

In-Silico Identification of Antibacterial Molecules in Cosmetic Ingredients: A QSAR and Mechanism-Driven AI Approach

Lin Li¹, Haijun Hong¹, Yan Liang^{*1} Yun Hu²

1 Research & Development, Unilever(China) Limited Shanghai Branch, Shanghai, China

2 DP technology, Beijing, 100080, China, huyun@dp.tech

*Corresponding author: Yan Liang, Research & Development, Unilever(China) Limited Shanghai Branch, 200050, Shanghai, China, +86-15901956414,
Erica.Liang@unilever.com

1. Introduction

Cosmetics products that are not completely sterile are susceptible to microbial contamination, which can pose significant safety risks to consumers. In most cases, such contamination is attributed to Gram Negative (GN) bacteria. In the years 2005-2025, there have been 207 cases on microbiological contaminations in cosmetics reported in the European Rapid Alert system for dangerous products (RAPEX)[1]. As statistically reviewed, GN bacteria are the primary culprits causing contamination (56.04%), where *Pseudomonas.aeruginosa* (*P.aeruginosa*) (31.88%) and Enterobacter species (12.08%) account for the majority. To ensure the microbiological safety of cosmetics products, preservatives are commonly employed world widely, described in the regulation lists include phenoxyethanol, sodium benzoate, benzyl alcohol, etc[2].

However, disputes regarding the safety of some certain conventional preservative and potential for allergic reactions have led to increased scrutiny and stricter regulations requirement. The demand for mild preservatives has accordingly emerged as an industrial trend to meet consumer needs[3]. As previously studied, preservatives could be reduced through enhanced antimicrobial efficacy working with multifunctional ingredients (MFIs) which primarily serve as humectants, antioxidants, emulsifiers, etc, 1,2-diols for example[4]. Polyols are another popular options for their ability to lower water activity to inhibit the growth of microorganisms[2, 4].

But formulation instability or discoloration do sometimes occur when incorporating MFIs. Besides, excess use of MFIs could raise skin irritation problems. Rapidly innovated formulation calls for more comprehensive and effective preservation strategies.

To screen out effective antimicrobial actives from the existing approved ingredients to achieve mild preservation, conventional practice including extensive scouting work and wet lab validation, which is time-consuming and resource-intensive. There comes an urgent need to enhance the overall efficiency of active screening.

With advancements in technology, Artificial intelligence (AI), including deep learning (DL) and machine learning (ML) algorithms, has been revolutionizing biomedical research by tackling the challenges in traditional drug discovery, design and development[5]. Based on this foundation, the AI approach emerges as a powerful tool to aid novel actives finding.

Quantitative Structure Activity Relationship (QSAR) is a computational method that uses mathematical models to predict the biological activity of compounds[6]. It is based on the relationship between the molecular structure of compounds and their biological activity. QSAR has made significant progress in the field of antimicrobial molecular prediction[7].

While QSAR models effectively predict compounds structurally similar to training data, addressing limitations for structural outliers is essential.

Structure-based virtual screening (SBVS) discovers and constructs the 3D structure of target proteins through in-depth exploration of the mechanism of action, and screens potential bioactive molecules by analyzing binding interactions. SBVS does not rely on known active compounds, making it possible to discover novel ligands with significant differences in structure from existing ones. In the discovery of antibacterial drugs for GN bacteria, SBVS usually targets what are crucial for bacterial growth and survival, such as fatty acid synthase [8], peptidoglycan synthase[9], DNA gyrase B[10], Dihydropteroate synthase (DHPS) [11], LasR[12], etc. Based on the screening of these targets, some new active compounds were discovered, such as N- (3- (5-bromo-2-hydroxybenzylideneamino) propyl) -2-hydroxybenzamide (MIC 0.39-3.13 lg/mL) [13], zidovudine[11], rosmarinic acid, naringin, chlorogenic acid, morin and mangiferin[12], etc

In the light of the facts above, a unique AI model combining QSAR model with Mechanism-driven virtual screening model was established. It identified 124 non-preservative molecules, from the Inventory of Cosmetic Ingredients in China, that were predicted to demonstrate notable antibacterial efficacy, out of which 3 molecules were validated to show significant antimicrobial efficacy.

2. Materials and Methods

2.1. Materials

Gram negative bacteria: *Pseudomonas.aeruginosa* (*P. aeruginosa*) ATCC 9027, *Burkholderia cepacian* (*B. cepacian*) ATCC 25416, pooled as Non-Fermentative GN (NFGN) bacteria; *Pluralibacter gergoviae* (*P. gergoviae*) ATCC 33028, *Klebsiella pneumoniae* (*K. pneumoniae*) ATCC 10031, pooled as GN Fermentative bacteria (GNF).

Gram positive bacteria: *Staphylococcus aureus* (*S. aureus*) ATCC 6538.

All the strains were aquired from the American Type Culture Collection. Tryptic Soy Broth for bacteria growth and suspension is from Oxoid (CM:0129B).

2.2. Methods

2.2.1. Construction of data-driven QSAR model

We collected 2 datasets with one containing about 17k pieces of data[7] and the other containing 120 pieces of data related to cosmetic ingredients. These molecules exhibited antibacterial activity against *Escherichia coli* (*E.coli*). Chemical structures were retrieved from PubChem dataset using Chemical Abstracts Service (CAS) registry numbers and chemical names, then removed duplicate structures using DataWarrior to avoid the impact of

duplicate data on model performance[14]. The t-SNE plot[14] was used to visualize the chemical spatial distribution of the two datasets.

In terms of dataset splitting, we randomly split 80% of the data as the training dataset and 20% as the test set. The 5-fold cross-validation procedure was used to train the model.

2.2.1.1. QSAR Modelling

Uni-QSAR[15] is an automated Machine Learning (Auto-ML) tool for building molecular property prediction model, which integrates multidimensional (1D, 2D, 3D) molecular representations with multiple algorithms, performs automated hyperparameter tuning, and finally generates the model based on the stacking strategy. The structures and antibacterial properties of these molecules above were uploaded to the Uni-QSAR platform to construct a model of molecular antibacterial properties. 7 basic models were selected to build the model as shown in Table 1 and the default parameters were used to train the model as shown in Table 2.

Table 1 Base models selected to build the model

Models	Explanation	Functions
Uni-Mol-All_H	A kind of 3D pretrained model that incorporates hydrogen atoms in its molecular representation.	Leveraging 3D spatial information to enhance predictive accuracy for molecular properties.
UniMol-No_H	Using 3D molecular representations but excludes hydrogen atoms from its pretrained model.	Simplifying the representation while still capturing essential spatial information.
BERT-SMILES	BERT: Bidirectional Encoder Representations from Transformers, an advanced model originally designed for natural language processing.	Typically using SMILES strings as input to capture sequential patterns in molecular representations when adapted for molecular property prediction.
LR-FP	LR: Logistic Regression FP: Molecular Fingerprint	LR: classifying binary FP: simplifying complex chemical information and encodes molecular structure into binary numerical strings, based on which molecular similarity to be assessed.
GBDT-MD	GBDT: an ensemble learning method that builds multiple decision trees sequentially, each trying to correct errors from the previous one.	GBDT: highly effective for both regression and classification tasks MD: quantify the physicochemical properties and structural features of molecules.

	MD: Molecular Descriptors, a set of numerical values	
ET-MD	ET: Extremely Randomized Trees, an ensemble method similar to Random Forests but with more randomness.	ET: better generalization and reduced overfitting.
SVM-FP	SVM: a powerful model for classification and regression tasks.	Finding the hyperplane that best separates the data into different classes or predicts continuous values. Kernel functions can be used to handle non-linear relationships.

Table 2 Default parameters

Hyperparameter	Explanation
Epoch	Increasing the number of epochs can enhance the training of a model, but it also increases the risk of overfitting.
Batch Size	Smaller: more frequent updates and good model generalization, but potentially reducing computational efficiency
Range	Lager: more computational efficiency, but training instability or negatively impact generalization.
Learning Rate	Smaller: converges stably, but slower training process
Range	Larger: rapid convergence, but unstable training process or miss the optimal solution.

2.2.1.2. QSAR Model for prediction

The constructed QSAR model was utilized to predict potential antimicrobial active ingredients from the Inventory of Cosmetics Ingredients in China. The cosmetic ingredients were first filtered to remove mixtures, proteins, enzymes, polysaccharides, inorganic salts, organometallic compounds, and polymers, retaining only monomer compounds. Subsequently, the chemical structures of these compounds were identified in the PubChem database using their CAS numbers and molecular names then excluded the unfound corresponding structure. Finally, the chemical structures of the compounds were input into the model to predict their antimicrobial activity.

2.2.2. Mechanism-based model

2.2.2.1. Target selection and structure preparation

After comprehensively searching the existing literature, we identified 5 targets covering different mechanisms: fatty acid synthase, the synthesis of cell walls, DNA replication, transcription and repair, folic acid synthesis, and quorum sensing.

Fatty acid biosynthesis is essential to GN pathogens growth, FabH (β -ketoacyl-acyl carrier protein synthase III) is one of the key enzymes[16]. Due to the lack of FabH crystal structure of *P. aeruginosa*, protein sequence (Uniprot database, Entry ID Q9HYR2) of *E.coli* on AlphaFold3 was performed as it's been widely used as a target. We assessed the

AlphaFold3-predicted FabH structure by docking the *E. coli* FabH inhibitor (1HNJ) into its binding pocket, observed collision, and subsequently optimized the pocket conformation via Hermit's "Binding Stability" molecular dynamics simulations.

Degradation of structure and function of cell wall can inhibit GN bacteria growth. MurG (PDB ID: 3S2U) is an essential bacterial glycosyltransferase enzyme in *P.aeruginosa* performing peptidoglycan synthesis[17].

The DNA gyrase B subunit (gyrB) serves as a key antibiotic target, particularly for fluoroquinolones. The crystal structure of the *P.aeruginosa* gyrB (PDB ID 7PTF) was selected for further study[18].

Folic acid is necessary for bacterial proliferation, with dihydropteroate synthase (DHPS) playing a crucial role in its synthesis. Although no suitable DHPS crystal structure was observed for *P.aeruginosa*, drugs virtually screened against *E.coli* (PDB ID: 5U12) also showed cross-activity experimentally, suggesting its utility as target for *P.aeruginosa*[19].

GN bacteria employ autoinducer-mediated quorum sensing (QS) to regulate virulence and antimicrobial resistance[20]. QS inhibitors shown to attenuate *P. aeruginosa* pathogenicity, prompting our structural study using the LasR target (PDB: 2UV0).

2.2.2. Molecular docking study

Virtual screening was performed using the Hermite computational platform. Potential antibacterial ingredients are screened from the monomer compounds pretreated. The target structure is prepared using the "Protein Preparation" module, including adding hydrogen, energy minimization, optimization of hydrogen bond network, protonation state optimized at pH 7.4. The molecular structure is prepared using the Ligand Preparation module, including generating isomers, adjusting protonation state at pH 7.4 and generating multiple conformations.

Molecular docking was conducted using the Vina scoring. Full docking parameters are documented in Table 3

Table 3 Pocket configuration

Pocket Config	Center_X	Center_Y	Center_Z	Size_X	Size_Y	Size_Z
FabH_Q9HYR_2	35.025000 Å	20.205000 Å	34.380000 Å	19.469997	29.290000	17.379999
3S2U	15.050000 Å	-17.433000 Å	-19.837500 Å	19.089999	15.278001	24.813000
7PTF	17.508000 Å	-9.061500 Å	51.864500 Å	17.825999	22.979001	21.118999
5U12	-1.924000 Å	-7.256500 Å	69.500000 Å	13.842000	18.809000	14.895996

2UV0	35.025000 Å	20.205000 Å	34.380000 Å	19.469997	29.290000	17.379999
-------------	-------------	-------------	-------------	-----------	-----------	-----------

2.3. In vitro antimicrobial activity

The antimicrobial activity of the compounds screened by the model previously established were tested against bacteria: dilutions of test (20–0.1mg/mL) compounds were prepared and the Minimum inhibitory concentration (MIC) determination was carried out using the 2-fold broth dilution method as described previously[21]. The samples were incubated at 37°C for 24h, then recorded the results.

3. Result

3.1. Data-driven QSAR model

The t-SNE is a non-linear dimensionality reduction technique that can map high-dimensional data to low-dimensional spaces (such as 2D and 3D). According to the t-SNE graph (Fig 1) the chemical space of the 2 datasets overlaps well, which means that incorporation makes it more applicable.

The QSAR model for anti GN bacterial activity demonstrated exceptional predictive capability. As shown in Table 4 and Fig 2 the model achieved an area under the receiver operating characteristic curve (AUC) exceeding 0.98 in both the validation and test datasets, indicating its effectiveness in accurately identifying molecules with anti-GN bacterial potential by using precise molecular features and algorithms to capture chemical-biological activity relationship[7].

The predictive outcomes identified 976 cosmetic ingredients as having potential anti-bacterial activity against GN bacteria, guiding subsequent *vitro* experimental validation.

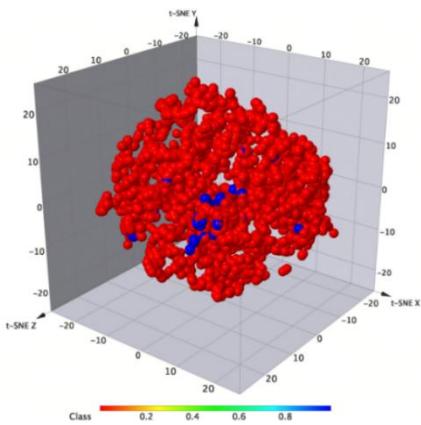


Figure 1 t-SEN plot

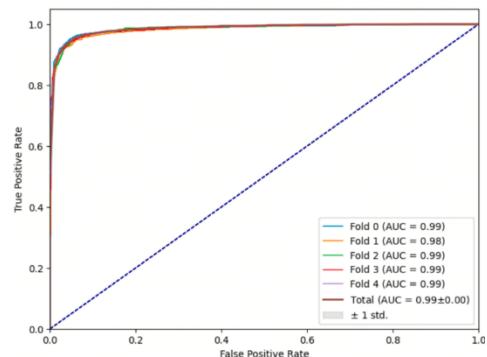


Figure 2 ROC curve

Table 4 Model Performance

Dataset	AUC	AUPRC	F1-Score	ACC	Precision	Recall
Validation	0.986	0.985	0.939	0.949	0.947	0.932
Test	0.990	0.989	0.941	0.950	0.947	0.935

3.2. Mechanism-based model

Molecules from cosmetic ingredients library were processed as described in 2.2.1.3.

Structural superposition of FabH-Q9HYR2 predicted by AlphaFold3 with 1HNJ revealed steric clashes (Fig. 3a), demonstrating the need for pocket conformation optimization to improve stability (Fig. 3b). Redocking validation confirmed ligand pose reproducibility (Fig. 3c), verifying the system's reliability for subsequent virtual screening. The suitability of Uni-docking was also confirmed by the good overlap between original and redocked ligands of MurG, GryB, DHPS and LasR (Fig.4a, 4b, 4c, 4d).

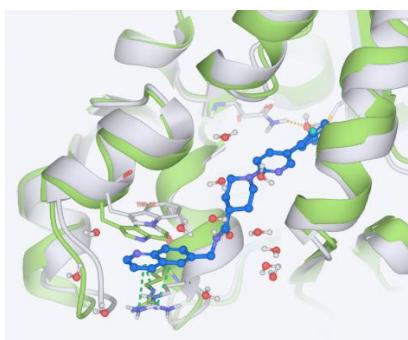


Figure 3a. The ligand collides with

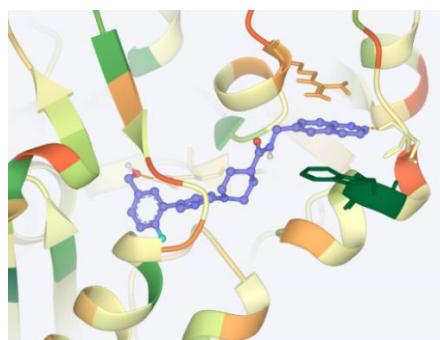


Figure 3b. Obtaining stable binding conformations using molecular dynamics

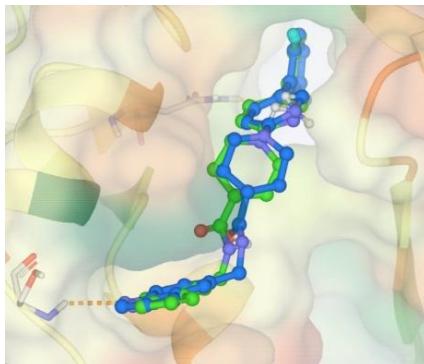


Figure 3c Redocking results

Virtual screening from the based on the 5 target proteins was performed, retaining the top 50% of the calculation results. 764 molecules for FabH-Q9HYR2 (Vina score <-7), 192 molecules for MurG (Vina < -8, Ginna > 0.8), 221 molecules for GryB (Vina < -7, Ginna > 0.8), 212 molecules for DHPS (Vina < -6, Ginna > 0.8) and 1722 molecules for LasR (Vina<-7) were obtained. Then conducted visually inspecting to remove molecules far from the key binding regions, assess structural complementarity and interaction patterns, evaluate binding

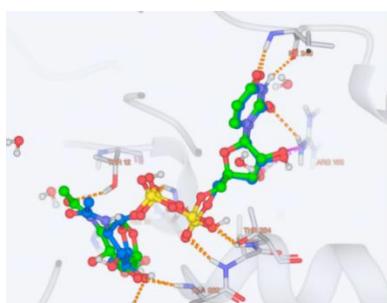


Figure 4a. MurG redocking

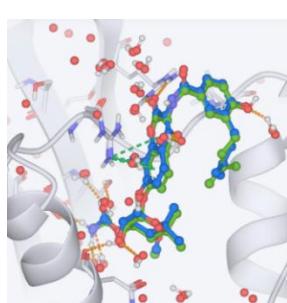


Figure 4b. GryB redocking

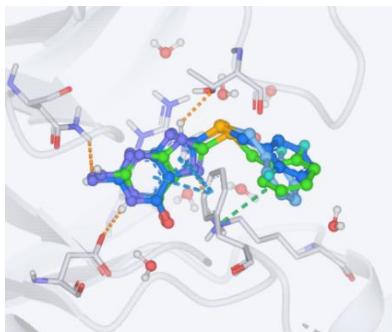


Figure 4c. DHPS redocking

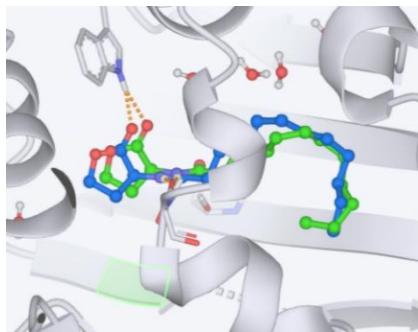


Figure 4d. LasR redocking

modes, etc., 127 molecules for FabH-Q9HYR2, 32 molecules for MurG, 25 molecules for GryB, 77 molecules for DHPS, and 72 molecules for LasR were ultimately screened out.

The screening results from the 5 targets were consolidated, yielding 155 non-redundant candidates. QSAR prediction of these 155 molecules identified 124 with antimicrobial potential, demonstrating good consistency between the 2 independent models.

3.3. In vitro antimicrobial activity

Compound X1, X2, X3 screened by the established AI model were selected to be validated. Phenoxyethanol was selected as positive control for commonly used as conventional preservative.

As shown in table 5, compound X1,X2,X3 all demonstrate outstanding antibacterial capability, range from 0.01-0.075%, much better than Phenoxyethanol as reference. Interestingly, all of these 3 compounds show an excellent inhibitory effect on *S.aureus*.

Table 5 MIC value of compound X1,X2,X3

MIC (%)	X1	X2	X3	Phenoxyethanol
Anti GNF	0.025	0.05	0.075	0.36
Anti NFGN	0.025	0.075	0.05	0.36
Anti <i>S.aureus</i>	0.03125	0.05	0.01	0.64

4. Discussion

Preservatives prevent microbial contamination in cosmetics while ensuring consumer safety, but safety concerns are driving demand for milder ones. Given the complexity of cosmetics products and the inherently low permeability of GN bacteria, finding novel antimicrobial actives is challenging. Traditional trial-and error approaches are inefficient, necessitating more intelligent and more precise methods for safer cosmetics products development.

In this work, an AI approach was employed to provide an innovative and powerful framework for anti GN bacteria discovery. The data-driven QSAR model predicted a total of 976 molecules from the Inventory of Cosmetic Ingredients in China, while mechanism-driven model predicted a total of 155 non-redundant molecules. Through cross-validation, 124 non-preservative molecules were identified as potential antimicrobial actives. 3 molecules selected from those experimentally validated to possess excellent antimicrobial properties, well confirming the reliability of this AI model, and greatly contributing to the development of milder-preserved or even preservative-free cosmetics.

Limitations remain, requiring for future validation of the specificity and potency of these inhibitors under physiologically relevant cosmetic conditions.

5. Conclusion

In this work, we identified 124 molecules within the Inventory of Cosmetic Ingredients in China by utilizing AI approach integrated data-driven QSAR model and mechanism-driven model. Firstly, we leveraged a large-scale pre-training model fine-tuned by antimicrobial active data to establish a model linking molecular structure to preservative efficacy. Secondly, we identified 5 key targets related to antimicrobial activity and utilized AI-optimized molecular docking technology to construct a virtual screening platform. Validated by wet lab experiments, we successfully identified 3 molecules that exhibited excellent antimicrobial efficacy. By adopting this AI approach, we accelerate novel antimicrobial actives development, providing a reliable technological guarantee for safer cosmetic products.

References

- [1] E.R.A.s.f.d.p. (RAPEX). <https://ec.europa.eu/safety-gate-alerts/screen/search>.
- [2] N. Halla, I.P. Fernandes, S.A. Heleno, P. Costa, Z. Boucherit-Otmani, K. Boucherit, A.E. Rodrigues, I. Ferreira, M.F. Barreiro, Cosmetics Preservation: A Review on Present Strategies, *Molecules* 23(7) (2018).
- [3] P. Rathee, R. Sehrawat, P. Rathee, A. Khatkar, E.K. Akkol, S. Khatkar, N. Redhu, G. Turkcanoglu, E. Sobarzo-Sanchez, Polyphenols: Natural Preservatives with Promising Applications in Food, Cosmetics and Pharma Industries; Problems and Toxicity Associated with Synthetic Preservatives; Impact of Misleading Advertisements; Recent Trends in Preservation and Legislation, *Materials (Basel)* 16(13) (2023).
- [4] A. Herman, Antimicrobial Ingredients as Preservative Booster and Components of Self-Preserving Cosmetic Products, *Curr Microbiol* 76(6) (2019) 744-754.
- [5] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R.K. Ambasta, P. Kumar, Artificial intelligence to deep learning: machine intelligence approach for drug discovery, *Mol Divers* 25(3) (2021) 1315-1360.
- [6] Y.L. Wang, F. Wang, X.X. Shi, C.Y. Jia, F.X. Wu, G.F. Hao, G.F. Yang, Cloud 3D-QSAR: a web tool for the development of quantitative structure-activity relationship models in drug discovery, *Brief Bioinform* 22(4) (2021).
- [7] Y.A. Ivanenkov, A. Zhavoronkov, R.S. Yamanov, I.A. Osterman, P.V. Sergiev, V.A. Aladinskiy, A.V. Aladinskaya, V.A. Terentiev, M.S. Veselov, A.A. Ayginin, V.G. Kartsev, D.A. Skvortsov, A.V. Chemeris, A.K. Baimiev, A.A. Sofronova, A.S. Malyshev, G.I. Filkov, D.S. Bezrukov, B.A. Zagribelnyy, E.O. Putin, M.M. Puchinina, O.A. Dontsova, Identification of Novel Antibacterials Using Machine Learning Techniques, *Front Pharmacol* 10 (2019) 913.
- [8] H.H. Amer, E.H. Eldrehmy, S.M. Abdel-Hafez, Y.S. Alghamdi, M.Y. Hassan, S.H. Alotaibi, Antibacterial and molecular docking studies of newly synthesized nucleosides and Schiff bases derived from sulfadimidines, *Sci Rep* 11(1) (2021) 17953.

- [9] P. Mahur, A.K. Singh, J. Muthukumaran, M. Jain, Targeting MurG enzyme in *Klebsiella pneumoniae*: An in silico approach to novel antimicrobial discovery, *Research in Microbiology* 176(1-2) (2025) 104257.
- [10] N.G. Bush, I. Diez-Santos, L.R. Abbott, A. Maxwell, Quinolones: mechanism, lethality and their contributions to antibiotic resistance, *Molecules* 25(23) (2020) 5662.
- [11] M. Shaw, A. Petzer, J.P. Petzer, T.T. Cloete, The pterin binding site of dihydropteroate synthase (DHPS): In silico screening and in vitro antibacterial activity of existing drugs, *Results in Chemistry* 5 (2023) 100863.
- [12] A. Annappoorani, V. Umamageswaran, R. Parameswari, S.K. Pandian, A.V. Ravi, Computational discovery of putative quorum sensing inhibitors against LasR and RhIR receptor proteins of *Pseudomonas aeruginosa*, *Journal of computer-aided molecular design* 26 (2012) 1067-1077.
- [13] A.M. El-Saghier, L. Abosella, E.O. Elakesh, G.E.-D.A. Abuo-Rahma, A. Abdou, A.M. Hamed, Synthesis, characterization, molecular docking, and antimicrobial activities of some new sulfur containing norfloxacin analogues, *Journal of Molecular Structure* 1307 (2024) 137916.
- [14] E. López-López, J.J. Naveja, J.L. Medina-Franco, DataWarrior: An evaluation of the open-source drug discovery tool, *Expert Opinion on Drug Discovery* 14(4) (2019) 335-341.
- [15] Z. Gao, X. Ji, G. Zhao, H. Wang, H. Zheng, G. Ke, L. Zhang, Uni-qsar: an auto-ml tool for molecular property prediction, *arXiv preprint arXiv:2304.12239* (2023).
- [16] G. Sabbagh, N. Berakdar, Molecular docking study of flavonoid compounds as inhibitors of β -ketoacyl acyl carrier proteinsynthase ii (kas ii) of *Pseudomonas aeruginosa*, *Int J Pharm Pharm Sci* 8(1) (2016) 52-61.
- [17] K. Brown, S. CM Vial, N. Dedi, J. Westcott, S. Scally, T. DH Bugg, P. A Charlton, G. MT Cheetham, Crystal structure of the *Pseudomonas aeruginosa* MurG: UDP-GlcNAc substrate complex, *Protein and Peptide Letters* 20(9) (2013) 1002-1008.
- [18] A.E. Cotman, M. Durcik, D. Benedetto Tiz, F. Fulgheri, D. Secci, M. Sterle, S. Mozina, Z. Skok, N. Zidar, A. Zega, Discovery and hit-to-lead optimization of benzothiazole scaffold-based DNA gyrase inhibitors with potent activity against *Acinetobacter baumannii* and *Pseudomonas aeruginosa*, *Journal of medicinal chemistry* 66(2) (2023) 1380-1425.
- [19] M.L. Dennis, M.D. Lee, J.R. Harjani, M. Ahmed, A.J. DeBono, N.P. Pitcher, Z.C. Wang, S. Chhabra, N. Barlow, R. Rahmani, 8-mercaptoguanine derivatives as inhibitors of dihydropteroate synthase, *Chemistry—A European Journal* 24(8) (2018) 1922-1930.
- [20] M.J. Bottomley, E. Muraglia, R. Bazzo, A. Carfi, Molecular insights into quorum sensing in the human pathogen *Pseudomonas aeruginosa* from the structure of the virulence regulator LasR bound to its autoinducer, *Journal of Biological Chemistry* 282(18) (2007) 13592-13600.
- [21] W. PA, Reference method for broth dilution antifungal susceptibility testing of yeasts, approved standard, CLSI document M27-A2 (2002).