KRAKENX: software for the generation of alignment-independent 3D descriptors

Article *in* Journal of Molecular Modeling · April 2016

DOI: 10.1007/s00894-016-2957-5

Bjørn Alsberg
Norwegian University of Science and Technology

J Mol Model
DOI 10.1007/s00894-016-2957-5

SOFTWARE REPORT

KRAKENX: software for the generation of alignment-independent 3D descriptors

Vishwesh Venkatraman[1] · Bjørn Kåre Alsberg[1]

Received: 4 November 2015 / Accepted: 7 March 2016

Abstract The KRAKENX software calculates a large variety of molecular descriptors based on quantum chemistry computations. The program supports over 2000 three-dimensional descriptors that are calculated from the output of different quantum chemistry packages. The current implementation supports semi-empirical MOPAC-based computations and primarily focuses on orientation-independent descriptors that have been discussed in the literature. The descriptor performance has been exemplified using a number of large and diverse datasets and can be seen to produce parsimonious linear models. The software can be run on multiple platforms and is available to academics free of charge.

Keywords Semi-empirical · MOPAC · Molecular descriptors · QSAR/QSPR · Alignment independence

Introduction

Understanding the factors underlying several physical, chemical, and biological phenomena has been a subject of keen interest in many fields such as toxicology, drug and materials design, and other areas of chemistry. In this respect, quantitative structure-activity/property relationship (QSAR/QSPR) models are being increasingly used to establish mathematical models correlating diverse molecular properties and descriptive parameters that can be derived from the structure of the molecule [1–3]. Central to the success of the modeling effort is the type and number of molecular descriptors being used. These descriptors can be sourced from different theories ranging from graph theory [4] to quantum chemistry [1] and vary in terms of information content [5, 6]. While in most cases no previous superposition of the structures is required, grid-based approaches such as comparative molecular field analysis [7] (CoMFA) and comparative molecular similarity index analysis [8] (CoMSIA) rely on a mandatory step of structural alignment (atom or field-based) to ensure comparability. However, suitable alignments are not always easily obtained and therefore superposition-free descriptors are actively sought.

In recent years, several commercial, free (for academics) and open-source software and libraries for descriptor calculation have been developed. Table 1 summarizes commonly used computational tools that provide a wide choice of descriptors calculated from different molecular representations. In addition, there are also web interfaces such as VCCLab [9], MODEL [10] and AMBIT [11] that provide descriptor calculation tools. The majority of the molecular descriptors are, however, derived from two-dimensional (2D) molecular graphs that capture bulk properties, atom types, connectivity indices, etc. More complex three-dimensional (3D) descriptors can be obtained from molecular orbital calculations that provide additional details about the geometrical and electronic properties of molecules. ParaSurf [12] for instance, uses semi-empirical methods to construct molecular surfaces and calculate local properties and descriptors. In software such as Open3DQSAR [13], molecular interaction fields (MIF) obtained using ab initio quantum chemistry calculations are used to define 3D-QSAR models.

✉ Bjørn Kåre Alsberg
alsberg@ntnu.no

[1] Department of Chemistry, Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway

Published online: 29 March 2016

Table 1 A summary of available molecular descriptor calculation software

Software	Comments	Availability	OS
DRAGON [14]	Calculates several constitutional, topological and geometrical descriptors from the 2D/3D structure	Commercial	Linux/Win
CODESSA-PRO [15]	various topological, geometrical and quantum chemical descriptors	Commercial	Win
MOLD^2 [16]	constitutional and topological descriptors from the 2D structure	Free	Win
PaDEL [17]	molecular descriptors and fingerprints	Open Source	Linux/Win/Mac
ChemoPy [18]	constitutional, topological and geometrical descriptors and molecular fingerprints	Open Source	Linux/Win
TMACC [19]	atomic property weighted topological maximum cross correlation descriptors	Open Source	Linux/Win/Mac
ISIDA [20]	derived from substructural and property labeled fragments	Free	Linux/Win/Mac
QuBiLS-MIDAS [21]	3D molecular descriptors based on multi-linear algebraic forms	Open Source	Linux/Win/Mac
JOELib [22]	constitutional and topological indices	Open Source	Linux/Win/Mac
Open3DQSAR [13]	descriptors based on molecular interaction fields	Open Source	Linux/Win/Mac
QuaSAR [23]	topological and conformation dependent descriptors	Commercial	Linux/Win/Mac
ParaSurf [12]	descriptors calculated from molecular surface properties	Commercial	Linux/Win/Mac
MOLCONN-Z [24]	connectivity, shape, and information Indices	Commercial	Linux/Win/Mac
VolSurf [25]	descriptors derived from 3D GRID based molecular interaction fields	Commercial	Linux/Win
PowerMV [26]	constitutional, atom pairs, Burden indices, pharmacophore	Free	Win
ChemAxon [27]	3D shape descriptors, Burden indices, fingerprints	Commercial	Linux/Win/Mac

”Win” and ”Mac” refer to Windows and Macintosh operating systems, respectively

Semi-empirical quantum chemical calculations are a rich source of molecular descriptors that can be obtained within a short computational timeframe and have a rich history of successful applications to QSAR/QSPR modeling [1, 28, 29]. While software such as CODESSA-PRO [15] and ParaSurf provide access to a limited set of descriptors, there is however, an evident lack of free software that can process the data. To address some of these limitations, we present KRAKENX, cross-platform open-source software that facilitates the calculation of various descriptors from semi-empirical MOPAC [30] calculations. The primary focus here is on whole molecule descriptors that do not require any molecular alignment. For a number of datasets tested, predictive and interpretable QSPR models have been obtained. The software is available to all academics for free and can be downloaded from http://www.krakenminer.com.

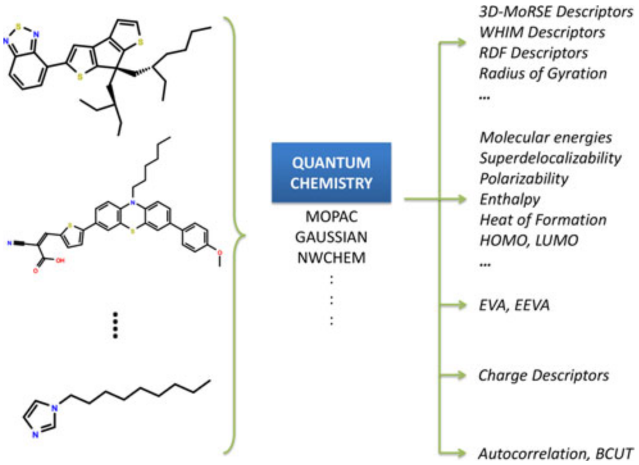


Fig. 1 Overall workflow of KRAKENX. The 3D structures are subjected to geometry optimization either at the semi-empirical level or using density functional theory. Molecular descriptors energies, charges, and other 3D indices are extracted from the output files. The program is written in Java and uses the CDK library routines and data structures for molecular manipulation

Software design

KRAKENX is written in the Java programming language and uses the Chemical Development Kit [31] (CDK) library for the manipulation of the molecular structures. A general workflow of the software is provided in Fig. 1. The program uses a command-line interface where the input is an ASCII file that has a shell script-like format (see example in Fig. 2). For descriptor calculation, KRAKENX requires the 3D molecular structures (in SDF format). The corresponding quantum chemistry output files can either be provided as a precomputed list of input files or alternatively calculated internally by specifying the location of the quantum chemistry software. Support currently only exists for semi-empirical MOPAC calculations but will be extended to other structure file formats and quantum chemistry programs in the future. The user is required to select the Yes or No option for the types of descriptors to be computed. Scripts for preparing and running MOPAC calculations are provided in the software.

Molecular descriptors

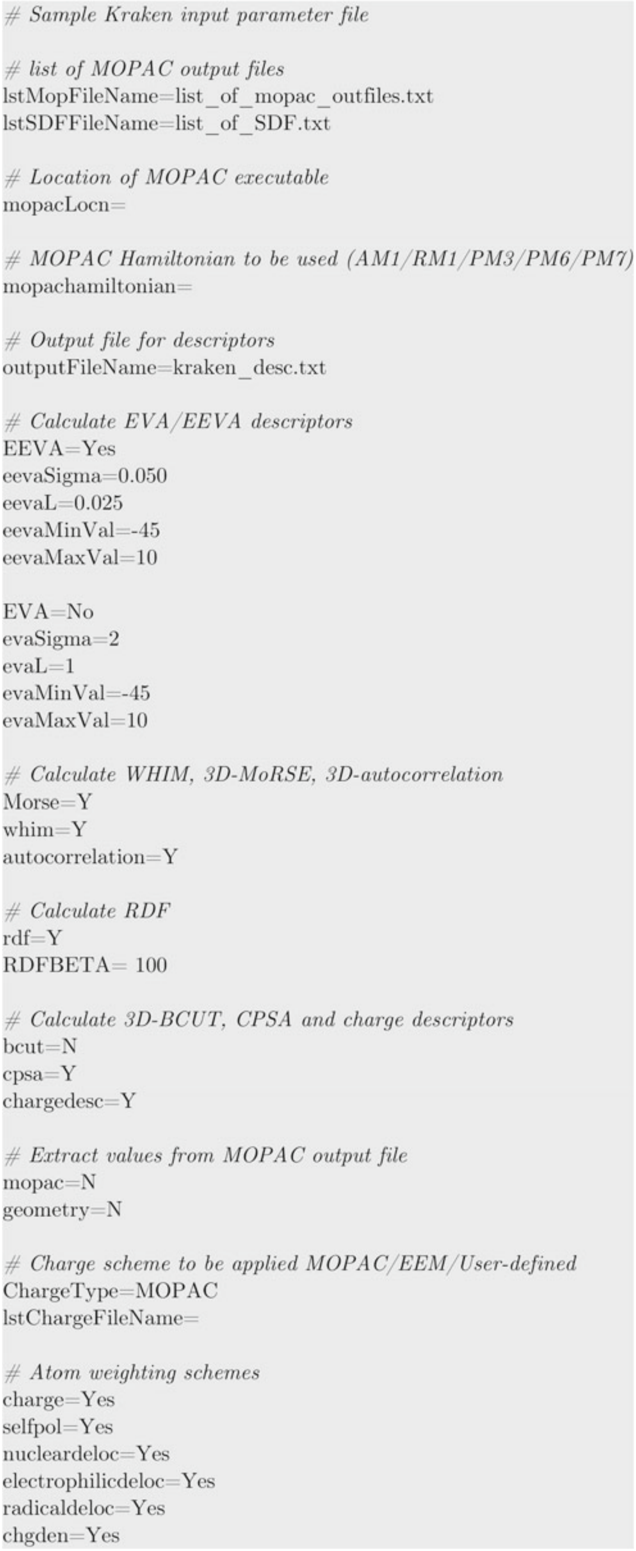
The descriptors computed by KRAKENX are divided into different classes/types. We have focused mainly on 3D and topographical descriptors (topological descriptors with 3D features included). Geometry optimization and force calculations in MOPAC are performed in two steps. The first optimization uses the keywords: ”HAMILTONIAN XYZ PRECISE MMOK SUPER ENPART ALLVEC LARGE ESP BONDS STATIC CYCLES=3000” where the HAMILTONIAN can be one of AM1/RM1/PM3/PM6/PM7. This is followed by a force calculation using the keywords: ”HAMILTONIAN OLDGEO FORCE THERMO”. The following set of descriptors (more details in the supplementary material) can be either derived or directly extracted from the MOPAC output files:

Quantum chemical: Quantities such as the heat of formation, area, volume, HOMO, LUMO, HOMO-LUMO gap, dipole moment, electronic energy, total self polarizability, electron-nuclear attraction and repulsion energies, total electrophilic and nucleophilic delocalizabilities, principal moments of inertia, polarizability and hyperpolarizability, the vibrational, rotational and translational values for the enthalpy, enthalpy and heat capacity [32–35].

Charge-related: Minimum and maximum atomic partial charge, polarity parameters [36], partial charge-weighted topological electronic descriptor [37], local dipole index [38], total squared charge, total absolute charge, total positive and negative charge [1].

CPSA: Charge partial surface area descriptors proposed by Stanton and Jurs [39] and Aptula et al. [40] combine partial charges and atomic solvent accessible surface areas (solvent radius set to 1.4Å).

EVA/EEVA: A density distribution-based eigenvalue [41] (EVA) and electronic eigenvalue [42] (EEVA) descriptors that are derived from the vibrational frequencies and molecular orbital energies of a geometry-optimized structure. These pseudo-spectral descriptors have been used to model different properties ranging from logP values of diverse organic chemicals [43] to power



```
# Sample Kraken input parameter file

# list of MOPAC output files
lstMopFileName=list_of_mopac_outfiles.txt
lstSDFFileName=list_of_SDF.txt

# Location of MOPAC executable
mopacLocn=

# MOPAC Hamiltonian to be used (AM1/RM1/PM3/PM6/PM7)
mopachamiltonian=

# Output file for descriptors
outputFileName=kraken_desc.txt

# Calculate EVA/EEVA descriptors
EEVA=Yes
eevaSigma=0.050
eevaL=0.025
eevaMinVal=-45
eevaMaxVal=10

EVA=No
evaSigma=2
evaL=1
evaMinVal=-45
evaMaxVal=10

# Calculate WHIM, 3D-MoRSE, 3D-autocorrelation
Morse=Y
whim=Y
autocorrelation=Y

# Calculate RDF
rdf=Y
RDFBETA= 100

# Calculate 3D-BCUT, CPSA and charge descriptors
bcut=N
cpsa=Y
chargedesc=Y

# Extract values from MOPAC output file
mopac=N
geometry=N

# Charge scheme to be applied MOPAC/EEM/User-defined
ChargeType=MOPAC
lstChargeFileName=

# Atom weighting schemes
charge=Yes
selfpol=Yes
nucleardeloc=Yes
electrophilicdeloc=Yes
radicaldeloc=Yes
chgden=Yes
```

Fig. 2 A sample input file showing the parameters to be read in by KRAKENX

conversion efficiencies of dye-sensitized solar cells [44–46].

Shape and geometry: This class of descriptors includes 3D Wiener index [47], inertial shape factor, radius of gyration, ovality, molecular eccentricity, asphericity, globularity [48], radial distribution function [49] (RDF), 3D Molecule Representation of Structure based on Electron diffraction [50] (MoRSE), Weighted Holistic Invariant Molecular [51] (WHIM).

Topographical: 3D topological distance-based autocorrelation [52] descriptors can be calculated by considering the average Euclidean distance between all atoms located at a given topological distance d. 3D BCUT values [53] can be calculated by encoding 3D atomic properties (charge, self-polarizability etc.) on the diagonals of the molecular connection table with off-diagonals encoding interatomic distances.

For each atom, the charge, electrophilic and nucleophilic frontier electron density, the electrophilic, nucleophilic and radical superdelocalizability, and the atom self-polarizability are used as atom-level weights in the calculation of descriptors such as the RDF, MoRSE, WHIM, and autocorrelation vectors. The atom-centered charges can be set to either the default MOPAC derived values, electrostatic potential (ESP), or user-defined charges (UDF) such as those based on the electronegativity equalization method [54] (EEM) (Table 2).

Results and discussion

To test the efficacy of the descriptors computed by KRAKENX, a number of publicly available datasets with different responses (boiling points, density, corrosion inhibition, viscosity, vapor pressure, water solubility, and biodegradability) have been analyzed in this study. Many of these properties are of particular relevance to post-combustion CO_2 capture absorbents. Here, the multivariate analysis has been performed using partial least squares regression [55] (PLSR) on molecular descriptors derived from a AM1 [56] Hamiltonian-based semi-empirical calculation. The pls [57] routines available in the R [58] statistical computing software were used to calculate the PLSR models. Prior to modeling, near-zero variance columns and those with missing values were removed. All variables were then autoscaled to zero mean and unit variance. The number of latent variables (N_{comp}) to include in the model was determined using tenfold cross-validation (CV). In addition, randomization tests (repeated 1000 times) were also carried out to guard against overfitting. In order to reduce model dimensionality and also to improve prediction performances (where possible), the number of variables was reduced using the variable importance in the projection [59] (VIP). These scores summarize the overall contribution of each variable to the model

Table 2 Prediction results for the different datasets

Property	Descriptors	Training		Testing	
		# Molecules	N_{comp}, R^2_{cv}	# Molecules	R^2_{test}
Density	EVA, EEVA, charge, autocorrelation, RDF, MoRSE, CPSA, MOPAC energies	7125	10, 0.85	1783	0.87
Vapor pressure	charge, autocorrelation, RDF, MoRSE, WHIM, CPSA, MOPAC energies,	2006	9, 0.84	504	0.85
Boiling point	charge, autocorrelation, RDF, MoRSE, WHIM, CPSA, MOPAC energies	4607	10, 0.84	1151	0.85
Viscosity	EVA, EEVA, charge, autocorrelation, RDF, MoRSE, CPSA, MOPAC energies	445	5, 0.68	113	0.68
Water solubility	charge, autocorrelation, RDF, MoRSE, WHIM, CPSA, MOPAC energies	4016	6, 0.75	1151	0.76
Corrosion inhibition	EVA, EEVA, charge, autocorrelation, RDF, MoRSE, WHIM, CPSA, MOPAC energies	131	4, 0.70	55	0.73

and a cut-off threshold of 1.0 was used. The predictive ability of the models was evaluated by the cross-validated correlation coefficient (R^2_{cv}) and root mean square error ($RMSE$):

$$R^2_{cv} = 1 - \frac{\sum(y_{obs,i} - \widehat{y_{cv,i}})^2}{\sum(y_{obs,i} - \overline{y_{obs}})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum(y^{obs} - y^{pred})^2}{n}} \quad (2)$$

where, $\widehat{y_{cv,i}}$ is the predicted value for the excluded (cross-validated) i^{th} compound, $y_{obs,i}$ is the corresponding experimental value and $\overline{y_{obs}}$ is the mean of the experimental values. To minimize any bias due to the random data splitting, the cross-validation was repeated 25 times with the training set instances shuffled during each run. All calculations were carried out on a desktop PC with Intel i5-2400 Quad-Core 3.10GHz CPU and 8GB RAM.

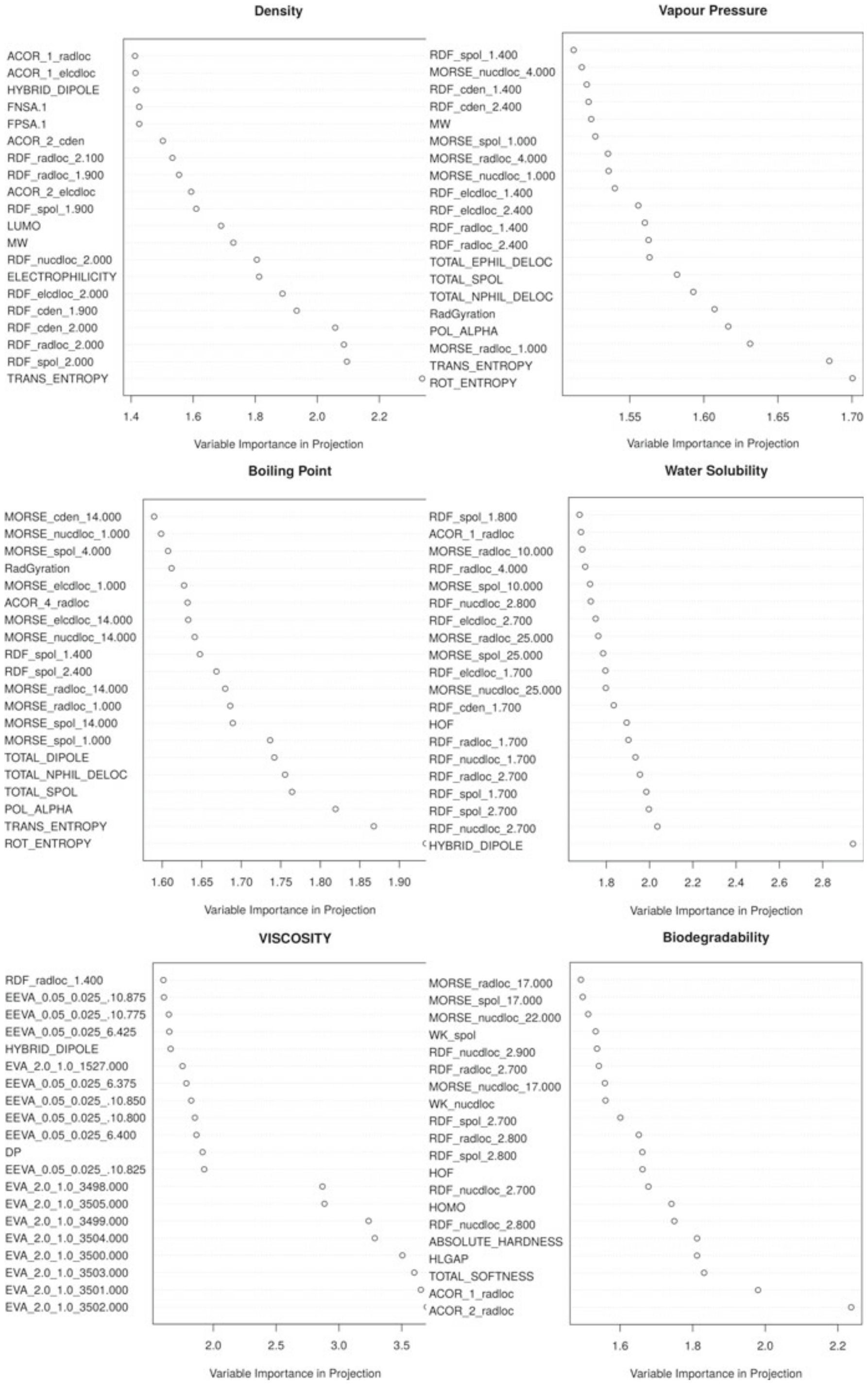
Reported models for the studied datasets have employed a broad range of 2D molecular descriptors and use an ensemble of models to combine predictions from nearest neighbors [60], multiple linear regression, clustering [61] and group contribution schemes [62]. Single models calculated for these 2D descriptors were found to have low predictive power. In comparison, for properties such as the density, viscosity, boiling point, and water solubility, results based on a single low complexity PLSR model were found to be comparable with the ensemble model approach used in previous studies. Table 3 shows variable importance plots for some of the models obtained.

Density: The dataset contains around 9000 diverse structures with densities in the range 0.5–5 g/cm^{-3}. For a training set of 7125 molecules, a ten-component PLSR model (based on 7000 descriptors comprising EVA, EEVA, RDF, MoRSE, CPSA, autocorrelation and charge-based indices) with $R^2_{cv} = 0.86$ was obtained. The test set R^2 for 1783 compounds was around 0.87. Important variables include the translational entropy, the LUMO energy, molecular weight (MW), and RDF descriptors weighted by the charge density, self-polarizability, and electrophilic delocalizability.

Vapor pressure: The $\log_{10}(vapor\ pressure\ mmHg)$ modeled for a training set of 2006 compounds yielded a nine-component PLSR model with $R^2_{cv} = 0.84$ and a test set squared correlation of 0.85. Here, important variables include static polarizability, the rotational and translational entropy, molecular weight, weighted RDF, and MoRSE descriptors.

Boiling point: For a set of 5759 chemicals containing boiling point data spanning the range -128 °C to 550 °C, modeling yielded training set statistics of $R^2_{cv} = 0.84$ for a ten-component PLSR model and corresponding test set

Table 3 The top 20 important variables (based on the variable importance in projection criterion) in the different PLSR models



$R^2 = 0.85$. The model shares a number of top-ranking descriptors with the vapor pressure model with add features such as autocorrelation and radius of gyration (gives the overall spread of the molecule).

Water solubility: The dataset comprising more than 5000 structures with water solubilities measured at around 25± 10° were modeled using a combination of EVA, EEVA, and other molecular descriptors. A six-component PLSR model yielded values of $R^2_{cv} = 0.75$ and $R^2_{test} = 0.76$. The model is dominated by weighted RDF and MoRSE descriptors and the heat of formation.

Viscosity: The dataset contains the viscosity values (measured at 25 °C) of 557 molecules [63]. For the modeled property $\log_{10}(viscosity\ cP)$, a five-component PLSR model with $R^2_{cv} = 0.68$ and $R^2_{test} = 0.68$ was obtained. The inclusion of EVA and EEVA descriptors in particular was seen to improve the model performance with numbers comparable to those of the consensus model. The use of random forest regression [64, 65] resulted in a marginal improvement in the test set results with $R^2_{cv} = 0.77$ and $R^2_{test} = 0.73$.

Corrosion inhibition: Data for the corrosion inhibition of steel in an acidic medium using organic inhibitors (such as triazole, oxadiazole aromatic hydrazides and thiadiazole derivatives) was taken from the literature [66]. A four-component PLSR model with training set $R^2_{cv} = 0.70$ and $R^2_{test} = 0.73$ was obtained.

Biodegradability: The biodegradability of chemicals is an important property, as accumulation of certain substances in the environment can be very harmful. A recent article by Mansouri et al. [67] investigated classification models to discriminate biodegradable and nonbiodegradable compounds. Here, chemicals with a biochemical oxygen demand (BOD) > 60 % are considered to be readily biodegradable (RB) while those with a BOD < 60 % are regarded as not readily biodegradable (NRB). The dataset used by Mansouri et al. [67] was additionally supplemented by compounds taken from the thesis by Eide-Haugmo [68], giving a total of 1746 compounds. A three-component partial least squares discriminant analysis [69, 70] (PLSDA) model yielded specificity and sensitivity values of 0.81 and 0.72, respectively, for a training set containing 858 compounds. Corresponding figures of merit for the test set (888 compounds) yielded values of 0.84 and 0.74, respectively. Prominent descriptors include the HOMO-LUMO gap (HLGap), weighted autocorrelation, WHIM, RDF, and MoRSE descriptors.

Conclusions

We have presented a standalone multi-platform molecular descriptor calculation software that provides a large number of alignment independent indices in one system. The script-based interface allows the user to select any combination of suitable descriptors. Multivariate analysis of large and diverse datasets has demonstrated performance comparable with published results (based on an ensemble of linear/non-linear approaches). A graphical user interface and extensions to other quantum chemical packages is currently underway. Details of the usage and availability of the software and associated scripts can be found at http://www.krakenminer.com.

Acknowledgments The Norwegian Research Council (NFR) is acknowledged for financial support from the CLIMIT (Grant No. 233776). We also thank ChemAxon (http://www.chemaxon.com) for free academic use of the Marvin package.

References

1. Karelson M, Lobanov VS, Katritzky AR (1996) Chem Rev 96(3):1027
2. Le T, Epa VC, Burden FR, Winkler DA (2012) Chem Rev 112(5):2889
3. Katritzky AR, Kuanar M, Slavov S, Hall CD, Karelson M, Kahn I, Dobchev DA (2010) Chem Rev 110(10):5714
4. Pogliani L (2000) Chem Rev 100(10):3827
5. Todeschini R, Consonni V (2010) Molecular Descriptors for Chemoinformatics, vol 41. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
6. Guha R, Willighagen E (2012) Curr Top Med Chem 12(18):1946
7. Cramer RD, Patterson DE, Bunce JD (1988) J Am Chem Soc 110(18):5959
8. Klebe G, Abraham U, Mietzner T (1994) J Med Chem 37(24):4130
9. Tetko I, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV (2005) J Comput-Aided Mol Des 19(6):453
10. Li Z, Han L, Xue Y, Yap C, Li H, Jiang L, Chen Y (2007) Biotechnol Bioeng 97(2):389
11. Jeliazkova N, Jeliazkov V (2011) J Cheminf 3(1):18
12. Parasurf'10 academic version (2010) CEPOS Insilico Ltd. Erlangen, Germany
13. Tosco P, Balle T (2011) J Mol Model 17(1):201
14. Talete srl, dragon (software for molecular descriptor calculation) (2012). Version 6.0, http://www.talete.mi.it
15. Codessa pro version 1.0 rc2 (2002). University of Florida: Gainesville, FL
16. Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, Su Z, Perkins R, Tong W, Chem J (2008) Inf Model 48(7):1337
17. Yap CW (2011) J Comp Chem 32(7):1466
18. Cao DS, Xu QS, Hu QN, Liang YZ (2013) Bioinformatics 29(8):1092
19. Melville JL, Hirst JD, Chem J (2007) Inf Model 47(2): 626
20. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov'ev V, Hoonakker F, Tetko IV, Marcou G (2008) Curr Comp Aided-Drug Des 4(3):191
21. García-Jacas CR, Marrero-Ponce Y, Acevedo-Martínez L, Barigye SJ, Valdés-Martinez JR, Contreras-Torres E (2014) J Comp Chem 35(18):1395
22. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen EL, Chem J (2006) Inf Model 46(3):991
23. Molecular operating environment (moe), 2013.08 (2015). Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7
24. Hall LH, Kellog GE, Haney DN (2002) Molconn-Z version 4.00 user guide. Edusoft LC, La Jolla, CA
25. Cruciani G, Crivori P, Carrupt PA, Testa B (2000) J Mol Struc-THEOCHEM 503(1-2):17
26. Liu J, Feng J, Young S, Chem J (2005) Inf Model 45:515
27. Marvin 5.9.3 (2012). ChemAxon (http://www.chemaxon.com)
28. Gece G (2008) Corros Sci 50(11):2981
29. Dehmer M, Varmuza K, Bonchev D (2012) Statistical Modelling of Molecular Descriptors in QSAR/QSPR. Wiley-VCH Verlag GmbH & Co, KGaA, Weinheim
30. Stewart JJP Mopac2012 version 14.1421 (2012). Stewart Computational Chemistry, Colorado Springs, CO, USA, (http://OpenMOPAC.net)
31. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen E (2006) Curr Pharm Des 12(17):2111

32. Csizmadia I (1976) Theory and Practice of MO Calculations on Organic Molecules. Progress in theoretical organic chemistry (Elsevier Scientific Pub. Co, Amsterdam, New York
33. Clementi E (1980) Computational Aspects for Large Chemical Systems Lecture Notes in Chemistry. Springer, Berlin Heidelberg
34. McQuarrie DA (1985) Statistical Thermodynamics, University Science Books
35. Akhiezer A. I., Peletminskii S. V., Ter Haar D. (eds) (1981) Methods of Statistical Physics, International Series in Natural Philosophy, vol 104. Pergamon
36. Osmiaowski K, Halkiewicz J, Radecki A, Kaliszan R (1985) J Chromatogr A 346:53
37. Katritzky AR, Mu L, Karelson M (1996) J Chem Inf Model 36(6):1162
38. Clare BW, Supuran CT (1994) J Pharm Sci 83(6):768
39. Stanton D, Jurs P (1990) Anal Chem 62:2323
40. Aptula A, Kühne R, Ebert RU, Cronin M, Netzeva T, Schüürmann G (2003) Mol Inf 22(1):113
41. Turner DB, Willett P (2000) Eur J Med Chem 35(4):367
42. Tuppurainen K (1999) SAR QSAR Environ Res 10(1):39
43. Heritage T, Ferguson A, Turner D, Willett P (1998) Perspect. Drug Discov 381:9–11
44. Venkatraman V, Åstrand PO, Alsberg BK (2014) J Comput Chem 35(3):214
45. Venkatraman V, Alsberg BK (2015) Dyes Pigment 114(0): 69
46. Venkatraman V, Foscato M, Jensen VR, Alsberg BK (2015) J Mater Chem A 3:9851
47. Bogdanov B, Nikolić S, Trinajstić N (1989) J Math Chem 3(3):299
48. Todeschini R, Consonni V (2003). In: Gasteiger J (ed) Handbook of Chemoinformatics: From Data to Knowledge. Wiley-VCH Verlag GmbH, Weinheim, Germany
49. Hemmer MC, Steinhauer V, Gasteiger J (1999) Vib Spectrosc 19(1):151
50. Schuur JH, Selzer P, Gasteiger J (1996) J Chem Inf Model 36(2):334
51. Todeschini R, Vighi M, Provenzani R, Finizio A, Gramatica P (1996) Chemosphere 32(8):1527
52. Klein CT, Kaiser D, Ecker G, Chem J (2004) Inf Model 44(1):200
53. Bajorath J (2004) Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery, Methods in Molecular Biology, vol 41. Humana Press
54. Bultinck P, Langenaeker W, Carbó-Dorca R, Tollenaere JP (2003) J Chem Inf Model 43(2):422
55. Geladi P, Kowalski BR (1986) Anal Chim Acta 185(0):1
56. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902
57. Mevik BH, Wehrens R (2007) J Stat Softw 18(2):1
58. Team RC (2015) R: A Language and Environment for Statistical Computing, Vienna, Austria. https://www.R-project.org/
59. Chong IG, Jun CH (2005) Chemomtr Intell Lab 78(1-2):103
60. Contrera JF, Matthews EJ, Benz RD (2003) Regul Toxicol Pharm 38(3):243
61. Martin TM, Harten P, Venkatapathy R, Das S, Young DM (2008) Toxicol Mech Methods 18(2-3):251
62. Martin TM, Young DM (2001) Chem Res Toxicol 14(10):1378
63. Viswanath D, Ghosh T, Prasad D, Dutt N, Rani K (2007) Viscosity of Liquids: Theory, Estimation, Experiment and Data. Springer, Netherlands
64. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and Regression Trees. Wadsworth
65. Liaw A, Wiener M (2002) R News 2(3):18
66. Melagraki G, Afantitis A (2013) Chemomtr Intell Lab 123:9
67. Mansouri K, Ringsted T, Ballabio D, Todeschini R, Consonni V (2013) J Chem Inf Model 53(4):867
68. Eide-Haugmo I (2011) Environmental impacts and aspects of absorbents used for co_2 capture. Ph.D. thesis, Norges Teknisk-Naturvitenskapelige Universitet, Norway
69. Brown BWE, Steven D, Tauler R (2009) Comprehensive chemometrics Chemical and biochemical data analysis. Elsevier
70. Sanchez G (2013) DiscriMiner: Tools of the Trade for Discriminant Analysis. http://CRAN.R-project.org/package=DiscriMiner. R package version 0.1-29