Research article

Natural products can be potential inhibitors of metalloproteinase II from *Bacteroides fragilis* to intervene colorectal cancer

Bushra Arif [a], Saba Yasir [b,*], Muhammad Saeed [a], M. Qaiser Fatmi [a,**]

[a] *Department of Biosciences, COMSATS University Islamabad, Islamabad Campus, Pakistan*
[b] *Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA*

A B S T R A C T

Bacteroides fragilis, a gram negative and obligate anaerobe bacterium, is a member of normal gut microbiota and facilitates many essential roles being performed in human body in normal circumstances specifically in Gastrointestinal or GI tract. Sometimes, due to genetics, epigenetics, and environmental factors, *Bacteroides fragilis* and their protein(s) start interacting with intestinal epithelium thus damaging the lining leading to colorectal cancers (CRC). To identify these protein (s), we incorporated a novel subtractive proteomics approach in the study. Metalloproteinase II (MPII), a *Bacteroides fragilis* toxin (*bft*), was investigated for its virulence and unique pathways to demonstrate its specificity and uniqueness in pathogenicity followed by molecular docking against a set of small drug-like natural molecules to discover potential inhibitors against the toxin. All these identified inhibitor-like molecules were analyzed for their ADMET calculations and detailed physiochemical properties to predict their druggability, GI absorption, blood brain barrier and skin permeation, and others. Resultantly, a total of ten compounds with the least binding energies were obtained and were subjected to protein-compound interaction analysis. Interaction analysis revealed the most common ligand-interacting residues in MPII are His 345, Glu 346, His 339, Gly 310, Tyr 341, Pro 340, Asp 187, Phe 309, Lys 307, Ile 185, Thr 308, and Pro 184. Therefore, top three compounds complexed with MPII having best binding energies were selected in order to analyze their trajectories. RMSD, RMSF, Rg and MMPBSA analysis revealed that all compounds showed good binding and keeping the complex stable and compact throughout the simulation time in addition to all properties and qualities of being a potential inhibitor against MPII.

1. Introduction

Colorectal Cancer (or CRC), also known as rectal, colon or bowel cancer, is among the well-studied and commonly known malignancies worldwide [1]. CRC is the most occurring cancer in both, men and women [2] and the third most commonly occurring cancer leading to more than million cases and ~500,000 deaths every year around the world [3]. It is benignly originated in colon epithelium as lesions or adenomas which ultimately develop into cancer with continuous progression over the period of time [4]. CRC

---

development is a multistep process which involves a series of morphological, histological [1] and genetic alternations that are accumulated with time [5]. In particular, there is no specific cause of CRC, however, several risk factors are reported that are involved in the initiation of carcinoma, such as age and growth [6], unhealthy dietary habits, smoking [7], and less physical activities [8]. In addition to all mentioned factors, obesity is an important factor playing a vital role in causation of CRC *via* a multi-staged process *i.e.* initiation, promotion, progression and metastasis [9]. As a result, inflammation is initiated in colonic epithelium by depositing visceral fat leading to metabolic disorders and altered production of cytokines and other immune cells [10]. Moreover, the environmental factors inducing genetic and epigenetic modifications are also an important reason to develop CRC [11,12]. These factors involve three distinct pathways *i.e.* chromosomal abnormalities or disabilities, also known as chromosomal instability (CIN), microsatellite instability (MSI), and CpG-island methylator phenotype pathways (CIMP) [11]. Among these, CIN plays an important role in damaging the colonic epithelial lining thus results in tumorigenesis [13], and is involved in programmed de-regulation of various oncogenes and tumor suppressor genes [14]. Until recently, aforementioned factors were considered as the risk factors for CRC. However, recent studies revealed that the members of human microbiota [15] which include trillions of bacterial and archaeal species residing in and on the human body symbiotically [16] are another imprtant factors for CRC. These species are spread all over the body sites and are referred to as human microbiota [17], while their genes and products are known as human microbiome [18]. Human microbiota is symbiotically involved in performing many essential roles with in the body [19]. For example, food digestion and metabolism [20], gut-brain axis development [21], short chain fatty acids production (SCFA), vitamin B and K production [22], influencing mood, anxiety [23], depression [24], social behaviors [25], immune system maturation [26] and many others. In addition to microbiota, intestines also play important role in maintaining immunity, and are known as the largest immunity organ rather than just serve as an absorption organ as it helps in maintaining health and immune response [27] thus continue to grow and mature throughout the life. Among GI tract, especially colonic [28], and intestinal microbiota, microbial species are dominantly harbored with anaerobic bacteria [29]. For example, most of the bacterial species in colon belongs to certain genera such as *Bacteroides*, *Propionibacterium*, *Eubacterium*, and *Fusobacterium* [30].

Besides maintaining health and well-being, human microbiota is also responsible for causing many serious health issues and diseases [31]. Dysbiosis (or imbalance) in microbiota is associated with and/or leads to various social, immune and metabolic disorders [32,33]. Similarly, microbiota dysbiosis is dominantly involved in causing cancer particularly related to GI tract [34]. Some of the microbial species in disturbed conditions damage the intestinal lining [35] and trigger cytokines production [10]. This causes inflammation in epithelial lining and progresses towards polyp formation and CRC [36].

Bacteroides fragilis, a gram negative bacterium, is one of the dominant members of normal distal-end of intestinal microbiota [37]. Generally, *Bacteroides fragilis* evades the host immune system by producing a large number of capsular polysaccharides [38], extracellular polysaccharides [39], and glycoproteins [40] thus creates variable surface antigenicities. Their polysaccharides (PSA) are responsible for mediating and activating T-cell dependent immunological responses [41]. However, sometimes due to certain factors, instead of symbiosis [42], Enterotoxigenic *Bacteroides fragilis* (ETBF) [43] and their protein(s) starts interacting and damaging the epithelial lining of colon leading to CRC [44,45]. For example, fragilysin (FRA) and metalloproteinase II (MPII) are the two virulent proteins produced by the pathogenic island in the genome of Enterotoxigenic *Bacteroides fragilis*. Both proteins (MPII and FRA) are the secretory metalloproteinases specifically zinc dependent proteins [46]. These proteins interact with the epithelium of colon and mediate signal transduction [47] and inflammatory pathways [48].

In general, there are several strains of Enterotoxigenic *Bacteroides fragilis*, for example, strains BOB25, 86-5443-2-2, 2-078382-3, 20793-3, 20656-2-1, CL07T00C01, CL07T12C05, and BJH_183 are known and reported to produce toxins *i.e.* fragilysin and MPII [49] and have been reported to cause inflammation when interact with lining in GI tract [50]. Therefore, we aim to take an advantage of therapeutic approach to develop potential inhibitor(s) with novel scaffolds against Enterotoxigenic *Bacteroides fragilis* secretory protein (s). Herein, we incorporated *state of the art* subtractive proteomic approach to identify the virulent protein(s). By targeting these protein(s), we proposed to identify the potential compounds with the least binding energies when docked and also possess the drug-like properties when all the ADMET profiles were analyzed. Moreover, top compounds with the best binding energies were selected in order to analyze the conformational changes and dynamic behavior of MPII in the presence of drug-like compounds.

2. Materials and methods

2.1. Data collection

Complete proteome of Enterotoxigenic *Bacteroides fragilis* and Human were retrieved from a publicly available repository "UniProt Proteomes". Proteome for Enterotoxigenic *Bacteroides fragilis* strain (2-078382-3) was downloaded to align with the human proteome. Protein sequence was retrieved from an online repository "UniProt" and its predicted 3D structure was downloaded from a RSCB Protein databank (PDB). In addition, complete bacterial essential proteins were downloaded from "database of essential genes (DEG)" in order to identify the essential and non-essential yet important proteins.

Furthermore, the ligands datasets for virtual screening were downloaded from online available repository *i.e.* ZINC15.

2.2. Non-homologues identification using subtractive proteomics approach

2.2.1. Identification and removal of orthologs from bacterial proteome
To identify unique or non-orthologous proteins against human proteome, a local database of complete human proteome was created using a standalone NCBI blast (v2.5.0). Enterotoxigenic *Bacteroides fragilis* proteome (2-078382-3) was subjected to blastp against human proteome using default parameters. Selection criteria includes sequence identity percentage ≤25 %, E-value ≥10-3 (0.001) and Bit-Score ≤50.

2.2.2. Identification of bacterial essential proteins
Essential proteins serve as survival proteins for bacteria, which can be or cannot be virulent or pathogenic. Identification of these proteins can help to deal with bacterial virulence. Bacterial proteome contains essential (as well as non-essential) proteins that are important to cause pathogenicity and are involved in various diseases when produced in disturbed conditions. All non-orthologous sequences were subjected to blastp by creating a local database of bacterial essential protein sequences was created using an online database, the Database of Essential Proteins or DEG v15.2 (http://tubic.tju.edu.cn/deg/). Therefore, the bacterial essential proteins were screened under selection criteria (*i.e.* sequence identity percentage ≥30 % and bit score ≥25) and those proteins which fulfilled the criteria were termed as essential proteins. A set of proteins which were categorized as non-essential based on selection criteria were removed from analysis.

2.2.3. Identification of non-paralogs in bacterial proteome
Orthologs and paralogs genes/proteins are evolutionary related and retain similar functions across different species and also within the same organism respectively [51]. In order to avoid redundancy, duplicated proteins or those belong to same functional category and family, were eliminated using an online tool CD HIT. All the essential protein sequences were given to CD-HIT by keeping the default parameters while the identity cut off value was set to 0.6 (or 60 %). This tool clusters all the sequences together in a group, which are equal or greater than 60 % sequence identity within same proteome. Therefore, all the clustered proteins (paralogs) representing same family and having the same function were eliminated from the study.

2.3. Characterization of bacterial protein

This step involves the identification and classification of those proteins which serves as a potential targets for further study and analysis.

2.3.1. Bacterial virulent protein identification
Microbial proteins, specific to members of microbiota, are important in mediating many pathways and have influential effect on host functions and health. These proteins can be pathogenic and serve as virulence factors for the enterotoxic microbial species as well as they can direct pathogenic pathways and cause diseases in humans. This step was performed to identify the virulence factor(s) in *Bacteroides fragilis* non-paralogous essential proteins by subjecting them to blastp against virulence factors dataset available at Virulence Factor Database or VFDB. In VFDB, there are two datasets representing virulent factors *i.e.* Set A and Set B. Set A (also termed as core set) represent only experimentally verified genes/proteins for virulence. Set B (referred as full dataset) covers all genes/proteins which are predicted to be virulent and are related to virulence factors of pathogenic bacteria in the database [52]. The identified non-paralogous essential proteins were subjected to blastp against locally created database containing set B virulence proteins downloaded from VFDB. Therefore, proteins selected under defined criterion *i.e.* sequence identity percentage ≥30 % and Bit score ≥150 were termed as virulent proteins which can be a potential candidate(s) for drugs targets.

2.3.2. Unique proteins and pathways identification in Bacteroides fragilis
Microbial species perform various essential and unique pathways within the host having specific effect in terms of maintaining health or initiating diseases [53]. The main purpose to identify unique pathways of *Bacteroides fragilis* being performed in humans is to determine their association with cancer or other diseases. These specific pathways were identified from online available database of pathways, KEGG (or Kyoto Encyclopedia of Genes and Genomes). All pathways of *Bacteroides fragilis* and *Homo sapiens* were manually compared to categorize as unique and common pathways. Furthermore, selected non-paralog essential proteins sequences identified as

virulence factors were subjected to Kegg 2/KEGG Mapper to assign KO numbers. These KO numbers were used to identify the function of protein which are categorized and assigned according to their functionality.

2.3.3. Subcellular localization analysis of Bacteroides fragilis proteins

Subcellular localization of proteins is an important feature to identify druggable targets, as most of the activities are performed in the cell compartments [54]. Here, virulent proteins of *Bacteroides fragilis* were subjected to online servers PSORTb (v.3.0) and CELLO (v.2.5) and were searched for the proteins associated to extracellular activities such as proteolysis.

2.4. MPII structure and modeling

Three dimensional structure of MPII protein was downloaded from RSCB Protein Databank (PDB) in order to perform docking studies. For this, MPII (PDB ID: 4ON1) was downloaded in fasta sequence and pdb format files. Missing residues present in the pdb file were fixed using the self-template with the help of Swiss model available at (https://swissmodel.expasy.org/) [55]. Moreover, energy minimization of the obtained protein structure was performed to stabilize the modeled structure by using GROMOS96 force field [107]. Furthermore, the obtained 3D structure was also assessed to identify its structural stability. After quality assessment, coordinate or pdb file of modeled 3D structure was downloaded and was prepared for docking process. In addition, PROCHECK (https://saves. mbi.ucla.edu/) was used to evaluate the overall quality of modeled structure.

2.4.1. Structure preparation and target identification

Protein structure preparation for screening and docking is considered as a crucial step in structure-based drug design, as it ensures the accuracy and reliability in predicting the docked poses and conformations. After identification and rectification of missing residue regions, extraneous components were removed such as ions and water molecules present in the protein. In addition, hydrogen atoms were placed followed by binding sites prediction. Binding sites were selected on the basis of presence of the catalytic domain and Zn^{2+} ion, which is involved in secretion based E-cadherin cleavage [46]. Therefore, grids were generated incorporating binding sites to facilitate docking and explore the orientations of bound ligands.

2.5. Virtual screening of ZINC database against selected drug targets

Structure-based virtual screening (or SBVS) is one of the most effective and promising *in silico* technique to discover new drugs. SBVS sorts the top hits and identifies the best conformation between protein and ligands forming a stable complex by estimating non-covalent forces using score functions [56]. In order to identify some potential small compounds to act as inhibitors, ZINC dataset containing 0.2 million ligands were screened against the *Bacteroides fragilis* protein MPII using Autodock VINA tools.

2.6. ADMET calculation of identified ligands

The identified drug-like molecules, selected for further investigation of drug discovery process, may possibly have compromised absorption and excretion properties. Due to their compromised profiles, they can fail in clinical trials [57]. Reportedly, safety and efficacy are two important reasons for drug failure [58] and to avoid and counter with drug failure, we calculated the ADMET properties beyond the traditional physiochemical properties. The detailed pharmacokinetics study was performed for the top-ranked hit compounds. The study involved qualitative or quantitative analysis of absorption, distribution, metabolism, excretion, and toxicity [59] using an online tool SwissADME, a freely available online tool [60]. The compounds, which fulfilled all the criteria of ADMET characteristics, were subjected to further analysis.

2.7. Protein-ligand interaction analysis

Characterization of protein-ligand interactions provides deep insights to the non-covalent interactions within the complexes [61]. For this purpose, Discovery studio visualizer and Chimera were used to identify the essential interactions between the protein and drug-like molecules. These give a schematic representation of interactions in 2D and 3D showing the interacting residues, hydrophobic interactions, and hydrogen bonds [62,63].

2.8. Molecular dynamic simulations

After detailed protein-ligand interaction analysis, complexes were selected on the basis of lowest binding energies. The protein-ligand complex(s) were protonated at 7.4 pH where the protonation states of Asp, His, Arg, Lys, and Glu were adjusted when evaluated at pH 7.4 using freely available web application "protein prepare" (available at https://www.playmolecule.com/proteinPrepare/). Molecular dynamic simulations were performed using the GROMACS (GROningen Machine for Chemical Simulations) simulation software (v. 2021.4) with CHARMM36m (Chemistry at HARvard Macromolecular Mechanics 36m) force field and antechamber force field was used for selected compounds [108]. Protein-ligand complex(s) were solvated with TIP3P water molecules in an octahedral water box keeping the minimum distance 10 Å between protein-ligand system(s) and simulation box edges in order to maintain the minimum image convention during the simulation time. Ionic concentration of 0.15 M KCl was introduced in the system by adding 92 K^+ and 81 Cl^- ions and PBC (periodic bound conditions) were applied in x, y, and z directions. CHARMM-GUI (Chemistry

at HARvard Macromolecular Mechanics Graphic User Interface) was used to prepare selected protein-ligand complex(s) and their input files [109].

Each complex system was separately minimized using the steep descent technique for 5000 steps. To avoid any steric clashes, force limit of 1000 kJ/mol/nm was maintained. After minimization, systems were equilibrated at NVT and NPT ensemble for 500,000 steps with a time step of 1 fs. The sampling for the selected protein-ligand complexes were conducted at NPT for 100,000,000 steps (200ns) at a constant temperature (310.15K) [110]. VMD was used for molecular visualization and R codes were used to calculate RMSD (root mean square deviation), RMSF (root mean square fluctuation) and Rg (radius of gyration).

2.8.1. Binding free energies analysis

For MMPBSA analysis, first 100 ns were excluded and the trajectories with the last 100 ns were utilized to analyze the binding energies of the protein and ligand complexes. To calculate the binding energies, following formula was used:

$$\Delta E = \Delta E_{Complex} - (\Delta E_{Receptor} + \Delta E_{Ligand})$$

Grace (Graphing, Advanced Computational and Exploration of data) was utilized to plot all graphs/plots.

3. Results

3.1. Subtractive analysis

The main purpose of incorporating this approach was to identify or confirm the uniqueness of protein(s) produced by *Bacteroides fragilis*. For this, several steps were performed (such as identification of non-orthologs, essential proteins, and virulence factors) by
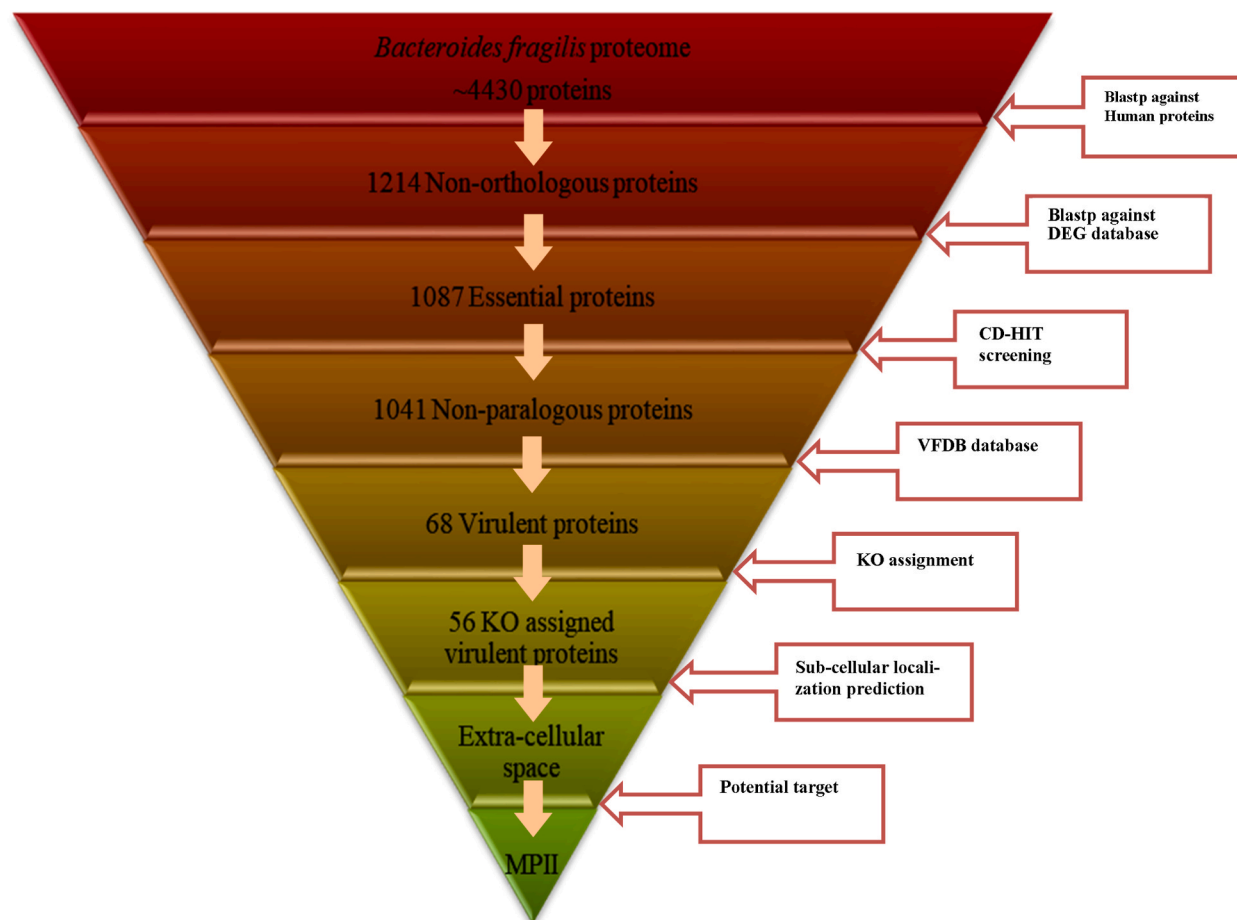
Fig. 1. Schematic overview of potential target identification in *Bacteroides fragilis* proteome. Following sequential elimination of proteins from *Bacteroides fragilis* proteome using subtractive proteomics, a unique set of 1214 non-orthologs, 1087 essential proteins, 1041 non-paralogs and 68 virulent proteins were identified. Fifty six virulent proteins had KEGG orthology (KO) number assigned. By localization prediction analysis, metalloproteinase II (MPII) was the only protein localized in extra-cellular space as a target protein.

using several databases respective to each step as mention in the methodology section. Fig. 1 shows the schematic work flow of subtractive proteomics approach to identify potential target in *Bacteroides fragilis* proteome.

3.1.1. Non-orthologs of Bacteroides fragilis proteome

Blastp of Enterotoxic *Bacteroides fragilis* (strain 2-078382-3) proteome against human proteome resulted in the identification of bacterial proteins non-homologues to human. This helps in preventing the binding of small compounds with active site present in human proteins. *Bacteroides fragilis* proteome contains a total of ~4430 proteins, which were subjected to Blastp against human proteome (contains ~82,493 proteins). A total of 1214 non-orthologous proteins were obtained after eliminating the proteins which did not meet the selection criteria and showed homology against human. Selection criteria of non-homologues proteins include sequence identity percentage ≤25 %, E-value ≥10-3 (0.001) and Bit-Score ≤50. Top 50 non-homologues sorted in ascending order on the basis of identity percentage are given in Table 1 representing their protein names. Selected non-orthologs have less than 25 % sequence similarity against human proteins.

Table 1
Top 50 Non-orthologs of *Bacteroides fragilis* proteome with respect to human proteome.

Bacterial Protein ID	*Bacteroides fragilis* Protein Names	Identity %	E-value	Bit-Score
WP_005795584	Thioredoxin family protein	15.28	7.5	27.7
OCR37541	Glycoside hydrolase	17.75	6.6	32
WP_009291817	Cadmium-translocating P-type ATPase	18.22	6.7	32.3
WP_069108608	Fimbrillin family protein	18.26	7.8	32.2
WP_005791859	IMP dehydrogenase	18.70	4.5	32.3
WP_009292809	Alpha-amylase	18.78	3.9	32.3
WP_005785270	S9 family peptidase	18.81	9.7	32
WP_005787238	Heavy-metal-associated domain-containing protein	19.30	8.5	27.7
WP_005794263	PAS domain-containing sensor histidine kinase	19.70	3.3	32.7
WP_065849766	MATE family efflux transporter	19.78	5.7	32
WP_005790353	Ion transporter	19.86	10	30.4
WP_005798974	RagB/SusD family nutrient uptake outer membrane protein	20.0	8.1	31.6
WP_005791317	Nucleoside kinase	20.0	0.08	37.4
WP_069108675	Amidoligase family protein	20.18	4.7	31.6
WP_065762777	HAMP domain-containing histidine kinase	20.34	0.89	34.3
WP_005789197	AhpC/TSA family protein	20.41	9.9	30
WP_005786014	LysE family transporter	20.42	8.4	29.6
WP_005787265	Sodium:solute symporter	20.42	9.7	31.2
WP_005786101	Hemolysin family protein	20.47	0.21	36.2
WP_009291372	Rhamnulokinase	20.49	0.013	40.4
WP_032532695	Polysaccharide pyruvyl transferase family protein	20.56	9.3	29.6
WP_005788055	2-iminoacetate synthase ThiH	20.64	9.7	30
WP_005803347	Aldo/keto reductase	20.71	1.4	32.7
WP_005795196	Arylsulfatase	20.73	0.2	36.6
WP_032561960	6-bladed beta-propeller	20.92	0.53	34.7
WP_009293167	Diaminopimelate decarboxylase	20.93	0.66	34.3
WP_009291431	Alginate export family protein	20.99	9.8	31.2
WP_005783471	Calcium/sodium antiporter	21.05	8.1	30.8
WP_005796577	Sodium-dependent transporter	21.07	0.019	39.7
WP_005787240	Heavy metal translocating P-type ATPase	21.22	1.9	34.3
WP_009292593	HAD family hydrolase	21.22	4.1	30.8
OCR43327	Immunoreactive antigen PG32	21.78	5.1	32
WP_065849773	Fimbrillin family protein	21.28	3.6	32
OCR44583	Heat-shock protein	21.30	7.9	28.1
WP_005798153	RagB/SusD family nutrient uptake outer membrane protein	21.30	2.6	33.1
OCR42631	Multidrug transporter MatE	21.51	2.2	33.1
WP_005784407	Potassium channel family protein	21.52	4.3	30.8
WP_005786111	Esterase family protein	21.69	9.6	30
WP_009292683	Virulence RhuM family protein	21.87	0.16	36.6
OCR44055	Transcriptional regulator	21.88	5	32
WP_005816525	Fic family protein	22.13	2.3	32
WP_005797764	SDR family oxidoreductase	22.22	2.6	31.6
WP_005793843	Phage holin family protein	22.22	3.5	30
OCR42372	Multidrug transporter AcrB	22.53	0.19	38.1
OCR44524	Radical SAM protein	22.58	9.6	27.3
WP_005797263	Fragilysin family metalloproteinase II (MPII)	22.727	0.16	36.6
OCR42316	ArsR family transcriptional regulator	22.73	9.8	31.2
WP_005786184	VOC family protein	22.77	7.8	28.5
WP_032568025	Reverse transcriptase family protein	22.77	2.2	33.1
WP_008769229	Response regulator	22.91	9.9	32.7

3.1.2. Bacterial essential proteins: important for survival

The prediction of essential bacterial proteins helps in identifying the important roles in their survival. As a result of blastp against locally created DEG database, 1087 essential proteins were obtained out of 1214 non-homologues. The predicted proteins are essential for the survival of bacteria and some of the proteins are known for causing pathogenicity and are actively involved in initiating immune responses. Top 20 bacterial essential proteins are given in Table 2 which is sorted in ascending order on the basis of identity percentage.

3.1.3. Bacteroides fragilis non-paralog proteins

Paralog proteins or redundant sequences are one of the major concerns in finding druggable targets or binding sites. These sequences are highly homologous and may have similar protein structures and binding pockets for drugs leading towards non-selective binding of drugs with paralog proteins [64]. Therefore, all paralogs were removed before proceeding towards drug(s) selection. All selected essential proteins were subjected to CD HIT, which clustered the paralogs on the basis of sequence similarity (*i.e.* 60 %). 1087 protein sequences were given to CD HIT and 21 different clusters (46 protein sequences) were identified as paralogs and were eliminated. As a result, a total of 1041 non-paralogs were selected for further analysis. List of all selected 1041 non-paralogs is given in Table S1 and list of identified paralog clusters is given in Table S2.

3.1.4. Virulence and pathogenicity posed by Bacteroides fragilis

Virulence factors in proteins are responsible for causing pathogenicity and diseases in hosts. After removing the paralogs, selected non-paralogous sequences were aligned against full dataset of virulent proteins. Unique 68 virulent proteins were obtained which met the selection criteria (sequence identity $\geq$30 % and Bit-Score $\geq$150). 68 virulent proteins from Enterotoxigenic *Bacteroides fragilis* (strain 2-078382-3) proteome are given in Table 3.

3.2. Uniqueness of Bacteroides fragilis in performing pathways

In nature, every organism is designed to perform pathways important for living which can be performed by hosts and also be performed in collaboration with human microbiota. Therefore, in order to identify common and unique pathways, KEGG database was used which provided insights of all pathways respective to both species. KEGG database provided a total of 345 pathways in humans (*hsa*) and 103 pathways in Enterotoxigenic *Bacteroides fragilis* (*bfb*). All pathways were manually compared in order to get unique and common pathways respective to human and *Bacteroides fragilis*. Enterotoxigenic *Bacteroides fragilis* has 27 unique pathways while 76

Table 2
Top 20 *Bacteroides fragilis* essential proteins searched against the dataset of DEG database with more than 30 % sequence identity.

Bacterial Protein ID	Bacterial Essential Protein Name	Identity %	E-value	Bit-Score
WP_032497126.1	6-bladed beta-propeller	32	0.071	32
WP_005788894.1	Zinc ribbon domain-containing protein	32	0.61	27.3
OCR44468.1	Glycosyl transferase	32	4.57E-09	55.8
WP_065849887.1	MBL fold metallo-hydrolase	32.08	4.43E-41	141
WP_005797467.1	Sugar porter family MFS transporter	32.1	0.002	38.9
WP_008769190.1	LPS biosynthesis protein	32.14	0.23	32.3
WP_065849886.1	Carbohydrate kinase	32.2	8.2	27.7
OCR44055.1	Transcriptional regulator	32.26	3.22E-04	41.6
WP_005785797.1	Iron ABC transporter permease	32.28	4.50E-40	142
WP_011201989.1	Serine hydrolase	32.31	0.59	32.3
WP_008769342.1	FecR family protein	32.37	2.02E-28	111
WP_005787435.1	Sigma-54 dependent transcriptional regulator	32.39	2.10E-72	234
WP_065849721.1	SDR family oxidoreductase	32.39	3.91E-27	107
WP_005790207.1	RagB/SusD family nutrient uptake outer membrane protein	32.4	8.5	28.1
WP_005783471.1	Calcium/sodium antiporter	32.4	0.097	32.7
OCR43616.1	Sensor histidine kinase	32.43	1.84E-25	111
WP_009291431.1	Alginate export family protein	32.5	9.2	27.3
WP_008657969.1	HAMP domain-containing histidine kinase	32.5	8.91E-08	53.5
WP_005794263.1	PAS domain-containing sensor histidine kinase	32.51	3.01E-28	118
WP_005789224.1	Septum formation initiator family protein	32.58	2.48E-10	52
WP_005803347.1	Aldo/keto reductase	32.61	0.89	30.4
WP_005802828.1	Acyl carrier protein	32.81	0.003	32.3
WP_032567884.1	Sodium:solute symporter	33.33	3.4	28.9
WP_005797263.1	Fragilysin family metalloproteinase II (MPII)	33.33	5.4	28.1
WP_005796883.1	Insulinase family protein	33.33	1.2	30.4
WP_005785270.1	S9 family peptidase	33.33	3.1	30
WP_005783851.1	Glycosyltransferase family 2 protein	33.33	9.74E-09	53.5
WP_032568565.1	SDR family oxidoreductase	33.47	1.01E-36	129
OCR42631.1	Multidrug transporter MatE	34.04	3.8	28.9
WP_009291748.1	Metallophosphoesterase	35	1	29.6
WP_005793843.1	Phage holin family protein	35	0.92	28.1
WP_069108608.1	Fimbrillin family protein	35.48	6.5	27.7

Table 3

Virulent proteins of *Bacteroides fragilis* against VFDB set B dataset with more than 30 % sequence similarity.

Bacterial Protein ID	Virulent Proteins Name	Identity %	E-value	Bit Score
WP_005789707	Acyl-CoA dehydrogenase C-terminal domain-containing protein	33.95	1.98E-43	163
WP_005813937	Gluconate 5-dehydrogenase	38.49	1.42E-49	163
WP_005797263	Fragilysin family metalloproteinase II (MPII)	30.63	4.59E-47	165
WP_005785797	Iron ABC transporter permease	36.16	7.92E-49	166
WP_005783787	Acyl-ACP–UDP-N-acetylglucosamine O-acyltransferase	37.89	1.23E-52	171
WP_009292952	Glycine betaine/L-proline ABC transporter ATP-binding protein	41.15	9.26E-50	171
WP_065849905	ATP-binding protein	42.15	6.28E-44	171
WP_053874191	Peptidase domain-containing ABC transporter	37.08	4.91E-46	173
WP_069108695	Response regulator transcription factor	37.84	7.04E-55	175
OCR39807	Peptide synthetase	34.44	3.37E-47	176
WP_005786511	Class I mannose-6-phosphate isomerase	33.84	3.02E-53	177
WP_005783851	Glycosyltransferase family 2 protein (267 a.a)	41.95	4.28E-52	177
WP_009291904	Glycosyltransferase family 2 protein (254 a.a)	42.86	3.11E-58	185
WP_005802507	TolC family protein	30.69	1.01E-53	186
WP_009292446	Phosphonopyruvate decarboxylase	34.73	5.17E-57	191
WP_011202702	Sigma-54 dependent transcriptional regulator	42.35	4.32E-55	191
WP_065849797	ATP-dependent Clp protease ATP-binding subunit	40.65	2.14E-51	194
WP_005788057	HesA/MoeB/ThiF family protein	46.96	2.24E-62	199
WP_005775379	Hsp 70 family protein	31.78	7.59E-58	203
WP_032575164	Glycosyltransferase family 2 protein (272 a.a)	40	9.11E-65	203
WP_005785118	Acyl-ACP–UDP-N-acetylglucosamine O-acyltransferase	40.08	6.28E-66	205
WP_122119233	Glycosyltransferase (271a.a)	41.13	9.06E-66	206
WP_005792012	3-oxoacyl-[acyl-carrier-protein] reductase	45.27	5.59E-68	209
WP_011202230	Pantoate–beta-alanine ligase	44.25	1.71E-67	212
WP_009293321	Sigma-54 dependent transcriptional regulator	44.26	2.10E-65	214
WP_005788554	Trypsin-like peptidase domain-containing protein	41.01	4.47E-65	215
OCR41021	Chemotaxis protein CheY	43.31	2.74E-69	224
WP_008769322	Amino acid permease	30.82	1.22E-67	225
OCR44682	Spermidine/putrescine ABC transporter ATP-binding protein	50.94	6.31E-70	226
WP_005783855	ABC transporter ATP-binding protein	50	1.74E-64	227
WP_005815453	Glutamine-hydrolyzing carbamoyl-phosphate synthase small subunit	36.01	4.41E-72	229
WP_069108703	Sigma-54 dependent transcriptional regulator	46.09	1.10E-70	231
WP_032568159	Nucleotidyltransferase family protein	41.69	4.08E-75	234
WP_009293195	Amino acid adenylation domain-containing protein	34.62	6.46E-65	237
WP_005785350	3-deoxy-8-phosphooctulonate synthase	52.44	5.13E-78	237
WP_005785154	Phosphoribosylaminoimidazolesuccinocarboxamide synthase	43.99	1.90E-74	243
WP_005787548	Aspartate carbamoyltransferase	48.19	5.35E-80	244
WP_005784372	Isoprenyl transferase	51.93	2.46E-81	244
WP_005787884	ATP-dependent Clp protease ATP-binding subunit	49.21	4.45E-69	246
WP_005787435	Sigma-54 dependent transcriptional regulator	42.27	9.12E-81	253
WP_005795491	Ferrous iron transport protein B	33.20	3.51E-73	254
OCR37850	Alginate O-acetyltransferase	41.6	2.66E-82	264
WP_065849951	DNA repair protein RecN	31.25	6.03E-83	269
WP_009291817	Heavy metal translocating P-type ATPase (648 a.a)	34.36	2.64E-90	296
OCR43114	Hypothetical protein AC239_11070	33.27	1.88E-95	300
WP_005781149	Glucose-1-phosphate cytidylyltransferase	58.69	2.37E-108	313
WP_005788150	UDP-glucose 4-epimerase GalE	49.25	1.14E-111	328
OCR37576	Nucleoside-diphosphate sugar epimerase	43.22	1.01E-111	333
WP_005790599	dTDP-glucose 4,6-dehydratase	52.96	6.11E-118	345
WP_009293271	Beta-N-acetylglucosaminidase domain-containing protein	40.72	1.65E-109	355
WP_065849657	Aminotransferase class I/II-fold pyridoxal phosphate-dependent enzyme	50.93	7.95E-128	372
WP_005790602	Glucose-1-phosphate thymidylyltransferase RfbA	62.76	8.60E-134	380
WP_065850067	CDP-glucose 4,6-dehydratase	52.96	8.09E-139	398
WP_008660312	NAD-dependent DNA ligase LigA	37.48	3.54E-135	412
WP_005796220	RelA/SpoT family protein	34.46	1.26E-134	416
WP_005786800	GDP-L-fucose synthase	58.59	2.42E-150	427
WP_005791918	UDP-N-acetyl-D-mannosamine dehydrogenase	55.70	6.76E-149	428
OCR43137	acriflavin resistance protein	31.64	6.33E-138	440
WP_008659262	UDP-glucose/GDP-mannose dehydrogenase family protein	51.36	1.87E-165	473
WP_005794816	Decarboxylating NADP (+)-dependent phosphogluconate dehydrogenase	50.11	4.14E-166	477
WP_005791919	UDP-N-acetylglucosamine 2-epimerase (non-hydrolyzing)	62.4	6.02E-170	478
WP_005786803	GDP-mannose 4,6-dehydratase	67.24	1.69E-173	486
OCR41890	peptidase M13	40.31	1.91E-165	489
WP_005787240	Heavy metal translocating P-type ATPase (736 a.a)	44.11	1.26E-166	498
WP_032529046	Lipopolysaccharide biosynthesis protein RfbH	56.43	2.19E-177	503
WP_009291901	Nucleotide sugar dehydrogenase	57.24	4.04E-179	506
WP_005785772	ATP-dependent chaperone ClpB	44.03	3.69E-168	509
WP_065849907	Cobaltochelatase subunit CobN	46.542	2.91E-165	532

Table 4

Top 10 *Bacteroides fragilis* unique pathways being performed in host/human body.

bfb code	Pathways unique to *Bacteroides fragilis*
00121	Secondary bile acid biosynthesis
00300	Lysine biosynthesis
00999	Biosynthesis of various plant secondary metabolites
01120	Microbial metabolism in diverse environments
01501	Beta-Lactam resistance
01502	Vancomycin resistance
01503	Cationic antimicrobial peptide (CAMP) resistance
02020	Two-component system
02040	Flagellar assembly
03070	Bacterial secretion system

Table 5

Bacteroides fragilis proteins with assigned KO numbers and their functional description.

Gene Name	KO Number	Description
WP_069108703.1	K02481	flgR; Two-component system, NtrC family, response regulator
WP_053874191.1	K06147	ABCB-BAC; ATP-binding cassette, subfamily B, bacterial
WP_009291817.1	K01534	zntA; Zn^{2+}/Cd^{2+}-exporting ATPase
WP_005787240.1	K01533	copB; P-type Cu^{2+} transporter
WP_005797263.1	K19959	bftP; Fragilysin
WP_005785772.1	K03695	clpB; ATP-dependent Clp protease ATP-binding subunit ClpB
WP_011202230.1	K01918	panC; Pantoate–beta-alanine ligase
WP_005787240.1	K01533	copB; P-type Cu^{2+} transporter

pathways were common with human thus were eliminated from the analysis. Unique pathways consist of mostly biosynthesis and degradation (for example, Monobactam biosynthesis, Carbapenem biosynthesis, Benzoate degradation, and others). Few pathways were involved in resistance system such as "beta-lactam resistance", "vancomycin resistance", and "cationic antimicrobial peptide (CAMP) resistance". In addition, *Bacteroides fragilis* is also uniquely involved in "two-component system", "quorum sensing", "Flagellar assembly", and "bacterial secretion system". 10 pathways that are unique to *Bacteroides fragilis* being performed in humans are mentioned in Table 4.

Out of 68, 56 virulent protein sequences were assigned KO numbers. Each protein was manually searched for the assigned KO and its pathways. The query protein of Enterotoxigenic *Bacteroides fragilis* (strain 2-078382-3) "Fragilysin family metalloproteinase II or MPII" (reference ID WP_005,797,263) was aligned against fragilysin of Enterotoxigenic *Bacteroides fragilis* strain available in KEGG database. MPII was assigned KO number *i.e.* K19959 (bftP; fragilysin). Reportedly, fragilysin is involved in causing the bacterial infection and inflammation [65], which results in bowel disease in humans while chronic infection can lead to cancers [45]. It is evident from the literature that blastp results of fragilysin and MPII have less than 25% sequence identity, while there is spatial structure similarity between fragilysin and MPII and therefore are involved in performing similar functions [46]. KO assigned virulent proteins with their function description are given in Table 5 and complete list is given in Table S3.

3.3. Localization of proteins, functional importance and target selection

Protein localization determines environments and also establishes and maintains protein at its specific location where protein can operate. Optimum functioning of a protein is closely associated to its subcellular locations thus can be targeted for drug discovery process. For this, many computational tools and methods have developed to predict subcellular locations of proteins [66]. CELLO, a multi-class SVM classification system, is a subCELlular LOcalization predictor for proteins. It uses various features of protein based on its amino acid sequence, such as amino acid composition, dipeptide composition, partitioned sequence composition, physiochemical properties and neighboring amino acids to train models that can then predict the subcellular localization of a given protein [67]. Majority of the 56 proteins were predicted to be localized in cytoplasm with significant reliable-index (*), while few of them were present in inner membrane or in periplasm. "Fragilysin family metalloproteinase II" (MPII, WP_005797263.1), a virulent factor of *Bacteroides fragilis* (strain 2-078382-3), was predicted to be localized in extracellular compartment which indicated it as suitable target for drug(s) [68]. In addition, it is reported that Enterotoxigenic *Bacteroides fragilis* toxin (*bft*) is involved in proteolysis and stimulating cleavage of E-cadherin present on colonic epithelial lining [69]. Moreover, binding of catalytic domain of MPII with E-cadherin [70] leads to increased cell permeability, damages colonic crypts and colonocytes surface [71], weakens cell to cell junction, and initiate disturbed immune pathways and cytokines production [96]. Detailed prediction, classification and reliability-index of seven proteins are given in Table 6.

Likewise, PSORTb (v3.0) is another tool to predict the subcellular localization of a protein based on its amino acid sequence. It uses a combination of algorithms and machine learning techniques [72]. Like CELLO, it considers various features of the protein sequence, including amino acid composition, signal peptides, transmembrane domains, and protein domains, to predict the subcellular localization of a protein to one or more subcellular compartments, such as the cytoplasm (CP), periplasm (PM), outer membrane (OM), inner

Table 6

Localization of *Bacteroides fragilis* proteins with their localization predicted by CELLO.

Protein Name	Predictor and Location, Reliable-Index					SVM classifier				
	Amino acid Composition	Dipeptide Composition	Partitioned sequence composition	Composition based on physiochemical properties	Neighboring amino acids	Extracellular Space (EC)	Outer Membrane (OM)	Periplasm (PP)	Inner Membrane (IM)	Cytoplasm (CP)
Sigma-54 dependent transcriptional regulator WP_069108703.1	CP 0.996	CP 0.997	CP 0.989	CP 0.903	CP 0.997	0.007	0.006	0.087	0.019	4.881*
DNA repair protein RecN WP_065849951.1	CP 0.98	CP 0.958	CP 0.961	CP 0.714	CP 0.993	0.045	0.132	0.156	0.067	4.6*
Peptidase domain-containing ABC transporter WP_053874191.1	OM 0.401	IM 0.635	IM 0.643	IM 0.377	IM 0.838	0.198	1.092	0.188	2.89*	0.693
Fragilysin family metalloproteinase MPII WP_005797263.1	EC 0.738	EC 0.865	OM 0.675	EC 0.55	EC 0.774	3.235*	1.114	0.388	0.034	0.23
Beta-N-acetylglucosaminidase domain-containing protein WP_009293271.1	PP 0.766	PP 0.402	PP 0.941	OM 0.326	PP 0.561	0.467	0.554	2.892*	0.395	0.692
Heavy metal translocating P-type ATPase WP_005787240.1	IM 0.934	IM 0.969	IM 0.986	IM 0.863	IM 0.984	0.027	0.057	0.049	4.735*	0.132
Phosphoribosyl aminoimidazole succinocarboxamide synthase WP_005785154.1	CP 0.98	CP 0.919	CP 0.921	CP 0.963	CP 0.986	0.02	0.024	0.136	0.052	4.786*

(Note: cytoplasm (CP), periplasm (PM), outer membrane (OM), inner membrane (IM), and extracellular space (EC). Proteins with (*) represent significant scores with their respective location).

Table 7
Localization of *Bacteroides fragilis* proteins with their localization predicted by PSORTb.

Protein Name	Localization				
	Extracellular Space (EC)	Outer Membrane (OM)	Periplasm (PP)	Cytoplasmic Membrane (CM)	Cytoplasm (CP)
Sigma-54 dependent transcriptional regulator WP_069108703.1	0.0	0.0	0.01	0.01	9.97*
DNA repair protein RecN WP_065849951.1	0.0	0.0	0.01	0.01	9.97*
Peptidase domain-containing ABC transporter WP_053874191.1	0.0	0.0	0.0	10.0*	0.0
Fragilysin family metalloproteinase MPII WP_005797263.1	9.72*	0.07	0.20	0.01	0.0
Beta-N-acetylglucosaminidase domain-containing protein WP_009293271.1	2.50	2.50	2.50	2.50	0.0
Heavy metal translocating P-type ATPase WP_005787240.1	0.0	0.0	0.0	10.0*	0.0
Phosphoribosyl aminoimidazole succinocarboxamide synthase, WP_005785154.1	0.0	0.0	0.01	0.01	9.97*

(Note: cytoplasm (CP), periplasm (PM), outer membrane (OM), inner membrane (IM), and extracellular space (EC). Proteins with (*) represent significant scores with their respective location).

membrane (IM), and extracellular space (EC). Table 7 displays the predicted localization with their respective scores when protein sequences of Enterotoxigenic gram negative *Bacteroides fragilis* were given to the PSORTb. Notably, fragilysin family metalloproteinase II (MPII) is localized in the extracellular domain with the significant score (mentioned with *).

Therefore, by considering MPII localization (*i.e.* extracellular), we selected MPII toxin or virulent protein as a potential target for further analysis and was used for virtual screening against small organic compounds from ZINC database.

3.4. Protein structure

In order to fix the missing residues present in MPII protein structure, self-template method was used to construct the missing region. The resultant protein structure was energy minimized by the steepest descent method using GROMOS96 force field [107] followed by structural assessment in order to evaluate its stability. Analysis revealed that 89.4 % of the residues fell in the most favorable region, 10.0 % in additional allowed region, 0.6 % in generously allowed region, and none of the residues fell in disallowed regions and is shown in Fig. S1 (supplementary material). Therefore, Ramachandran plot indicates that the modeled structure possess good quality and can be used for further analysis [106].

3.5. Virtual screening analysis and optimization of ligand-MPII interaction

A large dataset, containing 0.2 million ZINC molecules, was screened against the selected target virulent protein, *i.e.* MPII. The hit molecules were sorted according to their binding energies to extract complexes with binding energy ≤ −8 kcal/mol. A total of 249 compounds were extracted and subjected to ADMET properties analysis to identify the suitable candidates for protein-ligand inter-action analysis. Table 8 shows top 20 identified candidates with their respective least binding free energies forming stable complexes with MPII. ZINC compounds were shortlisted after molecular docking with MPII.

3.6. ADMET modulation approaches for drug-like-molecules in drug discovery

In drug discovery process, the calculation of ADMET properties is essential to estimate the efficacy and efficiency of drug-like molecules. In this study, SwissADME was used to shortlist a total of 249 unique compounds to predict their physiochemical proper-ties. Compounds observing the Lipinski's rules of five were sorted while the rest were discarded. According to the rule, compounds with molecular weight (MW) ≤ 500 g/mol, atom-based water/octanol partition coefficient (AlogP) ≤ 5, hydrogen bonds acceptors (HBA) and donors (HBD) ≤ 10 and ≤ 5 respectively are biologically and orally active. If a compound violates two or more rules, it is classified as orally inactive [58]. SwissADME predicted molecular weight, number of heavy atoms, aromatic heavy atoms, fraction Csp3, and number of rotatable bonds, hydrogen bond acceptors and donors, molecular refractivity, and topological polar surface area or TPSA for all molecules. According to the rule, for good permeation, TPSA should be less than 140 Å^2, and fraction Csp3 should be greater than 0.25. Out of 249 molecules, 10 compounds met the criteria and were selected for further analysis. Ten selected compounds are shown in Table 9 with their physiochemical properties.

3.6.1. Drug-likeliness

Drug likeliness is one of the important properties while assessing the drug-like properties in compounds as it refers to the druggability of compounds. Rules that are considered important in druggability assessment are shown in Table 10 for each molecule
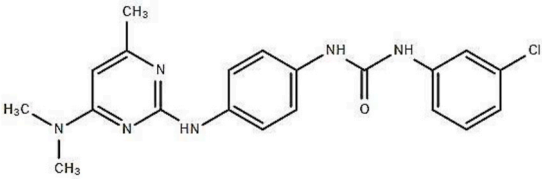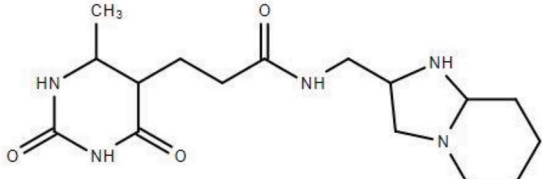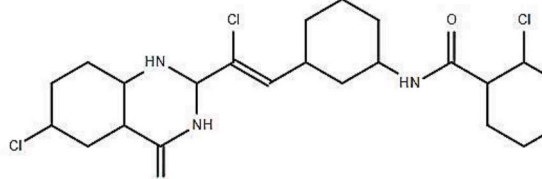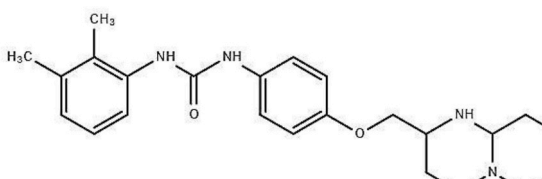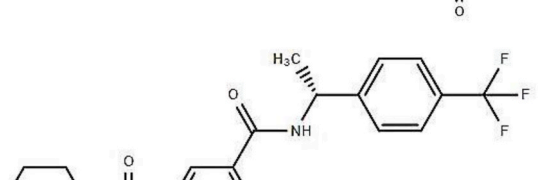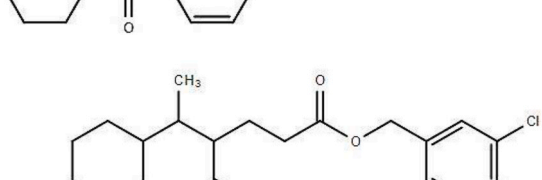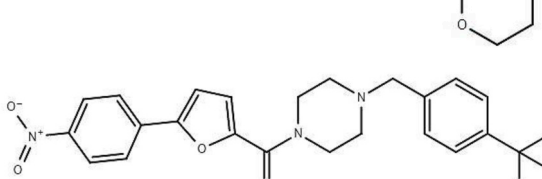
Table 8
Top best hit compounds with least binding energies of top five conformations, two dimensional structures along with their following docking with MPII.

No.	Zinc ID	2D Structure	Smiles	Energies of top five conformations (kcal/mol)
1.	ZINC000023353212		O=C(CCCN1C(=O) c2cccc3cccc (c23)C1=O) Nc1cnc2ccccc2c1	−12 −12 −9.6 −9.5 −9.5
2.	ZINC000017042744		Cc1ccc (Nc2cc(C)nc (Nc3ccc (NC(=O) CC4CCCC4)cc3)n2)cc1	−10.8 −10.3 −10.2 −9.9 −9.7
3.	ZINC000033068037		Cc1ccn2c (=O)cc (COc3ccc (NC(=O)Nc4cc(F)ccc4F) cc3)nc2c1	−10.8 −10.4 −10.1 −9.8 −9.7
4.	ZINC000017013121		Cc1cc (N2CCOCC2)nc (Nc2ccc (NC(=O)Nc3ccc (Cl)cc3)cc2)n1	−11.3 −10.2 −9.8 −9.8 −9.7
5.	ZINC000021797248		CCCn1cnc2nc (COc3cc (NC (=O)COc4ccc(C)c(C)c4) ccc3C)cc (=O)n21	−9.9 −9.6 −9.4 −9.4 −9.3
6.	ZINC000030479704		O=C(CCCN1C(=O)NC2 (CCCC2)C1=O)Nc1cccc (C (=O)N2CCCC2)c1	−10.2 −9.6 −9.3 −9.2 −9.2
7.	ZINC000016722540		Cc1ccc (-c2cc (C (=O)O [C@H](C)C (=O)NC(=O) NC3CCCC3)c3ccccc3n2)o1	−9.9 −9.4 −9.4 −9.4 −9.3

Table 8 (*continued*)

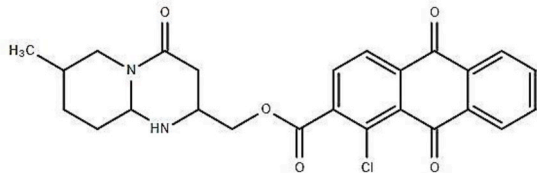| No.	Zinc ID	2D Structure	Smiles	Energies of top five conformations (kcal/mol)
8.	ZINC000017013141		Cc1cc (N(C)C)nc (Nc2ccc (NC(=O)Nc3cccc (Cl)c3) cc2)n1	−9.9 −9.4 −9.3 −9.2 −9.2
9.	ZINC000023754543		Cc1 [nH]c (=O)[nH]c (=O) c1CCC(=O) NCc1cn2ccccc2n1	−9.9 −9.4 −9.4 −9 −8.8
10.	ZINC000018913701		O=C(Nc1cccc (/C=C (\Cl) c2nc3ccc (Cl)cc3c (=O) [nH]2)c1)c1ccccc1Cl	−10.7 −10.7 −10.1 −9.5 −9.5
11.	ZINC000033067772		Cc1cccc (NC(=O)Nc2ccc (OCc3cc (=O)n4ccccc4n3) cc2)c1C	−10.7 −10.2 −9.8 −9.6 −9.5
12.	ZINC000029239937		C [C@@H](NC(=O)c1cccc (S (=O) (=O)N2CCCCC2) c1)c1ccc (C(F) (F)F)cc1	−9.9 −9.2 −9.2 −9.1 −9
13.	ZINC000032836915		Cc1ccc2c(C)c (CCC(=O) OCc3cc (Cl)c4c (c3) OCCO4)c (=O)oc2c1	−10.7 −9.3 −9 −8.7 −8.7
14.	ZINC000015767717		O=C (c1ccc (-c2ccc ([N+] (=O)[O-])cc2)o1)N1CCN (Cc2ccc (C(F) (F)F)cc2)CC1	−9.4 −9.3 −9.0 −9.0 −8.9

Table 8 (*continued*)
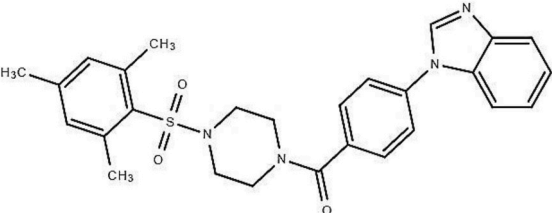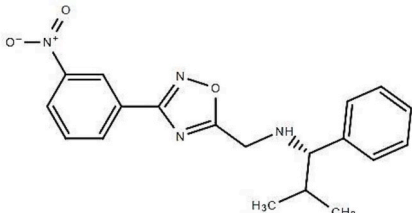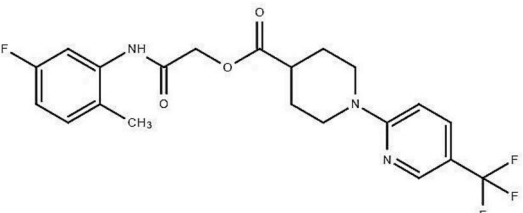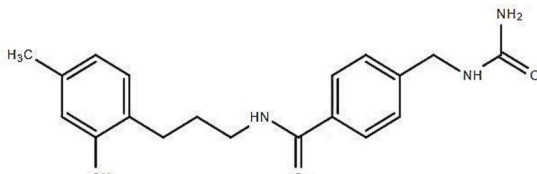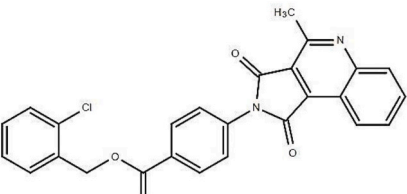
No.	Zinc ID	2D Structure	Smiles	Energies of top five conformations (kcal/mol)
15.	ZINC000032736238		Cc1cc(C)c (S (=O) (=O) N2CCN(C (=O)c3ccc (-n4cnc5ccccc54)cc3)CC2) c(C)c1	−9.2 −9.1 −9.1 −9.0 −9.0
16.	ZINC000025928534		CC(C)[C@H](NCc1nc (-c2cccc ([N+](=O)[O-]) c2)no1)c1ccccc1	−9.9 −9.3 −9.1 −8.8 −8.2
17.	ZINC000021942081		Cc1ccc2nc (COC(=O) c3ccc4c (c3Cl)C (=O) c3ccccc3C4=O)cc (=O) n2c1	−11.6 −9.8 −9.3 −9.3 −9.3
18.	ZINC000023804957		Cc1ccc(F)cc1NC(=O)COC (=O)C1CCN(c2ccc (C(F) (F)F)cn2)CC1	−10.6 −9.4 −8.9 −8.1 −8.1
19.	ZINC000024389255		Cc1ccc (CCCNC(=O)c2ccc (CNC(N)=O)cc2)c(C)c1	−10.5 −9.6 −9.3 −7.9 −7.8
20.	ZINC000016266116		Cc1nc2ccccc2c1C (=O)N (c1ccc (C (=O) OCc3ccccc3Cl)cc1)C2=O	−10.3 −10.2 −9.8 −9.2 −9.2

Table 9
Physiochemical properties of the selected ten ZINC compounds.

No.	Ligand ZINC ID	Molecular Formula	Mol. Wt. (g/mol)	Heavy Atoms	Arom. Heavy Atoms	Fraction Csp3	Rotatable Bonds	H-bond Acceptor	H-Donor	Molar Refractivity	TPSA (Å^2)
1.	ZINC000021797248	$C_{26}H_{29}N_5O_4$	475.54	35	21	0.31	10	6	1	134.33	99.75
2.	ZINC000016722540	$C_{24}H_{25}N_3O_5$	435.47	32	15	0.33	9	6	2	118.42	110.53
3.	ZINC000023754543	$C_{16}H_{17}N_5O_3$	327.34	24	15	0.25	6	4	3	88.28	112.21
4.	ZINC000029239937	$C_{21}H_{23}F_3N_2O_3S$	440.48	30	12	0.38	7	7	1	110.75	74.86
5.	ZINC000032836915	$C_{23}H_{21}ClO_6$	482.86	30	16	0.3	6	6	0	113.45	74.97
6.	ZINC000015767717	$C_{23}H_{20}F_3N_3O_4$	459.42	33	17	0.26	7	8	0	123.39	82.51
7.	ZINC000032736238	$C_{27}H_{28}N_4O_3S$	488.6	35	21	0.26	5	5	0	144.71	83.89
8.	ZINC000025928534	$C_{19}H_{20}N_4O_3$	352.39	26	17	0.26	7	6	1	100.04	96.77
9.	ZINC000023804957	$C_{21}H_{21}F_4N_3O_3$	439.4	31	12	0.38	8	8	1	108.34	71.53
10.	ZINC000024389255	$C_{20}H_{25}N_3O_2$	339.43	25	12	0.3	9	2	3	99.34	84.22

Table 10
Druggability assessments, number of violations and bioavailability score of selected compounds to be a good inhibitor.

No.	Ligand ZINC ID	Lipinski	Ghose	Veber	Egan	Muegge	Bioavailability score
1.	ZINC000021797248	0	1	0	0	0	0.55
2.	ZINC000016722540	0	0	0	0	0	0.55
3.	ZINC000023754543	0	0	0	0	0	0.55
4.	ZINC000029239937	0	1	0	1	0	0.55
5.	ZINC000032836915	0	0	0	0	0	0.55
6.	ZINC000015767717	0	0	0	0	0	0.55
7.	ZINC000032736238	0	2	0	0	0	0.55
8.	ZINC000025928534	0	0	0	0	0	0.55
9.	ZINC000023804957	0	0	0	0	0	0.55
10.	ZINC000024389255	0	0	0	0	0	0.55

with their predicted properties. All compounds showed minimal violations, thereby, could serve as potential candidates for further analysis. For Lipinski, MW ≤ 500, MLOGP ≤4.15, N or O ≤ 10, and NH or OH ≤ 5 [73], no compound mentioned below violated this rule. According to Ghose filter/rule, MW should be between 160 and 480, WLOGP range from −0.4 and 5.6, molar refractivity (MR) should be between 40 and 130 while atoms range is 20–70 [74]. Here, ZINC000032736238 showed 2 violations of Ghose rule (MW > 480 and MR > 130). Compounds should have rotatable bonds ≤10 and TPSA value ≤ 140 to follow veber rule [75] and all compounds are following this rule. In Egan rule/filter, WLOGP ≤5.88 and TPSA ≤131.6 [76] and no compound violated this rule. According to Muegge (Bayer) filter/rule, MW is between 200 and 600, XLOGP is between −2 and 5, TPSA ≤150, Num. rings <7, num. carbons >4, num. heteroatoms >1, num. rotatable bonds ≤15, H-bonds acceptors ≤10 and H-bond donors ≤5 [77]. All compounds also met these criteria. Bioavailability score (also known as Abbott bioavailability score) is the prediction of probability of compounds to contain at-least 10 % of oral bioavailability in rats or have measureable permeability of Caco-2 of compounds which should be greater than 10 % (F> 10 %). This is based on pre-dominant charges at biological pH in rat model [78]. According to Abbott rule, if a compound's bioavailability score is 0.55, this indicates that a compound passes the Lipinski rule of 5 [79]. Here, all compounds showed good bioavailability score which increase their chances to be a good inhibitor/drug candidate.

3.6.2. Lipophilicity

Lipophilicity is defined in terms of being dissolved in oils, fats or non-polar solvents and for this LogP is calculated to identify the co-efficient of partition between water and octanol. Higher value of consensus (ClogP) is important in order to determine the drug-like properties. Therefore, ClogP was calculated for compounds as shown in Table 11.

3.6.3. Water solubility

Solubility is also one of the important factors for a potential drug. Compounds with lower water solubility exhibit poor absorption and bioactivity with less oral bioavailability. Three different ways were used to determine water solubility of each molecule which includes Log S (ESOL), Log S (Ali), and Log S (SILICOS-IT) [60]. Log S (ESOL: Estimated SOLubility) is a simple method to estimate aqueous solubility of compounds from there structures. In ESOL, most important parameters which include $logP_{octanol}$, molecular weight, number of heavy atoms and rotatable bonds are calculated [80]. In Log S (Ali), general solubility equation (or GSE)-a QSPR model, is used to calculate topographical polar surface area (TPSA), melting point and log P value of non-ionizable chemical compounds [81]. Log S (SILICOS-IT) is third identifier to calculate solubility and was evolved by SILICOS-IT. This model uses molecular weight (*i.e.* $R^2 = 0.75$) of chemical compounds [60]. For each solubility class, there is a different way of solubility calculation which results in different values of class. There are certain parameters according to which these are classified *i.e.* insolubility < −10, poor solubility < −6, moderate solubility < −4, solubility < −2, very soluble <0, highly soluble >0 [82]. Results are given in Table 12. Literature shows that low/poor water solubility is an increasingly common property or characteristics of the lead compounds in the drug development process. One of the reasons is that most of the compounds in large chemical libraries are often screened in non-aqueous media. Moreover, these compounds are not analyzed when interacted with the receptor/target present in aqueous

Table 11
Determination of drug-like properties and lipophilicity along with ClogP analysis of Zinc molecules.

No.	Ligand ZINC ID	log P (o/w) (iLOGP)	log P (o/w) (XLOGP3)	Log P (o/w) (WLOGP)	Log P (o/w) (MLOGP)	Log P (o/w) (SILICOS-IT)	Consensus Log Po/w
1.	ZINC000021797248	4.1	3.64	3.48	3.08	3.8	3.62
2.	ZINC000016722540	2.91	4.38	4.12	1.97	3.65	3.4
3.	ZINC000023754543	2.14	0.29	0.12	0.3	2.33	1.03
4.	ZINC000029239937	3.62	3.97	5.9	3.25	3.67	4.08
5.	ZINC000032836915	3.99	4.12	4.36	3.31	6.35	4.42
6.	ZINC000015767717	3.54	4.46	5.07	2.42	2.65	3.63
7.	ZINC000032736238	3.65	4.27	4.42	3.15	3.29	3.75
8	ZINC000025928534	3.29	3.89	3.66	1.84	1.92	2.92
9.	ZINC000023804957	3.24	3.91	4.95	3.29	4.07	3.89
10.	ZINC000024389255	2.85	2.78	2.68	3.03	3.58	2.98

Table 12
Solubility classes and parameters for selected ten Zinc compounds.

No.	Ligand ZINC ID	Log S (ESOL)	Class	Log S (Ali)	Class
1.	ZINC000021797248	−4.87	Moderately soluble	−5.42	Moderately soluble
2.	ZINC000016722540	−5.05	Moderately soluble	−6.42	Poorly Soluble
'3.	ZINC000023754543	−2.12	Soluble	−2.21	Soluble
4.	ZINC000029239937	−4.91	Moderately soluble	−5.24	Moderately soluble
5.	ZINC000032836915	−5.09	Moderately soluble	−5.4	Moderately soluble
6.	ZINC000015767717	−5.42	Moderately soluble	−5.91	Moderately soluble
7.	ZINC000032736238	−5.67	Moderately soluble	−5.74	Moderately soluble
8.	ZINC000025928534	−4.5	Moderately soluble	−5.62	Moderately soluble
9.	ZINC000023804957	−4.79	Moderately soluble	−5.11	Moderately soluble
10.	ZINC000024389255	−3.46	Soluble	−4.2	Moderately soluble

(Note: Certain parameters according to which these are classified *i.e.* insolubility < −10, poor solubility < −6, moderate solubility < −4, solubility < −2, very soluble <0, highly soluble >0).

environment. The drug targets are somehow associated with intra-cellular signaling pathways, and lipid processing for poor soluble drug compounds [83]. ~40 % of the new chemical drugs/entities manufactured in pharmaceutical companies are insoluble or poorly soluble. To address this challenge, various strategies have been introduced to improve the solubility, for example, nanosuspension, micronization (physical modification techniques), pH change, buffer introduction, complexation, salt formation (chemical modification techniques), use of surfactants, solubilizers, hydrotrophy etc. (miscellaneous methods) [84]. However, besides improving the physiochemical properties, best property to be considered for a good drug candidate is Lipinski rule of 5 and others. Therefore, moderate solubility can be focused in the solubility classes [83]. Thus LogS (ESOL) and LogS (Ali) were considered for all the compounds solubility.

3.6.4. Pharmacokinetics

Pharmacokinetics deals with the kinetics of drug within the body and focuses on the ADME properties that include absorption, distribution, metabolism and excretion. The most desirable properties of drug-like molecules include their higher gastrointestinal (GI) absorption and lower permeability through blood-brain barrier (BBB). Considerably, P-glycoprotein (P-gp) substrate is also one of the important properties for drug design, as it is involved in actively transporting compounds outside of the cell with the help of ATP hydrolysis. Thus maintain balance and equilibrium of by-products inside the cells. In addition, P-gp is also involved in multi-drug resistance as drugs or their products are pumped out and concentrations of drugs become less at target site. This leads to failure of drugs chemotherapeutically in cancers treatments [85]. Taking this phenomenon into account to avoid multi-drug resistance, which is crucial with respect to bacteria, all selected compounds are not substrates of P-gp. In addition to P-gp, no molecule may cross BBB and the Log Kp value should be less than −2.5 cm/s to be permeable through skin. Detailed analysis of pharmacokinetics for all compounds is given in Table 13.

3.6.5. Medicinal chemistry

Medicinal chemistry addresses the challenges faced during the process of drug discovery. A good drug candidate exhibits no or less violations of all parameters of medicinal chemistry. Pan-assay interference compounds (PAINS) and Brenk violations were examined in order to identify if the molecule exhibits sub-structure or a fragment leading to false positives in the drug discovery process. PAINS filter was used to identify compounds if it is a response or they are in the favor of biological assays. Whereas, Brenk filter is used to identify the compounds that exhibit accepted toxic levels, chemicals reactivity and metabolic instability [86]. Unlike drug-likeness, which identifies molecules/fragments similar to those in known drugs, lead-likeness refers to the physicochemical properties of a molecule and structural similarity that resemble with known compounds as a starting point for drug development process [87]. Moreover, synthetic accessibility is a method of estimation of the ease of synthesis of drug like compounds in the process of drug

Table 13
Detailed pharmacokinetic analyses of ten selected ZINC compounds.

No.	Ligand ZINC ID	GI absorption	BBB permeation	P-gp substrate	CYP1A2 inhibitor	CYP2C19 inhibitor	CYP2C9 inhibitor	CYP2D6 inhibitor	CYP3A4 inhibitor	Log Kp (cm/s)
1.	ZINC000021797248	High	No	No	No	Yes	Yes	Yes	Yes	−6.62
2.	ZINC000016722540	High	No	No	No	Yes	Yes	No	Yes	−5.85
3.	ZINC000023754543	High	No	No	No	No	No	No	No	−8.09
4.	ZINC000029239937	High	No	No	No	Yes	Yes	Yes	Yes	−6.17
5.	ZINC000032836915	High	No	No	Yes	Yes	Yes	No	Yes	−5.99
6.	ZINC000015767717	High	No	No	No	Yes	Yes	Yes	Yes	−5.94
7.	ZINC000032736238	High	No	No	No	Yes	Yes	No	Yes	−6.25
8.	ZINC000025928534	High	No	No	Yes	Yes	Yes	No	Yes	−5.69
9.	ZINC000023804957	High	No	No	No	Yes	Yes	Yes	Yes	−6.2
10.	ZINC000024389255	High	No	No	No	Yes	No	Yes	Yes	−6.4

(Note: Gastrointestinal (GI) absorption, Blood Brain Barrier (BBB), P-glycoprotein (P-gp), and log Kp < −2.5 cm/s).

Table 14

PAINS assay, Brenk rule, leadlikeliness and synthetic ease scoring for drug developmental process.

No.	Ligand ZINC ID	PAINS	Brenk	Lead-likeness	Synthetic accessibility
1.	ZINC000021797248	0	0	3	3.74
2.	ZINC000016722540	0	0	3	4.12
3.	ZINC000023754543	0	0	0	2.7
4.	ZINC000029239937	0	0	2	3.35
5.	ZINC000032836915	0	1	2	3.78
6.	ZINC000015767717	0	2	2	3.44
7.	ZINC000032736238	0	0	2	3.41
8.	ZINC000025928534	0	2	2	3.64
9.	ZINC000023804957	0	0	3	3.5
10.	ZINC000024389255	0	0	1	2.12

discovery. Synthetic accessibility characterizes the ease of compounds synthesis in the form of score (also known as synthetic accessibility score or SA score). This score ranges from 1 (easy to synthesize) to 10 (very difficult to synthesize) [88]. Here, all 10 molecules did not show any alert for PAINS, while only 3 alerts were observed in Brenk. There are several lead-likeness alerts (which are MW ≥ 350 and XLOGP3>3.5). These alerts are acceptable because they are following the rule of Lipinski. Furthermore, all compounds shows the very good synthetic score as these are less than 6 and are easy to synthesize [88]. Table 14 describes violations of PAINS, Brenk, lead-likeness and synthetic accessibility score for all selected compounds.

3.7. Bioavailability graphs

In addition to ADMET calculation of ten selected ZINC Compounds, bioavailability graphs were also analyzed. There are six important properties shown in the graph *i.e.* lipophilicity, flexibility, insaturation, insolubility, polarization and size of compounds. Pink outline indicates that Csp3 is equal to 0.25 and inside pink outline is area with Csp3 more than 0.25. Graphs show that selected ZINC compounds are present within the Csp3 area. Bioavailability radar graphs of the selected compounds are given in Table 15. IUPAC names and ZINC IDs are also mentioned with each respective graph.

3.8. Characterization of protein-ligand interactions

All the selected compounds were subjected to detailed protein-ligand interaction analysis. Various ligand-protein interactions were studied in detail, for example, hydrogen bonds, *van der* Waal's, arene-arene, π-cation etc. In Fig. 2, all ligands are shown bound within the binding pocket(s) of metalloproteinase II (MPII).

3.8.1. Interaction of ZINC000021797248 with MPII

The IUPAC name of ZINC000021797248 is "2-(3,4-dimethylphenoxy)-N-{4-methyl-3-[(7-oxo-1-propyl-1,7-dihydro[1,2,4]triazolo [1,5-a]pyrimidin-5-yl)methoxy]phenyl}acetamide", which meets the criteria for ADMET properties for a good drug candidate as well as exhibits the lowest binding energies with MPII. In Fig. 3(A), the docked ligand inside the pocket of MPII is shown and Fig. 3(B) displays the two dimensional interaction of MPII with ligand. The triazolopyrimidine moiety of ZINC000021797248 forms hydrogen bond with His 345 and His 339 with a bond length of 3.39 Å and 3.19 Å, respectively. Lys 186 exhibits π-cation interaction with the phenyl group. Leu 362, Pro 340, Ile 175, Pro 369, Ser 368, Glu 346, Gly 310, Asp 187, Thr 308, and Phe 180 interact through *van der* Waals' interaction with ligand.

3.8.2. Interaction of ZINC000016722540 with MPII

IUPAC nomenclature of ZINC000016722540 is "[(2R)-1-(cyclopentylcarbamoylamino)-1-oxopropan-2-yl]-2-(5-methylfuran-2-yl)-quinoline-4-carboxylate" and this compound meets the criteria for all ADMET and physiochemical properties of a being a potential candidate for drug. Binding of ZINC000016722540 with MPII have the lowest binding energies. Thr 308 and Gly 310 are the MPII residues which form separate hydrogen bonds with furan moiety and ester's carbonyl group of the ligand exhibiting a bond length of 5.04 and 3.81 Å, respectively. In addition, several residues in the binding pocket of MPII display *van der* Waals' interaction with the ligand. Interactions and residues are shown in Fig. 3(C and D).
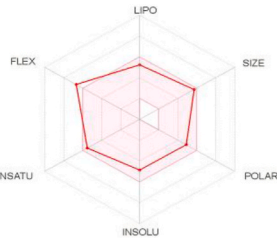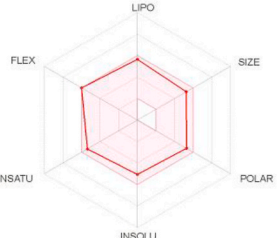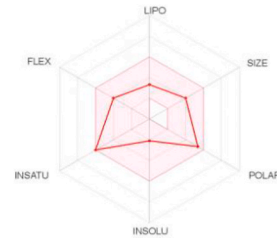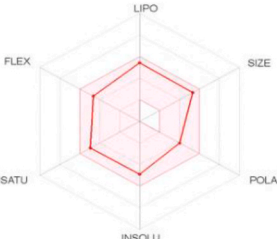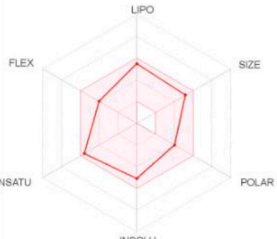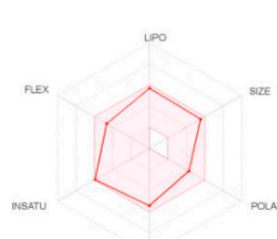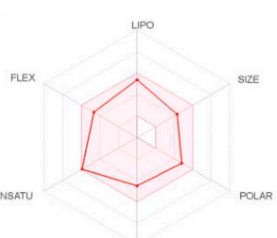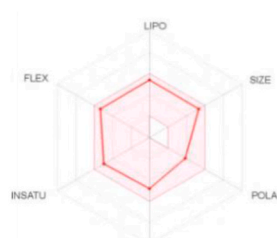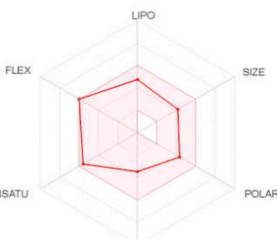
3.8.3. Interaction of ZINC000023754543 with MPII

In ZINC000023754543 (IUPAC nomenclature: N-({imidazo[1,2-a]pyridin-2-yl}methyl)-3-(6-methyl-2,4-dioxo-1,2,3,4-tetrahy-dropyrimidin-5-yl)propanamide), three residues of MPII namely, Gly 310, Thr 342, and Lys 186 form separate hydrogen bonds with amide carbonyl, imidazole's nitrogen and pyrimidine moiety of the ligand presenting bond lengths of 3.20, 2.52 and 3.33 Å, respectively. Interactions and residues are shown in Fig. 3(E and F).

3.8.4. Interaction of ZINC000029239937 with MPII

The sulfonyl and amide groups of ZINC000029239937 (IUPAC nomenclature: 3-piperidin-1-ylsulfonyl-N-[(1R)-1-[4-(trifluoro-methyl)phenyl]ethyl]benzamide) interact with Tyr 367 and Gly 310 of MPII forming hydrogen bonds with bond length of 5.59 and

Table 15
Bioavailability Radar graphs of Compounds.

| 2-(3,4-dimethylphenoxy)-N-{4-methyl-3-[(7-oxo-1-propyl-1,7-dihydro[1,2,4]triazolo[1,5-a]pyrimidin-5-yl)methoxy]phenyl}acetamide **ZINC000021797248**	[(2R)-1-(cyclopentylcarbamoylamino)-1-oxopropan-2-yl] 2-(5-methylfuran-2-yl)quinoline-4-carboxylate **ZINC000016722540**	N-({imidazo[1,2-a]pyridin-2-yl}methyl)-3-(6-methyl-2,4-dioxo-1,2,3,4-tetrahydropyrimidin-5-yl)propanamide **ZINC000023754543**
3-piperidin-1-ylsulfonyl-N-[(1R)-1-[4-(trifluoromethyl)phenyl]ethyl]benzamide **ZINC000029239937**	(5-Chloro-2,3-dihydro-1,4-benzodioxin-7-yl)methyl 3-(4,7-dimethyl-2-oxochromen-3-yl)propanoate **ZINC000032836915**	[5-(4-Nitrophenyl)furan-2-yl]-[4-[[4-(trifluoromethyl)phenyl]methyl]piperazin-1-yl]methanone **ZINC000015767717**
[4-(Benzimidazol-1-yl)phenyl]-[4-(2,4,6-trimethylphenyl)sulfonylpiperazin-1-yl]methanone **ZINC000032736238**	(1S)-2-methyl-N-[[3-(3-nitrophenyl)-1,2,4-oxadiazol-5-yl]methyl]-1-phenylpropan-1-amine **ZINC000025928534)**	[(5-Fluoro-2-methylphenyl)carbamoyl]methyl 1-[5-(trifluoromethyl)pyridin-2-yl]piperidine-4-carboxylate **ZINC000023804957**
4-[(carbamoylamino)methyl]-N-[3-(2,4-dimethylphenyl)propyl]benzamide **ZINC000024389255**		
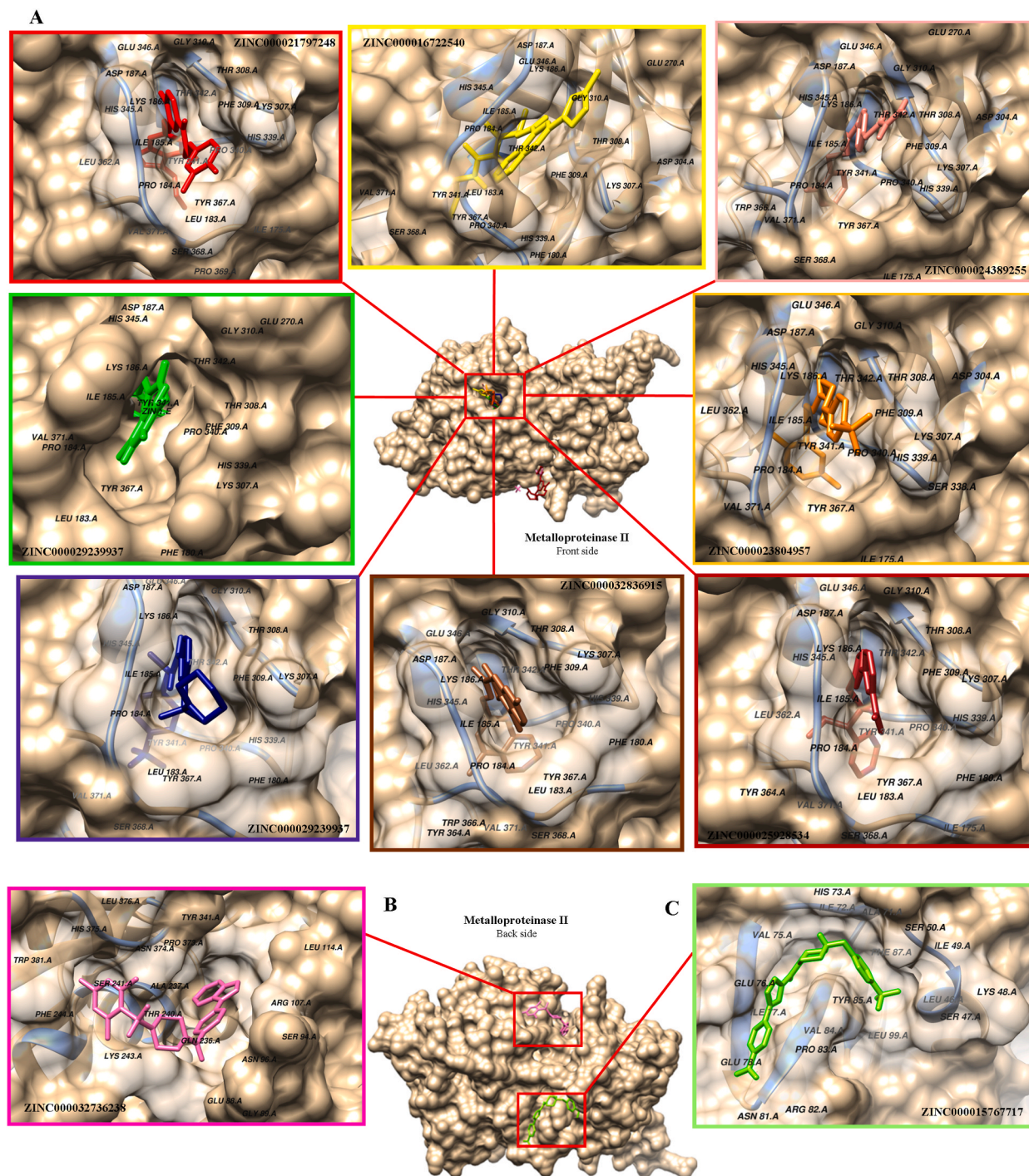
Fig. 2. Three-dimensional representation of metalloproteinase II docked structure with selected ten ZINC compounds and the interacting residues. (A) Eight compounds bound with binding pocket of MPII. ZINC000021797248 is represented in red, ZINC000016722540 in yellow, ZINC000023754543 in green, ZINC000029239937 in dark blue, ZINC000032836915 in brown, ZINC000025928534 in maroon, ZINC000023804957 in orange and ZINC000024389255 in coral pink color. **(B)** ZINC000032736238 in pink and **(C)** ZINC000015767717 in light green. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)
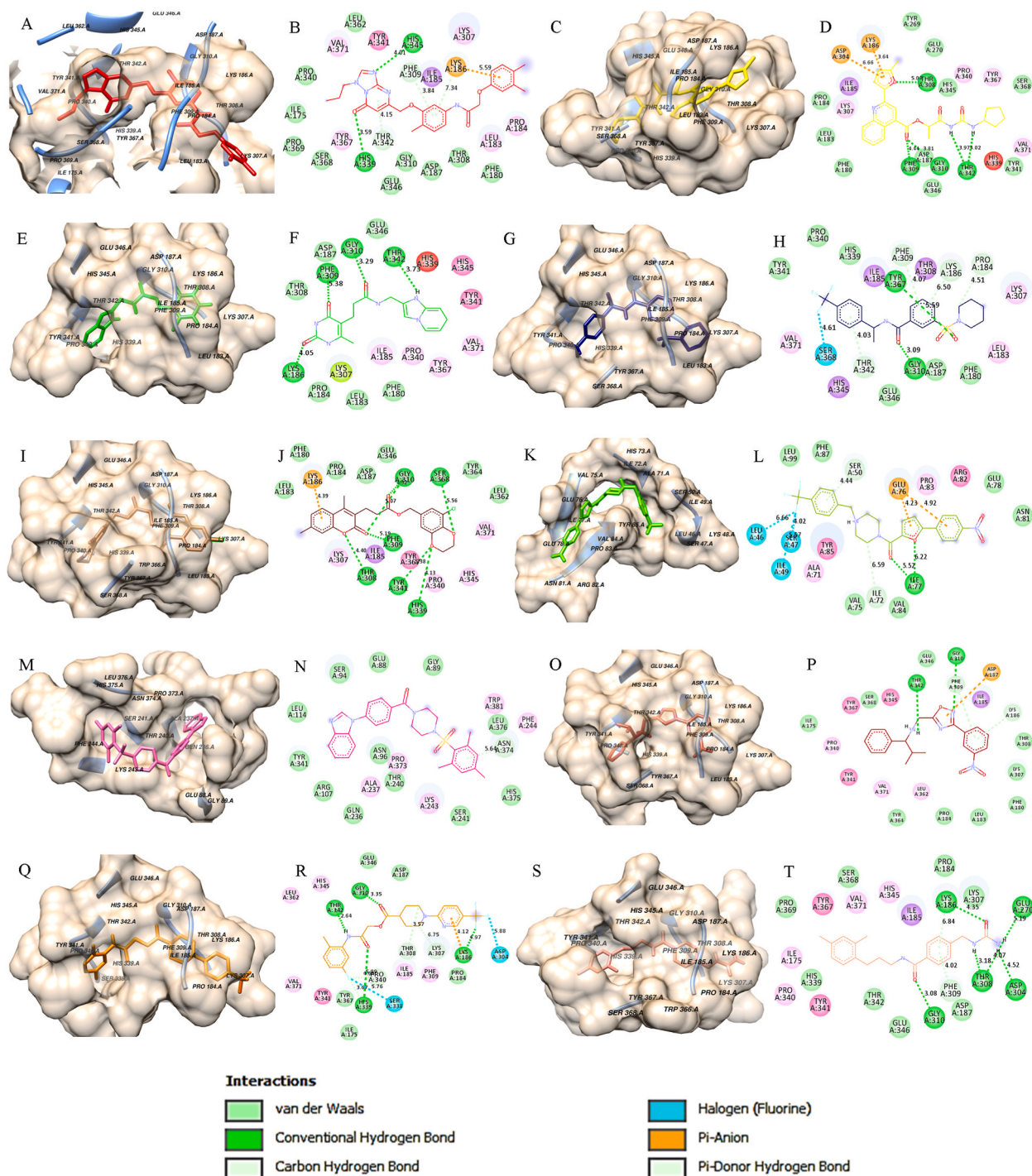
Fig. 3. Three-dimensional representation of molecular docking analysis between metalloproteinase II (MPII) and ZINC compound, involved interacting residues, and two-dimensional representation of interactions and bonds. (A–B) Binding mode of ZINC000021797248 in red, **(C–D)** ZINC000016722540 in yellow, **(E–F)** ZINC000023754543 in green, **(G–H)** ZINC000029239937 in dark blue, **(I–J)** ZINC000032836915 in brown, **(K–L)** ZINC000015767717 in light green, **(M–N)** ZINC000032736238 in pink, **(O–P)** ZINC000025928534 in maroon, **(Q–R)** ZINC000023804957 in orange, and **(S–T)** ZINC000024389255 is represented in coral pink. MPII residues are involved in hydrogen bonding and *van der* Waals, π-cationic, π-anionic interactions where hydrogen bonds are displayed in green, π-cationic and π-anionic in orange, halogen interaction in cyan, and *van der* Waals is displayed in light green color. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

2.90 Å, respectively. Moreover, fluorine present in the ligand further interacts with Ser 368 with a bond length 3.21 Å. Pro 184 shares a unique carbon hydrogen bond with piperidine moiety of ligand, while Lys 186, Phe 309, Thr 342 develop π-donor hydrogen bond with aromatic rings. There are also several residues of the protein present in the binding pocket, which have *van der* Waals' interaction with the ligand as shown in Fig. 3(G and H).

3.8.5. Interaction of ZINC000032836915 with MPII

ZINC000032836915 (IUPAC name: (5-Chloro-2,3-dihydro-1,4-benzodioxin-7-yl)methyl-3-(4,7-dimethyl-2-oxochromen-3-yl) propanoate) interacts with six residues of MPII namely, Thr 308, Phe 309, Gly 310, Tyr 341, Ser 368, and His 339 to form hydrogen bonds of various lengths. Dioxane moiety interacts with Tyr 341, Ser 368, and His 339, while ester groups link with Thr 308, Phe 309, and Gly 310. A π-cationic interaction of aromatic ring with Lys 186 can also be observed. Besides, few *van der* Waal's interactions between ligand and protein within the binding pocket are also visible in Fig. 3(I and J).

3.8.6. Interaction of ZINC000015767717 with MPII

The furan and carbonyl moieties of ZINC000015767717 (IUPAC nomenclature: "[5-(4-Nitrophenyl) furan-2-yl]-[4-[[4-(trifluoromethyl) phenyl] methyl] piperazin-1-yl] methanone") form two hydrogen bonds with Ile 77 of MPII with bond lengths 4.33 and 4.17 Å, respectively. In addition, Leu 46, Ser 47, and Ile 49 of MPII interact with fluorine of the ligand. The aromatic rings of furan and nitrophenyl groups also exhibit π-anionic interactions with Glu 76. Ser 50 develops π-donor hydrogen bond with aromatic ring, whereas, Ile 72 forms unique carbon hydrogen bond with piperazine moiety. Other residues present in the protein binding pocket also present various hydrophobic interaction with the ligand. Interactions and residues are shown in Fig. 3(K, L).

3.8.7. Interaction of ZINC000032736238 with MPII

The aromatic ring of trimethylphenyl group in ZINC000032736238 (IUPAC name: [4-(Benzimidazol-1-yl) phenyl]-[4-(2,4,6-trimethylphenyl)sulfonylpiperazin-1-yl]methanone) forms π-donor hydrogen bond with the Asn 347 exhibiting bond length of 3.94 Å. As evident from the interaction plots in Fig. 3(M, N), majority of the interactions are governed by the hydrophobic/*van der* Waals' interaction within the binding pocket.

3.8.8. Interaction of ZINC000025928534 with MPII

The oxadiazole in ZINC000025928534 (IUPAC name: (1S)-2-methyl-N-[[3-(3-nitrophenyl)-1,2,4-oxadiazol-5-yl]methyl]-1-phenylpropan-1-amine), Thr 342 and Gly 310 are involved in hydrogen bonding with the bond length of 2.51 and 3.08 Å. Asp 187 shared a π-anionic bond with the ligand. This compound has also several hydrophobic interactions with the docked protein residues present in the binding pocket as shown in Fig. 3(O, P).

3.8.9. Interaction of ZINC000023804957 with MPII

In ZINC000023804957, IUPAC name: [(5-Fluoro-2-methylphenyl)carbamoyl]methyl 1-[5-(trifluoromethyl)pyridin-2-yl]piperidine-4-carboxylate, Thr 342, Gly 310, His 339 and Lys 186 are the interacting residues of protein which made hydrogen bonds with the ligand in the binding pocket. Pro 340, His 339, Thr 308, and Lys 307 are involved in carbon hydrogen bond. Asp 304 and Ser 338 are involved in making the bond with the Halogen groups present in ligand. Other residues are also present in the protein binding pocket having the hydrophobic interaction towards the ligand docked in the binding pocket. Interactions are shown in Fig. 3(Q, R).

3.8.10. Interaction of ZINC000024389255 with MPII

In ZINC000024389255, IUPAC name: 4-[(carbamoylamino) methyl]-N-[3-(2, 4-dimethylphenyl) propyl] benzamide, interacting residues of protein made hydrogen bonds with the ligand in the binding pocket. In the case, there are six residues which make hydrogen bond with ligand with different bond lengths. For example, Thr 342 hydrogen bond length is 5.07 Å, Gly 310 with 2.81 Å, Thr 308 with 2.81 and 2.96 Å. Glu 270 with 4.47 and 3.36 Å, Asp 304 with 3.21 Å, Lys 186 with 3.79 and 4.38 Å. Other residues are also present in the protein binding pocket having the hydrophobic interaction towards the ligand docked in the binding pocket. Their interactions are shown in Fig. 3(S, T).

3.9. MD simulations

After the detailed protein-ligand interaction analysis, top three ligands were selected based on lowest binding free energies averaged over first five docked conformations, as given in Supplementary Table S4. On the basis of these scores, ZINC000021797248, ZINC000016722540, and ZINC000023754543 were selected for MD simulations and trajectory analysis in terms of RMSD (root mean square deviation), RMSF (root mean square fluctuation), and Rg (radius of gyration). In addition, the number of H-bonds formed between protein and ligands and MM/PBSA protein-ligand binding energies were also calculated.

Fig. 4 shows the plots, including RMSD of protein, RMSD of ligands, RMSF, and Rg of MPII protein with the three selected ZINC compounds. Complex of MPII with compound 1, compound 2 and compound 3 is shown in black, red and in green line, respectively. RMSD values of protein (shown in Fig. 4(A)) of all three compounds exhibited stability throughout the simulation time. However, mean values of RMSD of protein for compound 3 (*i.e.* 2.33 ± 0.27 Å) is slightly higher as compared to compound 1 and 2 (*i.e.* 2.16 ± 0.18 Å and 2.1 ± 0.21 Å, respectively). RMSD values (of protein) for compound 3 increased between 45 and 130 ns. Overall, no drastic deviation was observed as the fluctuation was with in ~3 Å. Hence, protein structure did not go under any major conformational changes in the presence of these compounds. Fig. 4(B) displays RMSD (ligand) values for all three compounds. RMSD plot of compound
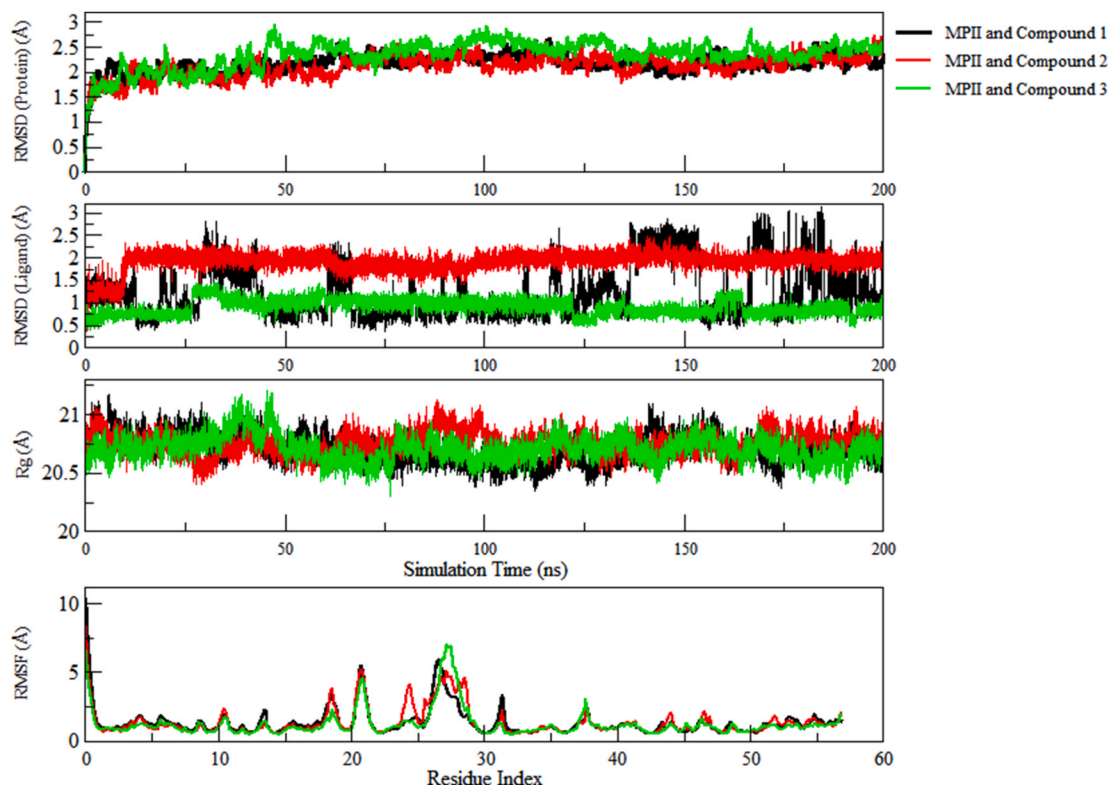
Fig. 4. Various Trajectory analysis of MPII complexed with selected ZINC compounds (A)RMSD of protein, (B) RMSD of ligands, (C) Rg, and (D) RMSF of MPII complexed with selected ZINC compound 1 (black lines), compound 2 (red lines) and compound 3 (green color). C-alpha atoms were used to calculate RMSD of protein, RMSF and Rg with gmx in Gromacs. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

1 (black line) showed that it exists in two different conformations or binding poses and is stable reflected by mean RMSD values 1.25 ± 0.6 Å. Whereas compound 2 (red line) and compound 3 (green line) are present in single conformation in MPII binding pocket with the mean RMSD values for compound 2 (1.92 ± 0.21 Å) and compound 3 (0.91 ± 0.17 Å). Hence, there is no significant difference in RMSD values of the selected compounds.

Rg (radius of gyration) values are represented in graph (Fig. 4(C)) for MPII complex with compound 1 (black line), compound 2 (red line), and compound 3 (green line). All three systems remained compact throughout the simulation time. Mean values for compound 1 are 20.7 ± 0.12 Å), for compound 2 is 20.8 ± 0.1 Å, and for compound 3 is 20.7 ± 0.11 Å. Therefore, MPII complex did not under go any influential conformational changes induced by the selected compounds.

In Fig. 4(D), RMSF plot is shown for all three compounds complexed with MPII. Similar pattern of fluctuations was observed in all complexes. However, complex of MPII and compound 2 (red line) exhibited more fluctuation at residue index 18, 22, 24 to 27, 44, 52, and 57. Compound 1 complex (black line) fluctuated for residues 26, 32, and 53–55 whereas more fluctuations were observed at index 27 and 37.

In Fig. 5, graph shows the number of hydrogen bonds established between MPII and the selected compounds during the simulation time. Complex of MPII and compound 2 (red line) are involved in making consistent hydrogen bonds with the residues present in the binding pocket of MPII. On the other hand, compound 1 shows partial movement in the binding pocket after 25 ns and after 75 ns, this compound formed greater number of hydrogen bonds. As compared to other two compounds, compound 3 is making maximum hydrogen bonds with the protein throughout the simulation time.

In Fig. 6, MM/PBSA of protein-ligand binding energies are shown. Mean values of ΔE for compound 1 is -27.12 ± 3.88 kcal/mol, for compound 2 is -27.34 ± 3.51 kcal/mol, and for compound 3 is -24.47 ± 3.94 kcal/mol. Compound 1 showed better binding with MPII after 36 ns and compound 2 showed good and balanced binding with MPII during the simulation time. Compound 3 showed higher ΔE after 20ns however, after 40ns it was overlapping and equivalent to compound 1 and 2 ΔE. In order to analyze the movement and binding of compounds, frames were extracted out of trajectories from the last 100 ns respective to each compound and are shown in Fig. S2.
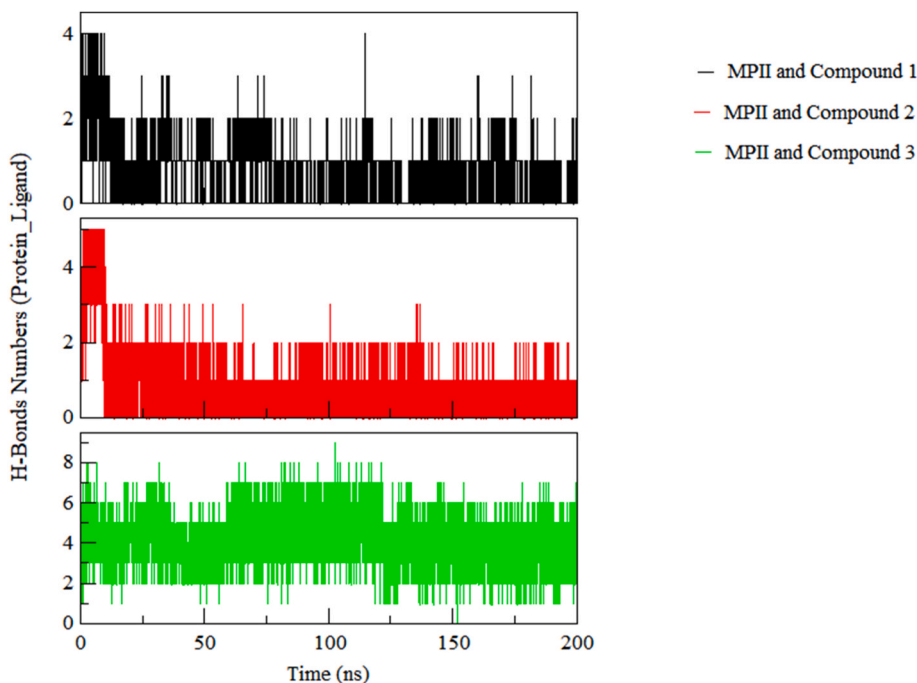
Fig. 5. Trajectory analysis of MPII complexed with selected ZINC compounds Hydrogen-bonds numbers between MPII and compound 1 (black line), compound 2 (red line), and compound 3 (green line). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)
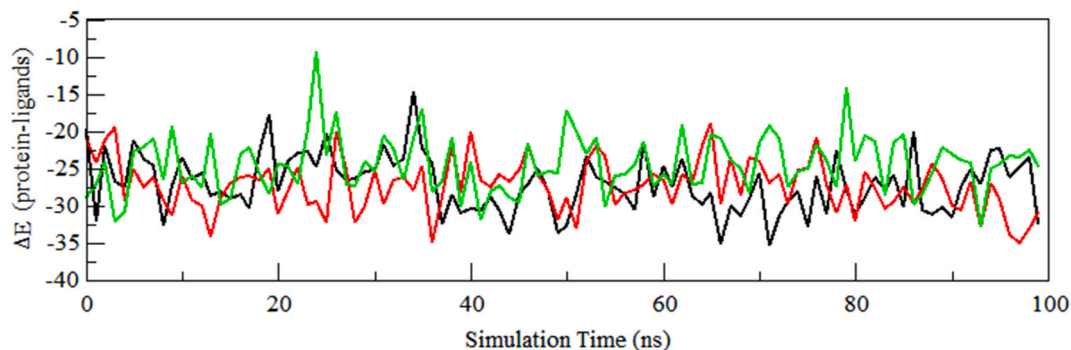

Fig. 6. MM/PBSA analysis of MPII complexed with selected ZINC compounds MM/PBSA (Molecular Mechanics/Poisson-Born Surface Area) protein-ligand(s) binding energies between MPII and compounds. Compound 1 is in black, compound 2 is in red, compound 3 is shown in green. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

4. Discussion

Colorectal cancer is known to be the commonly occurring cancer in both men and women across the world. Previously, the leading cause of CRC was considered to be genetics, epigenetics, and environmental factors which include diet habits, physical activity, age, smoking, consumption of alcohol etc. Amongst all, obesity has influential effect on human health and is directly associated with the pathways which initiate inflammation leading to the progression of CRC by depositing visceral fat and producing altered cytokines. Reportedly, pathogenesis factors of cancer (CRC) are, but not limited to, genetics, epigenetics, and environmental factors [89]. There are certain other factors, which play important roles in etiology of CRC such as dysbiosis in human microbiota which is actively involved in interacting with the epithelium and initiating CRC. Human microbiota, also known as an extended genome of human, were perceived to perform vital and beneficial roles and functions for the well-being of host [90], however, it is evident from the emerging research exploration that they are responsible for causing life-threatening diseases in human when disturbed [91]. Microbial dysbiosis, especially in gut microbiota, is capable of altering the normal epithelial lining state of colon by inducing inflammation, changing immune responses and by disturbing metabolic functions of host [92]. Out of all gut microbiota members, *Bacteroides* are of importance as it comprises of ~90 % of gut microbiota and are known to promote cancers [93,94].

Here, we revisited the microbial interactions with their host in colonic epithelium besides beneficial roles. Consequently, *Bacteroides fragilis* was taken into account due to its involvement in CRC pathogenesis. Herein, we incorporated a new evolving research area "subtractive proteomics" which revealed the non-human orthologous, essential and virulent protein, metalloproteinase II or MPII involved in causing CRC. Enterotoxigenic *Bacteroides fragilis* strain 2-078382-3 proteome was used to reveal the non-orthologs against human proteome and 1214 non-orthologs were obtained which were screened for essential proteins against dataset obtained from DEG database. 1087 essential genes were predicted in *Bacteroides fragilis* proteome and clustered paralogs or redundant protein sequences in essential proteins were eliminated from the study. As a result, 1041 unique *Bacteroides fragilis* proteins were identified and analyzed for their virulence and involvement in pathogenicity. Therefore, 68 unique virulent yet essential proteins were selected and were assigned KO number in order to identify pathological pathways showing involvement of these proteins. 56 genes were assigned the KO number and each protein was manually searched to identify their pathways. Therefore, we found out that the query protein of *Bacteroides fragilis* (strain 2-078382-3) protein "metalloproteinase II or MPII" with reference sequence ID "WP_005,797,263" was assigned the KO number (K19959: bftP; fragilysin) and was aligned against fragilysin. MPII and fragilysin are the two virulence factors or toxins that are produced by (Enterotoxigenic) *Bacteroides fragilis* strains [95].

Bacteroides fragilis are the members of gut microbiota [96] which in addition to play various important roles in human, are associated with and are responsible for various human diseases [97]. For example, Enterotoxigenic *Bacteroides fragilis* are mainly involved in causing diarrhea, inflammatory bowel disease (IBD) and other intestinal disorders [96]. The Enterotoxigenic strains of *Bacteroides fragilis* contain a pathogenicity island *i.e.* BFT PAI, known to produce toxins, the metalloproteinases (BFT or *Bacteroides fragilis* toxin) [50]. These toxins are fragilysin and metalloproteinase II (MPII), and are encoded by *bft* gene and *mpii* gene [98], respectively. These genes are associated with the colon mucosa [44] and inactivation of beta-catenin signaling pathways, cleave intestinal E-cadherin and promote tumor formation [100]. Therefore, the released toxins are known to be involved in proteolysis [69] and stimulating cleavage of E-cadherin present on colonic epithelial lining [99]. In addition, binding of MPII catalytic domain with E-cadherin [70] increases the cell permeability and damages colonic crypts and colonocytes surface [71]. Moreover, it weakens cell to cell junction and initiate disturbed immune pathways and cytokines production [96]. Thus, our analysis revealed that "fragilysin family metalloproteinase II" (MPII, WP_005797263.1), a virulent factor of *Bacteroides fragilis* (strain 2-078382-3), was predicted to be localized in extracellular compartment (or space) with significant score and is associated with colon mucosa [44].

Therefore, toxin Metalloproteinase II (MPII), encoded by the pathogenicity island of strain 2-078382-3, was selected to study its interactions with small organic compounds using molecular docking approach. As this protein is localized in extracellular matrix of the colon, thus, it can be a suitable target for drug(s). This screening study identified some potential compounds to be used as drug candidates, as they possess properties desired for a good drug *i.e.* low toxicity, high GI absorption, no blood-brain barrier permeability and have good synthetic accessibility. Moreover, they do not violate the Lipinski, veber, ghoose, egan, Muegge rules, PAINS, Brenk and leadlikeness suggesting that these compounds have drug-like properties. In addition to all properties, these potential compounds are not the substrate of P-glycoproteins. P-glycoproteins have an important role of taking in and out the by-products of the drugs of the epithelial membranes. However, these p-glycoproteins are also involved in causing drug resistance while pumping the by-products out [101,102] thus reducing the efficiency of drugs against cancer therapies [103]. To date, multidrug resistance or MDR is becoming a major challenge amongst microorganism [104]. To avoid the multidrug resistance, none of the selected compounds are substrate of P-glycoproteins (P-gp). This approach helps in enhancing the drug effect by increasing the interacting time thus weaken the MPII and E-cadherin interaction(s). Therefore, ten compounds were selected after detailed ADMET properties analysis and were analyzed for their interaction with MPII.

MPII, a zinc-dependent metalloproteinase, is encoded by *mpii* gene with 397 amino acids. Studies suggested that MPII and fragilysin (toxins) encoded by Pathogenicity Island in *Bacteroides fragilis* are structurally related toxins/enzymes. Therefore, a significant overlap among the E-cadherin cleavage preferences was noticed [99]. In the structure, there is a zinc binding site which is also known as catalytic site/domain. Here, Zn^{2+} ion is bounded by four residues, *i.e.* His 348, His 352, His 358 and Asp 194 [105]. There were eight ligands bound in the binding pocket with the Zn^{2+} metal ion catalytic site. Other sites contain two ligands on the other side of the protein as shown in Fig. 2 (A, B, and C) respectively. Interaction analysis revealed a set of conserved residues in MPII that are interacting with ligands. For example, His 345, Glu 346, His 339, Gly 310, Tyr 341, Pro 340, Asp 187, Phe 309, Lys 307, Ile 185, Thr 308, and Pro 184 are common interacting residues of MPII binding pocket with Zn^{2+} ion. Other binding site interacting residues are Leu 46, Ser 47, Lys 48, Ile 49, Ser 50, Ala 71, Ile 72, His 73, Val 75, Glu 76, Ile 77, and Glu 78 (Fig. 3)(K). Next ligand is bound in the binding pocket present at the back side of MPII where interacting residues are Glu 88, Gly 89, Gln 236, Ala 237, Thr 240, Ser 241, Lys 243, Phe 244, Pro 373 and Asn 374 as shown in Fig. 3(M).

Moreover, RMSD of protein, RMSF and Rg values of the three selected compounds (*i.e.* ZINC000021797248, ZINC000016722540, and ZINC000023754543) revealed that MPII protein remained stable and compact, and did not undergo any major conformational changes and residue fluctuations in the presence of these compounds. RMSD of ligands also showed that these compounds are bound well in the binding pocket of MPII. Moreover, mean Δ*E* values showed that there was no significant difference in binding energies which also suggest that these compounds showed good binding with the gut microbial protein MPII. Therefore, three selected compounds are the potential ligands against MPII, and can weaken its binding with E-cadherin and can inhibit the E-cadherin cleavage. Reduction in cleavage can also prevent loss of cell to cell connections and proteolysis associated with CRC. In order to analyze their behavior, interaction and to confirm the inhibitory effect of these potential compounds against MPII, *in vitro* experiments and testing is required.

5. Conclusion

In this study, one of the most common members of gut microbiota was identified to be involved in colorectal cancers. By incorporating *state of the art* approach "subtractive proteomics", we were able to extract Enterotoxigenic *Bacteroides fragilis* essential and virulent protein/toxin MPII. MPII, a Zn^{2+} ion dependent toxin, is involved in proteolysis and cleavage of colonic epithelium and trigger oncogenic pathways when disturbed. We identified ten compounds after detailed ADMET and physiochemical properties analysis which were used to study interactions when complexed with MPII. Interaction analysis showed some important residues of MPII making hydrogen bonds with the selected ligands. Top three candidates with least binding energies were selected to study their dynamic behavior and conformational changes when bound with MPII. This analysis revealed that all three compounds showed good binding with MPII possessing all properties of an inhibitor and can be used as potential drug candidates.

6. Ethics approval and consent to participate

Not applicable.

7. Funding resources

8. Data availability statement

Human proteome (UniProt accession ID: UP000005640) and Bacteroides fragilis (strain 2-078382-3, UniProt accession ID: UP000094213) are publicly available at repository "UniProt Proteomes" and were downloaded without any access restrictions. The datasets generated and analyzed in this study is included in article/supplementary material and are available from corresponding authors upon request.

CRediT authorship contribution statement

Bushra Arif: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Saba Yasir:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Formal analysis, Data curation. **Muhammad Saeed:** Writing – review & editing, Supervision, Formal analysis. **M. Qaiser Fatmi:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare no financial conflict of interest. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e32838.

References

[1] M. Van Engeland, S. Derks, K.M. Smits, G.A. Meijer, J.G. Herman, Colorectal cancer epigenetics: complex simplicity, J. Clin. Oncol. 29 (10) (2011) 1382–1391, https://doi.org/10.1200/JCO.2010.28.2319.

[2] T. Armaghany, J.D. Wilson, Q. Chu, G. Mills, Genetic alterations in colorectal cancer, Gastrointestinal cancer research: GCR 5 (1) (2012) 19. PMCID: PMC3348713.

[3] C. Tang, H. Sun, M. Kadoki, W. Han, X. Ye, Y. Makusheva, Y. Tan, Blocking Dectin-1 prevents colorectal tumorigenesis by suppressing prostaglandin E2 production in myeloid-derived suppressor cells and enhancing IL-22 binding protein expression, Nat. Commun. 14 (1) (2023) 1493, https://doi.org/10.1038/s41467-023-37229-x.

[4] M.K. Sethi, M. Thaysen-Andersen, H. Kim, C.K. Park, M.S. Baker, N.H. Packer, S. Fanayan, Quantitative proteomic analysis of paired colorectal cancer and non-tumorigenic tissues reveals signature proteins and perturbed pathways involved in CRC progression and metastasis, J. Proteomics 126 (2015) 54–67, https://doi.org/10.1016/j.jprot.2015.05.037.

[5] W.M. Grady, J.M. Carethers, Genomic and epigenetic instability in colorectal cancer pathogenesis, Gastroenterology 135 (4) (2008) 1079–1099, https://doi.org/10.1053/j.gastro.2008.07.076.

[6] L.A. Torre, F. Bray, R.L. Siegel, J. Ferlay, J. Lortet-Tieulent, A. Jemal, Global cancer statistics, 2012, CA: a cancer journal for clinicians 65 (2) (2015) 87–108, https://doi.org/10.3322/caac.21262.

[7] C.A. Doubeni, J.M. Major, A.O. Laiyemo, M. Schootman, A.G. Zauber, A.R. Hollenbeck, J. Allison, Contribution of behavioral risk factors and obesity to socioeconomic differences in colorectal cancer incidence, J. Natl. Cancer Inst. 104 (18) (2012) 1353–1362, https://doi.org/10.1093/jnci/djs410.

[8] K. Simon, Colorectal cancer development and advances in screening, Clin. Interv. Aging 11 (2016) 967, https://doi.org/10.2147/CIA.S109285.

[9] J. Terzić, S. Grivennikov, E. Karin, M. Karin, Inflammation and colon cancer, Gastroenterology 138 (6) (2010) 2101–2114. e2105, https://doi.org/10.1053/j.gastro.2010.01.058.

[10] K. Esposito, P. Chiodini, A. Colao, A. Lenzi, D. Giugliano, Metabolic syndrome and risk of cancer: a systematic review and meta-analysis, Diabetes Care 35 (11) (2012) 2402–2411, https://doi.org/10.2337/dc12-0336.

[11] D. Colussi, G. Brandi, F. Bazzoli, L. Ricciardiello, Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention, Int. J. Mol. Sci. 14 (8) (2013) 16365–16385, https://doi.org/10.3390/ijms140816365.

[12] N.B. Janakiram, C.V. Rao, The role of inflammation in colon cancer, in: Inflammation and Cancer, Springer, 2014, pp. 25–52, https://doi.org/10.1007/978-3-0348-0837-8_2.

[13] S. Al-Sohaily, A. Biankin, R. Leong, M. Kohonen-Corish, J. Warusavitarne, Molecular pathways in colorectal cancer, J. Gastroenterol. Hepatol. 27 (9) (2012) 1423–1431, https://doi.org/10.1111/j.1440-1746.2012.07200.x.

[14] M. Pancione, A. Remo, V. Colantuoni, Genetic and epigenetic events generate multiple pathways in colorectal cancer progression, Pathol. Res. Int. 2012 (2012), https://doi.org/10.1155/2012/509348.

[15] H. Tilg, T.E. Adolph, R.R. Gerner, A.R. Moschen, The intestinal microbiota in colorectal cancer, Cancer Cell 33 (6) (2018) 954–964, https://doi.org/10.1155/2012/509348.

[16] B. Wang, M. Yao, L. Lv, Z. Ling, L. Li, The human microbiota in health and disease, Engineering 3 (1) (2017) 71–82, https://doi.org/10.1016/J.ENG.2017.01.008.

[17] M.H. Raza, K. Gul, A. Arshad, N. Riaz, U. Waheed, A. Rauf, M. Abdullah, Microbiota in cancer development and treatment, J. Cancer Res. Clin. Oncol. 145 (1) (2019) 49–63, https://doi.org/10.1007/s00432-018-2816-0.

[18] E. Rosenberg, I. Zilber-Rosenberg, The hologenome concept, in: Beneficial Microorganisms in Multicellular Life Forms, Springer, 2012, pp. 323–340, https://doi.org/10.1007/978-3-642-21680-0_24.

[19] S. Conlan, H.H. Kong, J.A. Segre, Species-level analysis of DNA sequence data from the NIH human microbiome project, PLoS One 7 (10) (2012) e47075, https://doi.org/10.1371/journal.pone.0047075.

[20] L.K. Ursell, J.L. Metcalf, L.W. Parfrey, R. Knight, Defining the human microbiome, Nutr. Rev. 70 (suppl_1) (2012) S38–S44, https://doi.org/10.1111/j.1753-4887.2012.00493.x.

[21] E.A. Mayer, K. Tillisch, A. Gupta, Gut/brain axis and the microbiota, J. Clin. Invest. 125 (3) (2015) 926–938, https://doi.org/10.1172/JCI76304.

[22] G. Clarke, R.M. Stilling, P.J. Kennedy, C. Stanton, J.F. Cryan, T.G. Dinan, Minireview: gut microbiota: the neglected endocrine organ, Mol. Endocrinol. 28 (8) (2014) 1221–1238, https://doi.org/10.1210/me.2014-1108.

[23] E.A. Mayer, Gut feelings: the emerging biology of gut–brain communication, Nat. Rev. Neurosci. 12 (8) (2011) 453–466, https://doi.org/10.1038/nrn3071.

[24] J.A. Foster, K.A.M. Neufeld, Gut–brain axis: how the microbiome influences anxiety and depression, Trends Neurosci. 36 (5) (2013) 305–312, https://doi.org/10.1016/j.tins.2013.01.005.

[25] R.M. Stilling, T.G. Dinan, J.F. Cryan, Microbial genes, brain & behaviour–epigenetic regulation of the gut–brain axis, Gene Brain Behav. 13 (1) (2014) 69–86, https://doi.org/10.1111/gbb.12109.

[26] S. Shen, C.H. Wong, Bugging inflammation: role of the gut microbiota, Clinical & translational immunology 5 (4) (2016) e72, https://doi.org/10.1038/cti.2016.12.

[27] B. Zhou, Y. Yuan, S. Zhang, C. Guo, X. Li, G. Li, Z. Zeng, Intestinal flora and disease mutually shape the regional immune system in the intestinal tract, Front. Immunol. 11 (2020) 575, https://doi.org/10.3389/fimmu.2020.00575.

[28] I. Sekirov, S.L. Russell, L.C.M. Antunes, B.B. Finlay, Gut microbiota in health and disease, Physiol. Rev. 90 (3) (2010) 859–904, https://doi.org/10.1152/physrev.00045.2009.

[29] W. Dieterich, M. Schink, Y. Zopf, Microbiota in the gastrointestinal tract, Med. Sci. 6 (4) (2018) 116, https://doi.org/10.3390/medsci6040116.

[30] G.O. Canny, B.A. McCormick, Bacteria in the intestine, helpful residents or enemies from within? Infect. Immun. 76 (8) (2008) 3360–3373, https://doi.org/10.1128/iai.00187-08.

[31] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J.H. Badger, A.T. Chinwalla, R.S. Fulton, Structure, function and diversity of the healthy human microbiome, nature 486 (7402) (2012) 207, https://doi.org/10.1038/nature11234.

[32] C.L. Boulangé, A.L. Neves, J. Chilloux, J.K. Nicholson, M.-E. Dumas, Impact of the gut microbiota on inflammation, obesity, and metabolic disease, Genome Med. 8 (1) (2016) 1–12, https://doi.org/10.1186/s13073-016-0303-2.

[33] K. Brown, D. DeCoffe, E. Molcan, D.L. Gibson, Diet-induced dysbiosis of the intestinal microbiota and the effects on immunity and disease, Nutrients 4 (8) (2012) 1095–1119, https://doi.org/10.3390/nu4081095.

[34] J. Chen, J.C. Domingue, C.L. Sears, Microbiota dysbiosis in select human cancers: evidence of association and causality, Paper presented at the Seminars in immunology (2017), https://doi.org/10.1016/j.smim.2017.08.001.

[35] G.L. Hold, W.S. Garrett, Microbiota organization—a key to understanding CRC development, Nat. Rev. Gastroenterol. Hepatol. 12 (3) (2015) 128–129, https://doi.org/10.1038/nrgastro.2015.25.

[36] F. Mariani, P. Sena, L. Roncucci, Inflammatory pathways in the early steps of colorectal cancer development, World J. Gastroenterol.: WJG 20 (29) (2014) 9716, https://doi.org/10.3748/wjg.v20.i29.9716.

[37] H. Zafar, M.H. Saier Jr, Gut Bacteroides species in health and disease, Gut Microb. 13 (1) (2021) 1848158, https://doi.org/10.1080/19490976.2020.1848158.

[38] M.J. Coyne, M. Chatzidaki-Livanis, L.C. Paoletti, L.E. Comstock, Role of glycan synthesis in colonization of the mammalian gut by the bacterial symbiont Bacteroides fragilis, Proc. Natl. Acad. Sci. USA 105 (35) (2008) 13099–13104, https://doi.org/10.1073/pnas.0804220105.

[39] M. Chatzidaki-Livanis, M.J. Coyne, H. Roche-Hakansson, L.E. Comstock, Expression of a uniquely regulated extracellular polysaccharide confers a large-capsule phenotype to Bacteroides fragilis, J. Bacteriol. 190 (3) (2008) 1020–1026, https://doi.org/10.1128/jb.01519-07.

[40] C.M. Fletcher, M.J. Coyne, D.L. Bentley, O.F. Villa, L.E. Comstock, Phase-variable expression of a family of glycoproteins imparts a dynamic surface to a symbiont in its human intestinal ecosystem, Proc. Natl. Acad. Sci. USA 104 (7) (2007) 2413–2418, https://doi.org/10.1073/pnas.0608797104.

[41] E.B. Troy, D.L. Kasper, Beneficial effects of Bacteroides fragilis polysaccharides on the immune system, Frontiers in bioscience: J. Vis. Literacy 15 (2010) 25, https://doi.org/10.2741/3603.

[42] J.Y. Huang, S.M. Lee, S.K. Mazmanian, The human commensal Bacteroides fragilis binds intestinal mucin, Anaerobe 17 (4) (2011) 137–141, https://doi.org/10.1016/j.anaerobe.2011.05.017.

[43] C.L. Sears, A.L. Geis, F. Housseau, Bacteroides fragilis subverts mucosal biology: from symbiont to colon carcinogenesis, J. Clin. Invest. 124 (10) (2014) 4166–4172, https://doi.org/10.1172/JCI72334.

[44] A. Boleij, E.M. Hechenbleikner, A.C. Goodwin, R. Badani, E.M. Stein, M.G. Lazarev, E.C. Wick, The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients, Clin. Infect. Dis. 60 (2) (2015) 208–215, https://doi.org/10.1093/cid/ciu787.

[45] F. Haghi, E. Goli, B. Mirzaei, H. Zeighami, The association between fecal enterotoxigenic B. fragilis with colorectal cancer, BMC Cancer 19 (1) (2019) 879, https://doi.org/10.1186/s12885-019-6115-1.

[46] S.A. Shiryaev, A.E. Aleshin, N. Muranaka, M. Kukreja, D.A. Routenberg, A.G. Remacle, A.Y. Strongin, Structural and Functional Diversity of Metalloproteinases Encoded by the Bacteroides Fragilis Pathogenicity Island, Wiley Online Library, 2014, https://doi.org/10.1111/febs.12804.

[47] S. Grivennikov, E. Karin, J. Terzic, D. Mucida, G.-Y. Yu, S. Vallabhapurapu, L. Eckmann, IL-6 and Stat 3 are required for survival of intestinal epithelial cells and development of colitis-associated cancer, Cancer Cell 15 (2) (2009) 103–113, https://doi.org/10.1016/j.ccr.2009.01.001.

[48] E.T. Orberg, H. Fan, A.J. Tam, C.M. Dejea, C.D. Shields, S. Wu, P. Fathi, The myeloid immune signature of enterotoxigenic Bacteroides fragilis-induced murine colon tumorigenesis, Mucosal Immunol. 10 (2) (2017) 421–433, https://doi.org/10.1038/mi.2016.53.

[49] H.C. Carrow, L.E. Batachari, H. Chu, Strain diversity in the microbiome: lessons from Bacteroides fragilis, PLoS Pathog. 16 (12) (2020) e1009056, https://doi.org/10.1371/journal.ppat.1009056.

[50] J.V. Pierce, H.D. Bernstein, Genomic diversity of enterotoxigenic strains of Bacteroides fragilis, PLoS One 11 (6) (2016) e0158171, https://doi.org/10.1371/journal.pone.0158171.

[51] T. Gabaldón, E.V. Koonin, Functional and evolutionary implications of gene orthology, Nat. Rev. Genet. 14 (5) (2013) 360–366, https://doi.org/10.1038/nrg3456.

[52] B. Liu, D. Zheng, Q. Jin, L. Chen, J. Yang, Vfdb 2019: a comparative pathogenomic platform with an interactive web interface, Nucleic Acids Res. 47 (D1) (2019) D687–D692, https://doi.org/10.1093/nar/gky1080.

[53] M. Vital, A.C. Howe, J.M. Tiedje, Revealing the bacterial butyrate synthesis pathways by analyzing (meta) genomic data, mBio 5 (2) (2014), https://doi.org/10.1128/mBio.00889-14.

[54] Y. Hu, H. Lehrach, M. Janitz, Comparative analysis of an experimental subcellular protein localization assay and in silico prediction methods, J. Mol. Histol. 40 (5–6) (2009) 343–352, https://doi.org/10.1007/s10735-009-9247-9.

[55] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, L. Bordoli, SWISS-MODEL: homology modelling of protein structures and complexes, Nucleic Acids Res. 46 (W1) (2018) W296–W303, https://doi.org/10.1093/nar/gky427.

[56] E.H.B. Maia, L.C. Assis, T.A. de Oliveira, A.M. da Silva, A.G. Taranto, Structure-based virtual screening: from classical to artificial intelligence, Front. Chem. 8 (2020) 343, https://doi.org/10.3389/fchem.2020.00343.

[57] D. Lagorce, D. Douguet, M.A. Miteva, B.O. Villoutreix, Computational analysis of calculated physicochemical and ADMET properties of protein-protein interaction inhibitors, Sci. Rep. 7 (2017) 46277, https://doi.org/10.1038/srep46277.

[58] L. Guan, H. Yang, Y. Cai, L. Sun, P. Di, W. Li, Y. Tang, ADMET-score–a comprehensive scoring function for evaluation of chemical drug-likeness, Medchemcomm 10 (1) (2019) 148–157, https://doi.org/10.1039/C8MD00472B.

[59] L.L. Ferreira, A.D. Andricopulo, ADMET modeling approaches in drug discovery, Drug Discov. Today 24 (5) (2019) 1157–1165, https://doi.org/10.1016/j.drudis.2019.03.015.

[60] A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, Sci. Rep. 7 (2017) 42717, https://doi.org/10.1038/srep42717.

[61] S. Salentin, S. Schreiber, V.J. Haupt, M.F. Adasme, M. Schroeder, PLIP: fully automated protein–ligand interaction profiler, Nucleic Acids Res. 43 (W1) (2015) W443–W447, https://doi.org/10.1093/nar/gkv315.

[62] M.R. Islam, M.A. Awal, A. Khames, M.A. Abourehab, A. Samad, W.M. Hassan, B. Kim, Computational identification of druggable bioactive compounds from catharanthus roseus and avicennia marina against colorectal cancer by targeting thymidylate synthase, Molecules 27 (7) (2022) 2089, https://doi.org/10.3390/molecules27072089.

[63] Z. Yang, K. Lasker, D. Schneidman-Duhovny, B. Webb, C.C. Huang, E.F. Pettersen, T.E. Ferrin, UCSF Chimera, MODELLER, and IMP: an integrated modeling system, J. Struct. Biol. 179 (3) (2012) 269–278, https://doi.org/10.1016/j.jsb.2011.09.006.

[64] Z. Sa, J. Zhou, Y. Zou, Z. Su, X. Gu, Paralog-divergent features may help reduce off-target effects of drugs: hints from glucagon subfamily analysis, Dev. Reprod. Biol. 15 (4) (2017) 246–254, https://doi.org/10.1016/j.gpb.2017.03.004.

[65] W.T. Cheng, H.K. Kantilal, F. Davamani, The mechanism of Bacteroides fragilis toxin contributes to colon cancer formation, Malays. J. Med. Sci.: MJMS 27 (4) (2020) 9, https://doi.org/10.21315/mjms2020.27.4.2.

[66] X. Pan, H. Li, T. Zeng, Z. Li, L. Chen, T. Huang, Y.-D. Cai, Identification of protein subcellular localization with network and functional embeddings, Front. Genet. 11 (2021) 626500, https://doi.org/10.3389/fgene.2020.626500.

[67] C.S. Yu, C.J. Lin, J.K. Hwang, Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions, Protein Sci. 13 (5) (2004) 1402–1406, https://doi.org/10.1110/ps.03479604.

[68] T.M. Bakheet, A.J. Doig, Properties and identification of human protein drug targets, Bioinformatics 25 (4) (2009) 451–457, https://doi.org/10.1093/bioinformatics/btp002. DOI: 10.1093/bioinformatics/btp002.

[69] S. Wu, K.-J. Rhee, M. Zhang, A. Franco, C.L. Sears, Bacteroides fragilis toxin stimulates intestinal epithelial cell shedding and γ-secretase-dependent E-cadherin cleavage, J. Cell Sci. 120 (11) (2007) 1944–1952, https://doi.org/10.1242/jcs.03455.

[70] A.G. Remacle, S.A. Shiryaev, A.Y. Strongin, Distinct interactions with cellular E-cadherin of the two virulent metalloproteinases encoded by a Bacteroides fragilis pathogenicity island, PLoS One 9 (11) (2014) e113896, https://doi.org/10.1371/journal.pone.0113896.

[71] M. Riegler, M. Lotz, C. Sears, C. Pothoulakis, I. Castagliuolo, C. Wang, G. Bischof, Bacteroides fragilis toxin 2 damages human colonic mucosa in vitro, Gut 44 (4) (1999) 504–510, https://doi.org/10.1136/gut.44.4.504.

[72] N.Y. Yu, J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, L.J. Foster, PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics 26 (13) (2010) 1608–1615, https://doi.org/10.1093/bioinformatics/btq249.

[73] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 64 (2012) 4–17, https://doi.org/10.1016/j.addr.2012.09.019.

[74] A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski, A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases, J. Combin. Chem. 1 (1) (1999) 55–68, https://doi.org/10.1021/cc9800071.

[75] D.F. Veber, S.R. Johnson, H.-Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, Molecular properties that influence the oral bioavailability of drug candidates, J. Med. Chem. 45 (12) (2002) 2615–2623, https://doi.org/10.1021/jm020017n.

[76] W.J. Egan, K.M. Merz, J.J. Baldwin, Prediction of drug absorption using multivariate statistics, J. Med. Chem. 43 (21) (2000) 3867–3877, https://doi.org/10.1021/jm000292e.

[77] I. Muegge, S.L. Heald, D. Brittelli, Simple selection criteria for drug-like chemical matter, J. Med. Chem. 44 (12) (2001) 1841–1846, https://doi.org/10.1021/jm015507e.

[78] R. Brenk, A. Schipani, D. James, A. Krasowski, I.H. Gilbert, J. Frearson, P.G. Wyatt, Lessons learnt from assembling screening libraries for drug discovery for neglected diseases, ChemMedChem: Chemistry Enabling Drug Discovery 3 (3) (2008) 435–444, https://doi.org/10.1002/cmdc.200700139.

[79] Y.C. Martin, A bioavailability score, J. Med. Chem. 48 (9) (2005) 3164–3170, https://doi.org/10.1021/jm0492002.

[80] J.S. Delaney, ESOL: estimating aqueous solubility directly from molecular structure, J. Chem. Inf. Comput. Sci. 44 (3) (2004) 1000–1005, https://doi.org/10.1021/ci034243x.

[81] J. Ali, P. Camilleri, M.B. Brown, A.J. Hutt, S.B. Kirton, Revisiting the general solubility equation: in silico prediction of aqueous solubility incorporating the effect of topographical polar surface area, J. Chem. Inf. Model. 52 (2) (2012) 420–428, https://doi.org/10.1021/ci200387c.

[82] P. Riyadi, I. Sari, R. Kurniasih, T. Agustini, F. Swastawati, V. Herawati, W. Tanod, SwissADME predictions of pharmacokinetics and drug-likeness properties of small molecules present in spirulina platensis, Paper presented at the IOP Conference Series: Earth and Environmental Science (2021), https://doi.org/10.1088/1755-1315/890/1/012021.

[83] H.D. Williams, N.L. Trevaskis, S.A. Charman, R.M. Shanker, W.N. Charman, C.W. Pouton, C.J. Porter, Strategies to address low drug solubility in discovery and development, Pharmacol. Rev. 65 (1) (2013) 315–499, https://doi.org/10.1124/pr.112.005660.

[84] K.T. Savjani, A.K. Gajjar, J.K. Savjani, Drug solubility: importance and enhancement techniques, Int. Sch. Res. Notices 2012 (2012), https://doi.org/10.5402/2012/195727.

[85] L. Mora Lagares, N. Minovski, M. Novič, Multiclass classifier for P-glycoprotein substrates, inhibitors, and non-active compounds, Molecules 24 (10) (2019) 2006, https://doi.org/10.3390/molecules24102006.

[86] B. Prabha, M. Ezhilarasi, Synthesis, spectral characterization, in vitro and in silico studies of benzodioxin pyrazoline derivatives, Biointerface Res Appl Chem 11 (2020) 9126–9138, https://doi.org/10.33263/BRIAC112.91269138.

[87] M. Athar, A.N. Sona, B.D. Bekono, F. Ntie-Kang, Fundamental physical and chemical concepts behind "drug-likeness" and "natural product-likeness", Physical Sciences Reviews 4 (12) (2019) https://doi.org/10.1515/psr-2018-0101.

[88] P. Ertl, A. Schuffenhauer, Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, J. Cheminf. 1 (2009) 1–11, https://doi.org/10.1186/1758-2946-1-8.

[89] K. Bardhan, K. Liu, Epigenetics and colorectal cancer pathogenesis, Cancers 5 (2) (2013) 676–713, https://doi.org/10.3390/cancers5020676.

[90] E. Holmes, J.V. Li, T. Athanasiou, H. Ashrafian, J.K. Nicholson, Understanding the role of gut microbiome–host metabolic signal disruption in health and disease, Trends Microbiol. 19 (7) (2011) 349–359, https://doi.org/10.1016/j.tim.2011.05.006.

[91] F. Boccellato, T.F. Meyer, Bacteria moving into focus of human cancer, Cell Host Microbe 17 (6) (2015) 728–730, https://doi.org/10.1016/j.chom.2015.05.016.

[92] S. Zou, L. Fang, M.-H. Lee, Dysbiosis of gut microbiota in promoting the development of colorectal cancer, Gastroenterology report 6 (1) (2018) 1–12, https://doi.org/10.1093/gastro/gox031.

[93] J.J. Faith, J.L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A.L. Goodman, R.L. Leibel, The long-term stability of the human gut microbiota, Science 341 (6141) (2013), https://doi.org/10.1126/science.1237439.

[94] J. Gagnière, J. Raisch, J. Veziant, N. Barnich, R. Bonnet, E. Buc, M. Bonnet, Gut microbiota imbalance and colorectal cancer, World J. Gastroenterol. 22 (2) (2016) 501, https://doi.org/10.3748/wjg.v22.i2.501.

[95] E. Valguarnera, J.B. Wardenburg, Good gone bad: one toxin away from disease for Bacteroides fragilis, J. Mol. Biol. 432 (4) (2020) 765–785, https://doi.org/10.1016/j.jmb.2019.12.003.

[96] S. Zamani, R. Taslimi, A. Sarabi, S. Jasemi, L.A. Sechi, M.M. Feizabadi, Enterotoxigenic Bacteroides fragilis: a possible etiological candidate for bacterially-induced colorectal precancerous and cancerous lesions, Front. Cell. Infect. Microbiol. 9 (2020) 449, https://doi.org/10.3389/fcimb.2019.00449.

[97] J.L. Round, S.K. Mazmanian, The gut microbiota shapes intestinal immune responses during health and disease, Nat. Rev. Immunol. 9 (5) (2009) 313–323, https://doi.org/10.1038/nri2515.

[98] C.L. Sears, Enterotoxigenic Bacteroides fragilis: a rogue among symbiotes, Clin. Microbiol. Rev. 22 (2) (2009) 349–369, https://doi.org/10.1128/cmr.00053-08.

[99] S.A. Shiryaev, A.G. Remacle, P. Cieplak, A.Y. Strongin, Peptide sequence region that is essential for the interactions of the enterotoxigenic Bacteroides fragilis metalloproteinase II with E-cadherin, Journal of proteolysis 1 (1) (2014) 3. PMID: 25964952, PMCID: PMC4425422.

[100] S. Wu, K.-C. Lim, J. Huang, R.F. Saidi, C.L. Sears, Bacteroides fragilis enterotoxin cleaves the zonula adherens protein, E-cadherin, Proc. Natl. Acad. Sci. USA 95 (25) (1998) 14979–14984, https://doi.org/10.1073/pnas.95.25.14979.

[101] A. Seelig, P-Glycoprotein: one mechanism, many tasks and the consequences for pharmacotherapy of cancers, Front. Oncol. 10 (2020) 576559, https://doi.org/10.3389/fonc.2020.576559.

[102] T. Hu, Z. Li, C.-Y. Gao, C.H. Cho, Mechanisms of drug resistance in colon cancer and its therapeutic strategies, World J. Gastroenterol. 22 (30) (2016) 6876, https://doi.org/10.3748/wjg.v22.i30.6876.

[103] K. Robinson, V. Tiriveedhi, Perplexing role of P-glycoprotein in tumor microenvironment, Front. Oncol. 10 (2020) 265, https://doi.org/10.3389/fonc.2020.00265.

[104] M. Majid, S. Andleeb, Designing a multi-epitopic vaccine against the enterotoxigenic Bacteroides fragilis based on immunoinformatics approach, Sci. Rep. 9 (1) (2019) 1–15, https://doi.org/10.1038/s41598-019-55613-w.

[105] P. Metz, M.J. Tjan, S. Wu, M. Pervaiz, S. Hermans, A. Shettigar, A. Boleij, Drug discovery and repurposing inhibits a major gut pathogen-derived oncogenic toxin, Front. Cell. Infect. Microbiol. 9 (2019) 364, https://doi.org/10.3389/fcimb.2019.00364.

[106] N.M. Abd El-Aziz, B.E. Khalil, N.N. El-Gamal, Structure prediction, docking studies and molecular cloning of novel Pichia kudriavzevii YK46 metalloprotease (MetPr) for improvement of feather waste biodegradation, Sci. Rep. 13 (1) (2023) 19989, https://doi.org/10.1038/s41598-023-47179-5.

[107] T. Schwede, J. Kopp, N. Guex, M.C. Peitsch, SWISS-MODEL: an automated protein homology-modeling server, Nucleic Acids Res. 31 (13) (2003) 3381–3385, https://doi.org/10.1093/nar/gkg520.

[108] M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, E. Lindahl, GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers, SoftwareX 1 (2015) 19–25, https://doi.org/10.1016/j.softx.2015.06.001.

[109] J. Lee, X. Cheng, J.M. Swails, M.S. Yeom, P.K. Eastman, J.A. Lemkul, W. Im, CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field, J. Chem. Theor. Comput. 12 (1) (2016) 405–413, https://doi.org/10.1021/acs.jctc.5b00935.

[110] K.A. Qureshi, F. Azam, M.Q. Fatmi, M. Imtiaz, D.K. Prajapati, P.K. Rai, G.O. Elhassan, In vitro and in silico evaluations of actinomycin X2and actinomycin D as potent anti-tuberculosis agents, PeerJ 11 (2023) e14502, https://doi.org/10.7717/peerj.14502.