Chemometrics and Intelligent Laboratory Systems 169 (2017) 12–18

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

PyDescriptor: A new PyMOL plugin for calculating thousands of easily understandable molecular descriptors

Vijay H. Masand [a,*], Vesna Rastija [b]

[a] Department of Chemistry, Vidya Bharati Mahavidyalaya, Amravati, Maharashtra 444 602, India
[b] Department of Chemistry, Faculty of Agriculture, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia

ARTICLE INFO

Keywords:
Molecular descriptors
PyMOL
PyDescriptor
QSAR
Interpretation

ABSTRACT

The field of Quantitative Structure-Activity Relationship (QSAR) relies heavily on molecular descriptors. Among various guidelines suggested by Organisation for Economic Co-operation and Development (OECD), a very important guideline demands the mechanistic interpretation of a QSAR model. For this, a very attractive idea is to build a QSAR model using easily understandable molecular descriptors. To address this important issue, in the present work, we present an innovative chem-informatics tool, *PyDescriptor*. It can calculate a diverse pool of 11,145 molecular descriptors comprising easily understandable 1D- to 3D- descriptors encoding pharmacophoric patterns, atomic fragments and a variety of fingerprints. It is a new Python based plugin implemented within the commonly used visualization software PyMOL. *PyDescriptor* has several advantages like easy to install, open source, works on all major platforms (Windows, Linux, MacOS), easy to use through graphical user interface (GUI) and command-line, and output is saved in comma separated values (CSV) file format for further QSAR procedure. The plugin is freely available for academia.

1. Introduction

Computer Aided Drug Designing (CADD) has advanced with innovations in its thriving branches viz. Quantitative Structure-Activity Relationship (QSAR), molecular docking, pharmacophore modelling. The field of QSAR is among the oldest branches of CADD with its emphasis on prediction of activity/property (quantitative QSAR) and determination of pharmacophoric features or mechanistic interpretation (qualitative QSAR) [1–4].

Structure drawing and optimization, molecular descriptor calculations, model building and model validation are four basic steps of a typical QSAR analysis [5–8]. Molecular descriptors, which are used to represent the structural features in terms of numbers, encode valuable information about structure or patterns in the molecular structures [9–16].

Molecular descriptors have occupied unique place in chemistry, pharmaceutical sciences, quality control, etc. to provide valuable representation of molecular features in numerical and computational form for further evaluations [9–18]. With the progress of QSAR field, the types of descriptors have changed from simple and easily interpretative like number of carbon atoms, number of nitrogen atoms, log*P*, etc. to very complex descriptors like WHIM, BCUT, 3D-MoRSE, RDF, GETAWAY, and others [17,18]. These molecular descriptors are mostly classified as 1D-, 2D- and 3D- descriptors. The 1D- molecular descriptors represent bulk properties of compounds, such as the number of particular atoms, molecular weight, etc., and can be computed using molecular formula. 2D-molecular descriptors characterize structural information that can be calculated from 2D- structure of a molecule, such as the number of rings, the number of hydrogen bond acceptors, etc. 3D- molecular descriptors stand for structural information that has to be obtained from 3D- structure of a molecule, such as solvent accessible surface area with negative partial charge in the structure [17,18].

Manual calculation of descriptors like 3D-MoRSE, WHIM, BCUT, and similar complex (or esoteric [5]) descriptors was a very time consuming and laborious process [1,9–12,15,16]. To overcome this difficulty, computer programs were developed for computing descriptors either as independent software or as a part of QSAR software. The rapid developments in the field of computers and algorithms have made exact and

Abbreviations: OECD, Organisation for Economic Co-operation and Development; WHO, World health organisation; ADMET, Absorption, Distribution, Metabolism, Excretion and Toxicity; OLS, Ordinary Least Square; QSARINS-Chem, QSAR Insubria-Chemistry; GA, Genetic algorithm; MLR, Multiple linear Regression; OFS, Objective Feature Selection; SFS, Subjective Feature Selection; MMFF94, Molecular Mechanics Force Field 94; MAE, mean Absolute Error; CCC, Concordance Correlation Coefficient.

[*] Corresponding author.
E-mail address: vijaymasand@gmail.com (V.H. Masand).

http://dx.doi.org/10.1016/j.chemolab.2017.08.003
Received 11 May 2017; Received in revised form 15 July 2017; Accepted 10 August 2017
Available online 13 August 2017

precise calculations of theoretical molecular descriptors possible in shorter time and cost-effective [1,9–12,15,16]. At present, there are many free and commercial softwares like Dragon (Talete) [17,18], PaDEL [19], MOE [20], Schrodinger [21], ChemDes [22], etc. which can calculate a variety of molecular descriptors viz. 1D- to 3D-, constitutional, topological, fingerprints. Some of these have been developed exclusively for the calculation of molecular descriptors only such as PaDEL-Descriptor [19], ChemDes [22], etc. while others are QSAR softwares which have descriptor calculation as one of their features (e.g., CODESSA Pro [23], Accelrys Discovery Studio [24], Sybyl-x [25], MOE [20]). Also, there are some open source libraries, such as JOELib [26,27], Chemistry Development Kit [28], and Chemical Descriptors Library [29], to name a few, which have molecular descriptor calculation functionality. It is reasonable that a good descriptor calculation software should have following features [19]:

1. Free or low-priced so that it is easy to purchase it.
2. Open source so that researchers could introduce their specific molecular descriptor calculations.
3. Has an easy to use graphical user interface (GUI).
4. Independent of operating system.
5. Possibly processes different molecular file formats like mol2, mol, sdf, etc.
6. Ability to compute numerous types of molecular descriptors.

A careful analysis of various currently available molecular descriptor calculating softwares reveals that many softwares lack one or more above mentioned features, besides, having its own advantages and limitations. An important area of research in the field of molecular descriptors is introduction of new descriptors or improvements in existing descriptors with easy correlation in terms of structural and pharmacophoric patterns [1,10–16,22,28,29]. Therefore, the field of molecular descriptors is dynamic and open for future developments like introduction of new softwares with ease of use and better user control functionalities, new descriptors with enhanced abilities to capture structural features [1,10–16,22,28,29].

Among various guidelines suggested by Organisation for Economic Co-operation and Development (OECD), a very important guideline demands the mechanistic interpretation of a QSAR model. For this, a very attractive idea is to build a QSAR model using easily understandable molecular descriptors. Unfortunately, the physical correlation of esoteric descriptors like WHIM, GETAWAY, RDF, etc. with one or more structural features/patterns is very complicated and an active area of qualitative and quantitative QSAR [5]. Therefore, there is need for introduction of easily understandable molecular descriptors. In the present work, we present a new PyMOL plugin, *PyDescriptor*, which has capacity to calculate 11,145 easily understandable molecular descriptors. It is a new chem-informatics tool which transforms a variety of structural features and local environment of a molecule to understandable 1D- to 3D- descriptors, which include encoding pharmacophoric patterns, atom-centred descriptors and a variety of fingerprints. These descriptors are either available in costly commercial softwares or in operating system dependent free softwares, thereby restricting their wide use. *PyDescriptor* possesses many advantageous features and plethora molecular descriptors, which justify its usefulness and wide acceptance in the field of QSAR and allied areas.

2. Experimental details

2.1. Plugin design and availability

PyDescriptor has been written in the object-orientated programming language Python 2.7.10 (64 bit) as a plugin for the three-dimensional molecular viewer PyMOL 1.8.2 and higher versions (Schrödinger, LLC. http://www.pymol.org/). Therefore, the advantages and limitations of Python 2.7.10 and PyMOL are associated with this plugin also. PyMOL is a widely-used software proficient in rendering and ray-tracing high resolution molecular representations in publication quality [30]. Due to availability of an open-source version of PyMOL, it is an attractive choice for academic and educational use [30]. Apart from visualizations of molecular structures, PyMOL has emerged as a calculation software due to availability of different open source plugins for a variety of purposes for example APBS for electrostatic map calculation, CAVER for calculation and visualization of tunnels, MIPTOOL for LogP calculation, DYNAMICS for molecular dynamic simulations with Gromacs, a few to mention [30]. In addition, LIQUID is an open source plugin for PyMOL, which is capable of generating pharmacophore model for a molecule. The output of LIQUID is available in the form of spheres and ellipsoids in the 3D- viewer of PyMOL [31]. Though, *PyDescriptor* uses the framework of PyMOL, it has been fully coded by our group. Practical information, such as a user guide/manual and application notes, along with the plugin '*PyDescriptor*', are available free of charge from authors.

2.2. System requirements and installation

In order to use *PyDescriptor*, a working installation of PyMOL version ≥1.8.2 on a standard Linux or Windows or MacOS installations with Python 2.7.10 is essential. *PyDescriptor* can be used without any dependencies i.e. there is no need to install any other module or software. At present, the plugin has been built to use MOL2 file format containing single molecule only. MOL2 format has the benefit of storing all the essential information for atom type, position, partial charges, and connectivity. In addition, it is also a well-known standardized format that many programs can read. It is one of the few public formats capable of supporting both a chemically-accurate description of small organic molecules as well as protein or nucleic acid also. Other formats for representing molecular structure have to be converted to an MOL2 file format for use in *PyDescriptor*. For this purpose, users can use open-source programs (e.g. Open babel, Avogadro) to convert other file formats into MOL2 format. While using MOL2 file format, all atom-typing and atomic partial charges assignments need to be performed correctly with all hydrogen atoms added. After successful completion of the descriptor calculations, the molecular descriptor values are automatically saved in CSV file format.

2.3. Parsing and calculations

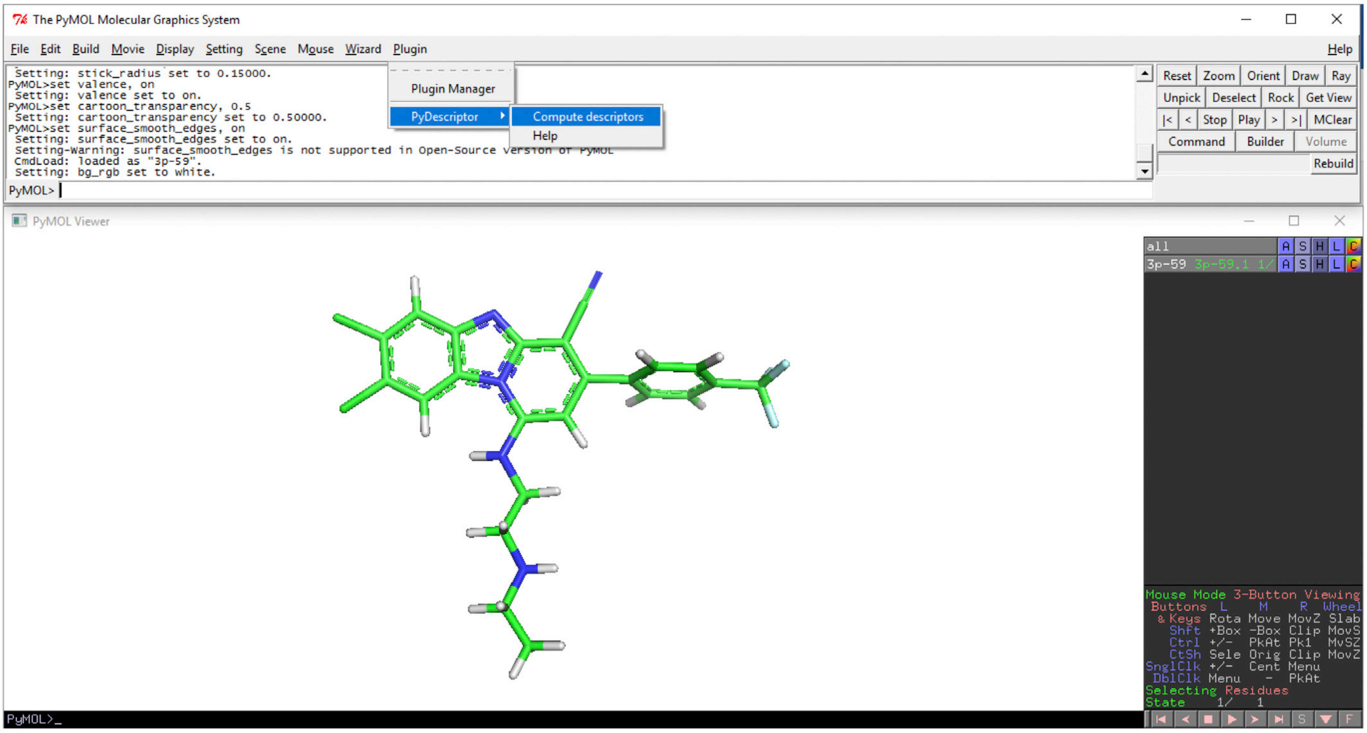
PyDescriptor performs the main task of reading the MOL2 files and calculating the molecular descriptor value for all the MOL2 files located in the folder (for windows users, C:\PyDescriptor). As shown in Fig. 1, when the user clicks 'Compute descriptors', the plugin executes the calculation of molecular descriptors. The values for all the molecular descriptors are entered and automatically saved iteratively into the CSV columns along with the name of MOL2 file in the first column.

The following set of codes is used to read MOL2 files:

```
import os, glob, csv, pymol
os.chdir('C:\PyDescriptor')
from pymol import cmd, stored, util
path = os.path.dirname(pymol.__script__)
cmd.delete('all')
mol_files = glob.glob(os.path.join(path, '*.mol2'))
```

***PyDescriptor* Protocol:**

- Read all the MOL2 files from a particular folder (for windows users, C:\PyDescriptor)
- Calculate the molecular descriptors for all the molecules in the given folder
- Read the name of the files and enter in the CSV file together with their corresponding descriptor values



```
 Setting: stick_radius set to 0.15000.
PyMOL>set valence, on
 Setting: valence set to on.
PyMOL>set cartoon_transparency, 0.5
 Setting: cartoon_transparency set to 0.50000.
PyMOL>set surface_smooth_edges, on
 Setting: surface_smooth_edges set to on.
 Setting-Warning: surface_smooth_edges is not supported in Open-Source version of PyMOL
 CmdLoad: loaded as "3p-59".
 Setting: bg_rgb set to white.
PyMOL>
```

Fig. 1. GUI for *PyDescriptor* plugin loaded with a 3D molecule.

2.3.1. *Descriptors calculation speed experiments*

A straightforward comparison of the descriptor calculation speed of different softwares in a strict way is hard as the number and types of descriptors calculated by each software are different [19]. Therefore, all experiments for calculating the speed of descriptor calculations were accomplished only for *PyDescriptor* using Windows 7 (64 bit) on two different computers with varying architectures. Each experiment was repeated five times and the average of the total time needed to complete the calculation has been reported. Python's “timeit” module was employed for the measurement of calculation time. The details of computers and data set are as following:

1. Computer-1: Windows 7 (64 bit) operating system installed on a Lenovo G560 system with Intel® Pentium® P6100 2.00 GHz processor with 3 GB RAM.

2. Computer-2: Windows 7 (64 bit) operating system installed on a Dell system with Intel i7 2.00 GHz processor with 8 GB RAM.

Data set: The data set contains a diverse set of 1290 molecules as reported by Xu et al. [32]. The data set was converted from the SMILES flat file representation to individual MOL2 file using OpenBabel 2.4.0 using MMFF94 for structure optimization.

2.3.2. *Derivation of interpretable QSAR*

As the present plugin calculates 11,145 molecular descriptors, a very logical question can arise about the possibility of using these descriptors for developing scientifically interesting new, better and interpretable models as well as about the diversity of pool of descriptors calculated by *PyDescriptor*.

In majority of situations, a small dataset is available to a QSAR researcher for building the models. Hence, to address these issues, new QSAR models were built and statistically compared for two datasets of small size using the molecular descriptors calculated by *PyDescriptor*.

Dataset 1. It comprises a small dataset of sixty phosphoramidate and phosphorothioamidate analogues of amiprophos methyl reported as antimalarial [33].

Dataset 2. This dataset encompasses ninety-seven substituted phenyl 4-(2-oxoimidazolidin-1-yl) benzenesulfonates exhibiting antiproliferative activity [4].

2.3.3. *Procedure for QSAR model development*

For QSAR model development, OECD guidelines were followed to ensure internal and external predictive ability with mechanistic interpretation. The procedure mentioned in development of QSAR model for dataset 1 and 2 has been followed to assure reproducibility of results and fair comparison [4,33]. That is, the training and prediction sets were kept identical with the training and prediction sets as in the respective original publication [4,33]. In addition, statistically robust multiple QSAR models were also developed by changing the composition of training and prediction sets. These multiple models are available in supporting information. In general, the structures were drawn and optimized using MMFF94, followed by calculation of molecular descriptors using PaDEL, e-Dragon and *PyDescriptor*. The next step comprises elimination of highly correlated ($|R| > 0.90$) and constant variables (>95%). Subjective feature selection was used to build the statistically robust OLS QSAR models using genetic algorithm (GA) in QSARINS-Chem 2.2.1 [34,35] using Q^2_{LOO} as the fitness function. The exhaustive search of optimum number and set of descriptors was performed till there was improvement in the value of Q^2_{LOO}. Th e GA module of QSARINS-Chem 2.2.1 does not require a prior knowledge of important descriptors, that is, an important descriptor may or may not be in the final QSAR model [34,35]. Exhaustive internal as well as external validation along with Y-scrambling and analysis of Applicability Domain (AD) by Williams plot [34,35] for all the developed models were performed using QSARINS-Chem 2.2.1 to reject over-fitting and spurious models. Various parameters for internal and external validation includes: determination coefficient R^2, leave-one-out (LOO) cross-validation Q^2, leave-many-out (LMO) Q^2_{LMO}, coefficient of determination for Y-scrambling R^2_{Yscr}, root mean squared error (RMSE), $RMSE_{ex}$, MAE_{ex}, R^2_{ex}, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, and CCC_{ex}. The mean value of Q^2_{LMO} has been reported.

3. Results and discussion

Molecular descriptors occupy a unique place in QSAR. The success of

QSAR models not only lies on accurate set and number of descriptors with proper validation but on correct correlation and interpretation of molecular descriptors in terms of structural features also [5]. Many a time, the QSAR models are derived using a set of esoteric descriptors only [5,7,33,36,37]. This substantially limits the use of a properly validated QSAR model by synthetic chemists, to whom the descriptor calculating software is not available, or he/she is unable to correlate structural feature with a specific descriptor, or has little knowledge of QSAR field [5,7,33,36,37]. Therefore, the molecular descriptors involved in an appropriately validated QSAR model must be understandable in terms of structural features and the descriptor calculating software must be available either free or at very low cost. To address this crucial issue, we have developed *PyDescriptor*. The sole purpose of *PyDescriptor* is to facilitate calculation of easily understandable molecular descriptors.

This PyMOL plugin possesses following merits: easy to operate, reproducible results, calculates thousands of molecular descriptors (11,145 descriptors), calculates unique molecular descriptors which are either available in commercial or operating system dependent free softwares, the results are directly saved in a CSV file, and free for academia. In addition, molecular descriptors are easily and rapidly calculated with no missing values, a common difficulty with many existing commercial systems.

3.1. PyDescriptor descriptors

PyDescriptor computes 11,145 easily understandable molecular descriptors using conventions and idioms used in PyMOL. The molecular descriptors that are calculated using this plugin possess a value that is independent of the particular characteristics of the molecular representation, such as atom numbering or labelling, spatial reference frame, translational invariance and rotational invariance, etc. The descriptors possess following additional advantages: easy interpretation in terms of structural moieties, applicable for representing local environment or structure, simple to understand, independent of experimental properties, efficient construction possible, use of familiar structural concepts, conformation dependent, and change according to continuing modification in structures. A majority of descriptors calculated by the present plugin are information-based descriptors i.e. encode the information stored in molecular structures. It can calculate 1D- descriptors like molecular weight, number of atoms, etc., 2D- descriptors like charge descriptors, H-bond donor acceptors, 2D- fingerprint, etc. and 3D-descriptors like charged partial surface area, three-dimensional autocorrelation (3DA) descriptors, etc. A majority of 2D- and 3D- descriptors calculated by *PyDescriptor* represent the relative position of atoms or atom properties by calculating the separation between atom pairs in terms of number of bonds (2DA) or Euclidean distance (3DA) [38].

A very important feature of *PyDescriptor* is its ability to calculate a good number of circular fingerprints (CFP) [39], extended connectivity fingerprints (ECFP) [40], and their variants. These fingerprints are extensively used in high-throughput screening (HTS), similarity searching, including chemical clustering and compound library analysis, etc. [39] [40] These fingerprints can capture rich local structural information available in a molecule. For example, *O_N_5A* is a circular fingerprint descriptor calculated by *PyDescriptor*. *O_N_5A*, which stands for the presence of N atoms within a spatial distance of 5 Å from O atom, looks for the N atom(s) within the radius of 5 Å whose center is O atom. *PyDescriptor* not only counts ECFP/FCFP/CFP but it can calculate several ECFP/PFP/Circular fingerprints inspired 'specific' descriptors containing additional features such as partial charges, frequency of connected or non-connected atoms or functional groups, etc. For example, *O_N_5Ac* stands for sum of partial charges on N atoms which are within 5 Å from O atom. Another example is *O_N_7Bc* which corresponds to sum of partial charges on N atoms which are within seven bonds from O atom.

As *PyDescriptor* is a software plugin dedicated for molecular descriptor calculations only, henceforth its comparisons shall only be made with other similar dedicated software instead of comparing it with general QSAR software. For comparison purpose, molecular weight, average molecular weight and number of atoms for simple organic molecules calculated by PaDEL, e-Dragon and *PyDescriptor* have been tabulated in Table 1. From Table 1, it is clear that the values of molecular descriptors calculated by *PyDescriptor* are in good agreement with the values for same descriptors calculated by PaDEL and e-Dragon.

PyDescriptor has numerous benefits that are generally associated with existing open and free dedicated molecular descriptor calculation software. Being free will broaden the easy availability of the software to a good number of users and open source will permit users to easily check the code and amend it to suit their requirements. This could possibly help in the recognition of errors/bugs and increase the number of molecular descriptor calculation abilities. Since *PyDescriptor* is a plugin built within the framework of PyMOL, the users of *PyDescriptor* must also agree with the respective licenses of PyMOL and Python. Another important advantage of *PyDescriptor* is that it can work on any platform on which PyMOL 1.8.2 and Python 2.7 have been installed. This allows it to run on the three major platforms, Windows, Linux, and MacOS.

In addition, *PyDescriptor* can be used not only through GUI but using command line (via PyMOL) also. Having both GUI and command line options for running *PyDescriptor* is important, as the GUI will cater the need of a large number of users while the command line is useful for those who need to run *PyDescriptor* in computer clusters for big databases.

At present, a major caveat of *PyDescriptor* is its inefficiency to calculate graph-based topological descriptors; work is in progress to overcome this limitation. However, to our knowledge, no simple, freely available Python and PyMOL tool is available that can easily perform molecular descriptor calculation using PyMOL (see Table 2).

3.2. Descriptor calculating speed

For a data set of 1290 molecules, computer-1 and computer-2 took 19406.00 (15.04 s per molecule) and 8845.56 (6.86 s per molecule) seconds, respectively. Thus, it is clear that *PyDescriptor* works well on a computer with high computational abilities. We clarify that *PyDescriptor* has not been optimized for speed. As *PyDescriptor* is open source, users can modify it for speed and their specific use.

3.3. Developing new QSAR models

According to OECD guideline, “mechanistic interpretations of (Q) SARs begin with the number and the nature of the molecular descriptors used in the model”. According to Johnson [41,42], a QSAR modeler must always keep in mind that mechanistically interpretable models are more likely to define causative relationships and are less liable to be the result of chance correlations. Therefore, understanding of the meaning of descriptors is very important during QSAR interpretation step. The mechanistic interpretation of a QSAR model helps to develop “action plan” by a decision maker, for example a medicinal chemist [43]. Since, many easily understandable descriptors calculated by *PyDescriptors* are able to provide useful information about local environment in the molecule and capture specific pharmacophoric patterns, deriving new QSAR models using descriptors calculated by *PyDescriptors* will be beneficial in mechanistic interpretation of QSAR model and in decision making.

1. QSAR modelling for anti-malarial activity of phosphoramidate and phosphorothioamidate analogues of amiprophos methyl [33]

Recently, our group published [33] multiple properly validated QSAR models for anti-malarial activity of phosphoramidate and phosphorothioamidate analogues of amiprophos methyl using understandable molecular descriptors for the best model (termed as Old Model 1 in the present work).

Table 1
Comparison of different molecular descriptors calculated by PaDEL, e-Dragon and *PyDescriptor*.

S.N.	Molecule	Molecular Weight			Number of rings			Number of Atoms		
		PaDEL	e-Dragon	*PyDescriptor*	PaDEL	e-Dragon	*PyDescriptor*	PaDEL	e-Dragon	*PyDescriptor*
1	Acetylene	26.01565	26.04	26.03728	0	0	0	4	4	4
2	Aniline	93.05785	93.14	93.12648	1	1	1	14	14	14
3	Benzene	78.04695	78.12	78.11184	1	1	1	12	12	12
4	Benzimidazole	117.04530	118.15	118.13590	2	2	2	14	15	15
5	Cyclohexane	84.09390	84.18	84.15948	1	1	1	18	18	18
6	Cyclopropane	42.04695	42.09	42.07974	1	1	1	9	9	9
7	Ethane	30.04695	30.08	30.06904	0	0	0	8	8	8
8	Ethene	28.03130	28.06	28.05316	0	0	0	6	6	6
9	Naphthalene	128.06260	128.18	128.1705	2	2	2	18	18	18
10	Phenol	94.04186	94.12	94.11124	1	1	1	13	13	13
11	Propane	44.06260	44.11	44.09562	0	0	0	11	11	11
12	Pyridine	80.05002	79.11	79.09990	1	1	1	12	11	11

Table 2
A representative list of different types of molecular descriptors calculated by *PyDescriptor* (see supporting information for complete list).

S.N.	Type of Molecular Descriptor	Some examples of Molecular Descriptor	Total Number
1	Constitutional • Functional groups • Molecular weight • Simple Atom counts • Ratio of various types of atoms	Molecular weight, Average Molecular weight, -OH, 3° Amine, number of atoms, total number of bonds, total number of rings, etc.	235
2	Geometric • Molecular surface area (MSA) • Solvent accessible molecular surface area (SASA) • Ratio of MSA and SASA of various types of atoms	Molecular Surface area and Solvent accessible molecular surface area of molecule, positively/negatively/neutral atoms, etc. Absolute Surface Area, MSA of C atoms, MSA of N atoms, SASA of C atoms, SASA of F atoms, etc.	212
3	Circular fingerprint • Presence/Absence of different types of atom pairs at specific spatial distance	Number of C atoms within 5 Å from ring atoms, etc.	2650
4	Quantum chemical • Charges	Sum of partial charges of C atoms within 4 bonds from O atoms, Sum of partial charges of C atoms within 4 Å from O atoms, etc.	3548
5	Topological • Atom-pairs	Number of C atoms within 9 bonds from O atoms, Number of N atoms within 5 bonds from Cl atoms, etc.	4500

Old Model-1: $pIC_{50} = 2.3367\ (\pm0.7641) + 1.5695\ (\pm1.7697)\ \mathit{R6p} - 0.0306\ (\pm0.0254)\ \mathit{nBT} + 0.4084\ (\pm0.1941)\ \mathit{nN} + 0.6338\ (\pm0.1605)\ \mathrm{ALogP}$
$R^2_{tr} = 0.79$, $Q^2 = 0.72$, $MAE_{cv} = 0.27$, $R^2_{ex} = 0.81$ and $CCC_{ex} = 0.89$

In the present work, the same dataset (Keto form) was used for developing new QSAR model using the molecular descriptors calculated by *PyDescriptor*, e-dragon and PaDEL with identical training and prediction sets as mentioned in our previous publication. The newly derived best four parametric QSAR model built for anti-malarial activity of phosphoramidate and phosphorothioamidate analogues of amiprophos methyl is as following:

New Model-1: $pIC_{50} = +2.682\ (\pm0.411) - 0.104\ (\pm0.071) * \mathit{N\_O\_3A} - 13.310\ (\pm9.785) * \mathit{all\_O\_8Ac} + 0.422\ (\pm0.168) * \mathit{plus\_N\_2A} + 0.434\ (\pm0.091) * \mathit{ALOGP}$
$R^2_{tr} = 0.83$, $RMSE_{tr} = 0.25$, $MAE_{tr} = 0.19$, $CCC_{tr} = 0.91$, $Q^2_{loo} = 0.79$, $RMSE_{cv} = 0.28$, $MAE_{cv} = 0.21$, $CCC_{cv} = 0.89$, $RMSE_{ex} = 0.32$, $MAE_{ex} = 0.28$, $Q^2_{F1} = 0.75$, $Q^2_{F2} = 0.73$, $Q^2_{F3} = 0.72$, $CCC_{ex} = 0.89$, $R^2_{ex} = 0.81$

The symbols used for various statistical parameters have their usual meaning and available in supporting information [34,35]. The descriptor ALOGP represents lipophilicity of the molecule. The descriptors *N_O_3A* stands for the presence of oxygen atom within a spatial distance of 3 Å from nitrogen atom. The descriptor *all_O_8Ac* corresponds to sum of partial charges of all atoms within 8 Å from oxygen atom. The descriptor *plus_N_2A* corresponds to the number of nitrogen atom present within 2 Å from positively charged atoms. The descriptors *N_O_3A*, *all_O_8Ac* and *plus_N_2A* have been calculated by *PyDescriptor* and represent local environment inside the molecule, while *ALOGP* is a property of whole molecule.

A simple comparison of statistical parameters of model-1 and old model-1 reveals that the new model has improved performance not only with respect to fitting but for cross-validation parameters like Q^2, MAE_{cv}, etc. also.

In addition, another adequately validated QSAR model was built using molecular descriptors calculated by *PyDescriptor* only (neither PaDEL nor e-Dragon descriptors were used) with identical training and prediction sets as stated in our previous publication. The newly derived best four parametric QSAR model is as following:

New Model-2: $pIC_{50} = +3.333\ (\pm0.356) + 5.159\ (\pm1.351) * \mathit{H\_S\_4Ac} - 0.123\ (\pm0.046) * \mathit{byring\ all\_S\_4A} + 0.096\ (\pm0.062) * \mathit{fHS6B} + 0.099\ (\pm0.038) * \mathit{C\_don\_6A}$
$R^2_{tr} = 0.82$, $RMSE_{tr} = 0.26$, $MAE_{tr} = 0.22$, $CCC_{tr} = 0.90$, $Q^2_{loo} = 0.75$, $RMSE_{cv} = 0.30$, $MAE_{cv} = 0.26$, $CCC_{cv} = 0.86$, $RMSE_{ex} = 0.33$, $MAE_{ex} = 0.27$, $Q^2_{F1} = 0.74$, $Q^2_{F2} = 0.72$, $Q^2_{F3} = 0.71$, $CCC_{ex} = 0.88$, $R^2_{ex} = 0.81$

The symbols used for various statistical parameters have their usual meaning and available in supporting information also [34,35]. The descriptor *H_S_4Ac* indicates sum of partial charges of sulphur atoms which are at a spatial distance of 4 Å from hydrogen atom. The descriptor *byring all_S_4A* stands for the presence of sulphur atoms which are at a spatial distance of 4 Å from ring atoms. The descriptor *fHS6B* corresponds to frequency of occurrence of hydrogen and sulphur atoms separated by six bonds. The descriptor *C_don_6A* resembles the number of presence of donor atom or group at a distance of 6 Å from carbon atom.

A comparison of statistical measures of model-2 with old model-1 indicates that the model-2 has outperformed the previously reported model. From model-2, it is clear that the model, derived using molecular descriptors calculated by *PyDescriptor* only, has better statistical performance and high degree of correlation of molecular descriptors with structure feature than the old model-1. This indicates that the molecular descriptors calculated by *PyDescriptor* could result in useful augmentation of statistical performance of the model and increase in mechanistic interpretation as well. It also indicates that scientifically interesting new and improved models could be built using descriptors calculated by *PyDescriptor*. In addition, the diversity of pool of descriptors calculated by *PyDescriptor* is also reflected.

A comparison of statistical parameters of model-1 and 2, derived in the present work, reveals that model-1 has better statistical performance than model-2. This indicates that a combination of molecular descriptors calculated by *PyDescriptor* with different types of descriptors generates a thriving QSAR model with easy interpretation and statistical robustness. Therefore, it is logical to use molecular descriptors calculated by *PyDescriptor* with different types of descriptors calculated by other softwares.

2. QSAR modelling for anti-proliferative activity of substituted phenyl 4-(2-oxoimidazolidin-1-yl) benzenesulfonates [4]

An appropriately validated QSAR model for undivided dataset of ninety-seven substituted phenyl 4-(2-oxoimidazolidin-1-yl) benzenesulfonates for anti-proliferative activity using three molecular descriptors was published by our group [24].

Old model-1b: $\log IC_{50} = -8.5590\ (\pm 4.0430) + 55.8097\ (\pm 23.1356) * Xt - 71.0572\ (\pm 18.4287) * VEA2 + 0.7420\ (\pm 0.2572) * nHDon.$
$R^2_{tr} = 0.87$, $Q^2 = 0.85$, $RMSE_{tr} = 0.50$, $RMSE_{cv} = 0.52$, $F = 205.12$, $CCC_{tr} = 0.93$, $CCC_{cv} = 0.92$

In the present work, the same dataset was used for developing new QSAR model using the molecular descriptors calculated by *PyDescriptor*, e-dragon and PaDEL. The newly constructed best three parametric QSAR model built for anti-proliferative activity of substituted phenyl 4-(2-oxoimidazolidin-1-yl) benzenesulfonates is as following:

New Model-1b: $\log IC_{50} = -29.646\ (\pm 3.549) + 1.043\ (\pm 0.135) * lipo\_N\_2B + 0.649\ (\pm 0.206) * all\_N\_6Ac + 115.995\ (\pm 19.541) * X3A$
$R^2_{tr} = 0.93$, $Q^2 = 0.92$, $RMSE_{tr} = 0.36$, $RMSE_{cv} = 0.38$, $F = 409.17$, $CCC_{tr} = 0.96$ and $CCC_{cv} = 0.96$

The descriptor *X3A* accounts for the multiplicity of the bond and for the presence of hetero atoms in the molecule, especially the hydrogen bond donor/acceptor atoms. The descriptor *lipo_N_2B* stands for number of lipophilic atoms separated by two bonds from nitrogen atoms. The third descriptor *all_N_6Ac* corresponds to sum of partial charges of all atoms present within a spatial distance of 6 Å from nitrogen atom. The descriptors *lipo_N_2B* and *all_N_6Ac* have been calculated by *PyDescriptor* and represent local environment of the molecule, whereas *X3A* is a property of whole molecule.

It is evident from the statistical parameters of **old model-1b** and **new model-1b** that the new model has superior statistical robustness not only with respect to fitting but also for cross-validation parameters like R^2, $RMSE_{cv}$, etc. Additionally, a different statistically validated QSAR model was built using molecular descriptors calculated by *PyDescriptor* only (neither PaDEL nor e-Dragon descriptors were used) with identical training set as specified in our previous publication. The newly derived best three parametric QSAR model is as following:

Model-2: $\log IC_{50} = -3.471\ (\pm 0.852) - 1.223\ (\pm 0.265) * N\_lipo\_3Bc - 5.082\ (\pm 2.065) * S\_all\_8Bc - 0.250\ (\pm 0.049) * S\_byring\ all\_9B$
$R^2_{tr} = 0.90$, $Q^2 = 0.89$, $RMSE_{tr} = 0.43$, $RMSE_{cv} = 0.45$, $F = 285.36$, $CCC_{tr} = 0.95$ and $CCC_{cv} = 0.94$

The descriptor *N_lipo_3Bc* corresponds to sum of partial charges of all lipophilic atoms which are separated from nitrogen atoms by three bonds. The descriptor *S_all_8Bc* represents sum of partial charges of all atoms separated from sulphur atom by eight bonds. The third descriptor *S_byring all_9B* stands for number of ring atoms which are separated from sulphur atom by nine bonds.

It is clear from the statistical measures of **model-2** that it has better statistical performance when compared with **old model 1b**. Thus, this again confirms that the molecular descriptors calculated by *PyDescriptor* are advantageous for increasing the statistical robustness of the model and mechanistic interpretation of the model in terms of structural features.

3.4. General comparison of newly developed models with old models

A comparison of newly developed models with the old models points out that the new QSAR models have better statistical performance and greater number of easily understandable molecular descriptors. The molecular descriptors used in the present models not only represent the local environment of the molecule but complete molecule also. This would have been difficult without the incorporation of new descriptors calculated by *PyDescriptor*. It appears that the use of esoteric descriptors along with the descriptors calculated by *PyDescriptor* significantly augment the statistical performance and mechanistic interpretation of the QSAR model. Therefore, a combination of e-Dragon, PaDEL and descriptors from *PyDescriptor* is useful for deriving quantitative and qualitative QSAR models with high statistical performance and mechanistic interpretation. A statistically best QSAR equation may have only complex descriptors which cannot be easily interpretable at the level of substructures of the molecules. In our opinion, a QSAR model should be selected which should be statistically sound and easier to relate to the structural features of the molecules under study.

4. Conclusions

The use of esoteric descriptors along with easily understandable descriptors provided models with better accuracy, fidelity and easy physical clarification in a biological perspective which, in turn, could yield perceptions of a causal mechanism of action, ways of decreasing a drug's toxicity or increasing its efficacy. In the present work, a PyMOL plugin *PyDescriptor* molecular descriptor calculator has been reported which possesses a good number of advantages. The plugin can be used on all the popular platforms (Windows, Linux, MacOS). The 11,145 *PyDescriptor* descriptors calculated here consists of 1D- to 3D- molecular descriptor. For QSAR community, it provides a zero-cost option for calculating a good number of easily understandable and informative molecular descriptors with broad applicability to various types of problems. To summarize, *PyDescriptor* is a useful addition to the currently existing molecular descriptor calculation software.

Acknowledgements

Authors are thankful to Josip Juraj Strossmayer University of Osijek, Osijek, Croatia for financial support to the project INGI-2015-20 to complete the present work. The authors are also thankful to anonymous reviewers of this manuscript for constructive observations and suggestions.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.chemolab.2017.08.003.

References

[1] M.H. Baig, K. Ahmad, S. Roy, J.M. Ashraf, M. Adil, M.H. Siddiqui, S. Khan, M.A. Kamal, I. Provaznik, I. Choi, Computer aided drug design: success and limitations, Curr. Pharm. Des. 22 (2016) 572–581.

[2] S.S. Imam, S.J. Gilani, Computer aided drug design: a novel loom to drug discovery, Org. Med. Chem. 1 (2017) 1–6.

[3] S.J.Y. Macalino, V. Gosu, S. Hong, S. Choi, Role of computer-aided drug design in modern drug discovery, Arch. Pharm. Res. 38 (2015) 1686–1701.

[4] V.H. Masand, D.T. Mahajan, A.M. Alafeefy, S.N. Bukhari, N.N. Elsayed, Optimization of antiproliferative activity of substituted phenyl 4-(2-oxoimidazolidin-1-yl) benzenesulfonates: QSAR and CoMFA analyses, Eur. J. Pharm. Sci. 77 (2015) 230–237.

[5] J.C. Dearden, M.T. Cronin, K.L. Kaiser, How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR), SAR QSAR Environ. Res. 20 (2009) 241–266.

[6] A. Tropsha, A. Golbraikh, Predictive QSAR modeling workflow, model applicability domains, and virtual screening, Curr. Pharm. Des. 13 (2007) 3494–3504.

[7] V.H. Masand, N.N.E. El-Sayed, D.T. Mahajan, V. Rastija, QSAR analysis for 6-arylpyrazine-2-carboxamides as Trypanosoma brucei inhibitors, SAR QSAR Environ. Res. 28 (2017) 165–177.

[8] V.H. Masand, D.T. Mahajan, G.M. Nazeruddin, T. Ben Hadda, V. Rastija, A.M. Alfeefy, Effect of information leakage and method of splitting (rational and random) on external predictive ability and behavior of different statistical parameters of QSAR model, Med. Chem. Res. 24 (2015) 1241–1264.

[9] Y. Marrero Ponce, Total and local (atom and atom type) molecular quadratic indices: significance interpretation, comparison to other molecular descriptors, and QSPR/QSAR applications, Bioorg. Med. Chem. 12 (2004) 6351–6369.

[10] J.L. Melville, J.D. Hirst, TMACC: interpretable correlation descriptors for quantitative structure−activity relationships, J. Chem. Inf. Model. 47 (2007) 626–634.

[11] F.R. Burden, M.J. Polley, D.A. Winkler, Toward novel universal descriptors: charge fingerprints, J. Chem. Inf. Model. 49 (2009) 710–715.

[12] C. Catana, Simple idea to generate fragment and pharmacophore descriptors and their implications in chemical informatics, J. Chem. Inf. Model. 49 (2009) 543–548.

[13] H.E.A. Ahmed, M. Vogt, J.r. Bajorath, Design and evaluation of bonded atom pair descriptors, J. Chem. Inf. Model. 50 (2010) 487–499.

[14] D.C. Kombo, K. Tallapragada, R. Jain, J. Chewning, A.A. Mazurov, J.D. Speake, T.A. Hauser, S. Toler, 3D molecular descriptors important for clinical success, J. Chem. Inf. Model. 53 (2013) 327–342.

[15] H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, W. Tong, Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics, J. Chem. Inf. Model. 48 (2008) 1337–1344.

[16] M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum-chemical descriptors in QSAR/QSPR studies, Chem. Rev. 96 (1996) 1027–1044.

[17] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, Weinheim, 2009.

[18] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.

[19] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474.

[20] https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm.

[21] https://www.schrodinger.com/.

[22] J. Dong, D.-S. Cao, H.-Y. Miao, S. Liu, B.-C. Deng, Y.-H. Yun, N.-N. Wang, A.-P. Lu, W.-B. Zeng, A.F. Chen, ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation, J. Cheminformatics 7 (2015).

[23] http://www.codessa-pro.com/.

[24] http://accelrys.com/products/collaborative-science/biovia-discovery-studio/.

[25] https://www.certara.com/software/molecular-modeling-and-simulation/sybyl-x-suite/.

[26] H. Fröhlich, J.K. Wegner, F. Sieker, A. Zell, Kernel functions for attributed molecular graphs – a new similarity-based approach to ADME prediction in classification and regression, QSAR Comb. Sci. 25 (2006) 317–326.

[27] J.K. Wegner, H. Fröhlich, H.M. Mielenz, A. Zell, Data and graph mining in chemical space for ADME and activity data sets, QSAR Comb. Sci. 25 (2006) 205–220.

[28] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen, The chemistry development Kit (CDK): an open-source Java library for chemo- and bioinformatics, J. Chem. Inf. Comput. Sci. 43 (2003) 493–500.

[29] V.J. Sykora, D.E. Leahy, Chemical descriptors library (CDL): a generic, open source software library for chemical informatics, J. Chem. Inf. Model. 48 (2008) 1931–1942.

[30] S. Yuan, H.C.S. Chan, Z. Hu, Using PyMOL as a platform for computational drug design, Wiley Interdiscip. Rev. Comput. Mol. Sci. 7 (2017) e1298.

[31] Y. Tanrikulu, M. Nietert, U. Scheffer, E. Proschak, K. Grabowski, P. Schneider, M. Weidlich, M. Karas, M. Gobel, G. Schneider, Scaffold hopping by "fuzzy" pharmacophores and its application to RNA targets, Chembiochem 8 (2007) 1932–1936.

[32] T.J. Hou, K. Xia, W. Zhang, X.J. Xu, ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach, J. Chem. Inf. Comput. Sci. 44 (2004) 266–275.

[33] V.H. Masand, D.T. Mahajan, P. Gramatica, J. Barlow, Tautomerism and multiple modelling enhance the efficacy of QSAR: antimalarial activity of phosphoramidate and phosphorothioamidate analogues of amiprophos methyl, Med. Chem. Res. 23 (2014) 4825–4835.

[34] P. Gramatica, S. Cassani, N. Chirico, QSARINS-chem: insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS, J. Comput. Chem. 35 (2014) 1036–1044.

[35] P. Gramatica, N. Chirico, E. Papa, S. Cassani, S. Kovarich, QSARINS: a new software for the development, analysis, and validation of QSAR MLR models, J. Comput. Chem. 34 (2013) 2121–2132.

[36] V.H. Masand, N.N.E. El-Sayed, D.T. Mahajan, A.G. Mercader, A.M. Alafeefy, I.G. Shibi, QSAR modeling for anti-human African trypanosomiasis activity of substituted 2-Phenylimidazopyridines, J. Mol. Struct. 1130 (2017) 711–718.

[37] V.H. Masand, D.T. Mahajan, A.K. Maldhure, V. Rastija, Quantitative structure–activity relationships (QSARs) and pharmacophore modeling for human African trypanosomiasis (HAT) activity of pyridyl benzamides and 3-(oxazolo[4,5-b]pyridin-2-yl)anilides, Med. Chem. Res. 25 (2016) 2324–2334.

[38] G. Sliwoski, J. Mendenhall, J. Meiler, Autocorrelation descriptor improvements for QSAR: 2DA_Sign and 3DA_Sign, J. Comput. Aided Mol. Des. 30 (2015) 209–217.

[39] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754.

[40] D. Zhou, Y. Alelyunas, R. Liu, Scores of extended connectivity fingerprint as descriptors in QSPR study of melting point and aqueous solubility, J. Chem. Inf. Model. 48 (2008) 981–987.

[41] S.R. Johnson, The trouble with QSAR (or how I learned to stop worrying and embrace fallacy), J. Chem. Inf. Model. 48 (2008) 25–26.

[42] T. Fujita, D.A. Winkler, Understanding the roles of the "two QSARs", J. Chem. Inf. Model. 56 (2016) 269–274.

[43] Y. Sushko, S. Novotarskyi, R. Körner, J. Vogt, A. Abdelaziz, I.V. Tetko, Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process, J. Cheminformatics 6 (2014).