Supplementary Material (ESI) for Chemical Society Reviews

Supporting Information for the article

Predictive modeling in homogeneous catalysis: A tutorial

Ana G. Maldonado, Gadi Rothenberg

Van't Hoff Institute of Molecular Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands.
Fax: (+31)-(0)20-525-5604; phone: (+31)-(0)20-525-6963; E-mail: g.rothenberg@ uva.nl

Supporting Information 1. Computational methods used in the case study

The original case study dataset of 115 bidentate ligand-Ni complexes consists on two-dimensional structures. Ligand geometry optimization for calculating three-dimensional descriptors was performed using Hyperchem.[1] We used the MM+ force field in combination with a conjugate gradient optimisation method (Polak-Ribiere). Then, the ligand descriptors were computed with the Codessa software package[2] and analysed using Matlab scripts.[3] A total of 168 (2D and 3D) descriptors were calculated (full list further in the Supporting Information 4).

The FOM and the descriptor data was analyzed using principal component analysis (PCA).[4] model constructions was done using a partial least squares (PLS) algorithm.[4, 5] The experimental FOM (expFOM) is the reference value. The predicted FOM (predFOM) is the value to be validated. The model validation error and the ΔFOM model validation, were computed using the following expressions:

$\Delta FOM = | expFOM - predFOM |$
$\%error = | \Delta FOM / expFOM | \times 100$

Supporting Information 2. List of selected software and algorithms for descriptor calculation and selection

The number and complexity of known molecular descriptors keep increasing with time. A full list of software for descriptor calculation and selection is thus out of the scope of this supplementary Material.

One of the most popular commercial software for descriptor calculation is Dragon.[6] It computes more than 3000 descriptors (2D, 3D, constitutional, and topological). Other software packages are: Codessa (~1500 descriptors. It calculates constitutional, topological, geometrical, charge-related, semi-empirical and thermodynamical descriptors),[2] ADMET Predictor (297 descriptors for absorption, distribution, metabolism, excretion and toxicity properties),[7] Adriana Code (1244 descriptors),[8]

Molgen (707 descriptors),[9] MOE (~300 descriptors including topological indices, structural keys, E-state indices and physical properties),[10, 11] Molconn-Z (40 descriptors including molecular connectivity, shape, and information indices),[12] ALMOND (computes GRid INdependent Descriptors)[13] and GRID (descriptors derived from molecular interaction fields).[14] The last two from Molecular Discovery Ltd.[15] If you fancy programming, then the Chemistry Development Kit (CDK)[16] offers great QSAR descriptor calculations.

For those who do not want to download and install software there is also the possibility to compute descriptors on line. E-DRAGON[17] is the on-line brother of Dragon.[6] It can compute up to 1600 descriptors in a maximum of 149 molecules per query. The JME molecular editor offers drawing or loading of your molecules and calculation of some important physicochemical descriptors.[18]

Supporting Information 3. List of selected software and algorithms for model building and data analysis

There are published studies on the use and comparison of different predictive models.[19] Often the software packages used to compute descriptors are also designed to build regression models. Among them we can mention: MOE[6] (inbuilt QSAR/QSPR model construction using linear, probabilistic and decision-tree methodologies), Molgen[9] (have a QSPR regression analysis engine), ALMOND[13] (computes QSAR, Quantitative Structure-Metabolism Relationships and Quantitative Structure-Transport Relationships), ADMET Predictor[7] (build predictive ADMET models), Codessa (computes QSPR regressions, multivariate methods and descriptor selection). Other software packages are exclusively design to build lineal, nonlinear or multivariate regressions. MobyDigs[20] calculates optimal subsets of regression models using genetic algorithms. Sarchitect[21] builds and select the 'best' model automatically. PowerMV is a software environment for statistical analysis, molecular viewing, descriptor generation, and similarity search.[22] JOELib is a platform independent open source computational chemistry package.[23, 24]

Algorithms are less numerous than software packages. They are often an important part on the model building and selection engines. The algorithms can be classified in two kinds:

For classification models (also know as discrete models)

- Naïve Bayesian[25]
- Classification Trees[26, 27]
- Neural Networks (NN)[28-30]
- Supported vector machines (SVM)[31, 32]
- Decision Forests[33]

For regression and predictive models (also know as continuous models)

- Multiple Linear Regression (MLR)[34]
- Neural Networks (NN) [28-30]
- Regression Forests[33]
- Partial Least Squares (PLS)[5,22]

Supporting Information 4. Complete list of the 168 computed descriptors

{1}=Number of atoms
{2}=Number of C atoms
{3}=Relative number of C atoms
{4}=Number of H atoms
{5}=Relative number of H atoms
{6}=Number of O atoms
{7}=Relative number of O atoms
{8}=Number of N atoms
{9}=Relative number of N atoms
{10}=Number of S atoms
{11}=Relative number of S atoms
{12}=Number of F atoms
{13}=Relative number of F atoms
{14}=Number of Cl atoms
{15}=Relative number of Cl atoms
{16}=Number of Br atoms
{17}=Relative number of Br atoms
{18}=Number of I atoms
{19}=Relative number of I atoms
{20}=Number of P atoms
{21}=Relative number of P atoms
{22}=Number of bonds
{23}=Number of single bonds
{24}=Relative number of single bonds
{25}=Number of double bonds
{26}=Relative number of double bonds
{27}=Number of triple bonds
{28}=Relative number of triple bonds
{29}=Number of aromatic bonds
{30}=Relative number of aromatic bonds
{31}=Number of rings
{32}=Relative number of rings
{33}=Number of benzene rings
{34}=Relative number of benzene rings
{35}=Molecular weight
{36}=Relative molecular weight
{37}=Gravitation index (all bonds)
{38}=Gravitation index (all pairs)
{39}=Wiener index
{40}=Randic index (order 0)
{41}=Randic index (order 1)
{42}=Randic index (order 2)

{43}=Randic index (order 3)
{44}=Kier&Hall index (order 0)
{45}=Kier&Hall index (order 1)
{46}=Kier&Hall index (order 2)
{47}=Kier&Hall index (order 3)
{48}=Kier shape index (order 1)
{49}=Kier shape index (order 2)
{51}=Kier shape index (order 3)
{52}=Kier flexibility index
{53}=Average Information content (order 0)
{54}=Information content (order 0)
{55}=Average Structural Information content (order 0)
{56}=Structural Information content (order 0)
{57}=Average Complementary Information content (order 0)
{58}=Complementary Information content (order 0)
{59}=Average Bonding Information content (order 0)
{60}=Bonding Information content (order 0)
{61}=Average Information content (order 1)
{62}=Information content (order 1)
{63}=Average Structural Information content (order 1)
{64}=Structural Information content (order 1)
{65}=Average Complementary Information content (order 1)
{66}=Complementary Information content (order 1)
{67}=Average Bonding Information content (order 1)
{68}=Bonding Information content (order 1)
{69}=Average Information content (order 2)
{70}=Information content (order 2)
{71}=Average Structural Information content (order 2)
{72}=Structural Information content (order 2)
{73}=Average Complementary Information content (order 2)
{74}=Complementary Information content (order 2)
{75}=Average Bonding Information content (order 2)
{76}=Bonding Information content (order 2)
{77}=Balaban index
{78}=Moment of inertia A
{79}=Moment of inertia B
{80}=Moment of inertia C
{81}=XY Shadow
{82}=XY Shadow / XY Rectangle
{83}=YZ Shadow
{84}=YZ Shadow / YZ Rectangle
{85}=ZX Shadow
{86}=ZX Shadow / ZX Rectangle
{87}=Molecular volume
{88}=Molecular volume / XYZ Box
{89}=Molecular surface area
{90}=Max partial charge for a C atom
{91}=Min partial charge for a C atom
{92}=Max partial charge for a H atom
{93}=Min partial charge for a H atom
{94}=Max partial charge for a O atom

{95}=Min partial charge for a O atom
{96}=Max partial charge for a N atom
{97}=Min partial charge for a N atom
{98}=Max partial charge for a F atom
{99}=Min partial charge for a F atom
{100}=Max partial charge for a Pd atom
{101}=Min partial charge for a Pd atom
{102}=Max partial charge (Qmax)
{103}=Min partial charge (Qmin)
{104}=Polarity parameter (Qmax-Qmin)
{105}=Polarity parameter / square distance
{106}=Topographic electronic index (all pairs)
{107}=Topographic electronic index (all bonds)
{108}=TMSA Total molecular surface area
{109}=PPSA-1 Partial positive surface area
{110}=PNSA-1 Partial negative surface area
{111}=DPSA-1 Difference in CPSAs (PPSA1-PNSA1)
{112}=FPSA-1 Fractional PPSA (PPSA-1/TMSA)
{113}=FNSA-1 Fractional PNSA (PNSA-1/TMSA)
{114}=WPSA-1 Weighted PPSA (PPSA1*TMSA/1000)
{115}=WNSA-1 Weighted PNSA (PNSA1*TMSA/1000)
{116}=PPSA-2 Total charge weighted PPSA
{117}=PNSA-2 Total charge weighted PNSA
{118}=DPSA-2 Difference in CPSAs (PPSA2-PNSA2)
{119}=FPSA-2 Fractional PPSA (PPSA-2/TMSA)
{120}=FNSA-2 Fractional PNSA (PNSA-2/TMSA)
{121}=WPSA-2 Weighted PPSA (PPSA2*TMSA/1000)
{122}=WNSA-2 Weighted PNSA (PNSA2*TMSA/1000)
{123}=PPSA-3 Atomic charge weighted PPSA
{124}=PNSA-3 Atomic charge weighted PNSA
{125}=DPSA-3 Difference in CPSAs (PPSA3-PNSA3)
{126}=FPSA-3 Fractional PPSA (PPSA-3/TMSA)
{127}=FNSA-3 Fractional PNSA (PNSA-3/TMSA)
{128}=WPSA-3 Weighted PPSA (PPSA3*TMSA/1000)
{129}=WNSA-3 Weighted PNSA (PNSA3*TMSA/1000)
{130}=RPCG Relative positive charge (QMPOS/QTPLUS)
{131}=RPCS Relative positive charged SA (SAMPOS*RPCG)
{132}=RNCG Relative negative charge (QMNEG/QTMINUS)
{133}=RNCS Relative negative charged SA (SAMNEG*RNCG)
{134}=HDSA H-donors surface area
{135}=FHDSA Fractional HDSA (HDSA/TMSA)
{136}=HASA H-acceptors surface area
{137}=FHASA Fractional HASA (HASA/TMSA)
{138}=HBSA H-bonding surface area
{139}=FHBSA Fractional HBSA (HBSA/TMSA)
{140}=HDCA H-donors charged surface area
{141}=FHDCA Fractional HDCA (HDCA/TMSA)
{142}=HACA H-acceptors charged surface area
{143}=FHACA Fractional HACA (HACA/TMSA)
{144}=HBCA H-bonding charged surface area
{145}=FHBCA Fractional HBSA (HBSA/TMSA)

{146}=Min (#HA, #HD)
{147}=Count of H-acceptor sites
{148}=Count of H-donors sites
{149}=HA dependent HDSA-1
{150}=HA dependent HDSA-1/TMSA
{151}=HA dependent HDSA-2
{152}=HA dependent HDSA-2/TMSA
{153}=HA dependent HDSA-2/SQRT (TMSA)
{154}=HA dependent HDCA-1
{155}=HA dependent HDCA-1/TMSA
{156}=HA dependent HDCA-2
{157}=HA dependent HDCA-2/TMSA
{158}=HA dependent HDCA-2/SQRT (TMSA)
{159}=HASA-1
{160}=HASA-1/TMSA
{161}=HASA-2
{162}=HASA-2/TMSA
{163}=HASA-2/SQRT (TMSA)
{164}=HACA-1
{165}=HACA-1/TMSA
{166}=HACA-2
{167}=HACA-2/TMSA
{168}=HACA-2/SQRT (TMSA)

References

1. *Hyperchem for Windows (Molecular Modeling System)*, http://www.hyper.com/.
2. *Codessa-Pro Software*, http://www.codessa-pro.com/.
3. *The Mathworks, Inc. Matlab*, http://www.mathworks.com/.
4. S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 37-52.
5. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, 1986, **185**, 1-17.
6. *Dragon Software*, http://www.talete.mi.it/dragon_net.htm.
7. *ADMET Predictor Software*, http://www.simulations-plus.com/.
8. *ADRIANA.code*, http://www.molecular-networks.com/products/adrianacode.
9. *MOLGEN*, http://www.molgen.de/?src=documents/molgenqspr.html.
10. *MOE*, http://www.chemcomp.com/.
11. P. Labute, *J. Mol. Graph Model.*, 2000, **18**, 464-477.
12. *Molconn-Z software*, http://www.edusoft-lc.com/molconn/.
13. *ALMOND*, http://www.moldiscovery.com/soft_almond.php.
14. M. Pastor, G. Cruciani and K. A. Watson, *Journal of Medicinal Chemistry*, 1997, **40**, 4089-4102.
15. *Molecular Discovery Ltd.*, http://www.moldiscovery.com/.
16. *Chemistry Development Kit (CDK)*, http://sourceforge.net/apps/mediawiki/cdk/index.php?title=Main_Page.
17. *E-Dragon*, http://www.vcclab.org/lab/edragon/.
18. *JME Molecular editor*, http://www.molinspiration.com/cgi-bin/properties.
19. M. Greener, *Drug Discovery and Development*, 2005.

20. *MobyDigs Software*, http://www.talete.mi.it/products/moby_description.htm.
21. *Sarchitect software*, http://www.strandls.com/sarchitect/features/modelbuilding.html.
22. *PowerMV*, http://nisla05.niss.org/PowerMV/.
23. F. Holger, K. W. Jörg and Z. Andreas, *QSAR & Combinatorial Science*, 2004, **23**, 311-318.
24. *JOELib package*, http://www.ra.cs.uni-tuebingen.de/software/joelib/index.html.
25. D. S. Silvia and J. Skilling, *Data analysis: a Bayesian tutorial*, Oxford University Press, 2006.
26. G. De'ath and K. E. Fabricius, *Ecology*, 2000, **81**, 3178-3192.
27. W. Buntine, *Stat. Comp.*, 1992, **2**, 63-73.
28. J. M. Serra, A. Corma, A. Chica, E. Argente and V. Botti, *Catalysis Today*, 2003, **81**, 393-403.
29. I. A. Basheer and M. Hajmeer, *J. Microbiol. Meth.* , 2000, **43**, 3-31.
30. S. Mazurek, T. R.Ward and M. Novic, *Mol. Divers.*, 2007, **11**, 141-152.
31. A. J. Smola and B. Schölkopf, *Stat. Comp.*, 2004, **14**, 199-222.
32. B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT Press, 2002.
33. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, *Journal of Chemical Information and Computer Sciences*, 2003, **43**, 1947-1958.
34. D. C.Montgomery, E. A. Peck and G. G. Vining., *Introduction to Linear Regression Analysis*, 3rd edn., Wiley, New York, 2001.