JURNAL FARMASI SAINS DAN KOMUNITAS
Journal of Pharmaceutical Sciences & Community

p-ISSN 1693-5683; e-ISSN 2527-7146

Vol. 23, No. 1, Mei 2026, pp. 65-74

Research Article

Employing Ensemble Protein-Ligand Interaction Fingerprints to Mimic Induced-Fit Theory in Structure-Based Virtual Screening Targeting Dipeptidyl Peptidase IV

Enade P. Istyastono[1,2], Bonifacius I. Wiranata[1,3], Florentinus D.O. Riswanto[4], Fransiska Kurniawan[5], Tasia Amelia[5], Nunung Yuniarti[6], Eko A. Prasetyanto[2,*]

[1] MOLMOD ID (Molmod Jaya Sejahtera Ltd.; https://molmod.id), Pogung Kidul, Sinduadi, Mlati, Sleman, Yogyakarta, 55244, Indonesia
[2] Department of Pharmacy, Faculty of Medicine and Health Sciences, Atma Jaya Catholic University of Indonesia, Jalan Pluit Raya 2, Jakarta 14440, Indonesia
[3] Pharmaceutical Sciences Department, Faculty of Pharmacy, Widya Mandala Catholic University, Surabaya, Indonesia
[4] Faculty of Pharmacy, Sanata Dharma University, Yogyakarta 55282, Indonesia
[5] School of Pharmacy, Bandung Institute of Technology, Jalan Ganesha 10, Bandung 40132, Indonesia
[6] Department of Pharmacology and Clinical Pharmacy, Faculty of Pharmacy, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia

https://doi.org/10.24071/jpsc.v23i1.1070

J. Pharm. Sci. Community, 2026, 23(1), 65-74

Article Info

Received: 2025-03-15
Revised: 2025-07-10
Accepted: 2025-07-12

***Corresponding author:**
Eko A. Prasetyanto
email:
prasetyanto@atmajaya.ac.id

Keywords:
AutoDock Vina; dipeptidyl peptidase IV; drug discovery; PyPLIF HIPPOS; YASARA-Structure

ABSTRACT

We have successfully employed PyPLIF HIPPOS in retrospective Structure-Based Virtual Screening (SBVS) campaigns targeting some G-protein coupled receptors (GPCRs), which could pinpoint the molecular determinants of the protein-ligand bindings and increase the quality of the SBVS protocols. We were then tempted to append with molecular dynamics simulations using YASARA-Structure to mimic the induced-fit theory in the construction of SBVS protocols targeting dipeptidyl peptidase IV (DPP4). The protocol was retrospectively validated by employing the DPP4 ligands and decoys provided by the Directory of Useful Decoys: Enhanced (DUDE). The best SBVS protocol from this research has the balanced accuracy (BA) value of 0.836, which could be used further in prospective screening campaigns.

INTRODUCTION

Structure-based virtual screening (SBVS) approaches have benefited from molecular interaction fingerprints (IFP) (Marcou and Rognan, 2007) to increase the prediction quality in both retrospective and prospective screening campaigns (Istyastono et al., 2015; Kooistra et al., 2015). In 2013, we introduced the Python implementation of the IFP named PyPLIF, which has successfully served our SBVS campaigns targeting estrogen receptor α (ERα) (Istyastono et al., 2017), cyclooxygenase-2 (COX2) (Istyastono, 2017), and acetylcholinesterase (AChE) (Istyastono and Prasasty, 2021; Prasasty et al., 2018; Prasasty and Istyastono, 2020; Riswanto et al., 2021, 2017). A descriptor derived from the protein-ligand interaction fingerprints (PLIF) identified by PyPLIF that combines all PLIF from the docking poses called ensemble PLIF (ensPLIF) was introduced and has played an essential role in increasing the prediction quality of the SBVS protocols (Istyastono, 2017; Istyastono et al., 2017; Riswanto et al., 2017) This ensPLIF resembles the lock-and-key theory incorporated in the SBVS approach (Istyastono et al., 2017). PyPLIF was recently upgraded to PyPLIF HIPPOS, which has more features and is 10 times faster compared to the predecessor (Istyastono et al., 2020). Notably, together with a machine learning technique called Recursive Partition and

Regression Trees (RPART) (Therneau et al., 2015), ensPLIF could identify the plausible molecular determinants of the ligand binding in retrospective SBVS campaigns (Istyastono, 2017; Istyastono et al., 2017; Riswanto et al., 2017). The ability to pinpoint the plausible molecular determinants of the ligand binding was retrospectively confirmed by employing ensPLIF derived from PLIF identified by PyPLIF HIPPOS in SBVS targeting several G protein-coupled receptors (GPCRs), i.e., Adenosine A2a receptor (AA2AR), β2 adrenergic receptor (ADRB2), C-X-C chemokine receptor type 4 (CXCR4), and Dopamine D3 receptor (DRD3) (Istyastono et al., 2021).

The previous uses of ensPLIF in SBVS campaigns have not considered protein flexibility (Istyastono, 2017; Istyastono et al., 2021, 2017; Riswanto et al., 2017). Thus, although the ensPLIF took into account multiple docking poses (Istyastono, 2017; Istyastono et al., 2021, 2017; Riswanto et al., 2017), the induced-fit theory (Koshland, 1994) was not fully implemented in the SBVS protocols. The two most popular approaches to implementing the induced-fit theory in SBVS protocols are induced-fit docking (IFD) and receptor ensemble docking (RED) (Wang et al., 2020). AutoDock Vina could perform IFD by making particular residues flexible in the docking simulations (Eberhardt et al., 2021). Unfortunately, PyPLIF HIPPOS was not designed to identify PLIF resulted from IFD using AutoDock Vina (Istyastono et al., 2020). Therefore, the RED approach is more suitable for PyPLIF HIPPOS-assisted mimicking induced-fit theory in SBVS. Moreover, there were several pieces of evidence showing that employing the RED approach resulted in a better prediction quality compared to the IFD approach (Antunes et al., 2015).

The research presented in this article aimed to mimic the induced-fit theory (Koshland, 1994) in SBVS targeting dipeptidyl peptidase IV (DPP4) (Cao et al., 2021) with the assistance of ensPLIF derived from PyPLIF HIPPOS (Istyastono et al., 2021, 2020). The enzyme DPP4 was selected since it is the enzyme of interest in the drug discovery for diabetes mellitus type 2 (DMT2) therapy (Li et al., 2018). Several natural compounds were reported as potent DPP4 inhibitors, which were identified both in silico and in vitro (Cao et al., 2021; Fan et al., 2013). These natural compounds could be potentially employed in the diet for diabetes management (Cao et al., 2021; Fan et al., 2013)The SBVS protocol, as presented in this article, could be prospectively employed to screen compounds, both naturally occurring and synthetically occurring, to identify DPP4 inhibitors.

METHODS

Materials

The DPP4 crystal structure with alogliptin as the co-crystalized ligand 2ONC.pdb was obtained from https://www.rcsb.org/structure/2ONC (accessed on 21 February 2022) (Feng et al., 2007). The active set of DPP4 inhibitors (actives_final.ism) and the decoy set (decoys_final.ism) in the SMILES form were downloaded from http://dude.docking.org/targets/dpp4 (accessed on 3 March 2022) (Mysinger et al., 2012). The computational simulations were performed in 3 work stations: (i) A 64-bit Windows 11 Home personal computer (PC-client) with Intel® Core™ i5-1135G7 @ 2.40GHz as the processor, NVIDIA® GeForce® MX350® as the graphical processing unit (GPU), and 8 GB of random-access memory (RAM). The main software in this machine involved in this research was YASARA-Structure version 21.12.19 (Krieger and Vriend, 2015) for visual inspections and preparations requiring visual inspections; (ii) A 64-bit Linux (Ubuntu 20.04.3 LTS) virtual private server (VPS-1) with 8 cores of Intel® Xeon® E5-2680 v3 @ 2.50GHz as the processors and 16 GB of RAM. The main software in this machine involved in this research was YASARA-Structure version 21.12.19 (Krieger and Vriend, 2015) for performing MD simulations and analysis; (iii) A 64-bit Linux (Ubuntu 18.04.6 LTS) virtual private server (VPS-2) with 32 cores of Intel® Xeon® Gold 5218 CPU @ 2.30GHz and 16 GB of RAM. The software involved in this research was YASARA-Structure version 21.12.19 (Krieger and Vriend, 2015), preparation scripts for AutoDockTools (Morris et al., 2009) (https://anaconda.org/InsiliChem/autodocktools-prepare; accessed on 21 March 2022), AutoDock Vina version 1.1.2, PLANTS version 1.2, SPORES, PyPLIF HIPPOS 0.1.2 (Istyastono et al., 2021, 2020), and R version 3.4.4 (R Core Team, 2019).

Procedures

Molecular dynamics simulations

The 2ONC.pdb was downloaded to the PC-client and uploaded to YASARA-Structure GUI. Only atoms in Chain A of the DPP4 and the SY1 800 residue of the Chain A were selected for further simulations. The modules “Edit > Build > N-terminal Loop” and “Edit > Build > Loop” were used to recover the missing DPP4 residues in 2ONC.pdb, which were identified by the “SeqRes” module. The pH system was set to 7.4 and the hydrogens and bond orders were then adjusted.

The receptor and the ligand were split into 2 objects, i.e., 2ONCR and 2ONCL, respectively. The system was subsequently energy minimized using the default setting. The simulation cell was then defined as 10 Å around all atoms. The cell boundaries and the Force Field (FF) were set to "Periodic" and AMBER14, respectively. The system was ready to be employed as the input for MD simulations. The system was stored as 2ONC.sce (Supporting File S1).

The file 2ONC.sce was copied to the VPS-1. In the same directory, a macro file md_run.mcr (Supporting File S2) was created. YASARA-Structure in the text mode was employed to run the md_run.mcr to perform the MD simulations for 50 ns with the 2ONC.sce as the starting point. The following is the summary of MD simulations by the md_run.mcr: "The simulation was run with YASARA. The setup included optimizing the hydrogen bonding network to increase the solute stability, and a pKa prediction to fine-tune the protonation states of protein residues at the chosen pH of 7.4. NaCl ions were added with a physiological concentration of 0.9%, with either Na or Cl excess to neutralize the cell. After the steepest descent and simulated annealing minimizations to remove clashes, the simulation was run for 50 ns using the AMBER14 force field for the solute, GAFF2, and AM1BCC for ligands, and TIP3P for water. The cut-off was 8 A for Van der Waals forces, no cut-off was applied to electrostatic forces (using the Particle Mesh Ewald algorithm). The equations of motion were integrated with a multiple timestep of 1.25 fs for bonded interactions and 2.5 fs for non-bonded interactions at a temperature of 310K and a pressure of 1 atm (NPT ensemble) using algorithms described in detail previously (Krieger and Vriend, 2015)." After inspection of the solute root-mean-squared deviation (RMSD) as a function of simulation time (Figures 1 and 2), the first five ns were considered equilibration time and excluded from further analysis. The average free energy of binding was calculated using the Molecular Mechanics/Poisson–Boltzmann Surface Area (MM/PBSA) using YASARA Structure (Patel et al., 2021). The snapshot files were converted from sim to pdb, which were subsequently subjected to PyPLIF HIPPOS (Istyastono et al., 2020) for PLIF identification during the MD simulations following the procedure published previously by Istyastono and Gani (Istyastono and Gani, 2021). The clustering analysis was performed with the minimum heavy-atom RMSD between different clusters of 2.0 Å. The resulting clusters were energy minimized and stored as pdb files (Supporting File S3).

Virtual targets preparation for retrospective SBVS targeting DPP4

The pdb files of the clusters were uploaded to the VPS-2. The module "splitpdb" from SPORES (ten Brink and Exner, 2009) was subjected to each cluster, which resulted in the following mol2 files: protein.mol2, water.mol2, and ligand_SY1800_0.mol2. The file protein.mol2 from each cluster was subjected to the preparation script prepare_receptor4.py (Morris et al., 2009) resulting in protein.pdbqt files readily for molecular docking simulations using AutoDock Vina (Eberhardt et al., 2021).

Ligands preparation for retrospective SBVS targeting DPP4

The preparation was performed in the VPS-2. The files actives_final.ism and decoys_final.ism were downloaded from http://dude.docking.org/targets/dpp4 (accessed on 3 March 2022) (Mysinger et al., 2012). The second column of the file actives_final.ism was removed. The rest of SMILES structure lines from actives_final.ism were appended to the file decoys_final.ism, and then the file was stored as dpp4-compounds.smi (Supporting File S4). A macro file smi2pdb.mcr (Supporting File S5) in the same directory as the file dpp4-compounds.smi was created to convert the structures from SMILES to pdb. The following is the summary of the smi2pdb.mcr: "The simulation was run with YASARA. The SMILES for each line in dpp4-compounds.smi was built into its three-dimensional (3D) form. The pH system was set to pH 7.4, and the hydrogens were updated. The compound was then energy minimized using NOVA as the force field (FF) followed by structure optimization using the semiempirical method AM1 (Krieger and Vriend, 2015). The optimized structure was saved as a pdb file." The resulting pdb files were then subjected to the preparation script prepare_ligand4.py (Morris et al., 2009) resulting in pdbqt files readily for molecular docking simulations using AutoDock Vina (Eberhardt et al., 2021). During this ligand preparation step, 44 SMILES structures of the decoys could not be converted into 3D forms in the pdbqt format. Since these compounds would not result in any docking poses, then they were predicted as inactive (N) in the further analysis.

Automated molecular docking simulations using AutoDock Vina

The docking simulations were performed in the VPS-2 using a similar method published by Istyastono et al. (2021). The generic configuration

for the docking simulations was set as follows: energy_range = 5 and cpu = 5, while the other options were left default. The XYZ coordinate position and the size of the docking box were specific for each virtual target (*vide supra*). The center of the SY1 residue of each virtual target was set as the XYZ coordinate position, and the distance of 5 Å from the surface of the residue was used to calculate the docking box size. The module "bind" from PLANTS was used to obtain the XYZ coordinate positions' values and the docking boxes' size. All prepared ligands were docked to all clusters using AutoDock Vina (Eberhardt et al., 2021) in parallel using the 32 processors from the VPS-2.

Ensemble docking scores and ensPLIF calculations

The ensemble docking scores and ensPLIF calculations were performed in the VPS-2 using a method similar to that published by Istyastono et al. (Istyastono et al., 2021). The configuration file to perform PLIF identification by PyPLIF HIPPOS required a list of residues in the binding pocket (Istyastono et al., 2020). Since the configuration file used for the docking results from different clusters, a consensus list of residues was created by combining all unique residues identified using the module "bind" from PLANTS (see 4.2.4). The following was the consensus list of residues used in the configuration file to run PyPLIF HIPPOS: Arg125, Trp201, Glu205, Glu206, Ser209, Tyr256, Phe357, Val546, Tyr547, Ala548, Gly549, Pro550, Cys551, Ser552, Gln553, Lys554, Leu598, Trp629, Ser630, Tyr631, Gly632, Gly633, Tyr634, Val653, Ala654, Pro655, Val656, Trp659, Tyr662, Asp663, Tyr666, Thr667, Arg669, Tyr670, Asp708, Asp709, Asn710, Val711, His740, and Gly741.

Employing the configuration files, the PLIF identifications were performed for all docking poses resulting from the retrospective SBVS (*vide supra*). The option "nobb" to neglect the interaction with the backbone atoms of the protein was used (Istyastono et al., 2020). Subsequently, employing the similar procedure presented by Istyastono et al. (Istyastono et al., 2021), ensPLIF values were calculated. The results were then arranged in a table for each receptor to quickly analyze using the RPART package in R (Supporting File S6). The tables started with the first column named "y" encoding the observed data ("1" for active; "0" for decoy), followed by "name" for the name of the corresponding ligand, "dg" for the average docking scores (kcal/mol) resulted from the docking simulations (*vide supra*), and then ensPLIF variables ("V1" for ensPLIF V1, "V2" for ensPLIF V2, until the whole ensPLIF values were covered).

Analysis using RPART in R

The analysis to provide the best decision tree with the highest BA value was performed using R version 3.4.4 in the VPS-2. The prior probabilities were optimized in this analysis. The best decision tree resulted from the RPART analysis was then examined for possibilities of overfitting, the cross-correlation between identified ensPLIF variables (Lanza and Waite, 2018), and chance-correlation (Istyastono et al., 2021).

RESULTS AND DISCUSSION

The DPP4-alogliptin complex remained stable during 50 ns MD simulations using YASARA Structure (Krieger and Vriend, 2015). The root-mean-square deviation (RMSD) and the RMSD deviation during five ns MD simulations (ΔRMSD) values of the DPP4 backbone atoms (RMSDBb) and the alogliptin heavy atoms (RMSDLigMove) presented in Figure 1 indicate the DPP4-alogliptin complex stability. According to Liu et al. (Liu et al., 2017), the complex could be considered stable if the ΔRMSD is lower than 2 Å, and Figure 2 shows that the complex was stable since the first five ns MD simulations. Nevertheless, the first five ns MD simulations of the DPP4-alogliptin complex here were considered as the equilibrium run (Istyastono and Gani, 2021; Krieger and Vriend, 2015).

All snapshots from the MD simulations were further analyzed using PyPLIF HIPPOS to identify the PLIFs of the DPP4-alogliptin complexes during the MD simulations. Based on the snapshots, the free energy of binding calculated from the simulations was -11.891 kcal/mol. The clustering of the production run snapshots resulted in 6 clusters, i.e., snapshots at 5.45, 18.93, 34.53, 43.66, 45.22, and 49.38 ns. The energy minimized snapshots were stored as cluster-1.pdb, cluster-2.pdb, cluster-3.pdb, cluster-4.pdb, cluster-5.pdb, and cluster-6.pdb, respectively. The files are provided in Supporting File S3. Together with the ensemble docking scores, the ensPLIF values for all docking poses resulted from the retrospective SBVS campaigns targeting DPP4 are supplied as the Supporting File S6. Fine-tuning the prior probabilities in running RPART on the dataset (Supporting File S6) by employing Balance Accuracy (BA) as the objective function (Figure 3) identified that prior probabilities of 0.28:0.72 resulted in the highest BA value (0.837). However, the ratio of the cross-validation error over the model error was more

than 1.5 (1.580), which was an indication of the overfitting of the model. The RPART was then re-run in the prior probabilities of 0.28:0.72 with the complexity parameter adjusted to 0.010675 from the default value (0.01) to avoid the overfitting. The ratio of the cross-validation error over the model error of the resulting regression tree (Figure 4) was 1.460.

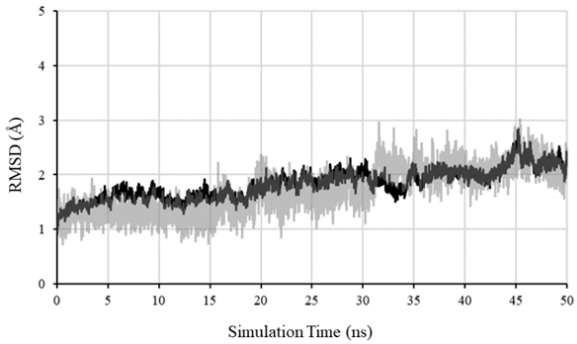

Figure 1. The RMSD values of the DPP4 backbone atoms (black) and the alogliptin heavy atoms (grey) during the 50 ns MD simulations.

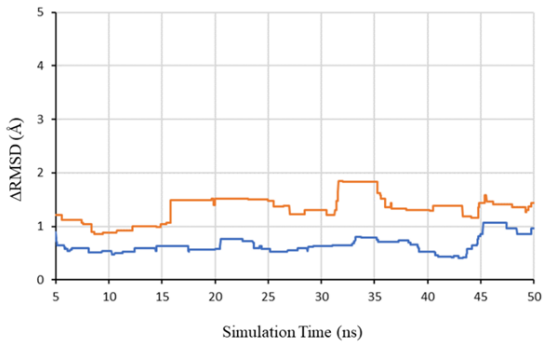

Figure 2. The ΔRMSD values at every five ns simulation time of the DPP4 backbone atoms (blue) and the alogliptin heavy atoms (orange) during the 50 ns MD simulations.

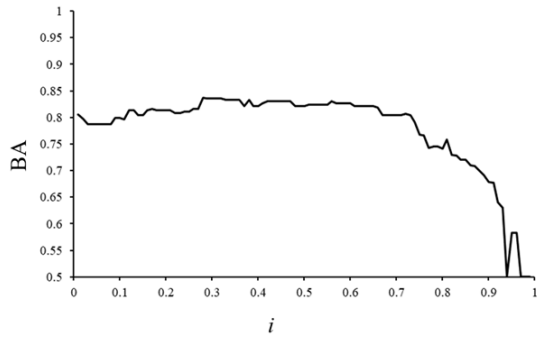

Figure 3. The Balance Accuracy (BA) optimization graph by fine-tuning the prior probabilities (i:1–i) in the RPART runs.

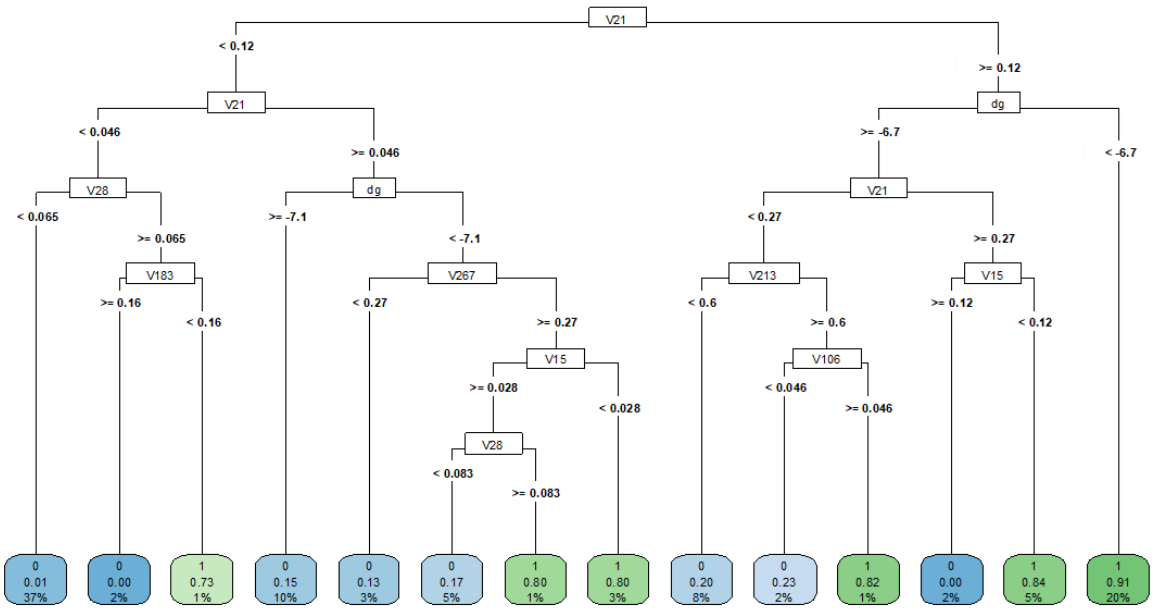

Figure 4. The regression tree resulted from the RPART run with the prior probabilities of 0.28:0.72 and the complexity parameter of 0.010675.

The DPP4-alogliptin complex with the PDB ID of 2ONC (Feng et al., 2007) was selected as the starting point in this research, despite the fact that the reference retrospective SBVS protocol used the PDB ID of 2I78 (Mysinger et al., 2012). There are 45 DPP4 crystal structures available at https://www.rcsb.org/ (accessed on 24 May 2022), and based on some of the crystal structures, the interactions of some commercially available DPP4 inhibitors, i.e., alogliptin, linagliptin, saxagliptin, sitagliptin, and vildagliptin in the DPP4 were analyzed and correlated to the inhibitory activity and glucose-lowering efficacy (Berger et al., 2018). Notably, alogliptin is a DPP4 commercially available non-covalent inhibitor that was reported to have the highest ligand efficiency (Berger et al., 2018). Moreover, saxagliptin and vildagliptin were identified as DPP4 covalent inhibitors (Berger et al., 2018), which in turn could create complications in performing MD simulations for the RED approach. Nath et al. (Nath et al., 2021) used also 2ONC as the main material in their SBVS campaigns and MD simulations to design potential DPP4 inhibitors.

During the preparation of the input file for the MD simulations, the module "SeqRes" in YASARA-Structure (Krieger and Vriend, 2015) identified some missing residues in 2ONC.pdb, i.e., His36, His37, Ala38, Ser39, Gln72, Glu73, and Asn74. The module "Build > N-terminal loop" and "Build > Loop" in YASARA-Structure (Krieger and Vriend, 2015) were used to complete the missing residues for further use. The DPP4 active site was categorized into two sub pockets, i.e., the S1 pocket (formed by Ser630, Tyr631, Val656, Trp659, Tyr662, Tyr666, Asn710, Val711, and His740) and the S2 pocket (formed by Glu205 and Glu206, Ser209, Phe357, Arg358, and Arg125) (Li et al., 2018). There were also 2 alogliptin structures in the selected Chain A of the crystal structures 2ONC.pdb, i.e., encoded as SY1 800 and SY1 801 (Feng et al., 2007). Since the alogliptin encoded as SY1 800 was located in the DPP4 active site (Berger et al., 2018; Li et al., 2018), this alogliptin was kept while the other was removed.

The cavity of DPP4 is considered large (diameter > 20 Å) (Li et al., 2018). The DPP4 ligand could therefore move around to find the most stable pose inside the cavity, which was indicated by the availability of some non-competitive inhibitors, e.g., luteolin and apigenin (Fan et al., 2013). Nevertheless, alogliptin stayed in the active site during the 50 ns MD simulations (Figures 1 and 2). Set aside the hydrophobic interactions, which have a little contribution to selective binding (Ferreira De Freitas and Schapira, 2017) alogliptin mostly performed ionic interaction with Glu205 in the S1 pocket and aromatic edge-to-face interaction with Tyr666. Notably, according to the regression tree presented in Figure 4, these interactions play an important role in the DPP4 competitive inhibitor binding.

The average value of the free energy of binding calculated from the snapshots during the MD production runs was 11.891 kcal/mol, equal to the inhibitory constant (Ki) value of 1.972 nM. This value is comparable to the in vitro IC_{50} value of 7.5 nM reported by Berger et al. (Berger et al.,

2018). Further MD studies to explore the correlation between the free energy binding values calculated from MD simulations and the in vitro IC_{50} values (Nath et al., 2021) would be performed in the near future post prospective SBVS campaigns.

The clusters from the 50 ns MD simulations (Supporting File S3) have offered opportunities to use the RED approach in the SBVS campaigns. The ensPLIF calculation takes into account all the docking poses, which represents the flexibility of the ligands (Istyastono et al., 2021)Therefore, the ensPLIF calculations from the RED simulation results represent the flexibility of the protein-ligand complexes. In other words, they implement the induced-fit theory in SBVS campaigns.

Instead of the commonly used enrichment factor (EF) (Mysinger et al., 2012) and F-measure (Istyastono et al., 2021, 2020), BA was selected as the objective function to measure the prediction quality of the SBVS protocols resulted from this research. The selection was based on the high imbalance of the dataset decoy/active ratio (40950/533) (Mysinger et al., 2012) and the possibility offered by RPART (Therneau et al., 2015) to fine-tune the prior probabilities to deal with such an imbalanced dataset. The fine-tuning resulted in the regression tree model resulted from the RPART run with the prior probabilities of 0.28:0.72 as the best model (Figure 3). Unfortunately, an indication of overfitting was identified in this model. Therefore, the complexity parameter of the RPART run was adjusted, resulting in the regression tree model presented in Figure 4. Overfitting, cross-correlation, and chance correlation were not observed in the model.

The BA value of the selected regression tree model (Figure 4) was 0.836, which outperformed the reference SBVS campaigns (0.652) (Mysinger et al., 2012). Based on the BA value, the SBVS protocol described in this article could be used further for prospective screening. However, the EF value of the model was 3.221, which was considered very low compared to the reference SBVS campaign (40.7) (Mysinger et al., 2012). This is an indication that there is plenty of room for improvements in the SBVS protocol. By examining Figure 4, it is known that a compound without at least a cation in pH 7.4 would not be predicted as a DPP4 inhibitor. To be predicted as a DPP4 inhibitor, a compound should be able to form ionic interaction either to Glu205 (V21) or Glu206 (V28) (Figure 4). Indeed, all commercially available DPP4 inhibitors and most of the recognized DPP4 inhibitors contain at least an amine that can act as a cation in physiological pH 7.4 (Li et al., 2018). Nonetheless, there were some flavonoids and phenolic compounds reported as marginal to potent DPP4 inhibitors, e.g., luteolin (IC_{50} = 120 nM), apigenin (IC_{50} = 140 nM), and resveratrol (IC_{50} = 0.6 nM) (Fan et al., 2013; Li et al., 2018). These compounds could perform ionic interaction with neither Glu205 nor Glu206. In fact, luteolin and apigenin were reported as DPP4 non-competitive inhibitors (Fan et al., 2013), which were an indication of allosteric sites in the large DPP4 cavity (diameter > 20 Å) (Li et al., 2018)Further research to explore the DPP4 allosteric sites should be performed, which could assist in improving SBVS protocols to identify DPP4 inhibitors.

Figure 4 shows that there are 6 possible paths for a compound to be identified as a DPP4 inhibitor. This is the main difference between the retrospective SBVS campaign presented here and the one presented by Istyastono et al. (Istyastono et al., 2021), which showed only 1 to 3 paths. The plausible reason is that the cavities of targets studied by Istyastono et al. (Istyastono et al., 2021) are smaller than DPP4, which decreases the flexibility of ligands during the docking simulations. Analyzing those 6 paths from the right part resulted in the following "keys" (from the lock-and-key theory) for a compound that would be identified as a DPP4 inhibitor: (i) the compound should have the ensPLIF V21 (the ionic interaction to Glu205) value of more than or equal to 0.12 and the ensemble docking score of less than -6.7 kcal/mol; (ii) the compound should have the ensPLIF V21 (the ionic interaction to Glu205) value of more than or equal to 0.27 and the ensPLIF V15 (the hydrophobic interaction to Glu205) of less than 0.12; (iii) the compound should have the ensPLIF V21 (the ionic interaction to Glu205) value of more than or equal to 0.12, the ensPLIF V213 (the aromatic edge-to-face to Tyr666) value of more than or equal to 0.6, and the ensPLIF V106 (the hydrophobic interaction to Lys554) of more than or equal to 0.046; (iv) the compound should have the ensPLIF V21 (the ionic interaction to Glu205) value of more than or equal to 0.046, the docking score of less than 7.1 kcal/mol, the ensPLIF V267 (the hydrophobic interaction to His740) of more than or equal to 0.27, and the ensPLIF V15 (the hydrophobic interaction to Glu205) of less than 0.028; (v) the compound should have the ensPLIF V21 (the ionic interaction to Glu205) value of more than or equal to 0.046, the ensemble docking score of less than 7.1 kcal/mol, the ensPLIF V267 (the hydrophobic interaction to His740) of more than or equal to 0.27, the ensPLIF V15 (the hydrophobic

interaction to Glu205) of more than or equal to 0.028, and the ensPLIF V28 (the ionic interaction to Glu206) value of more than or equal to 0.083; and (vi) the compound should have the ensPLIF V28 (the ionic interaction to Glu206) value of more than or equal to 0.065 and the ensPLIF V183 (the hydrophobic interaction to Val656) of less than 0.16. Most descriptors showed favorable conditions for identifying DPP4 inhibitors, except for the ensPLIF V15 (the hydrophobic interaction to Glu205) and the ensPLIF V183 (the hydrophobic interaction to Val656). A compound would have a chance to be identified as a DPP4 inhibitor if it has a lower value of either V15 or V183. This is in line with the findings of Istyastono et al. (Istyastono et al., 2021, 2017) that employing ensPLIF and RPART in retrospective SBVS campaigns could identify favorable and unfavorable interactions.

Based on Figure 4, further prospective SBVS campaigns would be performed on amine-containing compounds. The ionic interaction with either Glu205 or Glu206 requires a cation in the ligand, which the amine in the ligand would provide (Istyastono et al., 2015). The machine learning technique used here was the RPART approach (Therneau et al., 2015) since it was proven to be able to increase the prediction quality of the SBVS protocols as well as identifying the molecular determinant of the protein-ligand binding (Istyastono et al., 2021, 2017). This feature could be of invaluable assistance in the phase of hit or lead optimization (Istyastono et al., 2015). Nonetheless, the protein structure clusters resulted from the MD simulations (Supporting File S3) and the ensPLIF resulted from the SBVS campaigns (Supporting File S6) are provided here to allow reperforming SBVS campaigns or post-SBVS analysis using different settings, docking software, or other machine learning techniques.

CONCLUSIONS

Employing ensPLIF by using ensemble DPP4-alogliptin complexes clustered from 50 ns MD simulations has successfully mimicked the induced-fit theory in SBVS campaigns. The retrospective validation of the SBVS protocol resulted in the best protocol with a BA value of 0.836. The SBVS protocol could be employed further in prospective campaigns to discover potent DPP4 inhibitors from amine-containing compounds. Nonetheless, the protocol could be improved further by taking into account some flavonoids and phenolic compounds that showed high DPP4 inhibitory activities. The pieces of evidence of some non-competitive DPP4 inhibitors indicated the availability of the DPP4 allosteric site could also be further explored to improve the SBVS protocol.

ACKNOWLEDGEMENTS

Muhammad Radifar is acknowledged for technically maintaining PyPLIF HIPPOS.

DATA AVAILABILITY STATEMENT

All data (Supporting File 1, 2, 3, 4, 5, and 6) are available upon request.

CONFLICT OF INTEREST

E.P.I. is one of MOLMOD ID's co-founders, while B.I.W. is a technician and an associate researcher there. The institution provides a platform as a service (PaaS) in computer-aided drug design and discovery.

REFERENCES

Antunes, D.A., Devaurs, D., Kavraki, L.E., 2015. Understanding the challenges of protein flexibility in drug design. *Expert Opin. Drug Discov.*, 10(12):1301-1013.

Berger, J.P., SinhaRoy, R., Pocai, A., Kelly, T.M., Scapin, G., Gao, Y.-D., Pryor, K.A.D., Wu, J.K., Eiermann, G.J., Xu, S.S., Zhang, X., Tatosian, D.A., Weber, A.E., Thornberry, N.A., Carr, R.D., 2018. A comparative study of the binding properties, dipeptidyl peptidase-4 (DPP-4) inhibitory activity and glucose-lowering efficacy of the DPP-4 inhibitors alogliptin, linagliptin, saxagliptin, sitagliptin and vildagliptin in mice. *Endocrinol. Diabetes Metab.*, 1(1), e00002.

Cao, W., Chen, X., Chin, Y., Zheng, J., Lim, P.E., Xue, C., Tang, Q., 2021. Identification of curcumin as a potential α-glucosidase and dipeptidyl-peptidase 4 inhibitor: Molecular docking study, in vitro and in vivo biological evaluation. *J. Food Biochem.*, 2021(e13686), 1–11.

Eberhardt, J., Santos-Martins, D., Tillack, A.F., Forli, S., 2021. AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. *J. Chem. Inf. Model.*, 61(8), 3891-3898.

Fan, J., Johnson, M.H., Lila, M.A., Yousef, G., De Mejia, E.G., 2013. Berry and citrus phenolic compounds inhibit dipeptidyl peptidase IV: Implications in diabetes management. *Evid. Based Complement. Alternat. Med.*, 2013(479505), 1–13.

Feng, J., Zhang, Z., Wallace, M.B., Stafford, J.A., Kaldor, S.W., Kassel, D.B., Navre, M., Shi, L., Skene, R.J., Asakawa, T., Takeuchi, K., Xu, R., Webb, D.R., Gwaltney, S.L., 2007. Discovery of alogliptin: A potent, selective, bioavailable, and efficacious inhibitor of dipeptidyl

peptidase IV. *J. Med. Chem.*, 50(10), 2297–2300.

Ferreira De Freitas, R., Schapira, M., 2017. A systematic analysis of atomic protein-ligand interactions in the PDB. *Med. Chem. Commun.*, 8(10), 1970–1981.

Istyastono, E., Gani, M., 2021. Identification of Interactions of ABT-341 to Dipeptidyl Peptidase IV during Molecular Dynamics Simulations. *J. Farm. Galenika (Galenika J. Pharm.)*, 7(2), 91–98.

Istyastono, E.P., 2017. Binary quantitative structure-activity relationship analysis to increase the predictive ability of structure-based virtual screening campaigns targeting cyclooxygenase-2. *Indones. J. Chem.*, 17(2), 322–329.

Istyastono, E.P., Kooistra, A.J., Vischer, H., Kuijer, M., Roumen, L., Nijmeijer, S., Smits, R., de Esch, I., Leurs, R., de Graaf, C., 2015. Structure-Based Virtual Screening for Fragment-Like Ligands of the G Protein-Coupled Histamine H4 Receptor. *Med. Chem. Commun.*, 6(6), 1003–1017.

Istyastono, E.P., Prasasty, V.D., 2021. Computer-aided discovery of pentapeptide AEYTR as a potent acetylcholinesterase inhibitor. *Indones. J. Chem.*, 21(1), 243–350.

Istyastono, E.P., Radifar, M., Yuniarti, N., Prasasty, V.D., Mungkasi, S., 2020. PyPLIF HIPPOS: A molecular interaction fingerprinting tool for docking results of AutoDock Vina and PLANTS. *J. Chem. Inf. Model.*, 60(8), 3697–3702.

Istyastono, E.P., Yuniarti, N., Hariono, M., Yuliani, S.H., Riswanto, F.D.O., 2017. Binary quantitative structure-activity relationship analysis in retrospective structure based virtual screening campaigns targeting estrogen receptor alpha. *Asian J. Pharm. Clin. Res.*, 10(12), 206–211.

Istyastono, E.P., Yuniarti, N., Prasasty, V.D., Mungkasi, S., 2021. PyPLIF HIPPOS-assisted prediction of molecular determinants of ligand binding to receptors. *Molecules*, 26(2542), 1–12.

Kooistra, A.J., Leurs, R., de Esch, I.J.P., de Graaf, C., 2015. Structure-Based Prediction of G-Protein-Coupled Receptor Ligand Function: A β-Adrenoceptor Case Study. *J. Chem. Inf. Model.*, 55(5), 1045–1061.

Koshland, D.E., 1994. The key–lock theory and the induced fit theory. *Angew. Chem. Int. Ed. Engl.*, 33, 2375–2378.

Krieger, E., Vriend, G., 2015. New ways to boost molecular dynamics simulations. *J. Comput. Chem.*, 36(13), 996–1007.

Lanza, F., Waite, G.P., 2018. Nonlinear moment-tensor inversion of repetitive long-periods events recorded at pacaya volcano, Guatemala. *Front. Earth Sci.*, 6(139), 1–16.

Li, N., Wang, L.J., Jiang, B., Li, X., Guo, C., Guo, S., Shi, D., 2018. Recent progress of the development of dipeptidyl peptidase-4 inhibitors for the treatment of type 2 diabetes mellitus. *Eur. J. Med. Chem.*, 151, 145–157.

Liu, K., Watanabe, E., Kokubo, H., 2017. Exploring the stability of ligand binding modes to proteins by molecular dynamics simulations. *J. Comput. Aided Mol. Des.*, 31(2), 201–211.

Marcou, G., Rognan, D., 2007. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.*, 47(1), 195–207.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J., 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J. Comput. Chem., 30(16), 2785–2791.

Mysinger, M.M., Carchia, M., Irwin, J.J., Shoichet, B.K., 2012. Directory of Useful Decoys, Enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.*, 55(14), 6582–6594.

Nath, V., Ramchandani, M., Kumar, N., Agrawal, R., Kumar, V., 2021. Computational identification of potential dipeptidyl peptidase (DPP)-IV inhibitors: Structure based virtual screening, molecular dynamics simulation and knowledge based SAR studies. *J. Mol. Struct.*, 1224(129006), 1–13.

Patel, C.N., Kumar, S.P., Pandya, H.A., Rawal, R.M., 2021. Identification of potential inhibitors of coronavirus hemagglutinin-esterase using molecular docking, molecular dynamics simulation and binding free energy calculation. *Mol. Divers.*, 25(1), 421–433.

Prasasty, V., Radifar, M., Istyastono, E., 2018. natural peptides in drug discovery targeting acetylcholinesterase. *Molecules*, 23(9), 2344.

Prasasty, V.D., Istyastono, E.P., 2020. Structure-based design and molecular dynamics simulations of pentapeptide AEYTR as a potential acetylcholinesterase inhibitor. *Indones. J. Chem.*, 20(4), 953–959.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. Vienna. http://www.r-project.org.

Riswanto, F.D.O., Hariono, M., Yuliani, S.H., Istyastono, E.P., 2017. Computer-aided design of chalcone derivatives as lead compounds targeting acetylcholinesterase. *Indones. J. Pharm.*, 28(2), 100–111.

Riswanto, F.D.O., Rawa, M.S.A., Murugaiyah, V., Salin, N.H., Istyastono, E.P., Hariono, M., Wahab, H.A., 2021. Anti-cholinesterase activity of chalcone derivatives: synthesis, in vitro assay and molecular docking study. *Med. Chem.*, 17(5), 442–452.

Therneau, T., Atkinson, B., Ripley, B., 2015. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-9.

Wang, Z., Sun, H., Shen, C., Hu, X., Gao, J., Li, D., Cao, D., Hou, T., 2020. Combined strategies in structure-based virtual screening. Phys. Chem. Chem. Phys., 22: 1349-1359.