Drug Design, Development and Therapy

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Navigating the chemical space of dipeptidyl peptidase-4 inhibitors

This article was published in the following Dove Press journal:
Drug Design, Development and Therapy
10 August 2015
Number of times this article has been viewed

Watshara Shoombuatong[1]
Veda Prachayasittikul[1,2]
Nuttapat Anuwongcharoen[1]
Napat Songtawee[1]
Teerawat Monnor[1]
Supaluk Prachayasittikul[1]
Virapong Prachayasittikul[2]
Chanin Nantasenamat[1,2]

[1]Center of Data Mining and Biomedical Informatics, [2]Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand

Abstract: This study represents the first large-scale study on the chemical space of inhibitors of dipeptidyl peptidase-4 (DPP4), which is a potential therapeutic protein target for the treatment of diabetes mellitus. Herein, a large set of 2,937 compounds evaluated for their ability to inhibit DPP4 was compiled from the literature. Molecular descriptors were generated from the geometrically optimized low-energy conformers of these compounds at the semiempirical AM1 level. The origins of DPP4 inhibitory activity were elucidated from computed molecular descriptors that accounted for the unique physicochemical properties inherently present in the active and inactive sets of compounds as defined by their respective half maximal inhibitory concentration values of less than 1 μM and greater than 10 μM, respectively. Decision tree analysis revealed the importance of molecular weight, total energy of a molecule, topological polar surface area, lowest unoccupied molecular orbital, and number of hydrogen-bond donors, which correspond to molecular size, energy, surface polarity, electron acceptors, and hydrogen bond donors, respectively. The prediction model was subjected to rigorous independent testing via three external sets. Scaffold and chemical fragment analysis was also performed on these active and inactive sets of compounds to shed light on the distinguishing features of the functional moieties. Docking of representative active DPP4 inhibitors was also performed to unravel key interacting residues. The results of this study are anticipated to be useful in guiding the rational design of novel and robust DPP4 inhibitors for the treatment of diabetes.

Keywords: QSAR, decision tree, scaffold analysis, fragment analysis, antidiabetic, molecular docking, rational drug design

Introduction

Diabetes is a chronic disease and a major public health concern with an estimated global prevalence of 285 million.[1] In the United States, 29.1 million (or approximately 9.3% of the population) have diabetes, in which 21 million and 8.1 million are diagnosed and undiagnosed, respectively.[2] In fact, the estimated economic costs of diagnosed diabetes in the United States for 2012 was $245 billion, which increased from $174 billion in 2007.[3]

Given the multifaceted nature of diabetes, the search for robust drugs has been reported to entail a multitude of molecular targets.[4,5] Dipeptidyl peptidase-4 (DPP4) has emerged as a promising therapeutic route for the treatment of type 2 diabetes (T2D) because it regulates glucose homeostasis.[6] DPP4 is a serine protease that mediates the cleavage of two endogenous incretin hormones consisting of glucagon-like peptide and glucose-dependent insulinotropic polypeptide. Upon food ingestion, intestinal cells secrete these incretin hormones targeting pancreatic β-cells to stimulate insulin release. Generally, these two hormones exert a great effect on reducing blood glucose concentration; however, the rapid degradation of these hormones by DPP4 in T2D results in persistent high glucose level.[7] Therefore, the inhibition of

Correspondence: Chanin Nantasenamat
Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University
999 Phutthamonthon 4 Road, Salaya, Phutthamonthon,
Nakhon Pathom 73170, Thailand
Tel +66 2 441 4371
Fax +66 2 441 4380
Email chanin.nan@mahidol.ac.th

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/DDDT.S86529

DPP4 reduces blood glucose by preventing the degradation of these incretin hormones. Several DPP4 inhibitors have been released on the market, beginning with sitagliptin in 2006, vildagliptin in 2007, saxagliptin in 2009, alogliptin in 2010, linagliptin in 2011, and, finally, teneligliptin in 2012.[8] Generally, DPP4 inhibitors are considered to afford a favorable safety profile,[9,10] although rare side effects (ie, angioedema, hemolysis, leucopenia, rheumatoid arthritis, and drug-induced acute hepatic injury) have been documented but with low incidence.[11] Thus, there is ample room for additional improvement of the inhibitory and pharmacokinetic properties of DPP4 inhibitors. Medicinal chemistry approaches have been instrumental in the development of DPP4 inhibitors by facilitating the investigation of substituent effects in the quest for improved potency.[8,12] Complementing the effort of medicinal chemistry is computer-aided drug design, of which chemical space exploration and quantitative structure–activity relationship (QSAR) methods are employed in this study. The former entails exploration of the chemical space to gain insights on the molecular complexity of investigated compounds. The latter enables the correlation of molecular structure with its respective biological activity via multivariate learning methods.[13,14]

The availability of public databases of bioactivity significantly lowers the barriers for large-scale investigation of the structure–activity relationship for compounds of interest[15,16] and leads to accelerated drug discovery efforts. This study takes advantage of bioactivity data compilation of DPP4 inhibitors available from the BindingDB.[17] To the best of our knowledge, this study represents the first large-scale chemical space exploration and QSAR investigation of DPP4 inhibitory activity. Chemical space exploration was achieved by exploratory data analysis, cluster analysis, and chemical substructure analysis, whereas QSAR analysis was performed using decision tree (DT) analysis. A schematic representation of the computational workflow is summarized in Figure 1.

Material and methods

Compilation of the dataset

A large compilation of known compounds with inhibitory activity against DPP4 was extracted from the BindingDB,[17] which constituted 138 original articles. This nonredundant dataset comprises 2,937 compounds with the associated bioactivity reported as half maximal inhibitory concentration (IC_{50}) values. An IC_{50} cutoff value of ≤ 1 μM was employed to categorize compounds as "actives", whereas a cutoff value of ≥ 10 μM was utilized to categorize compounds as "inactives", which resulted in subsets of 2,075 and 534, respectively. The remaining 328 compounds exhibiting

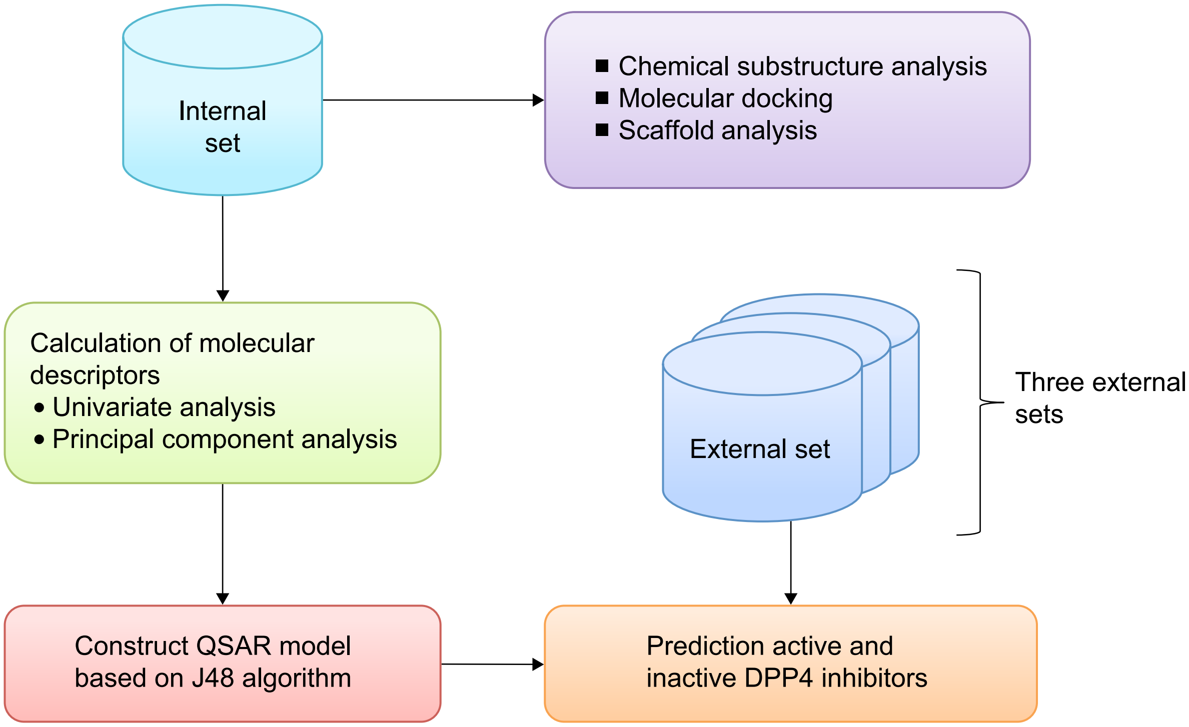

Internal set
■ Chemical substructure analysis
■ Molecular docking
■ Scaffold analysis
Calculation of molecular descriptors
• Univariate analysis
• Principal component analysis
External set
Three external sets
Construct QSAR model based on J48 algorithm
Prediction active and inactive DPP4 inhibitors

Figure 1 Schematic representation of the computational workflow.
Abbreviation: QSAR, quantitative structure–activity relationship.

intermediate bioactivity were not considered in this study due to their dubious nature while the subset of 2,609 was subjected to further investigations. Data imbalance observed for the active and inactive classes was addressed by subjecting the 2,075 actives to fuzzy C-means clustering,[18] which produced a final dataset consisting of 588 actives and 534 inactives (DPP4-TRN).

The constructed predictive model was rigorously validated against three external validation sets. To show the ability of predictive models for filtering inactives, in the present study, three external validation sets were employed as negative control and were compiled from the BindingDB as follows: 1) random selection of active and inactive inhibitors against a wide range of human target proteins (DPP4-TEST1); 2) random selection of active and inactive inhibitors against other human proteases (DPP4-TEST2); and 3) random selection of active and inactive inhibitors against other human DPP types such as DPP1, DPP2, and DPP7 (DPP4-TEST3).

According to the applicability domain, the robustness of a QSAR model applies well for predicting the activity of compounds belonging to similar chemotypes as those used as the training data for constructing the predictive model.[19] Thus, applicability domain was applied by selecting compounds to include in the external validation sets. Tanimoto coefficient is a commonly used metric for measuring the similarity between compounds of the internal and external sets, which varies between 0 (total lack of similarity) and 1 (compound from the internal set is identical to a compound in the external set). Herein, the average Tanimoto coefficient value was used as the cutoff for selection of compounds to include in the external validation sets.[20–22] Finally, the remaining DPP4-TEST1, DPP4-TEST2, and DPP4-TEST3 consisted of 149, 160, and 167 compounds, respectively.

Calculation of molecular descriptors

The molecular structures of the investigated compounds were converted to three-dimensional structures from their simplified molecular-input line-entry system notation using MarvinSketch, version 6.2.1, from ChemAxon (ChemAxon Ltd., Budapest, Hungary).[23] The file format of these structures was then converted to the appropriate file format using Babel, version 3.3,[24] for subsequent geometry optimization at the B3LYP/6-31G(d) level in Gaussian 09.[25] Our previous chemical space exploration of aromatase inhibitors was performed using a set of 13 descriptors selected to represent the general properties of a molecule.[26] Given the readily interpretative nature, this set of descriptors was also employed for this investigation. This set of descriptors included the following: 1) mean absolute charge (Q_m); 2) energy; 3) dipole moment; 4) highest occupied molecular orbital (HOMO); 5) lowest unoccupied molecular orbital (LUMO); 6) energy gap between the HOMO and LUMO states (HOMO–LUMO); 7) molecular weight (MW); 8) rotatable bond number (RBN); 9) number of rings (nCIC); 10) number of hydrogen bond donors (nHDon); 11) number of hydrogen bond acceptors (nHAcc); 12) Ghose–Crippen octanol–water partition coefficient (ALogP); and 13) topological polar surface area (TPSA).

Univariate analysis

Univariate statistical approaches were employed to perform exploratory data analysis. Specifically, six descriptive statistical parameters were used to summarize the aforementioned set of 13 descriptors. These parameters consisted of the minimum (Min), first quartile (Q1), median, mean, third quartile (Q3), and maximum (Max) of the dataset. Box plots were applied to visualize the relative distribution of the values for each investigated variable; this involved the analysis of a set of 13 descriptors to identify the descriptors that exert great influence on the active and inactive classes of DPP4 inhibitors. Histograms were used to visualize and estimate the distribution of active and inactive classes of DPP4 inhibitors. Furthermore, the *P*-value was used to assess whether active and inactive classes of DPP4 inhibitors were significantly different using Student's *t*-test.[27]

Principal component analysis

Principal component analysis (PCA) is an unsupervised learning approach that groups data into related clusters in an a priori fashion. Practically, the PCA approach reduces the dimensionality of the dataset, while most of the information of the original dataset is preserved.[28] This approach is performed by identifying directions, so-called principal components (PCs), along which variation in the data is maximal. In practice, PCs are obtained by calculating eigenvectors and eigenvalues of a data covariance (or correlation) matrix. The eigenvector associated with the largest eigenvalue has a direction that is identical to the first PC (PC1), whereas the eigenvector associated with the second largest eigenvalue determines the direction of the second PC (PC2) and so forth. In performing PCA analysis, a dataset is represented by a small number of PCs, in contrast to the initially large number of variables present in the original dataset.[29] In this study, PCA was performed on a set of 13 molecular descriptors, as described in the previous section. Prior to PCA analysis, all data were standardized to a comparable scale by transforming variables to zero mean and unit variance. Active and inactive classes of DPP4 inhibitors were individually calculated using the FactoMineR[30] package of the R statistical language.

DT analysis

A DT is composed of a hierarchical arrangement of nodes and branches in which the nodes represent the molecular descriptors, whereas the branches refer to decision rules to categorize compounds as actives and inactives. DT has been successfully applied in the analysis of various types of compounds, such as aromatase inhibitors,[26] volatile organic compounds,[31] and cytochrome P450-interacting compounds.[32] A DT was constructed with WEKA, version 3.6,[33] using the J48 algorithm (a Java implementation of the C4.5 algorithm). C4.5 establishes a DT by iteratively appending features having high information gains.[34] Finally, C4.5 automatically calculates the feature usage obtained from the full DT or collection of rules. Molecular descriptors having the highest feature usage are considered to be the most important features.

Chemical substructure analysis

In preparation for substructure analysis, the chemical structures of all DPP4 inhibitors were generated in structure-data file (SDF) format using MarvinSketch, followed by appending the bioactivity label to the SDF files using an in-house text processing tool coded in C^{++}. Substructure analysis was performed using the Fragmenter and FragmentStatistics components of JChem version 14.8.18.0.[35] Fragmenter processed the activity-tagged SDF file by generating molecular fragments according to the FragmenterAll protocol. Produced fragments were analyzed using the FragmentStatistics toolkit, whereby fragments were categorized as actives and inactives using pIC$_{50}$ cutoff values of 6 and 5, respectively. Subsequently, fragments were assigned molecular scores according to the following equation:

$$\text{Molecular score} = N_{atom} \times (N_{active} - N_{inactive}) \qquad (1)$$

where N_{atom} denotes the atom count of a given fragment of interest, whereas N_{active} and $N_{inactive}$ represent the number of occurrences of the fragment in the active and inactive classes, respectively.

Molecular docking and binding mode analysis

Molecular docking was performed to gain insights on how the inhibitors bind DPP4. Geometrically optimized structures of each compound were docked with the crystal structure of DPP4 catalytic domain (PDB code 3C45, resolution of 2.05 Å) using AutoDock version 4.2.6,[36] in which the rotational bonds of compounds were treated as flexible whereas those of DPP4 were rigid. United atom model was applied to both protein and ligand structures. Grid boxes were created to cover the inhibitor-binding site of the protein with the grid spacing of 0.375 Å while the co-crystalized ligand site was set as the center of the box. The Lamarckian genetic algorithm with 50 runs was used as the search parameter in which the population size was set at 150 and the Max number of energy evaluations was set to the high level. The anchor-binding mode of ligand docking poses with the lowest binding energy to the DPP4 active site was subsequently analyzed by the SiMMap server.[37] Three-dimensional models of the binding mode were visualized with PyMOL version 1.3.[38]

Results and discussion

Univariate analysis of active and inactive DPP4 inhibitors

The number of active and inactive DPP4 inhibitors compiled in this study was 2,075 and 534, respectively. Table 1 displays the six descriptive statistical parameters that offer the following advantages for summarizing the data: 1) the median and mean provide a measure of the centrality of the data; 2) the Min and Max indicate the data range; and 3) Q1 and Q3 provide the lower and upper boundaries, respectively, of the data. Furthermore, histograms shown in Figure 2 afford a graphical display of the data as tabulated frequencies of bars derived by binning continuous values into several data ranges. Figure 2A shows the distribution of active and inactive DPP4 inhibitors as red and blue bars, respectively, whereas the overlapping region is shown in purple. Figure 2B, which will be discussed in further details in the "Analysis of active DPP4 inhibitors" section, displays the distribution of two subsets of active DPP4 inhibitors that will be referred to as active I and active II.

MW is a general measure of the molecular size, and actives were found to be larger than inactives, with $P<0.001$, Q1 =340.9, median =386.5, mean =385.8, and Q3 =430.5 for actives, and Q1 =238.4, median =303.9, mean =315.1, and Q3 =359.5 for inactives (Table 1). As shown in Figure 2A, the distributions of actives and inactives were normal and positively skewed, respectively.

RBN is the number of rotatable bonds in a molecule and provides a relative measure of molecular flexibility. RBN is defined as any single bond, not in a ring, bound to a nonterminal heavy atom. Amide C–N bonds are excluded from the count because of their high rotational energy barrier. Actives

Table 1 Exploratory data analysis of actives and inactives using the six-term descriptive statistics

Statistics	MW	RBN	nCIC	nHDon	nHAcc	ALogP	TPSA	Q_m	Energy	Dipole moment	HOMO	LUMO	HOMO–LUMO
Actives													
Min	167.2	0.000	0.000	0.000	1.000	−2.936	29.260	0.137	−0.908	0.747	−0.572	−0.298	0.217
Q1	340.9	4.000	3.000	2.000	5.000	0.586	72.800	0.202	−0.260	4.075	−0.354	−0.039	0.301
Median	386.5	5.000	3.000	3.000	7.000	1.571	85.250	0.217	−0.123	5.831	−0.343	−0.025	0.314
Mean	385.8	5.008	3.155	2.735	6.897	1.566	89.490	0.222	−0.144	9.842	−0.352	−0.035	0.318
Q3	430.5	6.000	4.000	3.000	8.000	2.586	103.660	0.236	−0.017	8.111	−0.331	−0.008	0.332
Max	753.8	16.000	6.000	9.000	16.000	6.598	234.780	0.535	0.488	284.562	−0.286	0.047	0.386
Inactives													
Min	128.2	1.000	0.000	0.000	1.000	−2.485	3.240	0.142	−1.281	0.629	−0.490	−0.154	0.242
Q1	238.4	3.000	2.000	1.000	4.000	0.848	47.720	0.193	−0.172	2.890	−0.344	−0.022	0.310
Median	303.9	4.000	2.000	2.000	5.000	1.806	72.350	0.209	−0.097	3.961	−0.338	−0.005	0.331
Mean	315.1	4.605	2.609	2.375	4.991	1.859	72.002	0.213	−0.119	4.443	−0.337	−0.007	0.331
Q3	359.5	6.000	3.000	3.000	6.000	2.949	88.840	0.231	−0.048	5.262	−0.329	0.011	0.347
Max	1,174.6	36.000	6.000	11.000	25.000	7.528	351.810	0.346	0.139	42.433	−0.290	0.100	0.414

Abbreviations: ALogP, Ghose–Crippen octanol–water partition coefficient; HOMO, highest occupied molecular orbital; HOMO–LUMO, energy gap between the HOMO and LUMO states; LUMO, lowest unoccupied molecular orbital; Max, maximum; Min, minimum; MW, molecular weight; nCIC, number of rings; nHAcc, number of hydrogen bond acceptors; nHDon, number of hydrogen bond donors; Q1, first quartile; Q3, third quartile; Q_m, mean absolute charge; RBN, rotatable bond number; TPSA, topological polar surface area.

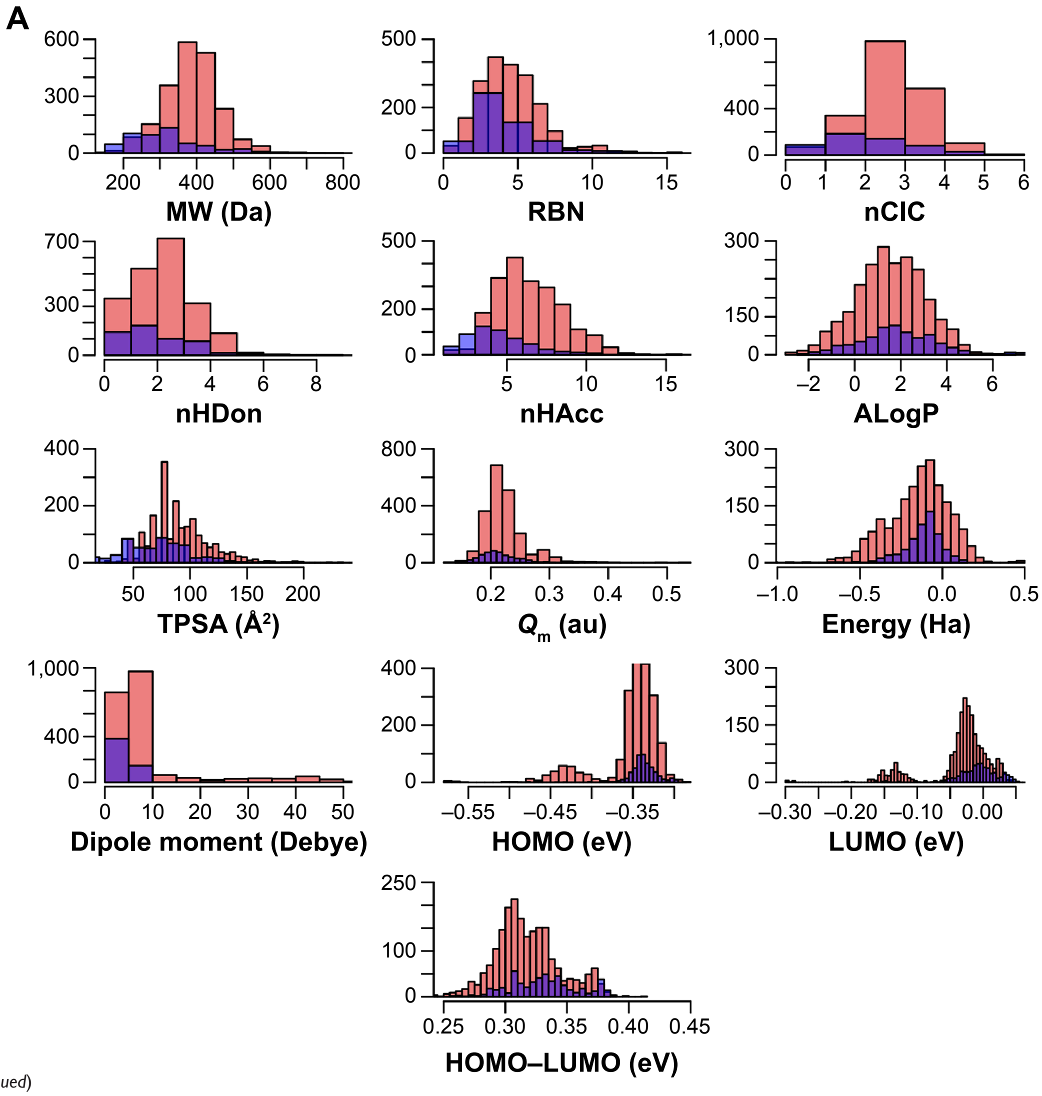


Figure 2 (*Continued*)

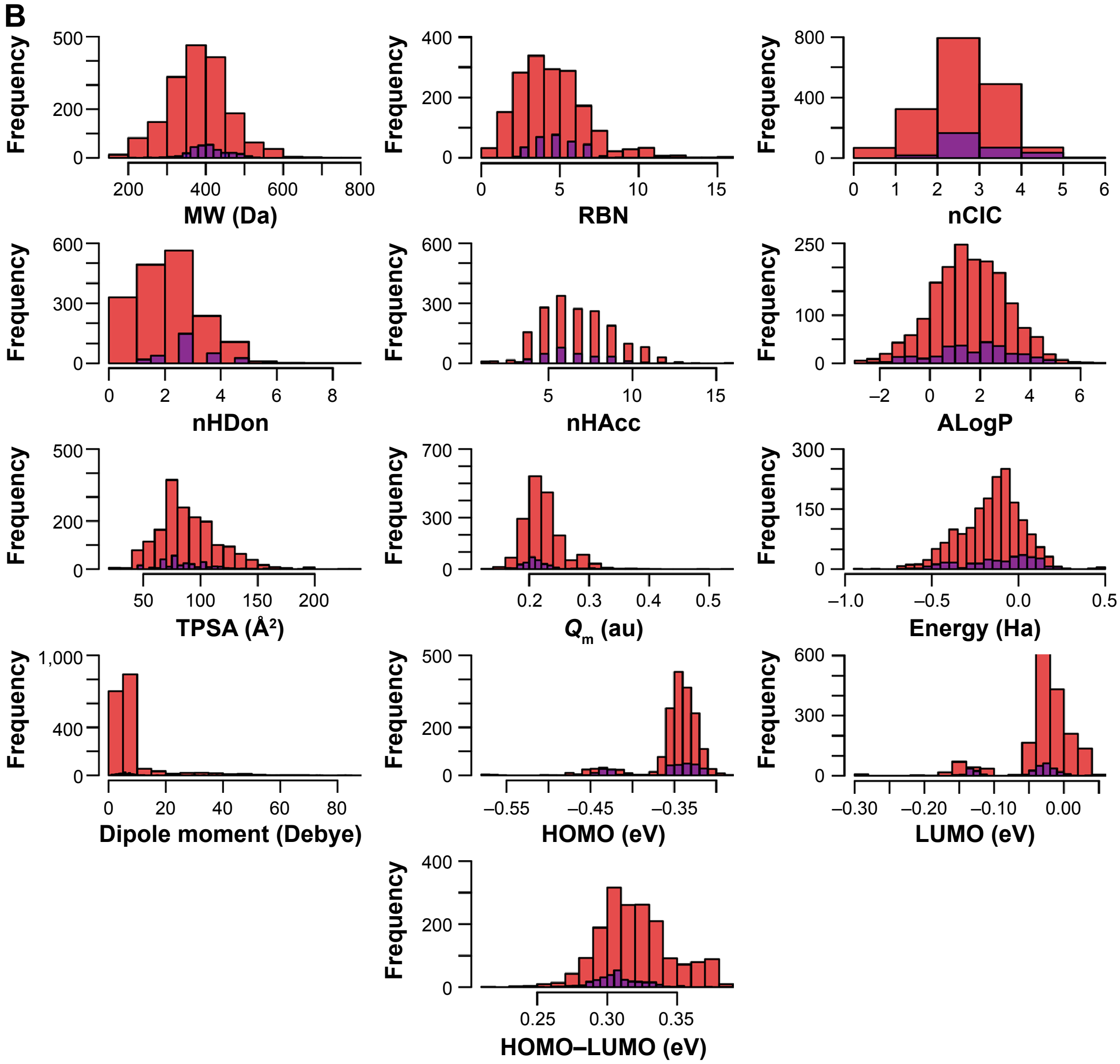


Figure 2 Histograms of the molecular descriptors for actives/inactives (**A**) and active I/active II DPP4 inhibitors (**B**).

Notes: Actives/active I and inactives/active II are shown in red and blue, respectively; purple regions represent their overlap.

Abbreviations: ALogP, Ghose–Crippen octanol–water partition coefficient; HOMO, highest occupied molecular orbital; HOMO–LUMO, energy gap between the HOMO and LUMO states; LUMO, lowest unoccupied molecular orbital; MW, molecular weight; nCIC, number of rings; nHAcc, number of hydrogen bond acceptors; nHDon, number of hydrogen bond donors; Q_m, mean absolute charge; RBN, rotatable bond number; TPSA, topological polar surface area.

were found to have higher RBN values than their inactive counterparts, thereby implying the importance of molecular flexibility for DPP4 inhibitory activity. Corresponding values of Q1 =4.0, median =5.0, mean =5.0, and Q3 =6.0 were obtained for actives, whereas values of Q1 =3.0, median =4.0, mean =4.6, and Q3 =6.0 were obtained for inactives. Although the distribution of actives and inactives are both positively skewed, the RBN values for actives are greater than those for inactives. Remarkably, all of these results indicated that the number of rotatable bonds in a molecule between active and inactive DPP4 inhibitors was slightly different, with P=0.001.

The nCIC is calculated as the cardinality of the set of independent rings known as the smallest set of smallest rings. The nCIC from actives was higher than that from inactives (P<0.001), affording values of Q1 =3.000, median =3.000, mean =3.155, and Q3 =4.000 for actives, and Q1 =2.000, median =2.000, mean =2.609, and Q3 =3.000 for inactives.

nHDon is the number of hydrogen bond donors present in a molecule. The mean of nHDon in actives (2.735±1.202) was higher than that in inactives (2.375±1.355). A six-number statistical descriptive confirmed that active and inactive DPP4 inhibitors differed from each other, with values in the range of [0.000, 9.000] and [0.000, 11.000], respectively; for active DPP4 inhibitors, median =3.000, mean =2.735, and Q3 =3.000, and for inactive DPP4 inhibitors, median =2.000, mean =2.373, and Q3 =3.000. Furthermore, the histogram for active DPP4 inhibitors does not differ from that of inactive

DPP4 inhibitors. Notably, all of these results indicated that the nHDon between active and inactive DPP4 inhibitors was significantly different, with $P<0.001$.

nHAcc represents the number of hydrogen bond acceptors present in a molecule. The mean values of rotatable bonds of DPP4 inhibitors are in the range of 6.897±2.122 (active) and 4.991±2.217 (inactive), whereas the values of descriptive statistics are Min =1.000, Q1 =5.000, median =7.000, mean =6.897, Q3 =8.000, and Max =16.000 for active DPP4 inhibitors, and Min =1.000, Q1 =4.000, median =5.000, mean =4.991, Q3 =6.000, and Max =25.000 for inactive DPP4 inhibitors. The histograms of these two inhibitor classes were found to differ from each other. These results indicated that the nHAcc for active and inactive DPP4 inhibitors was significantly different, with $P<0.001$.

ALogP is a computational estimation of the logarithm of the 1-octanol/water partition coefficient, and it is a well-known measure of molecular hydrophobicity. The mean values of ALogP are 1.556±1.488 and 1.859±1.691 for active and inactive DPP4 inhibitors, respectively, which are different, and the values of descriptive statistics confirm this finding, with values of Min =–2.936, Q1 =0.586, median =1.571, mean =1.566, Q3 =2.586, and Max =6.598 for active DPP4 inhibitors, and Min =–2.485, Q1 =0.848, median =1.806, mean =1.859, Q3 =2.949, and Max =7.528 for inactive DPP4 inhibitors. Additionally, the histograms of active and inactive DPP4 inhibitors were significantly different, with $P<0.001$.

TPSA is an empirical measure of the polar surface area of a molecule, and it describes the contribution of polar atoms to the molecular charge. TPSA is frequently used in the study of drug transport properties such as intestinal absorption[10] and blood–brain barrier permeability.[11] High TPSA, in addition to indicating that the molecule possesses a complex surface charge environment, also indicates that the molecule inherently possesses poor membrane permeability and would need to rely on active transport, such as membrane-bound receptors. The mean value of active DPP4 inhibitors (89.490±26.130) is greater than that of inactive DPP4 inhibitors (72.002±32.154); moreover, a six-number statistical descriptive confirms that the characteristics of active and inactive DPP4 inhibitors differ, with Min =29.260, Q1 =72.800, median =85.250, mean =89.490, Q3 =103.660, and Max =234.780 for active DPP4 inhibitors, and Min =3.240, Q1 =47.720, median =72.350, mean =72.002, Q3 =88.840, and Max =351.810 for inactive DPP4 inhibitors. These results indicated that the overall pattern of active and inactive DPP4 inhibitors, including the histogram shape in Figure 2A, were significantly different, with $P<0.001$.

Q_m is a global measure of the molecular charge. The mean values of active and inactive DPP4 inhibitors are 0.222±0.034 and 0.213±0.030, respectively. Histograms of these two inhibitor classes were significantly different, with $P<0.001$. A six-number statistical descriptive confirms this finding, with range values of [0.137, 0.535] for active DPP4 inhibitors and [0.142, 0.346] for inactive DPP4 inhibitors, whereas the top quartiles are [0.202, 0.236] for active DPP4 inhibitors and [0.193, 0.231] for inactive DPP4 inhibitors.

Energy is the sum of the atomic energy. The mean values of active and inactive DPP4 inhibitors are –0.144±0.183 and –0.119±0.129, respectively. Notably, the distributions of these two inhibitor classes are significantly different, with $P<0.001$. Furthermore, the six-number statistical descriptive indicates that active DPP4 inhibitors differ from inactive DPP4 inhibitors, ie, Min =–0.908, Q1 =–0.260, median =–0.123, mean =–0.144, Q3 =–0.017, and Max =0.488 for active DPP4 inhibitors, whereas Min =–1.281, Q1 =–0.172, median =–0.097, mean =–0.119, Q3 =–0.048, and Max =0.139 for inactive DPP4 inhibitors.

The dipole moment is a measure of the asymmetric distribution of charge in a molecule, where a low value suggests minimal charge distribution and vice versa. Table 1 indicates that the average value of active DPP4 inhibitors (9.842±15.038) is greater than that of inactive DPP4 inhibitors (4.443±3.303). The different patterns of these two DPP4 inhibitor classes are also indicated by a six-number statistical descriptive. The 6-number statistical descriptive of active DPP4 inhibitors consisted of Min =0.747, Q1 =4.075, median =5.831, mean =9.842, Q3 =8.111, and Max =284.562, whereas that of inactive DPP4 inhibitors consisted of Min =0.629, Q1 =2.890, median =3.961, mean =4.443, Q3 =5.262, and Max =42.433. The ranges of active and inactive DPP4 inhibitors were dramatically different, with values of [0.747, 284.562] and [0.629, 42.433], respectively, as shown in the corresponding histograms. Notably, these results indicated that the characteristics of active and inactive DPP4 inhibitors were significantly different, with $P<0.001$.

The HOMO and LUMO are the highest- and lowest-energy molecular orbitals that are occupied by electrons. The mean values of HOMO and LUMO in active and inactive DPP4 inhibitors are –0.352±0.038/–0.337±0.019 and –0.035±0.049/–0.007±0.031, respectively. The values of HOMO range from [–0.572, –0.286] for active DPP4 inhibitors and [–0.490, –0.290] for inactive DPP4 inhibitors, whereas the values of LUMO range from [–0.289, 0.047] for active DPP4 inhibitors and [–0.154, 0.100] for inactive DPP4

Table 2 Exploratory data analysis of subclasses of actives (I and II) using the six-term descriptive statistics

Statistics	MW	RBN	nCIC	nHDon	nHAcc	ALogP	TPSA	Q_m	Energy	Dipole moment	HOMO	LUMO	HOMO–LUMO
Actives I													
Min	167.2	0.000	0.000	0.000	1.000	–2.936	29.260	0.137	–0.908	0.747	–0.572	–0.298	0.217
Q1	333.5	3.000	3.000	2.000	5.000	0.637	73.250	0.203	–0.264	3.998	–0.353	–0.036	0.303
Median	381.0	5.000	3.000	3.000	7.000	1.587	85.250	0.219	–0.133	5.675	–0.343	–0.023	0.317
Mean	381.9	5.000	3.103	2.636	6.961	1.585	89.980	0.224	–0.154	7.976	–0.350	–0.029	0.320
Q3	429.4	6.000	4.000	3.000	8.000	2.577	104.670	0.237	–0.041	7.635	–0.331	–0.006	0.334
Max	753.8	16.000	6.000	9.000	16.000	6.598	234.780	0.535	0.488	80.233	–0.286	0.047	0.386
Actives II													
Min	202.4	2.000	1.000	1.000	3.000	–2.163	47.950	0.162	–0.695	1.035	–0.492	–0.155	0.243
Q1	375.5	4.000	3.000	3.000	5.000	0.350	69.460	0.200	–0.216	4.861	–0.411	–0.127	0.293
Median	403.4	5.000	3.000	3.000	6.000	1.504	84.660	0.210	–0.034	7.328	–0.349	–0.035	0.305
Mean	407.0	5.052	3.431	3.263	6.557	1.463	86.870	0.217	–0.093	19.814	–0.367	–0.062	0.304
Q3	435.4	6.000	4.000	4.000	8.000	2.641	101.040	0.224	0.068	33.395	–0.332	–0.023	0.316
Max	658.7	16.000	6.000	7.000	14.000	4.985	188.480	0.390	0.271	284.562	–0.307	0.037	0.386

Abbreviations: ALogP, Ghose–Crippen octanol–water partition coefficient; HOMO, highest occupied molecular orbital; HOMO–LUMO, energy gap between the HOMO and LUMO states; LUMO, lowest unoccupied molecular orbital; Max, maximum; Min, minimum; MW, molecular weight; nCIC, number of rings; nHAcc, number of hydrogen bond acceptors; nHDon, number of hydrogen bond donors; Q1, first quartile; Q3, third quartile; Q_m, mean absolute charge; RBN, rotatable bond number; TPSA, topological polar surface area.

inhibitors. The top quartiles for HOMO are [–0.354, –0.331] for active DPP4 inhibitors and [–0.344, –0.329] for inactive DPP4 inhibitors, whereas the top quartiles for LUMO are [–0.039, –0.008] for active DPP4 inhibitors and [–0.022, 0.011] for inactive DPP4 inhibitors. Remarkably, the histograms of HOMO and LUMO indicate that the distributions of active and inactive DPP4 inhibitors are significantly different, with $P<0.001$.

HOMO–LUMO is the energetic difference between the HOMO and LUMO states. HOMO–LUMO is a measure of kinetic stability and chemical reactivity, as HOMO and LUMO descriptors play fundamental roles in electron donation and acceptance. A large gap suggests high kinetic stability and low chemical reactivity because it is energetically unfavorable to add electrons to a high-lying LUMO or to extract electrons from a low-lying HOMO to form the activated complex of a potential reaction. Conversely, a molecule with a small or no HOMO–LUMO is chemically reactive. The mean values of HOMO–LUMO are 0.318±0.026 and 0.331±0.029 for active and inactive DPP4 inhibitors, respectively. The distributions of active and inactive DPP4 inhibitors are quite different. Additionally, the six-number statistical descriptive confirms this finding, with range values of [0.217, 0.386] for active DPP4 inhibitors and [0.242, 0.414] for inactive DPP4 inhibitors, whereas the lower and upper boundaries are [0.301, 0.332] for active DPP4 inhibitors and [0.310, 0.347] for inactive DPP4 inhibitors. These results indicate that the characteristics of active and inactive DP4 inhibitors were significantly different, with $P<0.001$.

All of these results indicated that nearly all of the 13 descriptors were significantly different between the two inhibitor classes at the level of $P<0.001$ except for RBN ($P=0.001$). With the exception of RBN descriptors, the remaining descriptors are significantly different for active and inactive DPP4 inhibitors and are efficient for discrimination.

PCA analysis of active and inactive DPP4 inhibitors

In this study, the 13 descriptors were analyzed by utilizing the first three PCs because the amount of cumulative variation of these PCs is as high as 70% of the original variance, as shown in Figure S1. Scores and loadings plots are presented in Figure 3A for actives (top row) and inactives (bottom row, bottom-left). Tables S1 and S2 show the loadings and contribution values, respectively, of each descriptor to the component. The contribution value of each descriptor can be obtained by the ratio of the squared factor score of this observation by the eigenvalue associated with that component.[12]

PC1 retained 27.93% and 35.06% of the original variance for active and inactive DPP4 inhibitors, respectively. Figure S1 indicates that the percentage of variance of inactive DPP4 inhibitors was higher than that of active DPP4 inhibitors. In Table S1 and Figure 3A (top-right), PC1 separates HOMO–LUMO from MW, RBN, nHDon, nHAcc, and TPSA for active DPP4 inhibitors, whereas in Figure 3A (bottom-right), PC1 separates energy from MW, RBN, nHAcc, TPSA, and Q_m for inactive DPP4 inhibitors. For loadings score analysis, PC1 highly correlated with MW (0.849), RBN (0.638), nHDon (0.578), nHAcc (0.637), TPSA (0.693), and HOMO–LUMO (–0.575) for active DPP4 inhibitors, whereas in inactive DPP4 inhibitors, PC1 highly correlated with MW (0.849), RBN (0.664), nHAcc (0.893), TPSA (0.815), Q_m (0.601), and energy (–0.654). These results indicated that PC1 correlated most strongly with MW and nHAcc for active and inactive DPP4 inhibitors, respectively. Furthermore, Table S2 also indicates that the MW descriptor highly contributes to PC1 for active DPP4 inhibitors, whereas the nHAcc descriptor highly contributes to PC1 for inactive DPP4 inhibitors. Descriptors consisting of nHDon, Q_m, energy, and HOMO–LUMO influenced PC1 for either active or inactive DPP4 inhibitors. Interestingly, the four differential descriptors are reported with $P<0.001$ and are significantly different between active and inactive inhibitor classes. It may be assumed that these four differential descriptors represent the informative features that discriminate between active and inactive DPP4 inhibitors.

PC2, which is the direction uncorrelated with PC1, retained 20.82% and 20.82% of the original variance for active and inactive DPP4 inhibitors, respectively. Figure S1 indicates that the first two components can preserve 51.29% and 55.88% of the original variance of active and inactive DPP4 inhibitors, respectively. The results indicated that the percentage and cumulative percentage of variance for inactive DPP4 inhibitors were greater than those for active DPP4 inhibitors. In Table S1 and Figure 3A (top-right), PC2 separates dipole moment and energy from HOMO and LUMO for active DPP4 inhibitors, whereas in Figure 3A (bottom-right), PC2 of inactive DPP4 inhibitors separates ALogP and nCIC from nHDon and HOMO–LUMO. The PCA loadings scores indicate that PC2 highly correlated with energy (–0.765), dipole moment (–0.617), HOMO

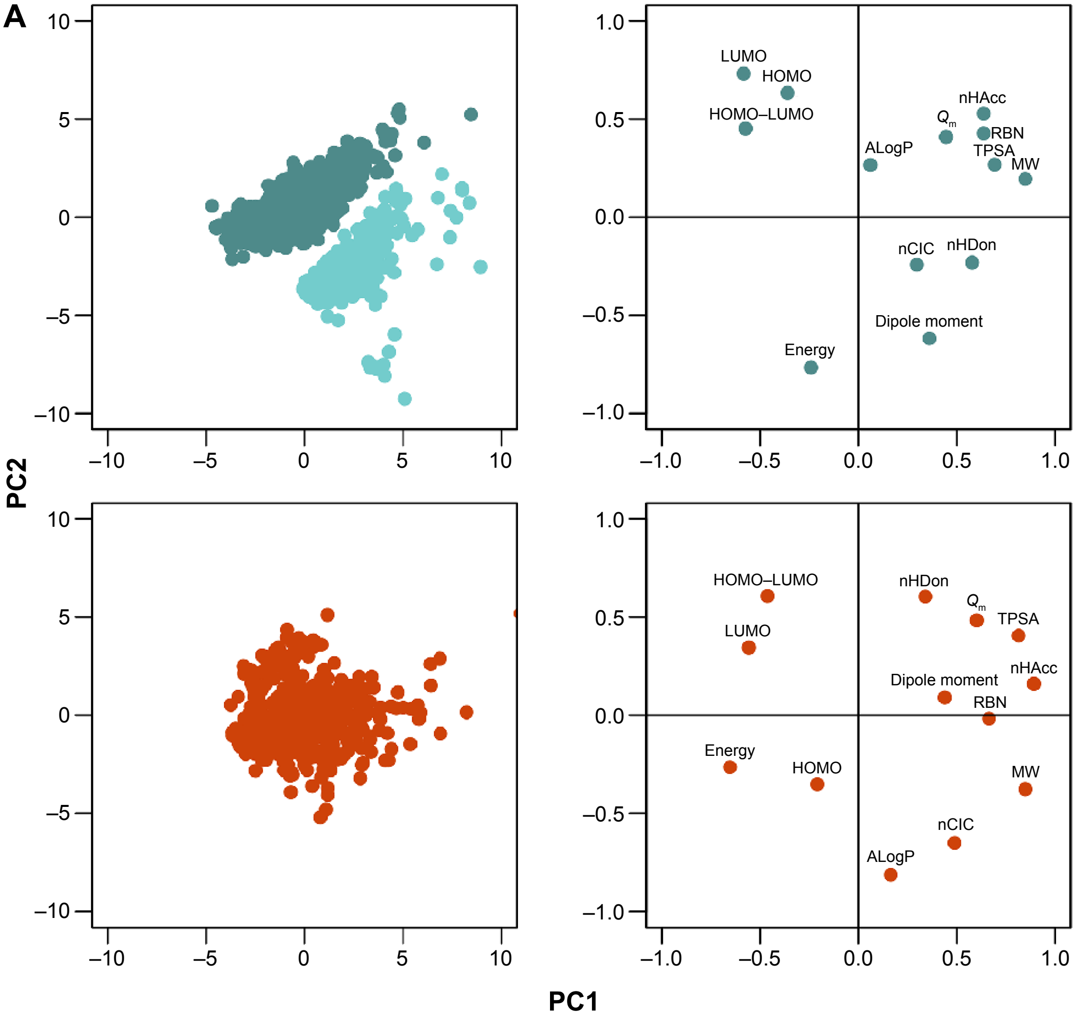


Figure 3 (*Continued*)

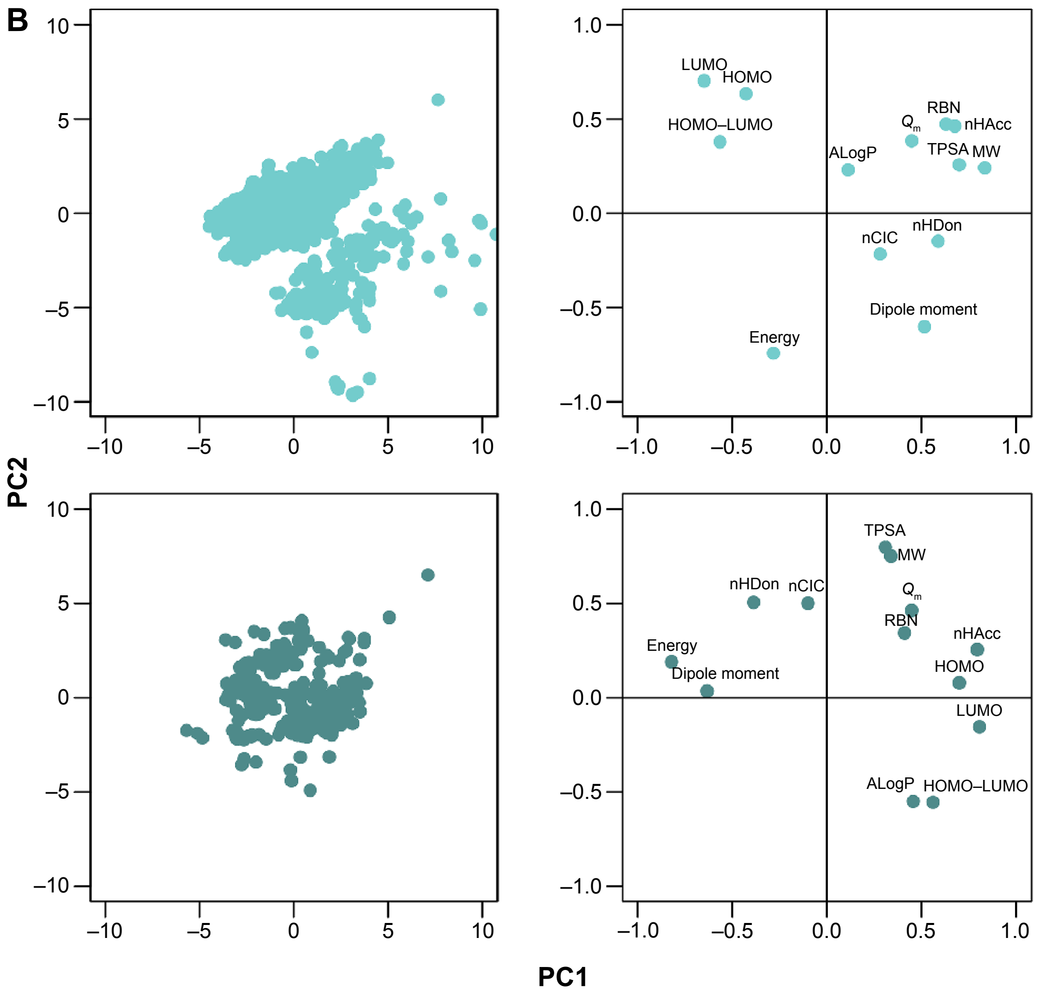


Figure 3 PCA scores plots of actives/inactives (**A**) and active I/active II (**B**) DPP4 inhibitors.

Note: The scores and loadings plots are shown in the left and right panels, respectively, where actives/active I and inactives/active II DPP4 inhibitors are shown in the top and bottom rows, respectively.

Abbreviations: ALogP, Ghose–Crippen octanol–water partition coefficient; HOMO, highest occupied molecular orbital; HOMO–LUMO, energy gap between the HOMO and LUMO states; LUMO, lowest unoccupied molecular orbital; MW, molecular weight; nClC, number of rings; nHAcc, number of hydrogen bond acceptors; nHDon, number of hydrogen bond donors; PCA, principle component analysis; Q_m, mean absolute charge; RBN, rotatable bond number; TPSA, topological polar surface area.

(0.637), and LUMO (0.732) for active DPP4 inhibitors, whereas for inactive DPP4 inhibitors, PC2 highly correlated with nClC (–0.651), nHDon (0.605), ALogP (–0.813), and HOMO–LUMO (0.608). Furthermore, Table S2 indicates that the LUMO (17.664) and ALogP (24.441) descriptors highly contribute to PC1 for active and inactive DPP4 inhibitors, respectively. Table S1 indicates that descriptors consisting of nClC, nHDon, ALogP, energy, dipole moment, HOMO, LUMO, and HOMO–LUMO influence PC2 in either active or inactive DPP4 inhibitors. Remarkably, the eight different descriptors are reported with $P<0.001$ and are significantly different between active and inactive DPP4 inhibitors. These eight different descriptors may represent the informative features that discriminate between active and inactive DPP4 inhibitors.

PC3, which is the direction that is orthogonal to both PC1 and PC2, accounted for 13.97% and 14.70% of the total variance for actives and inactives, respectively. Figure S1 indicates that the first three components can preserve 65.26% (active) and 70.58% (inactive) of the original variance. The results indicated that the percentage and cumulative percentage of variance of inactive DPP4 inhibitors remained larger than those of active DPP4 inhibitors. This result is consistent with the observation that the distribution of active DPP4 inhibitors can be further divided into two groups represented by the score plots in Figure 3A. In Table S1 and Figure 3A (top-right), it can be seen that PC3 separates Q_m from nClC and ALogP for active DPP4 inhibitors, whereas in Figure 3A (bottom-right), PC3 separates dipole moment from HOMO and LUMO for inactive DPP4 inhibitors. Table S1 indicates that PC3 highly correlated with nCIC (0.739), ALogP (0.564), and Q_m (–0.542) for active DPP4 inhibitors, whereas PC3 highly correlated with dipole moment (–0.634), HOMO (0.698), and LUMO (0.671) for inactive DPP4 inhibitors. nCIC and HOMO were the descriptors with the highest correlation with PC1 for active and inactive DPP4 inhibitors, respectively. Furthermore, Table S2 indicates that the nCIC (30.046) and HOMO (25.521) descriptors highly contribute to PC3 for active and inactive DPP4 inhibitors, respectively. Table S1 indicates that descriptors consisting of Q_m, nCIC, ALogP, dipole moment, HOMO, and LUMO influenced PC1 in either active or inactive DPP4 inhibitors. Interestingly, the six differential descriptors are reported with $P<0.001$ and are significantly different between active and inactive DPP4 inhibitors. These six different descriptors may represent the informative features that discriminate between active and inactive DPP4 inhibitors.

Analysis of active DPP4 inhibitors

Figure 3B indicates that the data points of the scores plots (top-left) of active DPP4 inhibitors can be well discriminated into two subclasses (called active I and active II DPP4 inhibitors). We assumed that the inhibitors in this class may be further separated into subclasses. Thus, in this section, the active DPP4 inhibitors were analyzed according to subclasses. Table 2 indicates that nine descriptors exhibit different patterns between active I and active II DPP4 inhibitors at the level of $P<0.001$ except for the five descriptors RBN ($P=0.593$), nCIC ($P=0.001$), ALogP ($P=0.208$), TPSA ($P=0.026$), and Q_m ($P=0.001$). These five descriptors have average values of 5.000±2.240 (RBN), 3.103±0.895 (nCIC), 1.585±1.461 (ALogP), 89.981±26.752 (TPSA), and 0.223±0.034 (Q_m) for active I DPP4 inhibitors, whereas active II DPP4 inhibitors have average values of 5.052±1.468 (RBN), 3.263±0.784 (nCIC), 1.463±1.628 (ALogP), 86.867±22.370 (TPSA), and 0.217±0.032 (Q_m). In Figure 2B, the histograms of active I and active II DPP4 inhibitors indicated that these five descriptors were not different between the two subclasses. Therefore, except for these five descriptors, the remaining descriptors are significantly different for active I and active II DPP4 inhibitors and are efficient for discrimination.

Figure 3B shows the scores and loadings plots for active I (top-left) and active II (bottom-left) DPP4 inhibitors. It is observed that the distribution of active I and active II DPP4 inhibitors cannot be further divided. The cumulative variances of the first three PCs of active I and active II DPP4 inhibitors were 66.63% and 68.21%, respectively, of the original variation and obtain 80.0% of the original variation performed on the first five PCs. To analyze the highest influence of each descriptor on PC, the loadings and contribution values are used, as shown in Tables S3 and S4, respectively. PC1 highly correlated with MW (0.834), RBN (0.629), nHDon (0.587), nHAcc (0.675), TPSA (0.698), and HOMO–LUMO (–0.565) for active I DPP4 inhibitors, whereas PC1 highly correlated with energy (–0.815), dipole moment (–0.634), HOMO (0.699), LUMO (0.807), and HOMO–LUMO (0.560) for active II DPP4 inhibitors. For PC2, the descriptors energy (–0.743), dipole moment (–0.602), HOMO (0.634), and LUMO (0.702) highly correlated with this component for active I DPP4 inhibitors, whereas MW (0.750), ALogP (–0.551), and TPSA (0.797) highly correlated with this component for active II DPP4 inhibitors. The third PC highly correlated with nCIC (0.734), ALogP (0.619), and Q_m (–0.540) for active I DPP4 inhibitors, whereas PC3 highly correlated with nCIC (0.698) for active II DPP4 inhibitors. The descriptors of MW, energy, nCIC, and TPSA provide the absolute highest loadings score values on PC1, PC2, and PC3, respectively, for active I DPP4 inhibitors, whereas the descriptors energy, TPSA, and nCIC provide the absolute highest loadings score values on PC1, PC2, and PC3, respectively, for active II DPP4 inhibitors. These result are consistent with the contribution score of MW (17.639), energy (19.829), and nCIC (27.810), providing the highest values on PC1, PC2, and PC3, respectively, for active I DPP4 inhibitors, whereas the descriptors energy (16.193), TPSA (22.921), and nCIC (24.428) provide the highest PCA loadings score values on PC1, PC2, and PC3, respectively, for active II DPP4 inhibitors, as shown in Table S3.

Prediction and identification of informative molecular descriptors for DPP4 inhibitors

In this study, a QSAR model based on the J48 algorithm is presented for discriminating DPP4 inhibitors as either actives or inactives. Each compound was calculated as an M-dimensional vector where $M=13$. The encoded compounds from the DPP4-TRN set were then used to construct a QSAR model, which was represented by a DT. To evaluate the internal prediction capacity of our proposed QSAR model on the DPP4-TRN set, two different experiments were performed: one experiment was performed on the full training data and one experiment was evaluated using a ten-fold cross validation (CV) procedure as shown in Table 3. The CV procedure was performed by firstly partitioning the data into ten equally-sized segments or folds; then, nine folds

Table 3 Summary of prediction performance of internal and external sets

Dataset	Details	N	Acc (%)	Sen (%)	Spec (%)	MCC
Internal set (DPP4-TRN)	Full training	1,122	96.43	98.30	94.38	0.929
	Ten-fold CV	1,122	82.26	84.69	79.59	0.644
External set 1 (DPP4-TEST1)	External validation	149	91.28	–	–	–
External set 2 (DPP4-TEST2)	External validation	160	95.63	–	–	–
External set 3 (DPP4-TEST3)	External validation	167	72.25	–	–	–

Note: N is the number of compounds.
Abbreviations: Acc, accuracy; CV, cross-validation; MCC, Matthews correlation coefficient; Sen, sensitivity; Spec, specificity.

were used as the training data while the remaining fold was used for validation. Finally, the results were then averaged across the ten experiments. Four measurements were used to assess the performance of the QSAR models, namely accuracy (Acc), sensitivity (Sen), specificity (Spec), and the Matthews correlation coefficient (MCC). Our proposed QSAR model yielded 96.43% Acc, 98.30% Sen, 94.38% Spec, and 0.929 MCC as performed on the full training data. The prediction results from the tenfold CV procedure were 82.26% Acc, 84.69% Sen, 79.59% Spec, and 0.644 MCC. This result indicated the superiority of the 13 molecular descriptors in predicting DPP4 inhibitors to provide Acc higher than 80.0% and a MCC as high as 0.644.

Identification of informative molecular descriptors provided a better understanding of the different characteristics between active and inactive DPP4 inhibitors. After construction of the DT, the informative molecular descriptor could be identified using the feature usage score. A molecular descriptor having the highest feature usage is the most important feature because it contributes the most to prediction performances. Figure 4 shows the feature usage of each descriptor or descriptor usage by using the J48 algorithm on DPP4-TRN.[34] The top five informative molecular descriptors having a descriptor usage score larger than 30 were MW, LUMO, nHDon, nHAcc, and ALogP. Interestingly, for the five top-ranked and informative molecular descriptors, the distributions of active and inactive DPP4 inhibitors were significantly different, with $P<0.001$, as shown in Table 1. Furthermore, the three external validation sets were used for evaluating the robustness and generalization ability of the proposed QSAR model established from the DPP4-TRN. Figure S2 shows the overview of Tanimoto coefficient for the four dataset as a heatmap. For example, the top-right panel shows the heatmap of DPP4-TRN versus DPP4-TEST2. Prediction results for QSAR model of DPP4-TEST1, DPP4-TEST2, and DPP4-TEST3 achieved test accuracies of 91.28%, 95.63%, and 72.25%, respectively. Based on our results, it could be concluded that our proposed QSAR model was efficient in prediction of DPP4 inhibitors into either actives or inactives and filtration of inactive DPP4 inhibitors from active DPP4 inhibitors.

Chemical substructure analysis

Chemical substructure analysis of active and inactive DPP4 inhibitors is an effective approach to identify important chemical fragments that may govern the biological activity toward the DPP4 enzyme. Tables 4 and 5 summarize the top ten fragments of the active and inactive inhibitor classes,

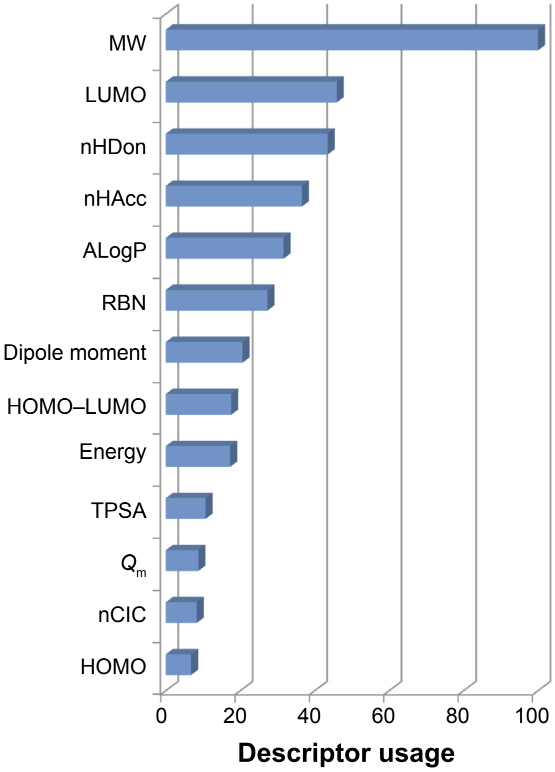
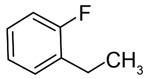
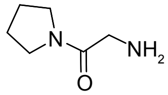
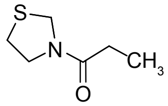
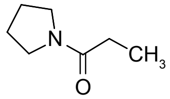
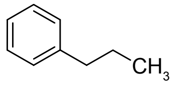
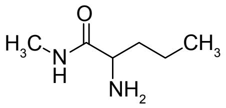
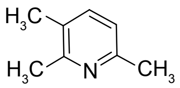
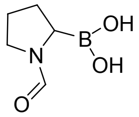
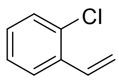
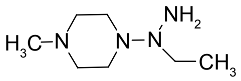


Figure 4 Plot of the descriptor usage derived from the J48 algorithm.
Note: The descriptor with the largest descriptor usage value is the most important.
Abbreviations: ALogP, Ghose–Crippen octanol–water partition coefficient; HOMO, highest occupied molecular orbital; HOMO–LUMO, energy gap between the HOMO and LUMO states; LUMO, lowest unoccupied molecular orbital; MW, molecular weight; nCIC, number of rings; nHAcc, number of hydrogen bond acceptors; nHDon, number of hydrogen bond donors; Q_m, mean absolute charge; RBN, rotatable bond number; TPSA, topological polar surface area.

Table 4 Summary of the top ten fragments in the active set of DPP4 inhibitors

Rank	IUPAC name	Structure	Fragment count
1	1-ethyl-2-fluorobenzene		617
2	2-amino-1-(pyrrolidin-1-yl)ethan-1-one		597
3	1-(1,3-thiazolidin-3-yl)propan-1-one		136
4	1-(pyrrolidin-1-yl)propan-1-one		101
5	Propylbenzene		52
6	2-amino-N-methylpentanamide		50
7	2,3,6-trimethylpyridine		45
8	(1-formylpyrrolidin-2-yl)boronic acid		43
9	1-chloro-2-ethenylbenzene		36
10	4-(1-ethylhydrazin-1-yl)-1-methylpiperazine		32

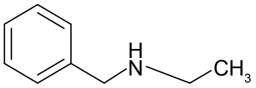
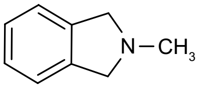
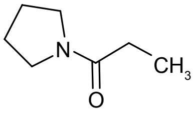
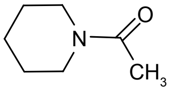
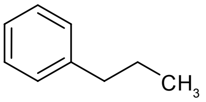
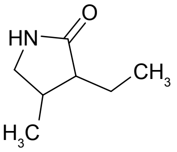
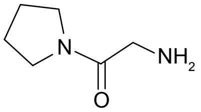
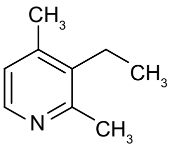
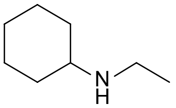
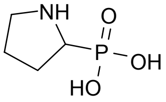
Abbreviation: IUPAC, International Union of Pure and Applied Chemistry.

respectively. The top ten fragments of active inhibitors indicated that pyrrolidine-based, thiazolidine-based, amino amide-based, pyridine-based, piperazine-based, and aromatic-based fragments are essential for DPP4 inhibition. The fragment 1-ethyl-2-fluorobenzene ranked first (617 counts), followed by 2-amino-1-(pyrrolidin-1-yl)ethan-1-one (597 counts). The occurrence of these top two fragments is clearly greater than that of the remaining fragments, as indicated by the fragment counts (Table 4), which indicate their important roles in DPP4 inhibition.

Because DPP4 prefers substrates containing proline or alanine at position 2 of the N-terminus, many inhibitors have been designed based on the peptidomimetic concept.[12] These peptidomimetic inhibitors are categorized as glycine-based and β-alanine-based types.[12] Pyrrolidine has been used as a core structure in the design of both inhibitor types with respect to its functional groups that play crucial roles for interaction at the active site of the enzyme. The DPP4 inhibitory activities of these inhibitors are similarly accomplished by hydrophobic and van der Waals interactions,[39] as well as hydrogen-bond and salt-bridge formation.[12] The active site of DPP4 consists of a catalytic triad (Ser630, H740, and D708), oxyanion hole, and specific residues, ie, S1 and S2 pockets.[12,39] All known DPP4 inhibitors have been reported to occupy these pockets for inhibition.[39]

The most frequently found fragment is 1-ethyl-2-fluorobenzene, which is a highly lipophilic aromatic-based fragment. The aryl substitution on the C-4 position of the pyrrolidine ring has been noted to improve the stability and duration of DPP4 inhibitors.[40] In addition, the fluorine

Table 5 Summary of the top ten fragments in the inactive set of DPP4 inhibitors

Rank	IUPAC name	Structure	Fragment count
1	Benzyl(ethyl)amine		102
2	2-methyl-2,3-dihydro-1H-isoindole		77
3	1-(pyrrolidin-1-yl)propan-1-one		64
4	1-(piperidin-1-yl)ethan-1-one		45
5	Propylbenzene		35
6	3-ethyl-4-methylpyrrolidin-2-one		19
7	2-amino-1-(pyrrolidin-1-yl)ethan-1-one		17
8	3-ethyl-2,4-dimethylpyridine		14
9	*N*-ethylcyclohexanamine		14
10	(Pyrrolidin-2-yl)phosphonic acid		13

substituent on the C-4 position of the pyrrolidine ring has been reported to provide good inhibitory properties, selectivity, and pharmacokinetic profiles.[41] The preferable pharmacokinetic profile may result from a lipophilic property governed by a planar aromatic ring and halogen atoms, which facilitates cell entry to the target site of action. In this study, a similar aromatic-based fragment containing a halogen atom, ie, 1-chloro-2-ethenylbenzene, was found as the ninth-ranked fragment. Additional aromatic-based fragments were also ranked as top ten fragments, such as propylbenzene and 2,3,6-trimethylpyridine. It could be hypothesized that the flexibility of the rotatable alkyl chain in the propylbenzene fragment may facilitate cell penetration and hydrophobic interactions at the active site, and the nitrogen atom in the pyridine ring of 2,3,6-trimethylpyridine may play a role in H-bond formation in the DPP4 active site.

The pyrrolidine amide is considered a key moiety in the design of DPP4 inhibitors.[12] Most of the potent inhibitors have been developed by substitution of the amide moiety of this core structure with an electrophile[42–44] that forms a covalent adduct with Ser630 of the DPP4 active site.[12] Therefore, it is not surprising that among the top ten fragments, the pyrrolidine-based fragments, ie, 2-amino-1-(pyrrolidin-1-yl)ethan-1-one, 1-(pyrrolidin-1-yl)propan-1-one, and (1-formylpyrrolidin-2-yl)boronic acid, appear to be the most frequently occurring fragments. Notably, the 2-amino-1-(pyrrolidin-1-yl)ethan-1-one fragment, which is presented in many compounds, has been used as a prototype for structural modification.[12] All of these fragments are amide derivatives of pyrrolidine. It is possible that the oxygen atom of the amide functional group may be essential for H-bond formation with the DPP4 active site.[12] In addition, the amine group has been noted for its role in forming a salt-bridge with Glu205 and/or Glu206 of DPP4.[12] Moreover, the boronic acid derivative of pyrrolidine amide, (1-formylpyrrolidin-2-yl)boronic acid, ranked eighth. This finding supported the fact that substitution of boronic acid at the 2-position of the pyrrolidine ring is effective for DPP4 inhibition, as observed from the progress of talabostat into Phase III clinical trials.[12,44]

The thiazolidine derivative fragment, 1-(1,3-thiazolidin-3-yl)propan-1-one, ranked third. Clearly, the shape of this fragment is similar to that of pyrrolidine amide derivatives (ie, 2-amino-1-(pyrrolidin-1-yl)ethan-1-one and 1-(pyrrolidin-1-yl)propan-1-one) except for the presence of a sulfur atom in the five-membered ring. The thiazolidine analog of pyrrolidine-based compounds has been noted for its stability, potency, selectivity, and oral bioavailability.[45–47]

The amide-based fragment, ie, 2-amino-*N*-methylpentanamide, was found to be the sixth-ranked fragment. The X-ray crystal structure indicated that the amide moiety is essential for a key interaction in DPP4 inhibition.[12] The amino group (-NH_2) forms a salt-bridge with Glu205, and the O atom of the carbonyl group (-C=O) forms an H-bond with Arg125 in the DPP4 active site.[12] In addition, the piperazine-based fragment, ie, 4-(1-ethylhydrazin-1-yl)-1-methylpiperazine, was found as the tenth-ranked fragment. DPP4 inhibitors containing a piperazine substituent have been reported to exhibit high potency.[48]

Notably, some fragments of active inhibitors, ie, 2-amino-1-(pyrrolidin-1-yl)ethan-1-one, 1-(pyrrolidin-1-yl)propan-1-one, and propylbenzene, were also found in the top ten fragments of the inactive inhibitor class. This finding may indicate that the inhibitory activities of DPP4 inhibitors are influenced by additional factors. The results of the inactive inhibitors (Table 5) indicated that the type and position of the substituents, type of functional groups, appropriate size, and arrangement of substructures may be crucial for DPP4 inhibition. For example, the effect of the position of substituents and the length of the alkyl chain were found when comparing 2,3,6-trimethylpyridine (active) and 3-ethyl-2,4-dimethylpyridine (inactive).

Scaffold analysis

Analysis of the molecular scaffold of DPP4 inhibitor was performed in order to discern important core structures giving rise to their bioactivity. Datasets of both active and inactive DPP4 inhibitors were subjected to molecular scaffold analysis using the Bemis–Murcko framework clustering method as implemented by JKlustor version 0.07.[49] In brief, this clustering method initially generates molecular frameworks representing molecular scaffolds as derived from compounds in datasets by removing side chain atoms from the main structures and finally presenting them in the form of a molecular graph, which is subsequently clustered based on the Bemis–Murcko framework algorithm.[50] A total of 332 and 152 scaffolds were obtained for actives and inactives, respectively. The large number of molecular scaffolds that were obtained is indicative of the higher diversity of molecular patterns presented in the dataset. Herein, this result suggests that molecular patterns in active DPP4 inhibitors are more diverse than their inactive counterpart. Further in-depth analysis of scaffolds from both active and inactive classes was performed by comparing members of each molecular scaffold from both classes. It was found that there were no significant differences in the molecular frameworks for both classes as can be seen in Tables S5 and S6 and Figure 5. This suggested that the important structures responsible for the bioactivity were functional groups as well as substructures of molecules.

In order to elucidate such important substructures, Klekota–Roth fingerprints consisting of 4,860 descriptors were generated by the PaDEL-Descriptor software on DPP4-TRN.[51,52] Consequently, a mean decrease of the Gini index (MDGI) as derived from random forest[53,54] was used as the basis for selecting the most important feature from the initial set of 4,860 descriptors. The descriptor having the highest MDGI value was deemed to be the most important feature because it affords the most influence to the prediction performance. The set of 30 top-ranked fingerprints having the largest MDGI values are summarized in Figure S3 and Table S7. It can be seen that the most important structural fingerprint is piperazine-1-carbaldehyde (KRFP4541) with a MDGI value as high as 9.197. Meanwhile, the second most important structural fingerprint with a MDGI value of 3.610 is the piperazine ring (KRFP2428). Interestingly, the significance of piperazine is supported by the fact that it is an important structural part of oral antihyperglycemic agents called gliptins, which target DPP4 receptors and have

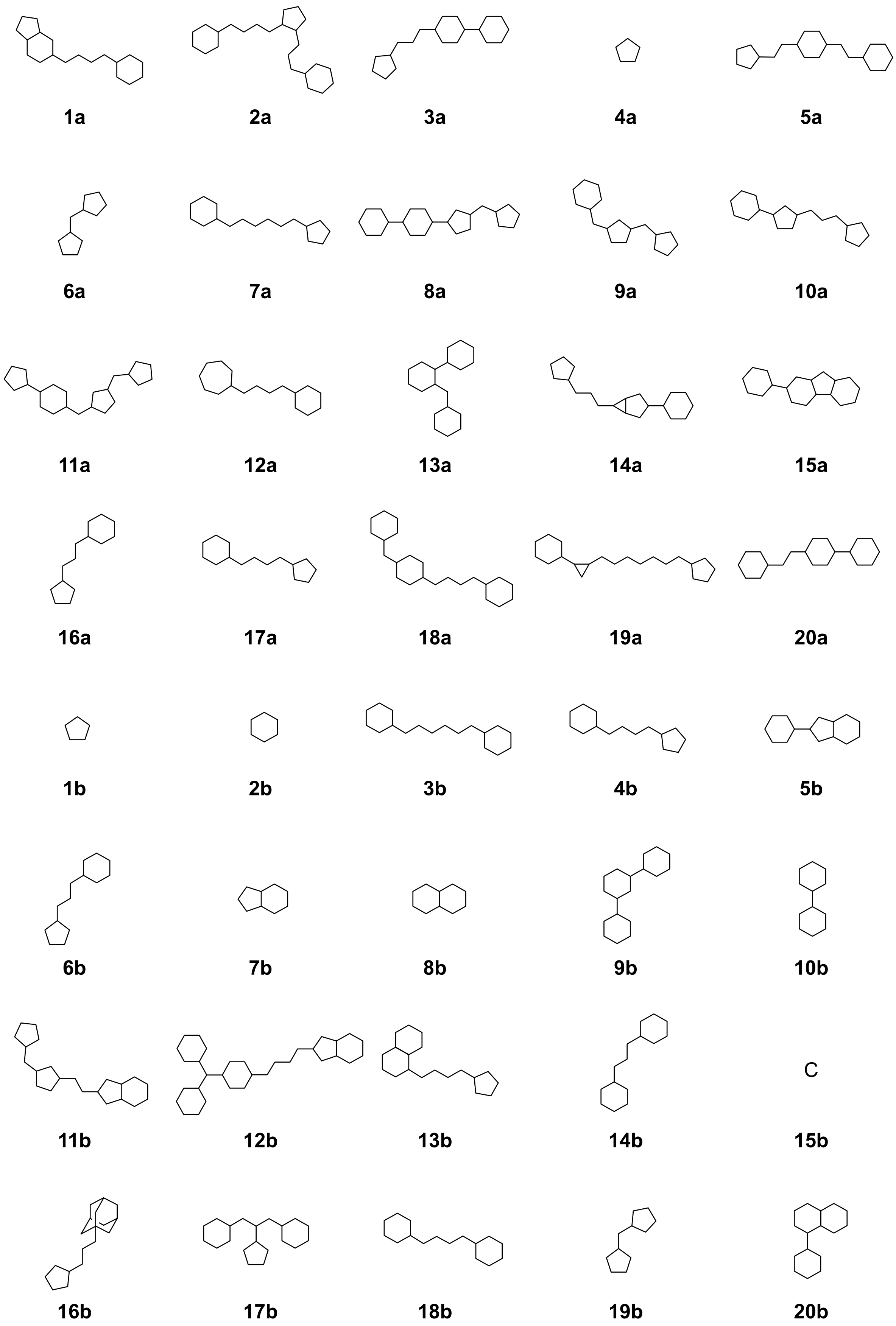


Figure 5 Summary of top 20 molecular frameworks for actives (1a–20a) and inactives (1b–20b).

been approved by the US Food and Drug Administration (FDA) for use in T2D treatment. Particularly, sitagliptin and teneligliptin, which are piperazine containing gliptins, have shown an additional mode of binding with the DPP4 receptor. In brief, the DPP4 inhibitors can be categorized into three classes according to their binding subsites.[55] Class I DPP4 inhibitors (ie, vildagliptin and saxagliptin) employed cyanopyrrolidine and hydroxyl adamantyl moieties to bind to S1 and S2 subsites of the DPP4 active site, respectively. In addition to the binding mode of class I, class II DPP4 inhibitors (ie, two recently released DPP4 inhibitors alogliptin and linagliptin) can further engage in π–π interaction with S'$_1$ and S'$_2$ subsites. As for class III DPP4 inhibitors (ie, sitagliptin and teneligliptin), the presence of the piperazine ring at the P2 position engages in interaction with the S2 extensive subsite and introduces the "anchor lock domain" resulting in an increase of the binding activity owing to the stronger hydrophobic interactions mediated by this domain.[56–59] In addition, results of contact area calculation of this domain also revealed correlation between the binding surface and the inhibitory activity against DPP4 receptor, further emphasizing the importance of this domain.[55] Nevertheless, the role of piperazine derivatives in DPP4 inhibitory activity is not only found in these two drugs but is also reported in various DPP4 inhibitors that are under active development.[60–62]

Binding mode of DPP4 inhibitors

Molecular docking and subsequent post-docking analyses using the SiMMap server identified the common binding mode of DPP4 inhibitors as well as key interactions with the enzyme. The SiMMap server provided a site-moiety map of the binding pocket along with details on conserved interacting residues, moiety preferences, and interaction types.[37] Analyses based on 100 active DPP4 inhibitors revealed three different binding anchors (*HB1*, *HB2*, and *vdW*) and their moiety preferences (Figure 6). The anchor *HB1* comprised side chains of Arg125, Glu205, Glu206, and Tyr662 while anchor *HB2* contained only the hydroxyl side chain of Tyr547. Both anchors were found to make hydrogen bonds with several nitrogen functional groups (ie, amine-, amide-, imine-, and nitrile-based) as well as ketone-based moieties of the inhibitors. In contrast, the anchor *vdW* consisted primarily of hydrophobic side chains of Tyr547, Tyr631, Trp659, Tyr662, and Tyr666 as well as the hydroxyl group of the catalytic residue Ser630. This pocket formed van der Waals contacts with aromatic, heterocyclic, and aliphatic moieties of DPP4 inhibitors. It should be noted that from our SiMMap analyses, the anchor *HB1* has been known as the S2 pocket, which is involved in key salt bridge interactions of either the free amino terminus of a peptide substrate or the cationic groups of an inhibitor with the carboxylate side chains of Glu205 (and/or Glu206) as well as the guanidinium side chain of Arg125, which also helps stabilize either the amide carbonyl group of a substrate or the ketone moiety of an inhibitor.[7,12] The anchor *vdW* corresponds to the S1 selectivity pocket of the enzyme that has been shown to be occupied with specific benzene- and pyrrolidine-based moieties of the DPP4 inhibitors.[7,12]

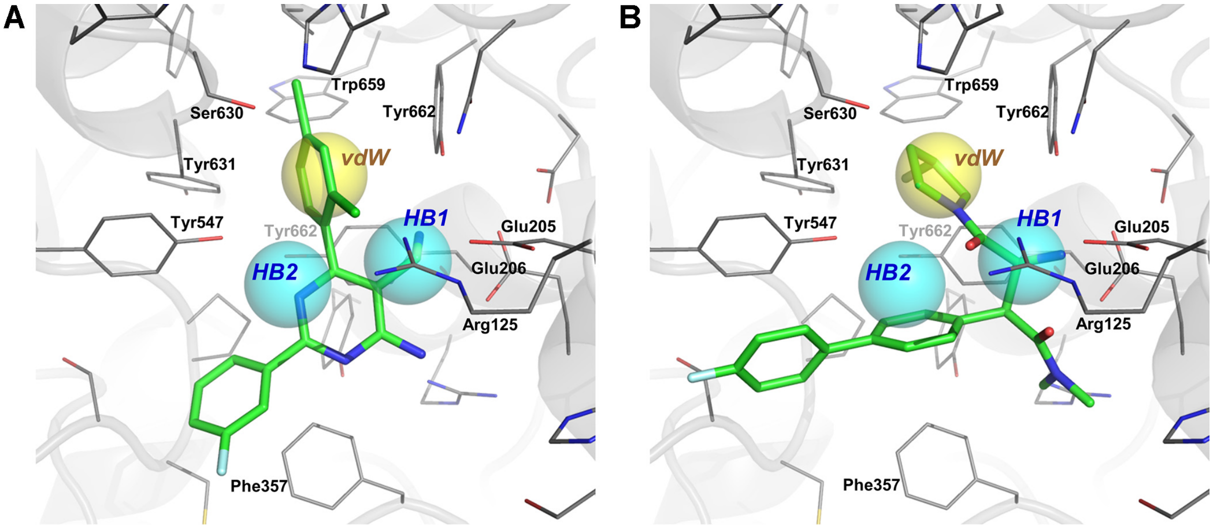


Figure 6 Three different binding modes of interaction of DPP4 inhibitors in the active site of the enzyme.

Notes: The identified anchors *HB1*, *HB2*, and *vdW* from the SiMMap server are labeled and shown in cyan and yellow spheres, respectively. Docking poses of two selected inhibitors are visualized herein: the compound with the best SiMMap score (**A**) and the compound with the lowest half maximal inhibitory concentration values (**B**). Residues at the active site are shown in green sticks while key interacting residues are labeled and shown in dark grey lines.

It should be noted that at least the first five DPP4 inhibitors with the best SiMMap score contained the amine-, amide-, and aromatic moieties for making interactions with all three different binding anchors (*HB1*, *HB2*, and *vdW*) of the enzyme. These findings suggested the significance of moiety preferences of inhibitors for binding and inhibiting DPP4 as well as serve as a general guideline for the design of novel inhibitors towards DPP4.

Comparison with FDA-approved drugs

In order to investigate the similarity between compounds investigated herein with those of FDA-approved DPP4 inhibitors, Tanimoto coefficient was computed for each compound in the dataset as well as six FDA-approved DPP4 inhibitors (ie, sitagliptin, vildagliptin, saxagliptin, alogliptin, linagliptin, and teneligliptin). The Tanimoto coefficient is a well-known metric for assessing the pairwise similarity between two molecules in which higher score represents high similarity. Results revealed that four of six DPP4 inhibitors (ie, sitagliptin, vildagliptin, saxagliptin, and linagliptin) were included in our curated dataset as observed from a Tanimoto coefficient of 1.000. The closest analog in our dataset to alogliptin and teneligliptin had Tanimoto coefficients of 0.819 and 0.602, respectively. Manual inspection of the pairwise Tanimoto coefficients between each compound of the dataset and the six FDA-approved drugs revealed that there were indeed several analogs of FDA-approved drugs present in the dataset. Such presence of analogs of FDA-approved drugs may densely populate the dataset and possibly mask the effect of less densely populated compounds. Concomitant with this issue is the observed imbalance in size of actives and inactives. Particularly, the rather small size of inactives may arise from the possibility that poor results for DPP4 inhibitory assays may not be published as often and therefore may contribute to the lower number of inactives. As fuzzy C-means clustering was applied in sampling the dataset for QSAR modeling, such aforementioned chemical space bias would not exert its influence on the constructed QSAR models.

A further look at the bioactivity of compounds exhibiting Tanimoto coefficient ≥ 0.5 to FDA-approved DPP4 inhibitors was performed. It was observed that the number of highly similar compounds with sitagliptin, vildagliptin, saxagliptin, alogliptin, linagliptin, and teneligliptin were 131, 273, 266, 87, 76, and 60 compounds, respectively. Of these compounds, a total of 130, 214, 192, 86, 76, and 60 compounds were classified as actives (IC_{50} less than 1 μM) for sitagliptin, vildagliptin, saxagliptin, alogliptin, linagliptin, and teneligliptin, respectively. Interestingly, a total of 59 and 74 compounds exhibiting high similarity with vildagliptin and saxagliptin, respectively, were classified as inactive. The R-group analysis of pyrrolidine as privileged structure of these molecules revealed pertinent insight of important substituent at positions 1, 2, and/or 5 on this ring. Alkyl group connected with nitrogen atom at position 1 seemed to be an important position since many structural modifications were observed at this position, which is followed by positions 2 and/or 5 where active moiety is usually nitrile. Herein, functional group and molecular fragment modifications based on commercially available DPP4 inhibitors could be a potent initial structure for further improving its bioactivity. Nevertheless, the agreement of binding mode to DPP4 receptor of any modified structures should be considered at the same time in order to abstain from steric effects that could lead to lowered bioactivity.

Furthermore, the Lipinski's rule of five was applied to the compiled compounds from all datasets and results are summarized in Table S8. Interestingly, it can be seen that compounds belonging to the internal set (DPP4-TRN) along with the external set (DPP4-TEST3) afforded roughly similar percentages of compounds passing the rule of five at approximately 90%, while DPP4-TEST1 and DPP4-TEST2 afforded close to 70%. The former sets contained primarily proteins belonging to the DPP family while the latter sets represent random proteins and proteases. Furthermore, actives (~94%) from DPP4-TRN provided higher percentages than their inactive counterpart (~84%–89%).

Limitations

In exploring the chemical space of DPP4 inhibitors through various means, an issue arises pertaining to the possibility of chemical space bias that may be inherently present in the compiled datasets. It should be noted that compounds were derived from the BindingDB, and although it is assumed to house nearly all (if not all) bioactivity data of DPP4, there is a possibility that some negative results for investigated compound series against DPP4 may not be published, while those that are published are those reporting favorable results for compounds affording nanomolar potency or those that further optimize lead compounds undergoing clinical trials. Bias may arise from medicinal chemists who may have inherent preference for certain chemical scaffolds, which could be attributed to the existence of common chemistry or the use of known fragments commonly found in drugs called privileged structures.[63] Thus, great caution should be taken in evaluating the essential functionality giving rise to potent bioactivity.

Conclusion

The search for novel antidiabetic agents has become increasingly important in drug design and development in light of the continual increase in the prevalence of diabetes worldwide. The inhibition of DPP4 is one strategy to combat diabetes. This study reports the large-scale chemical space exploration and QSAR investigation of DPP4 inhibitors. The QSAR model constructed by 13 descriptors provided good predictive performance as represented by an Acc close to 83.0% and a MCC as high as 0.644 for tenfold CV. In addition, a set of descriptors was identified as informative features influencing the predictive performance. The univariate analysis revealed the inherent physicochemical properties and important substructures governing inhibitory activity. The active inhibitors were found to be larger and more charged, polar, flexible, and stable than the inactive inhibitors. Furthermore, the chemical substructure analysis suggested that highly lipophilic aromatic-based and pyrrolidine-based fragments may be essential for DPP4 inhibition. Furthermore, the scaffold analysis revealed piperazine to be a privileged structure affording DPP4 inhibitory activity. Finally, our findings may provide a deeper understanding and pertinent knowledge for the design and development of DPP4 inhibitors.

Acknowledgments

This work was supported by the following funding agencies: Mahidol University Postdoctoral Fellowship (to W Shoombuatong under the supervision of C Nantasenamat and V Prachayasittikul) and an Annual Budget Grant (BE 2556-2558) from Mahidol University.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Clin Pract*. 2010;87:4–14.
2. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2014. 2014. Available from: http://www.cdc.gov/diabetes/data/statistics/2014statisticsreport.html. Accessed Juyly 7, 2015.
3. American Diabetes Association. Economic costs of diabetes in the U.S. in 2012. *Diabetes Care*. 2013;36:1033–1046.
4. Moller DE. New drug targets for type 2 diabetes and the metabolic syndrome. *Nature*. 2001;414:821–827.
5. Moneva MH, Dagogo-Jack S. Multiple drug targets in the management of type 2 diabetes. *Curr Drug Targets*. 2002;3:203–221.
6. Holst JJ, Deacon CF. Inhibition of the activity of dipeptidyl-peptidase IV as a treatment for type 2 diabetes. *Diabetes*. 1998;47:1663–1670.
7. Juillerat-Jeanneret L. Dipeptidyl peptidase IV and its inhibitors: therapeutics for type 2 diabetes and what else? *J Med Chem*. 2014;57:2197–2212.
8. Patel BD, Ghate MD. Recent approaches to medicinal chemistry and therapeutic potential of dipeptidyl peptidase-4 (DPP-4) inhibitors. *Eur J Med Chem*. 2014;74:574–605.
9. Kawalec P, Mikrut A, Łopuch S. The safety of dipeptidyl peptidase-4 (DPP-4) inhibitors or sodium-glucose cotransporter 2 (SGLT-2) inhibitors added to metformin background therapy in patients with type 2 diabetes mellitus: a systematic review and meta-analysis. *Diabetes Metab Res Rev*. 2014;30:269–283.
10. Karagiannis T, Boura P, Tsapas A. Safety of dipeptidyl peptidase 4 inhibitors: a perspective review. *Ther Adv Drug Saf*. 2014;5:138–146.
11. Zhao Y, Yang L, Zhou Z. Dipeptidyl peptidase-4 inhibitors: multitarget drugs, not only antidiabetes drugs. *J Diabetes*. 2014;6:21–29.
12. Havale SH, Pal M. Medicinal chemistry approaches to the inhibition of dipeptidyl peptidase-4 for the treatment of type 2 diabetes. *Bioorg Med Chem*. 2009;17:1783–1802.
13. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, et al. A practical overview of quantitative structure-activity relationship. *Excli J*. 2009;8:74–88.
14. Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. Advances in computational methods to predict the biological activity of compounds. *Expert Opin Drug Discov*. 2010;5:633–654.
15. Nicola G, Liu T, Gilson MK. Public domain databases for medicinal chemistry. *J Med Chem*. 2012;55:6987–7002.
16. Scior T, Bernard P, Medina-Franco JL, Maggiora GM. Large compound databases for structure-activity relationships studies in drug discovery. *Mini Rev Med Chem*. 2007;7:851–860.
17. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*. 2007;35:D198–D201.
18. Zhou B, Ha M, Wang C. An Improved Algorithm of Unbalanced Data SVM. In: Cao BY, Wang G, Chen S, Guo S, editors. *Fuzzy Information and Engineering 2010*. Vol 1. Berlin, Heidelberg: Springer; 2010:549–555.
19. Funatsu K, Miyao T, Arakawa M. Systematic generation of chemical structures for rational drug design based on QSAR models. *Curr Comput Aided Drug Des*. 2011;7:1–9.
20. Tetko IV, Sushko I, Pandey AK, et al. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model*. 2008;48:1733–1746.
21. Liu R, Tawa G, Wallqvist A. Locally weighted learning methods for predicting dose-dependent toxicity with application to the human maximum recommended daily dose. *Chem Res Toxicol*. 2012;25:2216–2226.
22. Osoda T, Miyano S. 2D-Qsar for 450 types of amino acid induction peptides with a novel substructure pair descriptor having wider scope. *J Cheminform*. 2011;3:50.
23. MarvinSketch, Version 6.2.1. Budapest: ChemAxon Ltd.; 2014. Available from: http://www.chemaxon.com
24. OpenEye Scientific Software. Babel, Version 3.3.0. Santa Fe: OpenEye Scientific; 2014. Available from: http://www.eyesopen.com
25. Frisch MJ, Trucks GW, Schlegel HB, et al. Gaussian 09, Revision A.1; Wallingford, Connecticut, 2009.
26. Nantasenamat C, Li H, Mandi P, et al. Exploring the chemical space of aromatase inhibitors. *Mol Divers*. 2013;17:661–677.
27. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med*. 1999;130:995–1004.
28. Jolliffe IT. Principal component analysis. In: Everitt BS, Howell DC, editors. *Encyclopedia of Statistics in Behavioral Science*. John Wiley and Sons, Inc.; 2005;3:1580–1584.
29. Ringnér M. What is principal component analysis? *Nat Biotechnol*. 2008;26:303–304.
30. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *J Stat Softw*. 2008;25:1–18.
31. Palomba D, Martínez MJ, Ponzoni I, Díaz MF, Vazquez GE, Soto AJ. QSPR models for predicting log P(liver) values for volatile organic compounds combining statistical methods and domain knowledge. *Molecules*. 2012;17:14937–14953.
32. Hammann F, Gutmann H, Baumann U, Helma C, Drewe J. Classification of cytochrome p(450) activities using machine learning methods. *Mol Pharm*. 2009;6:1920–1926.

33. Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. Amsterdam: Morgan Kaufmann; 2011.
34. Quinlan JR. *C4.5: programs for machine learning*. San Francisco: Morgan Kaufmann; 1993.
35. JChem, Version 14.8.18.0. Budapest: ChemAxon Ltd.; 2014. Available from: http://www.chemaxon.com
36. Morris GM, Huey R, Lindstrom W, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*. 2009;30:2785–2791.
37. Chen YF, Hsu KC, Lin SR, Wang WC, Huang YC, Yang JM. SiMMap: a web server for inferring site-moiety map to recognize interaction preferences between protein pockets and compound moieties. *Nucleic Acids Res*. 2010;38:W424–W430.
38. DeLano WL. PyMol, version 0.99; Schrödinger, LLC; 2010.
39. Metzler WJ, Yanchunas J, Weigelt C, et al. Involvement of DPP-IV catalytic residues in enzyme–saxagliptin complex formation. *Protein Sci*. 2008;17:240–250.
40. Kondo T, Sugimoto I, Nekado T, et al. Design and synthesis of long-acting inhibitors of dipeptidyl peptidase IV. *Bioorg Med Chem*. 2007; 15:2715–2735.
41. Haffner CD, McDougald DL, Reister SM, et al. 2-Cyano-4-fluoro-1-thiovalylpyrrolidine analogues as potent inhibitors of DPP-IV. *Bioorg Med Chem Lett*. 2005;15:5257–5261.
42. Senten K, Daniëls L, Van der Veken P, et al. Rapid parallel synthesis of dipeptide diphenyl phosphonate esters as inhibitors of dipeptidyl peptidases. *J Comb Chem*. 2003;5:336–344.
43. Demuth HU, Baumgrass R, Schaper C, Fischer G, Barth A. Dipeptidyl-peptidase IV – inactivation with N-peptidyl-O-aroyl hydroxylamines. *J Enzyme Inhib*. 1988;2:129–142.
44. Snow RJ, Bachovchin WW. Boronic acid inhibitors of dipeptidyl peptidase IV: A new class of immunosuppressive agents. In: Maryanoff BE, Maryanoff CA, editors. *Advances in Medicinal Chemistry*. Vol 3. Elsevier; 1995:149–177.
45. Sorbera LA, Revel L, Castañer J. P32/98: Antidiabetic Dipeptidyl-Peptidase IV Inhibitor. *Drugs Future*. 2001;26:859–864.
46. Epstein BJ. Drug evaluation: PSN-9301, a short-acting inhibitor of dipeptidyl peptidase IV. *Curr Opin Investig Drugs*. 2007;8:331–337.
47. Parmee ER, He J, Mastracchio A, et al. 4-Amino cyclohexylglycine analogues as potent dipeptidyl peptidase IV inhibitors. *Bioorg Med Chem Lett*. 2004;14:43–46.
48. Brockunier LL, He J, Colwell LF Jr, et al. Substituted piperazines as novel dipeptidyl peptidase IV inhibitors. *Bioorg Med Chem Lett*. 2004; 14:4763–4766.
49. JKlustor, Version 14.8.18.0. Budapest: ChemAxon Ltd.; 2014. Available from: http://www.chemaxon.com
50. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem*. 1996;39:2887–2893.
51. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. 2011;32: 1466–1474.
52. Klekota J, Roth FP. Chemical substructures that enrich for biological activity. *Bioinformatics*. 2008;24:2518–2525.
53. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
54. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and regression trees*. CRC press, New York; 1984.
55. Nabeno M, Akahoshi F, Kishida H, et al. A comparative study of the binding modes of recently launched dipeptidyl peptidase IV inhibitors in the active site. *Biochem Biophys Res Commun*. 2013;434:191–196.
56. Yoshida T, Akahoshi F, Sakashita H, et al. Discovery and preclinical profile of teneligliptin (3-[(2S,4S)-4-[4-(3-methyl-1-phenyl-1*H*-pyrazol-5-yl)piperazin-1-yl]pyrrolidin-2-ylcarbonyl]thiazolidine): A highly potent, selective, long-lasting and orally active dipeptidyl peptidase IV inhibitor for the treatment of type 2 diabetes. *Bioorg Med Chem*. 2012;20:5705–5719.
57. Yoshida T, Sakashita H, Akahoshi F, Hayashi Y. [(S)-gamma-(4-Aryl-1-piperazinyl)-l-prolyl]thiazolidines as a novel series of highly potent and long-lasting DPP-IV inhibitors. *Bioorg Med Chem Lett*. 2007;17:2618–2621.
58. Kim D, Wang L, Beconi M, et al. (2R)-4-oxo-4-[3-(trifluoromethyl)-5,6-dihydro[1,2,4]triazolo[4,3-a]pyrazin-7(8H)- yl]-1-(2,4,5-trifluorophenyl)butan-2-amine: a potent, orally active dipeptidyl peptidase IV inhibitor for the treatment of type 2 diabetes. *J Med Chem*. 2005; 48:141–151.
59. Xu J, Ok HO, Gonzalez EJ, et al. Discovery of potent and selective beta-homophenylalanine based dipeptidyl peptidase IV inhibitors. *Bioorg Med Chem Lett*. 2004;14:4759–4762.
60. Kim HJ, Kwak WY, Min JP, et al. Dipeptidyl peptidase-4 inhibitor with β-amino amide scaffold: synthesis, SAR and biological evaluation. *Bioorg Med Chem Lett*. 2012;22:5545–5549.
61. Kim MK, Chae YN, Kim HD, et al. DA-1229, a novel and potent DPP4 inhibitor, improves insulin resistance and delays the onset of diabetes. *Life Sci*. 2012;90:21–29.
62. Kim D, Kowalchick JE, Brockunier LL, et al. Discovery of potent and selective dipeptidyl peptidase IV inhibitors derived from beta-aminoamides bearing subsituted triazolopiperazines. *J Med Chem*. 2008; 51:589–602.
63. DeSimone RW, Currie KS, Mitchell SA, Darrow JW, Pippin DA. Privileged structures: applications in drug discovery. *Comb Chem High Throughput Screen*. 2004;7:473–494.

Supplementary materials

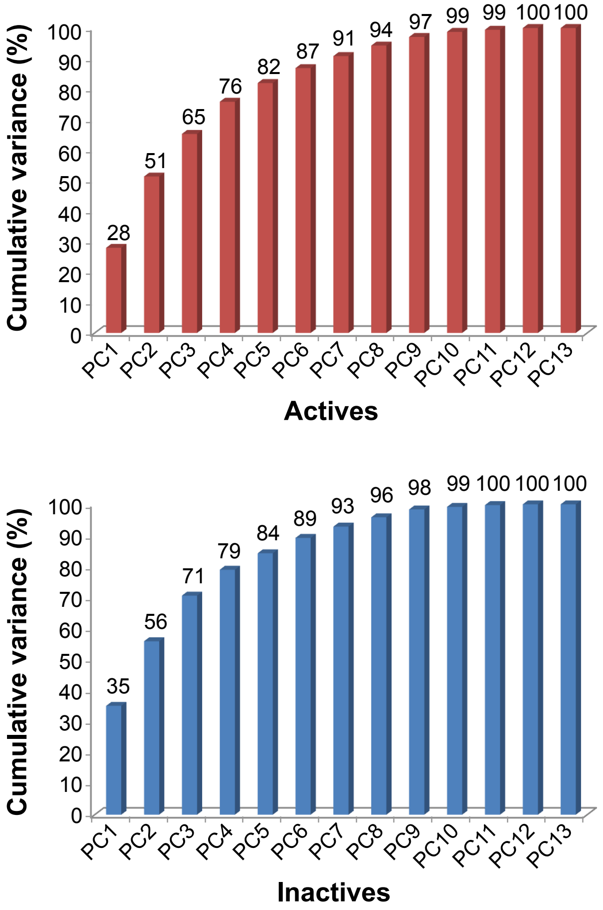


Figure S1 Cumulative variance from PCA analysis of active and inactive DPP4 inhibitors.
Abbreviation: PCA, principal component analysis.

Table S1 PCA loadings score for active and inactive DPP4 inhibitors

Descriptor	Active			Inactive		
	PC1	PC2	PC3	PC1	PC2	PC3
MW	**0.849**	0.195	0.402	**0.849**	–0.376	0.224
RBN	**0.638**	0.431	–0.025	**0.664**	–0.017	0.476
nCIC	0.297	–0.241	**0.739**	0.489	**–0.651**	0.077
nHDon	**0.578**	–0.231	–0.204	0.341	**0.605**	0.289
nHAcc	**0.637**	0.527	0.013	**0.893**	0.160	0.112
ALogP	0.061	0.267	**0.564**	0.164	**–0.813**	0.203
TPSA	**0.693**	0.268	–0.239	**0.815**	0.405	0.094
Q_m	0.445	0.410	**–0.542**	**0.601**	0.483	–0.144
Energy	–0.242	**–0.765**	0.102	**–0.654**	–0.264	–0.234
Dipole moment	0.361	**–0.617**	–0.261	0.439	0.091	**–0.634**
HOMO	–0.361	**0.637**	0.380	–0.211	–0.352	**0.698**
LUMO	–0.585	**0.732**	0.073	–0.559	0.344	**0.671**
HOMO–LUMO	**–0.575**	0.452	–0.413	–0.463	**0.608**	0.258

Notes: The bold values represent the highest loadings scores at the current PC, compared to other PCs. For instance, MW has a higher loading score of 0.849 at PC1, compared to PC2 (0.195), and PC3 (0.402).

Abbreviations: ALogP, Ghose–Crippen octanol–water partition coefficient; HOMO, highest occupied molecular orbital; HOMO–LUMO, energy gap between the HOMO and LUMO states; LUMO, lowest unoccupied molecular orbital; MW, molecular weight; nCIC, number of rings; nHAcc, number of hydrogen bond acceptors; nHDon, number of hydrogen bond donors; PCA, principal component analysis; Q_m, mean absolute charge; RBN, rotatable bond number; TPSA, topological polar surface area.

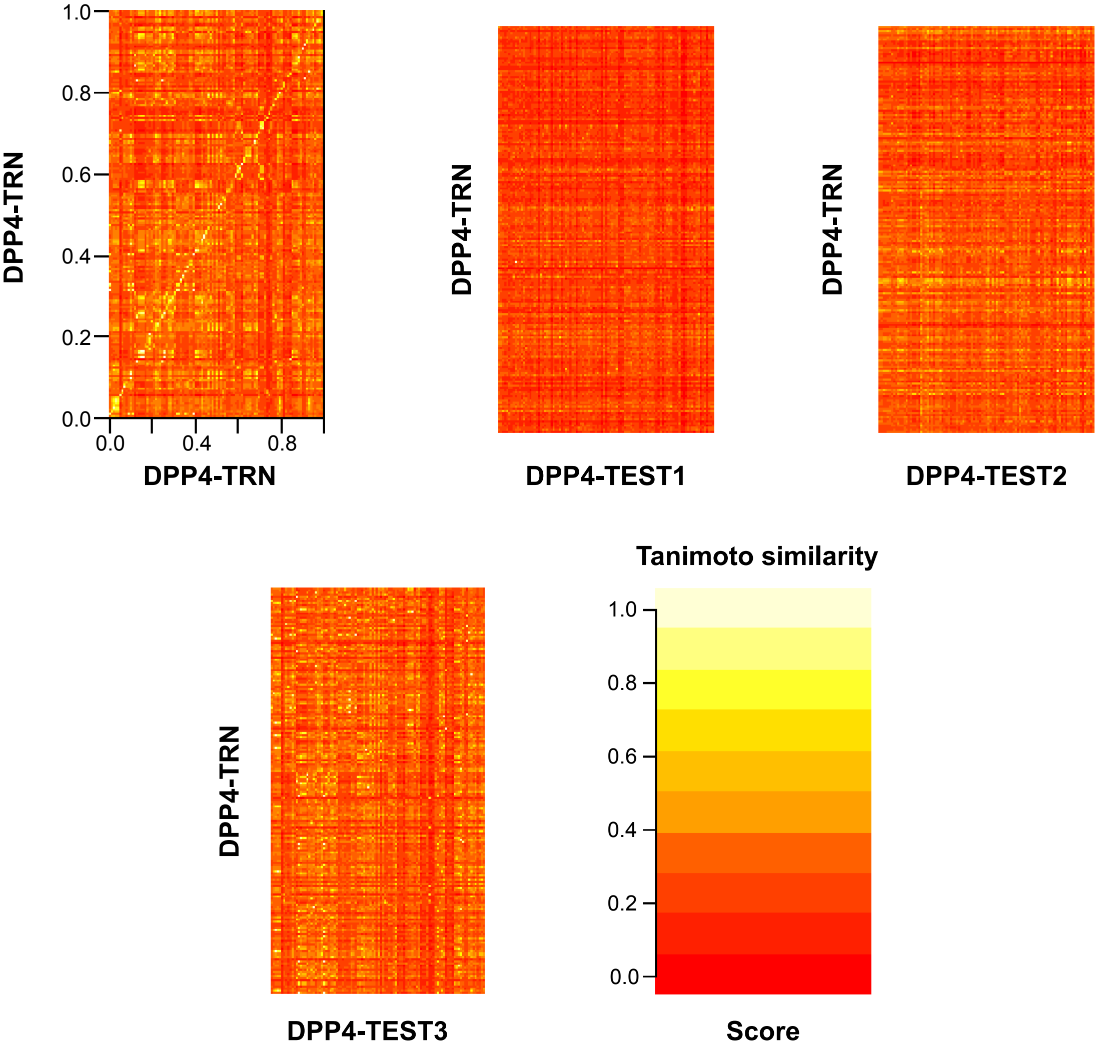


Figure S2 Heatmap of Tanimoto coefficient on five DPP4 datasets consisting of one internal set and four external validation sets. Tanimoto coefficient varies between 0 (total lack of similarity) to 1 (a compound has an identical constitution to a reference).

Table S2 Contribution value of each descriptor to principal component for active and inactive DPP4 inhibitors

Descriptor	Active			Inactive		
	PC1	PC2	PC3	PC1	PC2	PC3
MW	**19.851**	1.257	8.880	**15.820**	5.236	2.627
RBN	**11.202**	6.106	0.035	9.681	0.011	**11.868**
nCIC	2.421	1.914	**30.046**	5.246	**15.672**	0.308
nHDon	**9.214**	1.751	2.302	2.544	**13.523**	4.380
nHAcc	**11.191**	9.159	0.009	**17.499**	0.947	0.655
ALogP	0.101	2.341	**17.513**	0.589	**24.441**	2.160
TPSA(Tot)	**13.235**	2.371	3.139	**14.558**	6.069	0.466
Q_m	5.461	5.526	**16.146**	7.933	**8.633**	1.091
Energy	1.612	**19.254**	0.569	**9.382**	2.579	2.876
Dipole	3.594	**12.540**	3.757	4.220	0.306	**21.012**
HOMO	3.585	**13.378**	7.945	0.974	4.565	**25.521**
LUMO	9.419	**17.664**	0.291	6.851	4.380	**23.563**
HOMO–LUMO	9.114	6.738	**9.369**	4.701	**13.639**	3.472

Note: The bold values show the highest loadings scores at the current PC, compared to other PCs.

Abbreviations: ALogP, Ghose–Crippen octanol–water partition coefficient; HOMO, highest occupied molecular orbital; HOMO–LUMO, energy gap between the HOMO and LUMO states; LUMO, lowest unoccupied molecular orbital; MW, molecular weight; nCIC, number of rings; nHAcc, number of hydrogen bond acceptors; nHDon, number of hydrogen bond donors; PC, principal component; Q_m, mean absolute charge; RBN, rotatable bond number; TPSA, topological polar surface area.

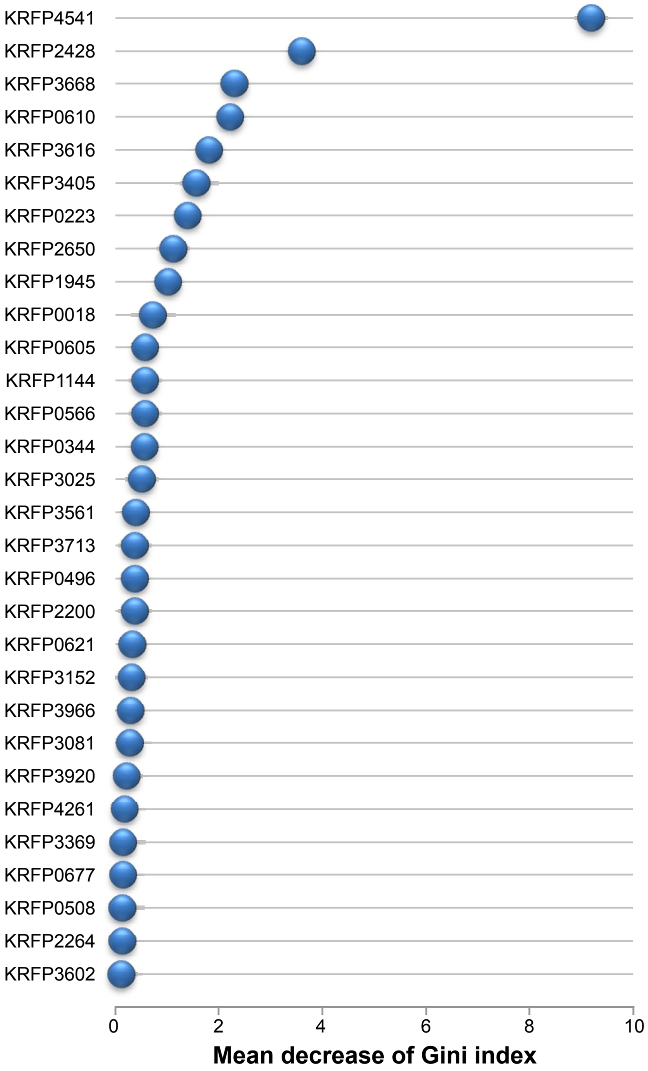


Figure S3 Important fingerprints of DPP4 inhibitors as ranked by the MDGI. The fingerprint with the largest MDGI value is deemed to be the most important.
Abbreviation: MDGI, mean decrease of Gini index.

Table S3 PCA loadings score for active I and active II DPP4 inhibitors

Descriptor	Active I					Active II				
	PC1	PC2	PC3	PC4	PC5	PC1	PC2	PC3	PC4	PC5
MW	**0.834**	0.241	0.412	–0.002	–0.047	0.338	**0.750**	–0.026	0.458	0.154
RBN	**0.629**	0.473	–0.001	–0.166	0.319	0.410	0.344	–0.493	0.266	–0.192
nCIC	0.281	–0.217	**0.734**	0.332	–0.267	–0.101	0.501	**0.698**	0.382	–0.066
nHDon	**0.587**	–0.148	–0.213	0.113	0.492	–0.388	0.506	–0.010	–0.257	**0.673**
nHAcc	**0.675**	0.463	0.007	–0.111	–0.487	0.796	0.253	–0.327	0.203	0.002
ALogP	0.112	0.230	**0.619**	–0.581	0.329	0.456	**–0.551**	–0.067	0.367	0.332
TPSA	**0.698**	0.257	–0.257	0.488	0.055	0.309	**0.797**	0.063	–0.322	–0.089
Q_m	0.448	0.383	**–0.540**	0.243	0.023	0.449	0.460	–0.439	**–0.472**	–0.097
Energy	–0.284	**–0.743**	0.104	0.389	0.184	**–0.815**	0.183	0.408	–0.076	–0.174
Dipole moment	0.515	**–0.602**	–0.285	–0.235	–0.101	**–0.634**	0.035	–0.443	0.182	–0.196
HOMO	–0.428	**0.634**	0.362	0.427	0.136	**0.699**	0.078	0.594	–0.073	–0.073
LUMO	–0.649	**0.702**	0.014	0.169	0.045	**0.807**	–0.155	0.488	–0.146	–0.046
HOMO–LUMO	**–0.565**	0.378	–0.469	–0.284	–0.105	**0.560**	–0.555	–0.022	–0.215	0.037

Note: The bold values show the highest loadings scores at the current PC, compared to other PCs.
Abbreviations: ALogP, Ghose–Crippen octanol–water partition coefficient; HOMO, highest occupied molecular orbital; HOMO–LUMO, energy gap between the HOMO and LUMO states; LUMO, lowest unoccupied molecular orbital; MW, molecular weight; nCIC, number of rings; nHAcc, number of hydrogen bond acceptors; nHDon, number of hydrogen bond donors; PCA, principal component analysis; Qm, mean absolute charge; RBN, rotatable bond number; TPSA, topological polar surface area.

Table S4 Contribution value of each descriptor to principal components for active I and active II DPP4 inhibitors

Descriptor	Active I					Active II				
	PC1	PC2	PC3	PC4	PC5	PC1	PC2	PC3	PC4	PC5
MW	**17.639**	2.080	8.743	0.000	0.259	2.787	**20.257**	0.035	18.959	3.269
RBN	10.046	8.042	0.000	2.121	**12.118**	4.100	4.257	**12.198**	6.411	5.086
nCIC	2.000	1.688	**27.810**	8.507	8.443	0.247	9.049	**24.428**	13.157	0.612
nHDon	8.749	0.790	2.352	0.989	**28.725**	3.664	9.219	0.005	5.945	**62.621**
nHAcc	11.573	7.711	0.003	0.945	**28.226**	**15.464**	2.310	5.355	3.723	0.000
ALogP	0.316	1.898	19.748	**26.044**	12.864	5.070	10.926	0.225	12.185	**15.284**
TPSA	12.370	2.369	3.408	**18.385**	0.356	2.336	**22.921**	0.201	9.384	1.108
Q_m	5.091	5.278	**15.070**	4.540	0.065	4.910	7.636	9.651	**20.154**	1.291
Energy	2.045	**19.829**	0.563	11.688	4.003	**16.193**	1.202	8.359	0.518	4.212
Dipole moment	6.731	**13.051**	4.200	4.271	1.203	9.800	0.044	**9.868**	2.981	5.291
HOMO	4.648	**14.434**	6.761	14.085	2.189	11.906	0.220	**17.690**	0.477	0.737
LUMO	10.690	**17.698**	0.010	2.190	0.243	**15.875**	0.863	11.962	1.929	0.296
HOMO–LUMO	8.103	5.131	**11.332**	6.234	1.308	7.647	**11.096**	0.023	4.179	0.193

Note: The bold values show the highest loadings scores at the current PC, compared to other PCs.

Abbreviations: ALogP, Ghose–Crippen octanol–water partition coefficient; HOMO, highest occupied molecular orbital; HOMO–LUMO, energy gap between the HOMO and LUMO states; LUMO, lowest unoccupied molecular orbital; MW, molecular weight; nCIC, number of rings; nHAcc, number of hydrogen bond acceptors; nHDon, number of hydrogen bond donors; PC, principal component; Q_m, mean absolute charge; RBN, rotatable bond number; TPSA, topological polar surface area.

Table S5 Summary of molecular framework generated from active DPP4 inhibitors

Number	SMILES	Member size
1	C(CCC1CCC2CCCC2C1)CC1CCCCC1	93
2	C(CCC1CCCCC1)CC1CCCC1CCCC1CCCCC1	87
3	C(CC1CCCC1)CC1CCC(CC1)C1CCCCC1	85
4	C1CCCC1	65
5	C(CC1CCC(CCC2CCCCC2)CC1)C1CCCC1	50
6	C(C1CCCC1)C1CCCC1	43
7	C(CCCC1CCCCC1)CCC1CCCC1	37
8	C(C1CCCC1)C1CCC(C1)C1CCC(CC1)C1CCCCC1	33
9	C(C1CCCC1)C1CCC(CC2CCCCC2)C1	32
10	C(CC1CCCC1)CC1CCC(C1)C1CCCCC1	32
11	C(C1CCCC1)C1CCC(CC2CCC(CC2)C2CCCC2)C1	31
12	C(CCC1CCCCCC1)CC1CCCCC1	31
13	C(C1CCCCC1)C1CCCCC1C1CCCCC1	30
14	C(CC1C2CC(CC12)C1CCCCC1)CC1CCCC1	29
15	C1C2CCCCC2C2CCC(CC12)C1CCCCC1	27
16	C(CC1CCCC1)CC1CCCCC1	27
17	C(CCC1CCC(CC2CCCCC2)CC1)CC1CCCCC1	27
18	C(CCC1CCCCC1)CC1CCCC1	25
19	C(CC1CCC(CC1)C1CCCCC1)C1CCCCC1	24
20	C(CCCC1CC1C1CCCCC1)CCCC1CCCC1	24
21	C(C1CCCC1)C1CCC(C1)C1CCCCC1	23
22	C1CCC(CC1)C1CCC2C(CCC3CCCCC23)C1	22
23	C(CCCC1CCCC1)CCCC1CCCCC1	22
24	C1CCC(CC1)C1CCCC2CCCCC12	21
25	C1CCC(CC1)C1CCCCC1	21
26	C(CC1CCCC1)CC1CCC(CC1)C1CCC2CCCC2C1	20
27	C1CCC(CC1)C1CCCC(C1)C1CCCCC1	20
28	C(C1C2CCCCC2CC1C1CCCCC1)C1CCCCC1	19
29	C(CC1CCCC1)CC1CC2CCCC2C1	18
30	C(C1CCC2CC(CC2C1)C1CCCCC1)C1CCCC2CCCCC12	18
31	C1CC2CCC(CC2C1)C1CCCCC1	18
32	C(CC1CCCC1)CC12CC3CC(CC(C3)C1)C2	17
33	C(CC1CCCC1)CC1CCC(CCC2CCCCC2)CC1	17
34	C(CC1CCCCC1)C1CCC(CC2CCCC2)C1	17
35	C(CCC1CCC2CCC(C2C1)C1CCCCC1)CC1CCCCC1	16
36	C(CC1CCC2CC(CC2C1)C1CCCCC1)C1CCCCC1	16
37	C(CC1CCC(CCC2CCCCC2)CC1)C1CCCCC1	16
38	C(C1CCCC1)C1CCC(CC2CC3CCCCC3C2)C1	15

(*Continued*)

Table S5 (*Continued*)

Number	SMILES	Member size
39	C(CC1CCCC1)CC1CCC(CCCC2CCCCC2)CC1	15
40	C(CCC1CCCCC1)CC1CCCC1CCC1CCCCC1	15
41	C(CC1CCCCC1)C1CCCCC1	14
42	C(CC1CCCC1)CC1CCCC1	14
43	C(CCC1CCCCC1)CCC1(CCCCC1)C1CC2CCCCC2C1	14
44	C(C1CCCCC1)C1CC2CCCCC2CC1C1CCCCC1	13
45	C(CC1CCC(CCC2CCCCC2)C1)C1CCCC1	13
46	C(CCC1CCCC(CC2CCCCC2)CC1)CC1CCCCC1	13
47	C(CCC1CCCCC1)CC1CCCC1C1CCC(C1)C1CCCCC1	13
48	C(CCC1CCC2CCCC2C1CC1CCCCC1)CC1CCCCC1	13
49	C1CCC(CC1)C1CCC(CC1)C1CCCCC1	13
50	C(CC1CCCCC1)C1CCCC1	12
51	C(CCC1CCCC1)CCC1CCC(C1)C1CCCCC1	12
52	C(C1CCCC1)C1CCC(C1)C1CCC(CC1)C1CC2CCCCC2C1	11
53	C(CCC1CCCCC1)CC1CCCCC1	11
54	C1CC2CCC(CC2C1)C1CCC(CC1)C1CCCCC1	11
55	C(CCC1CCC2CC(CC2C1)C1CCCCC1)CC1CCCCC1	11
56	C(CCCCC1CCCCC1)CCCC1CCCC1	11
57	C(C(CC1CCCCC1)C1CCCC1CC1CCCC1)C1CCCCC1	10
58	C(C1CCC2CC(CC2C1)C1CCCCC1)C1CCC2CCCCC2C1	10
59	C(CCC1CCC(CCC2CCCC2)CC1)CC1CCCCC1	10
60	C(CCCC1CCCC1)CCC1CCCC1	10
61	C(CCC1CCCC1)CCC1CCC2CCCC2C1	10
62	C(CCC1CCCCC1CC1CCCCC1)CC1CCCCC1	10
63	C(CCCC1CCCC1)CCCC1CCC2CCCC2C1	10
64	C1CCC(CC1)C1CCC(CC1)C1CCC2CCCCC2C1	9
65	C(CCCCC1CCCC1)CCCCC1CCCCC1	9
66	C(CCC1CCCC1)CCC1CCCC1	9
67	C(CC1CCC(CCC2CCCC2)C1)CC1CCCCC1	8
68	C(C1CCCC1)C1CCC(CC2CCC3CCCCC3C2)C1	8
69	C(CC1CCCC1)CC1CCC(CC1)C1CCCC(C1)C1CCCC1	8
70	C(CCC1CCCCC1)CC1CCCC1CCCC1CC1	8
71	C(C1CCCC1)C1CCC(C1)C1CCC(CC1)C1CCCC2CCCCC12	8
72	C(CC1CCCCC1C1CCCCC1)C1CCCC1	8
73	C(CCCC1CCC2CCCCC2C1)CCC1CCCC1	8
74	C(CCC1CCC(CC1)C1CCCCC1)CC1CCCCC1	8
75	C(C1CCCCC1)C1CC2CCCC2CC1C1CCCCC1	8
76	C(CC1CCCC1)CC1CCC(CC2CCCCC2)CC1	8
77	C1CCC(C1)C1CCC2C(CCC3CCCCC23)C1	7
78	C(CCC1CCCC1)CCC1CCCCC1	7
79	C(C1CCC2CC(CC2C1)C1CCCCCC1)C1CCC2CCCCC2C1	7
80	C(CCC1CCCCCC1CC1CCCCC1)CC1CCCCC1	7
81	C(CCC1CCC2CCCCC2C1)CC1CCCC1	7
82	C(CC1CCCCC1)CC1CCCCC1C1CCCCC1	7
83	C(C1CCCC1)C1CCC(C1)C1CCC2CCCCC12	7
84	C1CC(CC1C1CCCCC1)C1CCCC(C1)C1CCCCC1	6
85	C(CCC1CCC(CCCC2CCCCC2)CC1CC1CCCCC1)CC1CCCCC1	6
86	C(CC1CCCCC1)CC1CCCC(CCC2CCCC2)C1	6
87	C(CCC1CCCCC1)CC1CCCC1C1CCC(C1)C1CC1	6
88	C(C1CCC2CC(CC2C1)C1CCCCCC1)C1CCCC2CCCCC12	6
89	C(CC1CCCC1CC1CCCC1)CC12CC3CC(CC(C3)C1)C2	6
90	C(CC1CCCC1)C(CC1CC1)C1CCC(C1)C1CCCCC1	6
91	C(CCC1CCC2CCC(C3CCCC3)C2C1)CC1CCCCC1	5
92	C(CCC1CCC2CC(CC3CCCCC3)CC2C1)CC1CCCCC1	5
93	C1CCC(C1)C1CCCC(C1)C1CCC(C1)C1CCCCC1	5
94	C1CC2CCCC2C1	5
95	C(CC1CCC(C1)C1CCCCC1)CC1CCCC2CCCCC12	5

(*Continued*)

Table S5 (*Continued*)

Number	SMILES	Member size
96	C1C(CC2CCCCC12)C1CCCCC1	5
97	C(CCC1CCC2CC(CC2C1)C1CC1)CC1CCCCC1	5
98	C(CCC1CCCCC1)CC1CCCC1CC1CCCCC1	4
99	C(CC1CCCC1)CC1CCC(CC1)C1CCCC2CCCC12	4
100	C(CC1CCC(CC2CCCC2)C1)CC1CCCCC1	4
101	C1CC1C1CCCC2CCC(CC12)C1CCC(C1)C1CCCCC1	4
102	C(CCC1CCC2CCCC2C1CC1CCCC1)CC1CCCCC1	4
103	C(CC1CCC1)CC1CCC(CC1)C1CCCCC1	4
104	C(C1CCCCC1)C1CC(CCC1C1CCCCC1)C1CCCC1	4
105	C(CC1CCC(CC2CCCCC2)CC1)C1CCCCC1	4
106	C(CC1CCCC1)CC1C2CC3CC(C2)CC1C3	4
107	C(C1CCCCC1)C1CC(CCC1C1CCCCC1)C1CCCCC1	4
108	C(CCC1CCC2C(CCC2C2CCCCC2)C1)CC1CCCCC1	4
109	C(CCC1CCC(CC2CC3CCCCC3C2)CC1)CC1CCCCC1	4
110	C(CCCC1CCCCC1)CCC1CCCCC1	4
111	C(CC1CCC2CC(C(CC3CCCCC3)C2C1)C1CCCCC1)C1CCCCC1	4
112	C1CCC(CC1)C1CCC2CCCC(C3CCCCC3)C2C1	4
113	C(CC1CCC(CCC2CCCC2)C1)CC1CCCC2CCCCC12	4
114	C(CC1CCCCC1)CC1CCC(CCC2CCCC2)CC1	4
115	C(CCC1CCCC1)CCC1CC2CCCCC2C1	4
116	C(CC1CCC(C(CC2CCCCC2)C1)C1CCCCC1)C1CCCCC1	3
117	C(CCC1CC2CCCCC2C1)CC1CCCC1	3
118	C(C1CCC1)C1CCC(C1)C1CCC(CC1)C1CCCCC1	3
119	C(CCC1CCC(CC2CCC3CCCCC3C2)CC1)CC1CCCCC1	3
120	C1CC2CC(CC2C1)C1CCCCC1	3
121	C(CCC1CCCC1)CCC1CCC2CCCCC2C1	3
122	C(CCC1CCCCC1)CCC1(CCCCC1)C1CCCCC1	3
123	C(CCC1CCCCC1)CC1CCCC1C1CCCC1	3
124	C(C1CCCCC1)C1CCC2CC(CC2C1)C1CCCCC1	3
125	C(CC1CCCCC1)CC1CCC(CC1)C1CCCCC1	3
126	C(CCC1CCCCC1)CC1CCCC1C1CCC(C1)C1CCC1	3
127	C(CC1CCCC1)CC12CC3CC(C1)CC(CCCC1CCCCC1)(C3)C2	3
128	C(CC1CCCC1)C1CCCC1	3
129	C(CC1CC1)CC1CCCC1	3
130	C(CC1CCC(CCC2CCCC3CCCCC23)CC1)C1CCCC1	3
131	C(CC1CCC(CCC2CCC3CCCCC3C2)CC1)C1CCCC1	3
132	C	3
133	C(CC1CCC1)CC1CCCC1	3
134	C(CC1CCCC1)CC1CCCCCCC1	3
135	C(CCCC1CC2CCCCC2C1)CCC1CCCC1	3
136	C(C1CCCC1)C1CCCC(C1)C1CCCCC1	2
137	C(CCC1CCCC(CC1)C1CCCCC1)CC1CCCCC1	2
138	C(C1CC1)C1CCCC(C1)C1CCC(C1)C1CCCCC1	2
139	C(CCCC1CCCC1)CCCC1CCC2CC(CC2C1)C1CCCCC1	2
140	C(CC1CCCC(C1)C1CCCCC1)C1CC1	2
141	C(CCC1CCCC(CC1)C1CC1)CC1CCCCC1	2
142	C(CC1C2CC(CC12)C1CC2CCCCC2C1)CC1CCCC1	2
143	C(CC1CCC(C1)C1CCCCC1)CC1CCCCC1	2
144	C1CCC(C1)C1CCC2CCCC(C3CCCCC3)C2C1	2
145	C(CCCC1CCCC1)CCCC1CCCC1	2
146	C(CCCCCC1CCCCC1)CCCCC1CCCC1	2
147	C(CC1CCCC1)CC1(CCC2CCCCC2)CCCC1	2
148	C(CC1CCC(CCC2CCCC3CCCCC23)C1)C1CCCC1	2
149	C(CCC1CCC(CC2CCC(CC2)C2CCCCC2)CC1)CC1CCCCC1	2
150	C(CCC1CCCC1)CC1CCCC1	2
151	C(CC1CCCC1)CC1CCCCC1C1CCCCC1	2
152	C(CCC1CCCCC1)CC1CCCC1CCCC1CCCC1	2

(*Continued*)

Table S5 (*Continued*)

Number	SMILES	Member size
153	C(C1CC2CCCC2C1)C1CCCCC1	2
154	C1CC2CCCCCCCCCCCCCCC3CCCC(C3)CCC2C1	2
155	C(CC1CCCC1)CC1CC2CCCCC2C1	2
156	C(CC1CC1)CC1CCC(CCCC2CCCC2)CC1	2
157	C(CC1CCCC1)CC1CCC(CCCC2CCCC2)CC1	2
158	C(CC1CCCC1)CC1CC2CC(CC3CCCCC3)CC2C1	2
159	C1CCC2CCCCC2C1	2
160	C(CC1CCC(CC1)C1CCC2CCCC2C1)C1CCCCC1	2
161	C(CC1CCC(CC1)C1CCCC1)C1CCCCC1	2
162	C1CC2CCC(CC2C1)C1CCCC(C1)C1CCC(C1)C1CCCCC1	2
163	C(CC1CCC(CCC2CCC(CC2)C2CCCC2)CC1)C1CCCCC1	2
164	C(CC1CCCC1)CC1CCCCCC1	2
165	C(C1CCCCC1)C1CCCC(C1)C1CCCCC1	2
166	C(CC1CCCC1)CC1CC2CCC1C2	2
167	C(CCC1CCCCC1)CC1CCCC1CCCC1CCC2CCCC2C1	2
168	C(CCC1CCCCC1)CC1CCCC1CCCC1CCC2CCCCC2C1	2
169	C(C1CCCC1)C1CCCCC1	2
170	C(C1CCCC1)C1CCC(C1)C1CCCC2CCCCC12	2
171	C(CCC1CCC2CCCC2C1CC1CC1)CC1CCCCC1	2
172	C(C1CCCC1)C1CCC(CC2CCCC2)C1	2
173	C(CCC1CCCC1CC1CCCC1)CCC1CCCCC1	2
174	C(C1CCC1)C1CCC(CC2CCCC2)C1	2
175	C(CCCC1CCCC1)CCCC1CCC2CCCCC2C1	2
176	C(CCC1CCC2C(CCC2C2CC2)C1)CC1CCCCC1	2
177	C(C1CCC2CCCC2C1)C1CCC2CCCCC2C1	2
178	C(C1CCC2CC(CC2C1)C1CCCCC1)C1CC2CCCCC2CC2CCCCC12	2
179	C1C(CC2CCCCC12)C1CCC(CC1)C1CCCCC1	2
180	C(C(CC1CCCCC1)C1CCC(CC2CCCC2)C1)C1CCCCC1	2
181	C(C1C2CC(CC3CCCC4CCCCC34)CCC2CC1C1CCCCC1)C1CCCCC1	2
182	C(CC1CCCC1)CC1CCC(CC1)C1CCCC(C1)C1CCCCC1	1
183	C(C1CCC1)C1CCCC1	1
184	C(C1CCCCC1)C1CC2CC(CC2CC1C1CCCCC1)C1CCCCC1	1
185	C(CC1CCC(CC1)C1CCC(CC1)C1CCC(CC1)C1CCCCC1)C1CCCCC1	1
186	C(C1CCCCC1)C1CCCCC1CC1CCCCCC1	1
187	C(C1CCCC1)C1CCCCC1C1CCCCC1	1
188	C(CC1CCC(CC2CCCCC2)C1)C1CCCC1	1
189	C(CC1CCCC1)CC1CCC(CCC2CC2)CC1	1
190	C(C1CCCCC1)C1CC2CCC(CC2CC1C1CCCCC1)C1CCCCC1	1
191	C(C1CCCCC1)C1CCC(CC1C1CCCCC1)C1CCCCC1	1
192	C(CCCC1CCCCCC1)CCC1CCCC1	1
193	C(CCCC1C2CCC1CC2)CCC1CCCC1	1
194	C(CCC1CCC(CC2CCCC(C2)C2CCCCC2)CC1)CC1CCCCC1	1
195	C(CCCC1CCCC1)CCC1CC1C1CCCCC1	1
196	C(CCC1CCC(CC1)C(C1CCCCC1)C1CCCCC1)CC1CCCCC1	1
197	C(CC1CCCC1)CC1CCC(CCCC2CCC3CCCCC3C2)CC1	1
198	C(CCC1CCC(CC1)C1CCCCC1C1CCCCC1)CC1CCCCC1	1
199	C(CC1CCCC1)CC1CCC(CCCC2CC3CCCCC3C2)CC1	1
200	C(CC1CCCC1)CC1CC2CC(CC3CCCC3)CC2C1	1
201	C1CC2CCCCCCCCCCCCCCCCCCCCC3CCCC(C3)CCC2C1	1
202	C(CCCC1CCC(CC1)C1CCCCC1)CCC1CCCC1	1
203	C(C1CCCC1)C1CCC(C1)C1CCC(CC1)C1CCC2CCCCC2C1	1
204	C(CCC1CCC(CC2CCCC2)CC1)CC1CCCCC1	1
205	C(CC1CCCC1)CC1CC2CC(CCCC3CCCCC3)CC2C1	1
206	C(CCCC1CCC(CC1)C1CC2CCCCC2C1)CCC1CCCC1	1
207	C(CCC1CCCCC1)CCC1CCC(CC(CC2CCC(CCCCCC3CCCCC3)CC2)C2CCCC2CC2CCCC2)CC1	1
208	C(CC1CC2CCCCC2C1)C1CCC(CC2CCCC2)C1	1

(*Continued*)

Table S5 (*Continued*)

Number	SMILES	Member size
209	C(C1CC2CCCCC2C1)C1CCC2CC(CC2C1)C1CCCCC1	1
210	C(CC1CCCCC1)C1CCCC1CC1CCCC1	1
211	C(C1CCCC1)C1CCCC1CC1CC2CCCCC2C1	1
212	C(C1CCCC1)C1CCC(CC2CCC(CC2)C2CCC(C2)C2CCCCC2)C1	1
213	C(CCC1CCC(CC2CCC3CCCCC23)CC1)CC1CCCCC1	1
214	C(CC1CCCCC1)C1CCCC1CC1CC2CCCCC2C1	1
215	C(CCC1CCCC1CCC1CCCCC1)CC1CCCC1	1
216	C(CC1CCCCC1)C1CCC1CC1CCCC1	1
217	C(CC1CCCCC1)C1CCC1CC1CC2CCCCC2C1	1
218	C(CCC1CCC(CCC2CCCCC2)CC1)CC1CCCCC1	1
219	C(C1C2CC(CC3CCCC4CCCCC34)CCC2CC1C1CCCCCC1)C1CCCCC1	1
220	C(CC1CCC2CC(CC2C1)C1CCCCCC1)C1CCCCC1	1
221	C(CCCC1CCC(CC2CCCCC2)CC1)CCC1CCCC1	1
222	C(CCC1CCCCCC1CC1CCCC1)CC1CCCCC1	1
223	C(CC1CCCC1)CC1CCC(CC1)C1CCCC(CC2CC2)C1	1
224	C(C1CCCC1)C1CCC(CC2CCCC(CC2)C2CCCC2)C1	1
225	C(C1CC2CCCCC2C1)C1CCC2CC(CC2C1)C1CCCCCC1	1
226	C(CCC1CCCCC1CCC1CCCC1)CC1CCCCC1	1
227	C(CC1CCC(CCC2CCCCC2)CC1)C1CCC2CCCCC12	1
228	C(CC1CCC(CC1)C(CC1CCCC1)CC1CCCCC1)C1CCCCC1	1
229	C1C(CC2CC(CCC12)C1CCCCC1)C1CCCCC1	1
230	C1CCC(C1)C1CCC(C1)C1CCCC(C1)C1CCCCC1	1
231	C(C1CCCC1)C1CCC2CC(CCC2C1)C1CCCC(C1)C1CCCC1	1
232	C(CCC1CCCCC1)CC1CCCC1CCCC1CCC(CCCC2CCCCC2)CC1	1
233	C(C1CCC2CCC(CC12)C1CCCCC1)C1CCCCC1	1
234	C(CCC1CCCC2CCCC12)CC1CCCCC1	1
235	C(CC1CCC1)CC1CCC(CC1)C1CCC2CCCC2C1	1
236	C(CC1CCCC(C1)C1CCCCC1)C1CCCC1	1
237	C(C1CC2CCCCC2C1)C1CCCC1CC1CC2CCCCC2C1	1
238	C(CCC1CCCC1CCCC1CCCCC1)CC1CCCC1	1
239	C(CCC1CCCC1)CC(CCC1CCCC1)CC1CCCCC1	1
240	C(CC1CCCC1)C1CC1	1
241	C1CCC(C1)C1CCC(C1)C1CCCC(C1)C1CCC2CCCC2C1	1
242	C(C1CCCC1)C1CC2CCCCC2C1	1
243	C(CCC1CCCC(CCCCC2CCCCC2)CC1)CC1CCCCC1	1
244	C(CCCC1CCCC1CCCCC1CCCCC1)CCC1CCCC1	1
245	C(CCC1CCC(CC2CCCCC2)CC1)CC1CCCC1	1
246	C(C1CCCC1)C1CCC(CC2CCC3CCCC3C2)C1	1
247	C(CCC1CCCCC1CC1CCCC1)CC1CCCCC1	1
248	C(CCC1CCC(CC2CCCCC2)CC1CC1CCCCC1)CC1CCCCC1	1
249	C(CC12CC3CC(CC(C3)C1)C2)C1CCC2CC12	1
250	C(CCC1CCC(CC1)C1CCCCC1)CC1CCCC1	1
251	C(CCC(CC1CCCCC1)C1CCCCC1)CCC1CCCC1	1
252	C(CCC1CCC2CCCC2C1C1CC1)CC1CCCCC1	1
253	C(CCC1CCC2C(CC3CCCCC3)CCC2C1)CC1CCCCC1	1
254	C(CCC1CCC2CCC(CC3CCCCC3)C2C1)CC1CCCCC1	1
255	C(CC1CCCC1)C(CC1CCC1)C1CCC(CC1)C1CCCCC1	1
256	C(CCCC(C1CCCCC1)C1CCCCC1)CCCC1CCCC1	1
257	C(CC1CCCC1)C(CCC1CCCC1)C1CCC(CC1)C1CCCCC1	1
258	C(C1CCCC1)C1CCC(C1)C1CCC(CC1)C1CCC2CCCCC12	1
259	C(CCCC1CC1C1CCCC1)CCCC1CCCC1	1
260	C1CCC(C1)C1CCC2CCCCC12	1
261	C(C1CCCC1)C1CCC2CCCCC2C1	1
262	C(CCC1CCCCC1C1CCCCC1)CC1CCCCC1	1
263	C(CC1CCCC1)C(CC1CCCC1)C1CCC(CC1)C1CCCCC1	1
264	C1C2CCCCC2C2CCCC(C12)C1CCCCC1	1
265	C(CCC(CC1CCCCC1)CC1CCCCC1)CCC1CCCC1	1

(*Continued*)

Table S5 (*Continued*)

Number	SMILES	Member size
266	C(C1CCCC1)C1CCC(C1)C1CCCC1	1
267	C(C1CCCC1)C1CCC(C1)C1CCC(CC1)C(C1CCCCC1)C1CCCCC1	1
268	C(C1CCCC1)C1CCC(C1)C1CCCC(CC1)C1CCCCC1	1
269	C(CC1CCCC1)CC1(CCC1)C1CCCCC1	1
270	C(C1CCCC1)C1CCC(C1)C1CCC(CC2CCCCC2)CC1	1
271	C(CCC1CCC2CCCCC2C1)CC1CCCCC1	1
272	C(CCCC1CCCC1)CCCC1CCC(CC1)C1CCCCC1	1
273	C(C1CCCC1)C1CCC(C1)C1CC2CCCCC2C1	1
274	C(C1CCCC1)C1CCC(C1)C1C2CC3CC(C2)CC1C3	1
275	C(C1CCCC1)C1CCC(CC2CCCCCCC2)C1	1
276	C(C1CCCC1)C1CCC(CC2CCCCCC2)C1	1
277	C1CC(C2CC(CCC12)C1CCC(CC1)C1CCCCC1)C1CCCCC1	1
278	C(CC1CCCC1)CC1CCCCCCCCC1	1
279	C1CC1C1CCCC2CCC(CC12)C1CCC(CC1)C1CCCCC1	1
280	C(CCC1CCC2CCCCC2C1CC1CCCCC1)CC1CCCC1	1
281	C1CC1C1CC2CCC(CC2C1)C1CCC(CC1)C1CCCCC1	1
282	C(CC1CCCC1)CC12CCC(CC1)CC2	1
283	C1CCC2C(C1)CCC1CCCCC21	1
284	C(CC1CCCC1)CC1CCC(CC1)C1CCCC(C1)C1CCC(C1)C1CC1	1
285	C(CC1CCCC1)CC12CC3CC1CC(C2)C3	1
286	C(CC1CCCC1)CC1CC2CC1CCC2	1
287	C1CC1C1CCC2CC(CC2C1)C1CCC(CC1)C1CCCCC1	1
288	C(CC1CC1)C(CCC1CCCC1)C1CCC(CC1)C1CCCCC1	1
289	C(CC1CCCCC1)C1CC2CCCC2C1	1
290	C1CCC(C1)C1CC2CCCC2C1	1
291	C(CC1CCCC1)C(C1CCCC1)C1CCC(CC1)C1CCCCC1	1
292	C(CCC1CCCCC1)CC1CCCC1CCCC1CCC1	1
293	C(CCC1CCCCC1)CC1CCCC1CCC1CC1	1
294	C(CC1CCCC1)CC1CC2CC(C2)C1	1
295	C(CC1CCC(CC2CCC3CCCCC23)CC1)C1CCCCC1	1
296	C(CC1CCCC1)CC1CCCCCCCCCCC1	1
297	C(CC1CCC(CC2CCC3CCCC3C2)CC1)C1CCCCC1	1
298	C(CC1CCC(CC2CCC3CCCCC3C2)CC1)C1CCCCC1	1
299	C(CCC1CC2CC(C2)C1)CC1CCCC1	1
300	C(CC1CCCC1)CC1CCC(CCCC23CC4CC(CC(C4)C2)C3)CC1	1
301	C1CC1C1CCCC(C1)C1CCC2C(CCC3CCCCC23)C1	1
302	C(CCC1CCCC1)CCC1CCC(CCCC2CCCC2)CC1	1
303	C(CC1CCCC1)CC1CCC(CCCC2CCC3CCCC3C2)CC1	1
304	C(CCC1CCC(CCCC2CCCC2)CC1)CC1CCCCC1	1
305	C1CCC2C(C1)CCC1CC(CCC21)C1CCCCCC1	1
306	C(CCC12CC3CC(CC(C3)C1)C2)CC1CCCC1	1
307	C(CC1CCCC1)CC1CCC(CC2CC3CCCCC3C2)CC1	1
308	C(CCC1CCCCC1)CC1CCCC1C1CCC(CC2CC2)C1	1
309	C(CC1CCCC1)CC1CCC(CC1)C1CC2CCCCC2C1	1
310	C(CCC1CCCCC1)CC1CCCC1C1CCC(C1)C1CCCC1	1
311	C(CCC1CCC2CCCCC2C1)CC1CCC(CCC2CCCC2)CC1	1
312	C(CCC1CCCC2CCCCC12)CC1CCC(CCC2CCCC2)CC1	1
313	C(CCC1CCCC1)CCC1CCC(CC1)C1CCCCC1	1
314	C(CCC1CCCC1)CCC1CCC2CC(CC2C1)C1CCCCC1	1
315	C(CCC(C1CCCCC1)C1CCCCC1)CCC1CCCC1	1
316	C(CCC1CCCC(CC2CCCC2)CC1)CC1CCCCC1	1
317	C(CC1C2CC(CC12)C1CCC2CCCCC2C1)CC1CCCC1	1
318	C(CC1C2CC(CC12)C1CCCC1)CC1CCCC1	1
319	C(CCC1CCCC(CC2CCC3CCCC3C2)CC1)CC1CCCCC1	1
320	C(CCC1CCCC(CCC2CCCCC2)CC1)CC1CCCCC1	1
321	C(CCC1CCCC(CCCC2CCCCC2)CC1)CC1CCCCC1	1
322	C(CCC1CCCC2CCCCC12)CC1CCCC1	1

(*Continued*)

Table S5 (*Continued*)

Number	SMILES	Member size
323	C(CCC1CCCCC1)CCC1(CCCCC1)C1CCCC1	1
324	C(CCC1CCCCC1)CCC1(CCCCC1)C1CC2CCC(CCCC3CCCCC3)CC2C1	1
325	C(CC1CCC2CC(CC2C1)C1CCC2CCCC12)C1CCCCC1	1
326	C(CC1CCCC1)C(C1 CCCCC1)C1CCCCC1	1
327	C(CCC1CCCCC1)CCC1(CCC(CC2CCCCC2)CC1)C1CCCCC1	1
328	C1CCCCCC1	1
329	C1CCCCC1	1
330	C(CC1CCC2CCCC2C1)C1CCCC1	1
331	C(C1CCCC1)C1CC2CCC1C2	1
332	C(CCC1CCCC1)CCC1CCC2CCCCC12	1

Abbreviation: SMILEs, simplified molecular-input line-entry system.

Table S6 Summary of molecular framework generated from inactive DPP4 inhibitors

Number	SMILES	Member size
1	C1CCCC1	43
2	C1CCCCC1	32
3	C(CCCC1CCCCC1)CCC1CCCCC1	29
4	C(CCC1CCCCC1)CC1CCCC1	24
5	C1C(CC2CCCCC12)C1CCCCC1	20
6	C(CC1CCCC1)CC1CCCCC1	20
7	C1CC2CCCCC2C1	18
8	C1CCC2CCCCC2C1	15
9	C1CCC(CC1)C1CCCC(C1)C1CCCCC1	14
10	C1CCC(CC1)C1CCCCC1	12
11	C(CC1CC2CCCCC2C1)C1CCC(CC2CCCC2)C1	12
12	C(CCC1CCC(CC1)C(C1CCCCC1)C1CCCCC1)CC1CC2CCCCC2C1	9
13	C(CCC1CCCC2CCCCC12)CC1CCCC1	9
14	C(CC1CCCCC1)CC1CCCCC1	8
15	C	8
16	C(CC1CCCC1)CC12CC3CC(CC(C3)C1)C2	8
17	C(CCC1CCCCC1)CC1CCCCC1	8
18	C(C(CC1CCCCC1)C1CCCC1)C1CCCCC1	7
19	C1CCC(CC1)C1CCCC2CCCCC12	7
20	C(C1CCCC1)C1CCCC1	6
21	C(CC1CCCCC1)C1CCCC1	6
22	C(C(CC1CCCCC1)C1CCC2CCCCC12)C1CCCCC1	6
23	C(CCC1CCCCC1)CCC1CCCCC1	6
24	C(CCC1CCCC1)CCC1CCCCC1	6
25	C(C1CCCCC1)C1CCCC(C1)C1CCCCC1	6
26	C(C1CCCC1)C1CCCC(C1)C1CCCCC1	6
27	C(CC1CCCCC1)C1CCCCC1	5
28	C1CCC(CC1)C1CCC2CCCCC2C1	5
29	C(CC1CCCC1)CC1CCCC1	4
30	C(C(CC1CCCCC1)C1CCCC1CC1CCCC1)C1CCCCC1	4
31	C(C1CCCCC1)C1CCCCC1	4
32	C(CCC1CCC2CCCC2C1)CC1CCCCC1	4
33	C(CC1CCC(CC1)C1CCCCC1)C1CCC(CC2CCCC2)C1	3
34	C(CC1CCCCC1)CC1CCCC(C1)C(CCC1CCCCC1)C1CCCCC1	3
35	C(CCC1CCC(CC2CCCCC2)CC1)CC1CC2CCCCC2C1	3
36	C(CCCC1CCCCC1)CCC1CCCC1	3
37	C1CCCCCC1	3
38	C(CC1CCCCC1)C(C1CCCCC1)C1CCCCC1CCC1CCCCC1	3
39	C(CCC1CCC(CCCC2CCCCC2)CC1CC1CCCCC1)CC1CCCCC1	3
40	C(CC1CCCC(C1)C1CCCCC1)C1CC1	3
41	C1CCC1	3
42	C(CC1CCC2CCCC2C1)C1CCC(CC2CCCC2)C1	3

(*Continued*)

Table S6 (*Continued*)

Number	SMILES	Member size
43	C(CC1CCCC1)CC1CCC(C1)C1CCCCC1	3
44	C(CCC1CCCCC1)CC1CCCC1CC1CCCC1	3
45	C(CCC1CCC2CCCC2C1)CC1CCCC1	3
46	C(CC1CCC2CCCCC2C1)C1CCCCC1	3
47	C(CCC1CCCC1)CC1CCCC1	2
48	C(CCC(C1CCCCC1)C1CCCCC1)CCC1CC2CCCCC2C1	2
49	C1C2CC(CCC2C2CCC3C(CCC4CCCCC34)C12)C1CCCCC1	2
50	C1CCC(CC1)C1CCC2C(CCC3CCCCC23)C1	2
51	C(CC1CCCCC1)C1CCC(CC2CC3CCCCC3C2)C1	2
52	C1CC2CCC(CC2C1)C1CCCCC1	2
53	C(CCCC1CCCCC1)CCC1CC2CCCCC2C1C(CC1CCCCC1)CC1CCCCC1	2
54	C(C1CCC(CC2CCC3CCCCC3C2)C1)C1CC2CCCCC2C1	2
55	C(CC1CCCC2CCCCC12)C1CCC(CC2CC3CCCCC3C2)C1	2
56	C(CCC1CCCC1)CCC1CCC(C1)C1CCCCC1	2
57	C(CC1CCCCC1)CC1CCC(CC2CCCCC2)CC1	2
58	C(CC1CCCC1CC1CCCC1)CC1CCCCC1	2
59	C(CCC1CCC2CCCCC2C1)CC1CC2CCCCC2C1	2
60	C(CC1CCCCC1)CC12CC3CC(CC(C3)C1)C2	2
61	C(CCC1CCCCC1)CC(CC1CCCCC1)CC1CCCCC1	2
62	C(CCC1CCC(CC2CCC3CCCCC3C2)CC1)CC1CC2CCCCC2C1	1
63	C(CCC1CCC(CC2CCC(CC2)C2CCCCC2)CC1)CC1CC2CCCCC2C1	1
64	C(CCC1CCC(CC2CCCC(C2)C2CCCCC2)CC1)CC1CC2CCCCC2C1	1
65	C(CCC1CC2CC1CC2C(C1CCCCC1)C1CCCCC1)CC1CC2CCCCC2C1	1
66	C(CCC1CCCC(CC1)C(C1CCCCC1)C1CCCCC1)CC1CC2CCCCC2C1	1
67	C(CCC1CCC(CC(C2CCCCC2)C2CCCCC2)CC1)CC1CC2CCCCC2C1	1
68	C(CC1CCCCC1)C1CC2CCCCC2C1	1
69	C(CCC1CCC(CC2C3CCCCC3C3CCCCC23)CC1)CC1CC2CCCCC2C1	1
70	C(CCC1CCC(CC1)C1CCCCC1)CC1CC2CCCCC2C1	1
71	C(CC1CCC2C1CCC1C2CCC2CCCCC12)CC1CCCCC1	1
72	C(CC1CCC(CC2CCCC2)C1)CC1CCCCC1	1
73	C(CC1CCCC1)CC1CCC(CC2CCCCC2)CC1	1
74	C(CC1CCCCC1)C1CCC(CC2CCCC2)C1	1
75	C(CC1CCCCC1)CC1CCC(CCCC2CCCCC2)CC1	1
76	C(CC1CCC(CC2CCCC2)C1)C1CCCC1	1
77	C1CC2CCC3C(CCC4CCCCC34)C2C1	1
78	C(CC1CCCCC1)C1CC2CCC3C(CCC4CCCCC34)C2C1	1
79	C(CCC1CCCC(C1)C1CCCCC1)CC1CCCCC1	1
80	C(CC1CCCCC1)CC1CCCC(C1)C1CCCCC1	1
81	C(CC1CCCC(C1)C1CCCCC1)C1CCCC1	1
82	C(CCC1CCC(CC1)C1CCCC1)CC1CCCC1	1
83	C(CCC1CCCC(C1)C1CCCCC1)CC1CCCC1	1
84	C(CCC1CCC2CCCCC2C1)CC1CCCC1	1
85	C(CCC1CCCC1)CCC1CCC(CC1)C(C1CCCCC1)C1CCCCC1	1
86	C(CCC1CCCC1)CCC1CCCC1	1
87	C(CCC1CCCC(CC2CCCC2)C1)CC1CCCC1	1
88	C(CCCC1CCCC1)CCCC1CCC(CC1)C(C1CCCCC1)C1CCCCC1	1
89	C1CCC(C1)C1CCCCC1	1
90	C(CCC1CCCC2CCCCC12)CC1CCCCC1	1
91	C(CCCCC1CCCC1)CCCCC1CCCCC1	1
92	C(CCCCC1CCCC1)CCCC(C1CCCCC1)C1CCCCC1	1
93	C(CCCC1CCCC1)CCCC1CCC(CC2CCCCC2)CC1	1
94	C(CC1CCC2CCCCC12)CC1CCC2CCCCC2C1	1
95	C(CCC1CCC(C1)C(C1CCCCC1)C1CCCCC1)CCC1CC2CCCCC2C1	1
96	C(C1CCCC1)C1CCC2CCCCC2C1	1
97	C(CCC1CC2CCCCC2C1)CCC1CCC(CC1)C(C1CCCCC1)C1CCCCC1	1
98	C(CCC1CCCC1CCC(C1CCCCC1)C1CCCCC1)CC1CC2CCCCC2C1	1
99	C(CCC1CC2CCCCC2C1)CCC1CCC(CC2CCCCC2)CC1	1

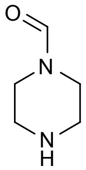
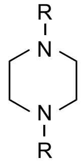
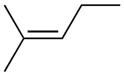
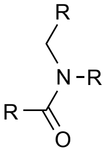
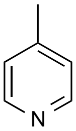
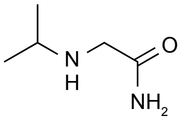
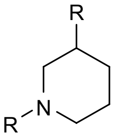
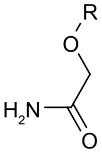
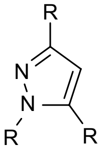
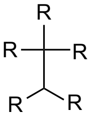
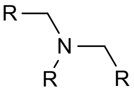
(*Continued*)

Table S6 (*Continued*)

Number	SMILES	Member size
100	C1CC2CCC(CC2C1)C1CCCC(C1)C1CCCCC1	1
101	C(CC1CCC(CC1)C1CCC(CC1)C1CCCC1)C1CCCCC1	1
102	C(CCC1CCCC1)CC(C1CCCCC1)C1CCCCC1	1
103	C(CCC1CCCC1CC1CCC(CC1)C(C1CCCCC1)C1CCCCC1)CC1CC2C1CCCC2C1	1
104	C(CCC1CC2CCCCC2C1)CCC12CC3CC(CC(C3)C1)C2	1
105	C(C1CCCC1)C1CCCC1CC1CC2CCCCC2C1	1
106	C(CC1CCC1)CC12CC3CC(CC(C3)C1)C2	1
107	C(CC1CCCC1)CC1CCCC1C(CC1CCCCC1)CC1CCCCC1	1
108	C(CC1CCCCC1)C(C1CCCCC1)C1CCC(CCC2CCCCC2)CC1	1
109	C(CC1CCCCC1)C(C1CCCCC1)C1CCC2CC(CCC2C1)C1CCCCC1	1
110	C(CC1CCCC1)CC1CC2CCCCC2C1	1
111	C(CCCC1CCCC1)CCCC1CCCCC1	1
112	C(C1CCCCC1)C1CCC2CCC(CC2C1)C1CCCCC1	1
113	C(CCCCCC1CCCCC1)CCCCC1CC2CCCCC2C1	1
114	C(CC1CCCC1C(CC1CCCCC1)CC1CCCCC1)CC1CCCCC1	1
115	C(C(CC1CCCCC1)C1CCCCC1)C1CCCC1	1
116	C(CC1CCC2CCCCC2C1)CC12CC3CC(CC(C3)C1)C2	1
117	C(CCC1CCCC1)CC(C(CC1CCCCC1)CC1CCCCC1)C1CCCCC1	1
118	C(CC1CCCCC1)CC1CCC(CCCC2CCCCC2)C(CC2CCCCC2)C1	1
119	C(CC1CCCC1)CC1CCCC1C1CCCCC1	1
120	C(CCCCC1CCCCC1)CCCC1CCCCC1	1
121	C(CCCC1CCCCC1CCC1CCCCC1)CCC1CCCCC1	1
122	C(CCCC1CCC2CCCCC2C1)CCC1CCCCC1	1
123	C1CC2CCCC2C1	1
124	C(CCCC1CCCCC1)CCCC1CCCCC1	1
125	C(CCCC1CCC(CCC2CCCCC2)CC1)CCC1CCCCC1	1
126	C(CCCC1CCCC2CCCCC12)CCC1CCCCC1	1
127	C1CCC(CC1)C(C1CCCCC1)C1CCCCC1	1
128	C(C1CCCCC1)C1CCC(CC1)C1CCCCC1	1
129	C(CCCCC1CCCCC1)CCCC1CC2CCCCC2C1	1
130	C(CC1CCCC(CC2CC3CCCCC3CC2C2CCCCC2)C1)C1CCCCC1	1
131	C(C1CCCC1)C1CC2CCCC2C1	1
132	C(CC1CCC2CCCCC2C1)C1CCCC1	1
133	C(CC1CC2CCCCC2C1)CC1CCC2CCCCC2C1	1
134	C(CCC1CCCCC1)CC1CC2CCCCC2C1	1
135	C(CC1CC2CCCC2C1)CC1CCCCC1	1
136	C(CC1CCC2CCCCC2C1)C1CCC(CC2CCCC2)C1	1
137	C(CCC1CCCCCC1)CC1CCCCC1	1
138	C(CC1CCCCC1)C1CCC(C1C1CCCCC1)C1CCCCC1	1
139	C1CCC(C1)C1CCC2C(CCC3CCCCC23)C1	1
140	C(CC1CCCC1)CC1CCC(CCCC2CCCCC2)CC1	1
141	C(CC1C2CC(CC12)C1CCCCC1)CC1CCCC1	1
142	C(CC1CCCCC1C1CCCCC1)C1CCCC1CC1CCCC1	1
143	C(C1CCCC1)C1CCCC1C1CC2CCCCC2C2CCCCC2C1	1
144	C(CCCC1CC2CCCCC2C1)CCC(C1CCC2CCCCC2C1)C1CCC2CCCC1C2C1	1
145	C(CCCC1CC2CCCCC2C1)CCC(CC1CCCCC1)CC1CCCCC1	1
146	C(CCCC1CC2CCCCC2C1)CCCC1CCC2CCCCC2C1	1
147	C(C1CCCC1)C1CC2CCCCC2C1	1
148	C(CCCCC(CCCCC1CCCCC1)CC1CCCC1)CCCCC1CCCCC1	1
149	C(CCCCCC1CCCC2CCCCC12)CCCCC1CCCCC1	1
150	C(CCCCCC1CCCCC1)CCCCC1CCCC1	1
151	C(C(CC1CCCCC1)C1CCC(CC2CC3CCCCC3C2)C1)C1CCCCC1	1
152	C(CC1CCCC1)CC1CCC2CCCCC12	1

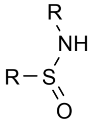
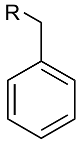
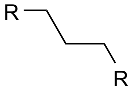
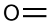
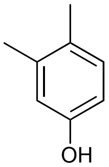
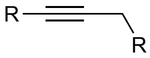
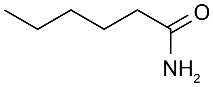
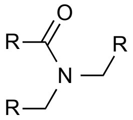
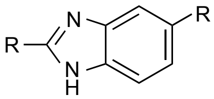
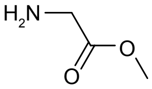
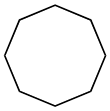
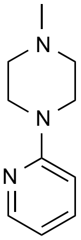
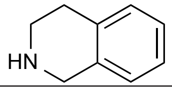
Abbreviation: SMILEs, simplified molecular-input line-entry system.

Table S7 Summary of important structural fingerprints ranked by the MDGI

Rank	Fingerprint	Structure	Fingerprint occurrence		MDGI
			Actives	Inactives	
1	KRFP4541		98	43	9.197
2	KRFP2428		84	56	3.610
3	KRFP3668		15	2	2.312
4	KRFP0610		362	1,685	2.227
5	KRFP3616		16	10	1.813
6	KRFP3405		31	332	1.563
7	KRFP0223		3	134	1.400
8	KRFP2650		5	0	1.119
9	KRFP1945		9	1	1.021
10	KRFP0018		7	29	0.727
11	KRFP0605		284	1,182	0.588

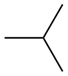
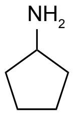
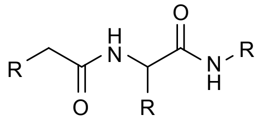
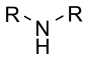
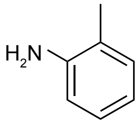
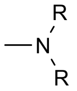
(*Continued*)

Table S7 (*Continued*)

Rank	Fingerprint	Structure	Fingerprint occurrence		MDGI
			Actives	Inactives	
12	KRFP1144		5	80	0.587
13	KRFP0566		55	84	0.581
14	KRFP0344		301	935	0.572
15	KRFP3025		458	1,874	0.511
16	KRFP3561		2	58	0.407
17	KRFP3713		29	247	0.391
18	KRFP0496		0	49	0.382
19	KRFP2200		2	0	0.381
20	KRFP0621		246	776	0.341
21	KRFP3152		1	62	0.320
22	KRFP3966		7	13	0.302
23	KRFP3081		0	70	0.278
24	KRFP3920		5	9	0.214

(*Continued*)

Table S7 (*Continued*)

Rank	Fingerprint	Structure	Fingerprint occurrence		MDGI
			Actives	Inactives	
25	KRFP4261		3	73	0.185
26	KRFP3369		189	789	0.161
27	KRFP0677		236	1,026	0.155
28	KRFP0508		1	1	0.137
29	KRFP2264		85	167	0.131
30	KRFP3602		4	52	0.123

Abbreviation: MDGI, mean decrease of Gini index.

Table S8 Applying Lipinski's rule of five on investigated data sets

Data sets	Total	Actives	Inactives
DPP4-TRN	2,339/2,609 (89.651%)	1,961/2,075 (94.506%)	478/534 (89.513%)
DPP4-TEST1	222/325 (68.308%)		
DPP4-TEST2	215/325 (66.154%)		
DPP4-TEST3	301/325 (92.615%)		

Note: Values shown are for compounds passing the Lipinski's rule of five/in relation to the total number of compounds (values in parentheses are percentages passing the Lipinski's rule of five).