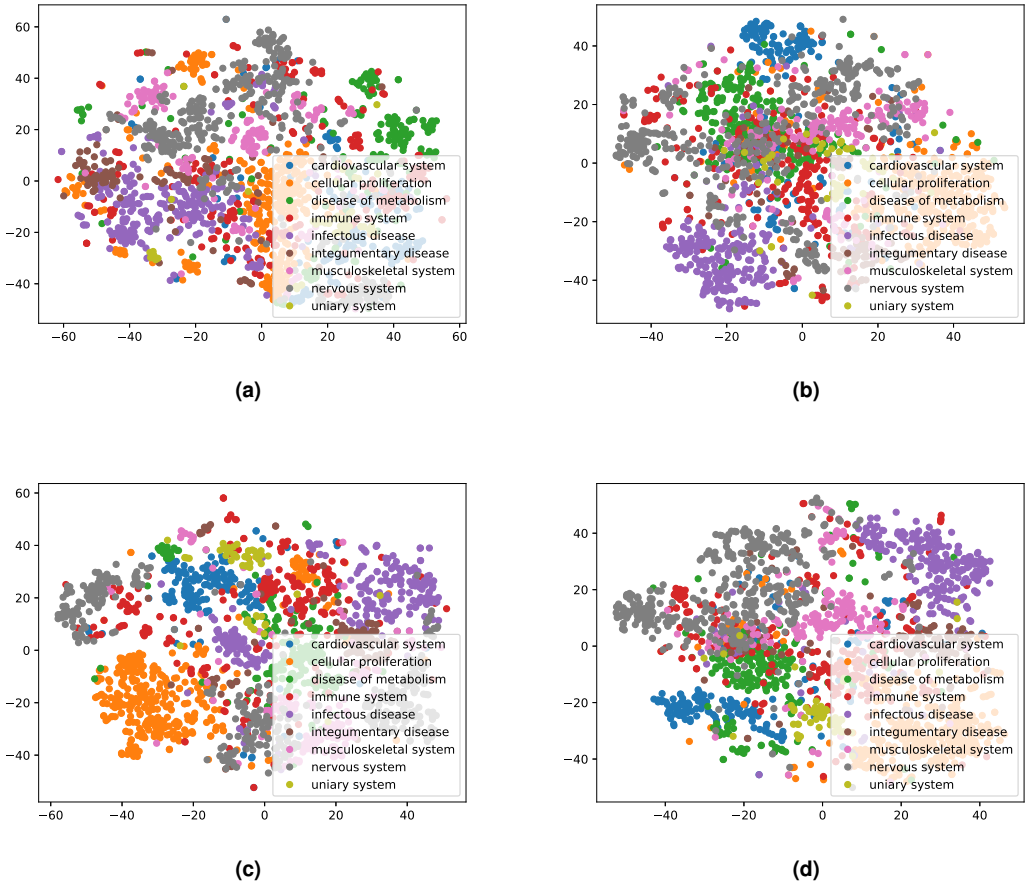


Figure 4 Illustrations of the 2D t-SNE plots for diseases based on different embeddings. (A) Knowledge graph. (B) MEDLINE abstracts. (C) Concatenated embeddings. (D) Concatenated corpora through jointly learned embeddings from literature and knowledge graph. The diseases are colored according to their top-level categories in the Disease Ontology.

Full-size DOI: 10.7717/peerj.13061/fig-4

from a massive number of negative samples that exist in the datasets. In this manner, we ensured a balanced dataset was used to develop the prediction models.

For predicting drug-target interaction, we used an evaluation set of 820 drugs that was mapped to 17,380 targets. For predicting drug indications, we used 754 drugs with one or more known indications and rank 2,552 diseases for each of the drugs to determine which disease it may treat (see Tables S4 and S5 for details about the counts in all resources). For each model, the input feature vector is the drug embedding concatenated with the target embedding, for the drug-target pairs. Similarly, for drug indications the input vector is the drug embedding concatenated with the disease embedding. The output indicates whether the drug interacts with the targets or the drug treats the disease. We evaluated the performance of each model, using 20% of associations left out of the training process. All three of our classification models can provide confidence values for a prediction, and we ranked predicted associations based on their confidence value. We then calculated the area

Table 3 AUROC results of comparisons with other methods on Yamanishi (Enzyme and Ion Channel) and Drugbank datasets. Bold indicates best performing model while underline indicates second best performing model.

Model	Datasets		
	Enzyme	Ion channel	Drugbank
Ours (KG)	0.900	0.970	0.840
Ours (PubMed abstracts)	0.950	0.970	0.880
Ours (Concatenated embeddings)	0.940	**0.990**	0.880
Ours (Concatenated corpus)	0.960	**0.990**	0.890
BioBERT embeddings	0.920	0.880	0.900
BLM-NII	0.950	0.900	0.940
DNILMF	0.950	0.930	0.940
KRONRLS-MKL	0.920	0.890	0.920
TriModel	**0.990**	**0.990**	**0.990**
DTiGEMS+	**0.990**	**0.990**	0.970

under the receiver operating characteristic (ROC) curve (AUROC) (*Fawcett, 2006*), as well as the recall of each drug averaged among all drugs.

Furthermore, we have compared the results of using the four variants of our models (the knowledge graph, PubMed abstracts, Concatenated embeddings, and Concatenated corpus) with other benchmark datasets. Tables 3 and 4 shows the results in terms of AUROC and AUPR. We compared the results of our approach with five state-of-the-art methods, namely: BLM-NII (*Mei et al., 2013*), KRONRLS-MKL (*Nascimento, Prudêncio & Costa, 2016*), DNILMF (*Li, Li & Bian, 2019*), and the latest two methods: TriModel (*Mohamed, Nováček & Nounu, 2020*) and DTiGEMS+ (*Thafar et al., 2020a*). We observe that our models' performance, which utilizes the multimodal approaches (Concatenated embeddings and Concatenated corpus), is not as competitive as the latest methods but shows comparable results (especially in Yamanishi datasets with the previous methods including BLM-NII, KRONRLS-MKL, and DNILMF), coming as the best and the second best performing models for the Ion Channel, while it comes as the second best in Enzyme dataset in the AUROC results analysis shown in Table 3.

Additionally, we have employed BioBERT (*Lee et al., 2020*) embeddings, which is a domain-specific language model based on the BERT model (*Devlin et al., 2018*), pre-trained on large-scale biomedical text (PubMed abstracts and PMC full-text articles). For each drug and gene name, we have extracted their BioBERT embeddings. We used each pair of interacting drug-gene BioBERT embeddings as inputs to the Siamese network. We have followed the same approach of training and testing as described in Training the prediction models Tables 3 and 4 show the ROCAUC and AUPR scores on the three datasets we used for benchmarking namely (Drugbank, Yamanishi Enzyme and Yaminishi Ion channel). Tables S7 and S8 (Supplementary) summarizes our results for the prediction of DTIs and drug indications using different machine learning models.

While other methods may be limited to certain sizes of the graph, the main advantages of our models canbe summarized in the following points. first, Its scalability to large and massive knowledge graphs, our graph used in this work is $\approx$ 500 times larger than the

Table 4 AUPR results of comparisons with other methods on Yamanishi (Enzyme and Ion Channel) and Drugbank datasets. Bold indicates best performing model while underline indicates second best performing model.

Model	Datasets		
	Enzyme	Ion channel	Drugbank
Ours (KG)	0.690	0.920	0.280
Ours (PubMed abstracts)	0.740	0.900	0.320
Ours (Concatenated embeddings)	0.740	0.950	0.340
Ours (Concatenated corpus)	0.760	0.950	0.320
BioBERT embeddings	0.908	0.870	**0.879**
BLM-NII	0.830	0.800	0.110
DNILMF	0.820	0.840	0.410
KRONRLS-MKL	0.800	0.820	0.340
TriModel	0.930	0.950	0.670
DTiGEMS+	**0.960**	**0.960**	0.610

Enzyme dataset, and ≈ 148 times larger than the Drugbank dataset. Second, Our models are generic and automatically learn the features, while other methods may rely on laborious feature extraction and manually engineered feature vectors. For example, DtiGems+ construct different types of graphs and compute many similarity scores such as drug–drug similarity, target–target similarity as well as adapting several techniques such as graph embeddings, graph mining as well as the use of machine learning models as downstream classifiers. Although these approaches resulted in improved prediction accuracy due the collective power of different types of features, they require domain-specific knowledge of manually-engineered features, incorporate complex processes of extraction and include many steps of data integration and graph infusions. This doesn't fully utilize feature learning as an optimal, efficient, and elegant way of finding the most relevant features. lastly, our proposed models attempt to resolve the issues related to the low coverage in the knowledge graph and the textual content by utilizing bio-ontologies for entity normalization.

We found that both ANN and RF classifiers were able to accurately predict both DTIs and drug indications, while the LR classifier results in relatively worse performance. An obvious explanation is that LR mainly assigns weights to individual features and cannot compare or match elements of the two input embedding vectors, while both the ANN and RF classifiers can provide a classification based on comparing elements of the two input embedding vectors. Furthermore, we found that, in general, using embeddings generated from literature results in higher predictive performance across all classifiers compared to embeddings generated from the knowledge graph alone. Also, combining the embeddings, or using jointly learned embeddings sometimes but not always improves or changes the predictive performance.

Additionally, we examine the actual performance in terms of the ranking given by each of our embeddings approaches for a sample of drug–targets and drug–indications pairs. Table S7 (refer to Supplementary) shows the predicted rank number given for each approach for the prediction of drug–targets, while Table S10 (refer to Supplementary) shows the predicted rank number in drug–indications. We find that the combined approaches

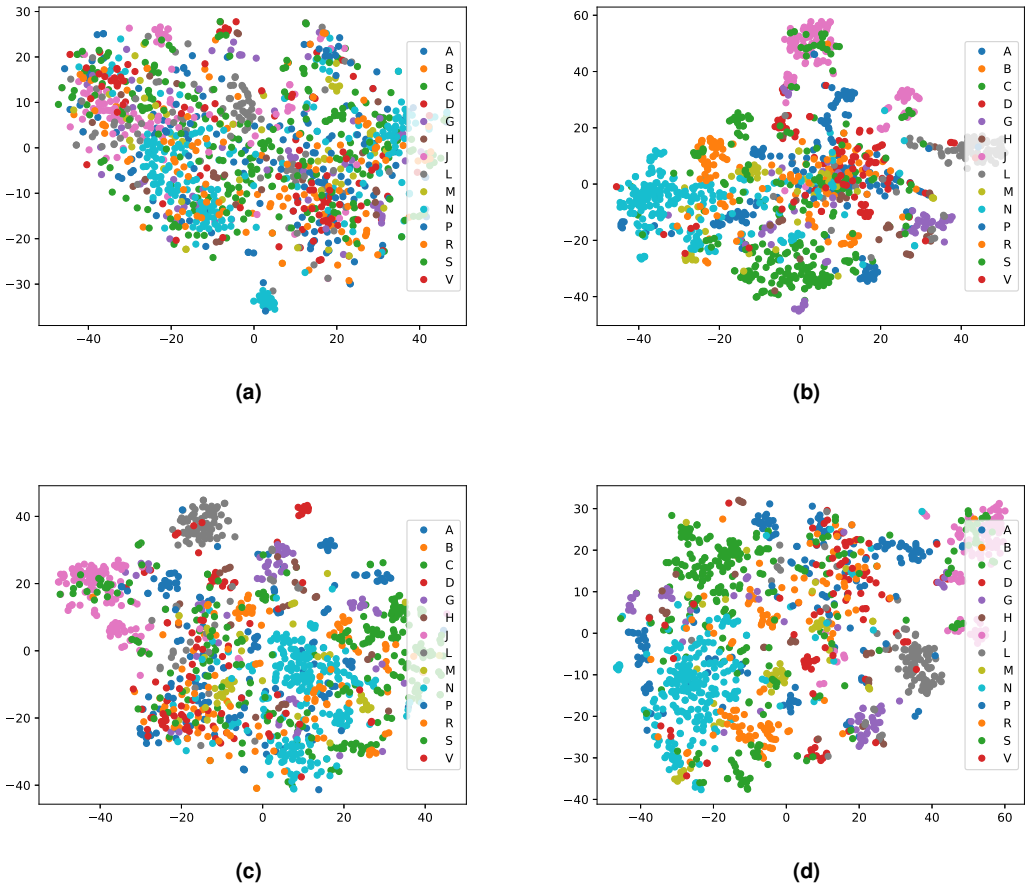


Figure 5 Illustrations of the 2D t-SNE plots for drugs based on different embeddings approaches. (A) Knowledge graph. (B) MEDLINE abstracts. (C) Concatenated embeddings. (D) Concatenated corpora through jointly learned embeddings from literature and knowledge graph. The drugs are colored according to their top-level ATC class.

Full-size DOI: 10.7717/peerj.13061/fig-5

(Concatenated embeddings and Concatenated corpus) improved the predicted ranks over the performance of the knowledge graph and the PubMed abstracts alone.

While our results indicate that both literature-derived and knowledge graph embeddings can be used to predict interactions, the main contribution of our multi-modal approach is the increased coverage through combining database content and literature (see Fig. S1). We used the common drug, target, and disease entities between the knowledge graph and literature in the previous experimental setups. Here, we further quantify fairly the impact of the information provided by each data modality on the prediction performance. We also demonstrate the broader application of our method by extending our evaluation set to contain all the drugs, genes, and diseases found in either our knowledge graph, literature abstracts, or the union of the entities in the knowledge graph and literature trained on the combined corpus. Figure 6 shows the ROC curves and the AUROC for predicting DTIs

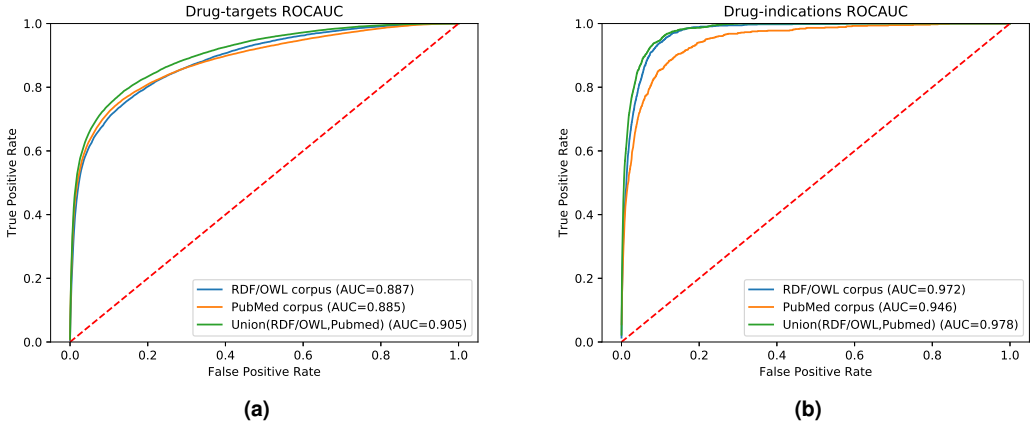


Figure 6 **ROC curve of our neural network for predicting drug targets in the union of associations present in the knowledge graph and PubMed abstracts (left); ROC curve of our neural network for predicting drug indications found in the union of knowledge graph and PubMed abstracts (right).**
Full-size DOI: 10.7717/peerj.13061/fig-6

and drug indications using ANN, based on a combination of the literature corpus and the random walk corpus.

Our knowledge graph contains a massive number of chemicals, many of which are not drug-like, and while the performance in predicting drug targets is somewhat higher when using the knowledge graph embeddings, the overall performance is still dominated by the literature-derived embedding vectors. However, when predicting indications for known drugs, both our graph and literature overlap more substantially while nevertheless containing complementary information. We observe a significant improvement in predicting drug indications when combining the information from literature and the knowledge graph. All DTI predictions, as well as the predictions for drug indications, are available at https://github.com/bio-ontology-research-group/multi-drug-embedding.

DISCUSSION

There are many scenarios in biological and biomedical research in which predictive models need to be built that can utilize information that is represented in different formats. Our key contribution is a method to integrate data represented in structured databases, in particular knowledge graphs represented in RDF and OWL, and integrate this information with information in literature. While we primarily focus on the prediction of DTIs and drug indications based on information in text and databases, our approach is generic and can serve as a paradigm for learning from multi-modal, heterogeneous data in biology and biomedicine.

Our method uses feature learning to project different types of data into a vector space, and combine data of different modes either within a single vector space (when mapping data of different modes to the same space, or to vector spaces of identical dimensions) or we combine the vector spaces themselves. We rely on the recent success of deep learning methods (*Raví et al., 2017*; *Angermueller et al., 2016*) which improved our ability to learn

relevant features from a data set and project them into a vector space. In particular, our approach relies on natural language models, in particular Word2Vec (*Mikolov et al., 2013*), and recent approaches to project information in knowledge graphs into vector spaces (*Nickel et al., 2016*; *Alshahrani et al., 2017*). Furthermore, the use of supervised learning on feature vectors has been shown to improve classification performance over traditional techniques as they become accessible to build task-specific machine learning models (*Smaili, Gao & Hoehndorf, 2018*; *Smaili, Gao & Hoehndorf, 2019*). In this work, the classifiers we used utilize the similarity-based embeddings to learn decision boundaries between the two classes(*i.e.,* interacting/non-interacting relations). These approaches are now increasingly applied in biological and biomedical research (*Alshahrani & Hoehndorf, 2018*) yet often restricted to single types of representation (such as images, genomic sequences, text, or knowledge graphs).

Our approach naturally builds on the significant efforts that have been invested in the development of named entity recognition and normalization methods for many different biological entities (*Rebholz-Schuhmann, Oellrich & Hoehndorf, 2012*) as well as the effort to formally represent and integrate biological data using Semantic Web technologies (*Jupp et al., 2014a*; *Callahan et al., 2013*). Several biological data providers now provide their data natively using RDF (*Jupp, Stevens & Hoehndorf, 2012*; *UniProt Consortium, 2018*). Furthermore, many methods and tools have been developed to normalize mentions of biological entities in text to biological databases, for example for mentions of genes and proteins, (*Leaman & Gonzalez, 2008*; *Wei, Kao & Lu, 2015*), chemicals (*Leaman, Wei & Lu, 2015*) as well as diseases (*Leaman, Islamaj Doğan & Lu, 2013*), and repositories have been developed to aggregate and integrate the annotations to literature abstracts or full-text articles (*Wei, Kao & Lu, 2013*; *Kim & Wang, 2012*). While these methods, tools, and repositories are not commonly designed to normalize mentions of biological entities to a knowledge graph, we demonstrate here how a normalization of text to a knowledge graph can be achieved, and subsequently use the combined information in our multi-modal machine learning approach. Consequently, our method has the potential to increase the value of freely available Linked Data resources and connect them directly to the methods and tools developed for natural language processing and text mining in biology and biomedicine.

One potential objection to using features generated from the biomedical literature is that the association between a drug and its target or indication may already be stated explicitly in the literature and could therefore be extracted more easily by methods relying on text mining and natural language processing. We tested how many drugs co-occur with their targets or indications in our literature-derived corpus compared to the total number of co-occurrences between mentions of drugs and proteins or diseases. Among all of the directly co-occurring mentions of drugs and proteins and drugs and diseases in the abstracts, 2.8% and 0.8% are positive pairs in our drug–target and drug–indication set, respectively. However, among the positive pairs that are both found in literature and the knowledge graph, the directly co-occurring drug–target pairs are 27.3% and drug–disease pairs 63.4%. We experimented with removing all abstracts in which the drug and protein or drug and disease pairs that are in our evaluation set co-occur. Table S6 shows the resulting

performance and demonstrates that removing the directly co-occurring pairs does not change results significantly.

CONCLUSION

We developed a generic method for combining information in knowledge graphs and natural language texts, and jointly learns both. This method is capable of utilizing information in a knowledge graph as background knowledge when ''reading'' text and vice versa when learning from structured information in a knowledge graph. We demonstrate that our method can be used to predict DTI and indications.

In the future, it would be beneficial to develop better entity normalization methods that can directly normalize entity mentions in text to a knowledge graph. We also intend to evaluate the success of our approach on full-text articles so that more information, in particular regarding methods and experimental protocols, can be utilized by our approach. Methodologically, we also intend to apply other knowledge graph embedding methods, in particular translational embeddings (*Bordes et al., 2013*; *Nickel, Rosasco & Poggio, 2016*; *Dai & Yeung, 2006*), that have previously been combined successfully with textual information (*Wang et al., 2014c*), and evaluate their performance for prediction of biological relations.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Mona Alshahrani, Abdullah Almansour and Asma Alkhaldi are supported by the National Center of Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA), Saudi Arabia. Magbubah Essack has been supported by King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) grant no. FCC/1/1976-20-01.

Grant Disclosures

The following grant information was disclosed by the authors:
National Center of Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA), Saudi Arabia.
King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR): FCC/1/1976-17-01.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Mona Alshahrani conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Abdullah Almansour, Asma Alkhaldi and Mahmut Uludag performed the experiments, prepared figures and/or tables, and approved the final draft.

- Maha A Thafar analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Magbubah Essack analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Robert Hoehndorf conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available in GitHub: https://github.com/bio-ontology-research-group/multi-drug-embedding.

The raw data is available at Zenodo: Alshahrani, Mona. (2022). Combining biomedical knowledge graphs and text to improve predictions for drug-target interactions and drug-indications. [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6148694.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.13061#supplemental-information.

REFERENCES

Agarwal P, Searls DB. 2008. Literature mining in support of drug discovery. *Briefings in Bioinformatics* **9(6)**:479–492 DOI 10.1093/bib/bbn035.

Ali M, Hoyt CT, Domingo-Fernandez D., Lehmann J, Jabeen H. 2018. BioKEEN: a library for learning and evaluating biological knowledge graph embeddings. *bioRxiv* 475202.

Alshahrani M, Hoehndorf R. 2018. Semantic disease gene embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes. *Bioinformatics* **34**:i901–i907 DOI 10.1093/bioinformatics/bty559.

Alshahrani M, Khan MA, Maddouri O, Kinjo AR, Queralt-Rosinach N, Hoehndorf R. 2017. Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics* **33(17)**:2723–2730 DOI 10.1093/bioinformatics/btx275.

Alshahrani M, Thafar MA, Essack M.. 2021. Application and evaluation of knowledge graph embeddings in biomedical data. *PeerJ Computer Science* **7**:e341 DOI 10.7717/peerj-cs.341.

Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A. 2011. Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics* **12(4)**:357–368 DOI 10.1093/bib/bbr005.

Angermueller C, Pärnamaa T, Parts L, Stegle O. 2016. Deep learning for computational biology. *Molecular Systems Biology* **12(7)**:878 DOI 10.15252/msb.20156651.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry MJ, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Tarver LI, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000.

Gene ontology: tool for the unification of biology. *Nature Genetics* **25(1)**:25–29 DOI 10.1038/75556.

Beckett D. 2004. RDF/XML Syntax Specification (Revised). W3C recommendation, World Wide Web Consortium (W3C).

Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. 2008. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics* **41(5)**:706–716 DOI 10.1016/j.jbi.2008.03.004.

Berners-Lee T, Hendler J, Lassila O. 2001. The semantic web. *Scientific American* **284(5)**:34–43.

Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PH. 2016. Fully-convolutional siamese networks for object tracking. In: *European conference on computer vision*. 850–865.

Bizer C, Heath T, Berners-Lee T. 2011. Linked data: The story so far. In: *Semantic services, interoperability and web applications: emerging concepts*. IGI Global, 205–227.

Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. 2013. Translating embeddings for modeling multi-relational data. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in neural information processing systems 26*. Red Hook: Curran Associates, Inc., 2787–2795.

Brickley D, Guha RV. 2004. RDF vocabulary description language 1.0: RDF schema. *Available at https://www.w3.org/2001/sw/RDFCore/Schema/200212bwm/*.

Callahan A, Cruz-Toledo J, Ansell P, Dumontier M. 2013. Bio2RDF release 2: improved coverage, interoperability and provenance of life science linked data. In: *Extended semantic web conference*. 200–212.

Chen L, Zeng W-M, Cai Y-D, Feng K-Y, Chou K-C. 2012. Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLOS ONE* **7(4)**:e35254 DOI 10.1371/journal.pone.0035254.

Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, Zhang Y. 2015. Drug–target interaction prediction: databases, web servers and computational models. *Briefings in Bioinformatics* **17(4)**:696–712.

Dai G, Yeung D-Y. 2006. Tensor embedding methods. In: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, AAAI'06*. Palo Alto: AAAI Press, 330–335.

Devlin J, Chang M-W, Lee K, Toutanova K. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. ArXiv preprint. arXiv:1810.04805.

Dietz L, Kotov A, Meij E. 2018. Utilizing knowledge graphs for text-centric information retrieval. In: *The 41st international ACM SIGIR conference on research & development in information retrieval*. New York: ACM, 1387–1390.

Ehrlinger L, Wöß W. 2016. Towards a definition of knowledge graphs. In: *SEMANTiCS (Posters, Demos, SuCCESS)*.

Ezzat A, Wu M, Li X-L, Kwoh C-K. 2018. Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Briefings in Bioinformatics* **20(4)**:337–1357 DOI 10.1093/bib/bby002.

Fawcett T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* **27(8)**:861–874 DOI 10.1016/j.patrec.2005.10.010.

Frijters R, Van Vugt M, Smeets R, Van Schaik R, De Vlieg J, Alkema W. 2010. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLOS Computational Biology* **6(9)**:e1000943 DOI 10.1371/journal.pcbi.1000943.

Fu G, Ding Y, Seal A, Chen B, Sun Y, Bolton E. 2016. Predicting drug target interactions using meta-path-based semantic network analysis. *BMC Bioinformatics* **17(1)**:160 DOI 10.1186/s12859-016-1005-x.

Gulli A, Pal S. 2017. *Deep learning with Keras*. Birmingham: Packt Publishing Ltd.

Gutiérrez-Basulto V, Schockaert S. 2018. From knowledge graph embedding to ontology embedding: region based representations of relational structures. ArXiv preprint. arXiv:1805.10461.

Gysi DM, Do Valle Í, Zitnik M, Ameli A, Gan X, Varol O, Ghiassian SD, Patten J, Davey Robert ALJ, Barabasi A-L. 2021. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences of the United States of America* **118(19)**.

Hinton G, Srivastava N, Swersky K. 2012. Lecture 6a overview of mini–batch gradient descent. *Available at https://www.youtube.com/watch?v=tCTfb6PAr4w*.

Hoehndorf R, Schofield PN, Gkoutos GV. 2015. Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases. *Scientific Reports* **5**:10888 DOI 10.1038/srep10888.

Hoffmann R, Zhang C, Ling X, Zettlemoyer L, Weld DS. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies-volume 1*. 541–550.

Ji S, Pan S, Cambria E, Marttinen P, Philip SY. 2021. A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems* **33(2)**:494–514 DOI 10.1109/TNNLS.2021.3070843.

Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, Gaulton A, Gehant S, Laibe C, Redaschi N, Wimalaratne SM, Martin M, Le Novre N, Parkinson H, Birney E, Jenkinson AM. 2014a. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics* **30(9)**:1338–1339 DOI 10.1093/bioinformatics/btt765.

Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, Gaulton A, Gehant S, Laibe C, Redaschi N, Wimalaratne SM, Martin M, Novére NL, Parkinson H, Birney E, Jenkinson AM. 2014b. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics* **30(9)**:1338–1339 DOI 10.1093/bioinformatics/btt765.

Jupp S, Stevens R, Hoehndorf R. 2012. Logical Gene Ontology Annotations (GOAL): exploring gene ontology annotations with OWL. *Journal of Biomedical Semantics* **3(Suppl 1)**:S3 DOI 10.1186/2041-1480-3-S1-S3.

Kim J-D, Wang Y. 2012. PubAnnotation: a persistent and sharable corpus and annotation repository. In: *Proceedings of the 2012 workshop on biomedical natural language processing, BioNLP '12*. Stroudsburg, PA, USA: Association for Computational Linguistics, 202–205.

Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GCM, Brown DL, Brudno M, Campbell J, FitzPatrick DR, Eppig JT, Jackson AP, Freson K, Girdea M, Helbig I, Hurst JA, Jähn J, Jackson LG, Kelly AM, Ledbetter DH, Mansour S, Martin CL, Moss C, Mumford A, Ouwehand WH, Park S-M, Riggs ER, Scott RH, Sisodiya S, Vooren SV, Wapner RJ, Wilkie AOM, Wright CF, Vulto-van Silfhout AT, Leeuw Nd, de Vries BBA, Washingthon NL, Smith CL, Westerfield M, Schofield P, Ruef BJ, Gkoutos GV, Haendel M, Smedley D, Lewis SE, Robinson PN. 2014. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research* **42(D1)**:D966–D974 DOI 10.1093/nar/gkt1026.

Kuhn M, Letunic I, Jensen LJ, Bork P. 2015. The SIDER database of drugs and side effects. *Nucleic Acids Research* **44(D1)**:D1075–D1079 DOI 10.1093/nar/gkv1075.

Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P. 2012. STITCH 3: zooming in on protein-chemical interactions. *Nucleic Acids Research* **40(D1)**:D876–D880 DOI 10.1093/nar/gkr1011.

Leaman R, Gonzalez G. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. In: *Pacific symposium on biocomputing*. 652–663.

Leaman R, Islamaj Doğan R, Lu Z. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* **29(22)**:2909–2917 DOI 10.1093/bioinformatics/btt474.

Leaman R, Wei C-H, Lu Z. 2015. tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics* **7(1)**:S3 DOI 10.1186/1758-2946-7-S1-S3.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* **521(7553)**:436 DOI 10.1038/nature14539.

Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36(4)**:1234–1240.

Li Y, Li J, Bian N. 2019. DNILMF-LDA: prediction of lncRNA-disease associations by dual-network integrated logistic matrix factorization and Bayesian optimization. *Genes* **10(8)**:608 DOI 10.3390/genes10080608.

Lin Y, Liu Z, Sun M, Liu Y, Zhu X. 2015. Learning entity and relation embeddings for knowledge graph completion. 2181–2187.

Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, Peng J, Chen L, Zeng J. 2017. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications* **8(1)**:573 DOI 10.1038/s41467-017-00680-8.

Mei J-P, Kwoh C-K, Yang P, Li X-L, Zheng J. 2013. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics* **29(2)**:238–245 DOI 10.1093/bioinformatics/bts670.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013. Distributed representations of words and phrases and their compositionality. ArXiv preprint. arXiv:10.4546.

Mohamed SK, Nováček V, Nounu A. 2019. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* **36(2)**:603–610 DOI 10.1093/bioinformatics/btz600.

Mohamed SK, Nováček V, Nounu A. 2020. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* **36(2)**:603–610.

Muñoz E, Nováček V, Vandenbussche P-Y. 2017. Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models. *Briefings in Bioinformatics* **20(1)**:190–202.

Nair V, Hinton GE. 2010. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

Nascimento AC, Prudêncio RB, Costa IG. 2016. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinformatics* **17(1)**:1–16 DOI 10.1186/s12859-015-0844-1.

Nelson W, Zitnik M, Wang B, Leskovec J, Goldenberg A, Sharan R. 2019. To embed or not: network embedding as a paradigm in computational biology. *Frontiers in Genetics* **10**.

Nickel M, Murphy K, Tresp V, Gabrilovich E. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE* **104(1)**:11–33 DOI 10.1109/JPROC.2015.2483592.

Nickel M, Rosasco L, Poggio T. 2016. Holographic embeddings of knowledge graphs. In: *Proceedings of the thirtieth AAAI conference on artificial intelligence, AAAI'16*. Palo Alto: AAAI Press, 1955–1961.

Paulheim H. 2017. Knowledge graph refinement: a survey of approaches and evaluation methods. *Semantic Web* **8(3)**:489–508.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* **12**:2825–2830.

Pennington J, Socher R, Manning C. 2014. Glove: global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

Percha B, Altman RB. 2018. A global network of biomedical relationships derived from text. *Bioinformatics* **34(15)**:2614–2624 DOI 10.1093/bioinformatics/bty114.

Perozzi B, Al-Rfou R, Skiena S. 2014. Deepwalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: ACM, 701–710.

Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. 2016. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* **45(D1)**:D833–D839 DOI 10.1093/nar/gkw943.

Pryor R, Cabreiro F. 2015. Repurposing metformin: an old drug with new tricks in its binding pockets. *Biochemical Journal* **471(3)**:307–322 DOI 10.1042/BJ20150497.

Raví D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, Yang GZ. 2017. Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics* **21(1)**:4–21 DOI 10.1109/JBHI.2016.2636665.

Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. 2012. Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics* **13(12)**:829–839 DOI 10.1038/nrg3337.

Ristoski P, Paulheim H. 2016. RDF2Vec: RDF graph embeddings for data mining. In: Groth P, Simperl E, Gray A, Sabou M, Krötzsch M, Lecue F, Flöck F, Gil Y, eds. *The Semantic Web –ISWC 2016*. Cham: Springer International Publishing, 498–514.

Sang S, Yang Z, Wang L, Liu X, Lin H, Wang J. 2018. SemaTyP: a knowledge graph based literature mining method for drug discovery. *BMC Bioinformatics* **19(1)**:193 DOI 10.1186/s12859-018-2167-5.

Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, Feng G, Kibbe WA. 2011. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Research* **40(D1)**:D940–D946.

Seal A, Ahn Y-Y, Wild DJ. 2015. Optimizing drug–target interaction prediction based on random walk on heterogeneous networks. *Journal of Cheminformatics* **7(1)**:40 DOI 10.1186/s13321-015-0089-z.

Smaili FZ, Gao X, Hoehndorf R. 2018. Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics* in press.

Smaili FZ, Gao X, Hoehndorf R. 2019. Opa2vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics* **35(12)**:2133–2140 DOI 10.1093/bioinformatics/bty933.

Swanson DR. 1990. Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association* **78(1)**:29.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, Mering Cv. 2010. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research* **39(suppl_1)**:D561–D568.

Thafar M, Raies AB, Albaradei S, Essack M, Bajic VB. 2019. Comparison study of computational prediction tools for drug-target binding affinities. *Frontiers in Chemistry* **7**.

Thafar MA, Albaradie S, Olayan RS, Ashoor H, Essack M, Bajic VB. 2020a. Computational drug-target interaction prediction based on graph embedding and graph mining. In: *Proceedings of the 2020 10th international conference on bioscience, biochemistry and bioinformatics*. 14–21.

Thafar MA, Olayan RS, Albaradei S, Bajic VB, Gojobori T, Essack M, Gao X. 2021. DTi2Vec: Drug–target interaction prediction using network embedding and ensemble learning. *Journal of Cheminformatics* **13(1)**:1–18 DOI 10.1186/s13321-020-00477-w.

Thafar MA, Olayan RS, Ashoor H, Albaradei S, Bajic VB, Gao X, Gojobori T, Essack M. 2020b. DTiGEMS+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *Journal of Cheminformatics* **12**.

The UniProt Consortium. 2016. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45(D1)**:D158–D169 DOI 10.1093/nar/gkw1099.

UniProt Consortium T. 2018. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **46(5)**:2699 DOI 10.1093/nar/gky092.

Van der Maaten L, Hinton G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**:2579–2605.

Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. 2014a. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11(3)**:333 DOI 10.1038/nmeth.2810.

Wang Z, Zhang J, Feng J, Chen Z. 2014b. Knowledge graph and text jointly embedding. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1591–1601.

Wang Z, Zhang J, Feng J, Chen Z. 2014c. Knowledge graph and text jointly embedding. In: *The 2014 conference on empirical methods on natural language processing*. Cedarville: ACL Association for Computational Linguistics.

Wei C-H, Kao H-Y, Lu Z. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research* **41(W1)**:W518–W522 DOI 10.1093/nar/gkt441.

Wei C-H, Kao H-Y, Lu Z. 2015. GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed Research International* **2015**.

Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, Mons B. 2012. Open PHACTS: semantic interoperability for drug discovery. *Drug Discovery Today* **17(2122)**:1188–1198 DOI 10.1016/j.drudis.2012.05.016.

Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* **36(suppl_1)**:D901–D906 DOI 10.1093/nar/gkm958.

Xie R, Liu Z, Jia J, Luan H, Sun M. 2016. Representation learning of knowledge graphs with entity descriptions. In: *Thirtieth AAAI conference on artificial intelligence*.

Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. 2008. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **24(13)**:i232–i240 DOI 10.1093/bioinformatics/btn162.


PeerJ
Life & Environment
PeerJ
Organic Chemistry