Supporting information for the article

New tricks by very old dogs: Predicting the catalytic hydrogenation of HMF derivatives using Slater-type orbitals

Erik-Jan Ras,*[a,b] Manuel J. Louwerse[b] and Gadi Rothenberg*[b]

[a] Avantium Technologies B.V. Zekeringstraat 29, 1014BV Amsterdam, The Netherlands. Tel. +31 (0)20 586 8080, Fax. +31 (0)20 586 8085, email: erikjan.ras@avantium.com; www.avantium.com

[b] Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Science Park 904, 1098XH Amsterdam, The Netherlands. Tel. +31 (0)20 525 6963, Fax.+31 (0)20 525 5604, e-mail: g.rothenberg@uva.nl; www.science.uva.nl/hims/hcsc

Summary

This supporting information contains a detailed description of the modeling methods, a description of the subsets data selection process, details of the data and methods used in the model validation process, a typical example of the model equations and their application, and seven additional references.

Modeling methods

The models were created using the Orthogonal Partial Least Squares (OPLS) method implemented in Simca-P+ 12 (Umetrics). This method was first introduced by Trygg et al.,[1-5] and is basically an extension of the Partial Least Squares (PLS) approach. It combines Orthogonal Signal Correction (OSC, an algorithm often applied in spectroscopic data analysis) with classical PLS regression. Using OPLS, the first step is orthogonalising the variables and responses, thus establishing a 1:1 relationship between the variables (x) and each response (y). This simplifies the interpretation of the results. When the responses are correlated, this will also show up as a result. We preferred using OPLS in this case over PLS because of the internal correlations in the descriptor matrix. Note, however, that using OPLS does not improve the model's performance – it only simplifies the interpretation. Figure S1 shows a block flowchart of the OPLS method. A detailed technical description of the OPLS algorithm is given in the work of Trygg and Wold.[6]

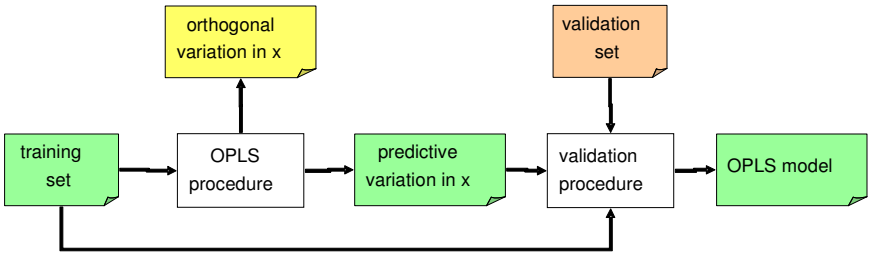


Figure S1 Block flowchart of the OPLS procedure, including model validation

Selecting the training sets and the validation sets

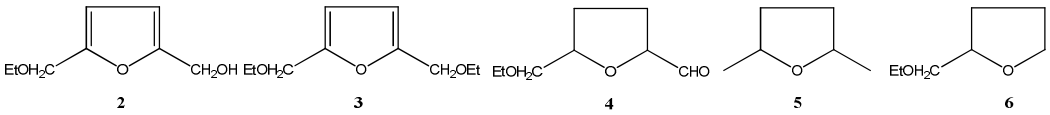
Selecting representative training and validation sets is highly important. The training set must be large enough and diverse enough so that the model can span the entire data set, yet small enough so that sufficient validation data remains. Usually, candidates for each set are chosen using random selection. However, this does not take into account the distribution of the variables in the data set. We believe that in order to select an appropriate candidate set, one must consider this distribution in the selection process. To do this, we used a combination of algorithms from statistical experimental design, as implemented in Matlab R2007b.

Prior to selection, the data is auto-scaled (zero mean, unit variance). This avoids bias in the selection as a result of scale differences. The first candidate points then are selected using a D-optimal method.[7] This ensures that the extremes and the center of the design space are sufficiently covered in all dimensions. As a rule of thumb, at least twice the number of variables plus one (for the center point) points should be selected using this method. Subsequently, additional points are selected using a space-filling method, ensuring an equally distributed design across all dimensions. After the selection is complete, the selected points form the training set, and the remaining points form the validation set. The model is then trained using the training set, and validated using the validation set.

Validation data associated with the full monometallics data set

The data set consists of 24 data points (8 catalysts x 3 temperatures). Due to the small data set size a proper partitioning in training set and validation set is not possible. Instead, we used a modification of the leave-one-out procedure. In eight independent validation rounds, the subset of data for each of the metals was excluded once and the model was refitted. Using this model, a prediction was made for the left-out metal. Table S1 shows the errors obtained for each response (the respective yields of products **2–6**) in each case. After completing all the eight validations, a prediction overview can be made. Note that the data is transformed to a log scale [$y^* = {}^{10}\log(y+1)$] prior to regression. Consequently, all the root-mean-square-error (RMSE) values and ranges in Table S1 are expressed in the same transformed units. The use of the log transform is justified for two reasons. One is the high skewness of the responses. The other is that a model based on untransformed data is clearly non-linear, whereas for a linear-in-parameters model such as OPLS one expects a linear relationship. OPLS, like any regression method, assumes the data set follows a Gaussian distribution. A data set that is skewed to one end of the scale will show a closer resemblance to a Gaussian (= normal) distribution after a log transform is applied.

Table S1 Root mean square errors in log units of estimate (n=21) and of prediction (n=3) for the 8 validation rounds



2 3 4 5 6

				Metal excluded from data set							
Response	**Data set range**		**Error measure**	**Au**	**Cu**	**Ir**	**Ni**	**Pd**	**Pt**	**Rh**	**Ru**
Yield of **2**	Min	0.00	RMSEE	0.43	0.40	0.41	0.41	0.21	0.38	0.40	0.41
	Max	1.84	RMSEP	0.23	0.17	0.30	0.88	0.92	0.42	0.36	0.06
Yield of **3**	Min	0.00	RMSEE	0.35	0.41	0.33	0.17	0.12	0.33	0.35	0.31
	Max	1.82	RMSEP	0.18	0.47	0.70	2.35	1.03	0.19	0.34	0.41
Yield of **4**	Min	0.00	RMSEE	0.46	0.57	0.37	0.32	0.20	0.49	0.55	0.40
	Max	1.90	RMSEP	0.47	0.34	1.53	2.84	1.30	0.16	0.29	1.00
Yield of **5**	Min	0.13	RMSEE	0.08	0.16	0.09	0.10	0.08	0.12	0.10	0.15
	Max	1.46	RMSEP	0.17	0.29	0.20	0.10	0.14	0.13	0.10	0.23
Yield of **6**	Min	0.00	RMSEE	0.34	0.38	0.26	0.34	0.19	0.41	0.36	0.29
	max	1.56	RMSEP	0.26	0.16	1.19	0.77	0.76	0.05	0.33	0.78

The model robustness suffers from the small data set size. This is indicated by the fact that the error measure for the training set (RMSEE) is often larger than the error in the validation set (RMSEP). Still, in most cases the error is of such a magnitude that, when compared with the range of the response, a prediction could be made for an untested metal that would classify that metal as likely to perform "good" or "bad" when actually tested. Of course, some responses have an overall better model performance (e.g. the yield of **5**) than others (e.g. the yield of **4**). The same holds for the metals. Excluding the Pd gives an overall better model for all responses. This agrees with our chemical intuition, as the Pd-containing catalyst is the only one that gives a substantial yield of the saturated aldehyde **4**. In essence, this implies a) that Pd is in fact an outlier in the data set and b) the response yield of **4** has little relevance for the remainder of the metals. When this model is regressed an improved fit is obtained with an R^2 of 0.85 and a Q^2 of 0.76, compared to an R^2 of 0.64 and a Q^2 of 0.41 for the original model. The R^2 values for individual responses in the original and the modified (excluding observations with Pd catalyst as well as the yield of 4 as a response) model are given in Table 2.

Table S2 Squared correlation coefficients R^2 for the original model (all metals, all responses) and the modified model (Pd catalyst and Yield of **4** excluded)

	R^2 for individual responses				
	Yield of **2**	Yield of **3**	Yield of **4**	Yield of **5**	Yield of **6**
Original model	0.28	0.72	0.53	0.95	0.63
Modified model	0.75	0.96	n.a.	0.96	0.74

Example of typical model equations and usage

The latent variables from the OPLS model, each consisting of a linear combination of the original variables, can be transformed to a regular polynomial. This polynomial allows for the calculation of the response (y) as a function of the original variables (x). As an example, Table S3 gives the coefficients for the polynomials for the monometallic catalysts model. Note that the polynomials yield the transformed response $y^* = {}^{10}\log(y+1)$ instead of the yield in real numbers.

Table S3 Model coefficients for the monometallic catalysts model reported in the main text.

	Model for			
Term	Yield of **2**	Yield of **3**	Yield of **5**	Yield of **6**
intercept	1.194E+00	2.582E+00	5.098E+00	1.154E+01
T	-4.249E-03	8.702E-03	7.601E-03	9.109E-03
$R(r)_{APEX}$	-7.937E-01	-7.125E-01	-2.218E+00	-2.647E+00
r_{APEX}	1.718E+00	6.999E-01	1.251E+00	-7.245E-01
FWHH	2.267E-01	-1.330E+00	-3.362E-02	-1.080E+00
SKEW	-1.438E-01	-2.387E-01	-9.461E-01	-1.117E+00
$R(r)_{APEX}^2$	-5.642E-01	-1.206E+00	-2.112E+00	-3.037E+00
r_{APEX}^2	1.549E+00	1.641E-01	4.266E-01	-1.941E+00
$FWHH^2$	9.627E-02	-9.498E-01	-4.007E-01	-1.243E+00
$SKEW^2$	-1.635E-03	5.173E-03	-1.734E-01	-2.139E-01
$T*R(r)_{APEX}$	1.173E-02	1.259E-03	1.890E-02	-4.165E-03
$T*r_{APEX}$	-1.715E-02	1.207E-02	-1.606E-02	4.221E-03
T*FWHH	-1.560E-02	2.515E-03	-1.247E-02	2.234E-03
T*SKEW	5.348E-03	1.114E-03	7.157E-03	-1.761E-03

As an example, the full polynomial describing the log transformed yield of **2** is as follows:

$$y^* = 1.495 - 4.249\cdot10^{-3}*T - 7.937\cdot10^{-1}*R(r)_{APEX} + 1.718*r_{APEX} + 2.267\cdot10^{-1}*FWHH - 1.438\cdot10^{-1}*SKEW$$
$$- 5.642\cdot10^{-1}*R(r)_{APEX}^2 + 1.549*r_{APEX}^2 + 9.627\cdot10^{-2}*FWHH^2 - 1.635\cdot10^{-3}*SKEW^2 + 1.173\cdot10^{-2}*T*R(r)_{APEX}$$
$$- 1.715\cdot10^{-2}*T*r_{APEX} - 1.560\cdot10^{-2}*T*FWHH + 5.348\cdot10^{-3}*T*SKEW$$

And the yield of **2** = $-1 + 10^{y^*}$

Using the tabulated descriptor values from the main text the yield values according to the model can be reproduced using above equations. Note that all terms are present in the models describing the four yields. This is because the equations are all derived from the same model. If we would model each of the yields separately, a slightly different equation would be obtained in each case. To facilitate comparison between calculated and experimental values, the values are given in Table S4.

Table S4 Overview of all experimental and calculated yield values in raw and transformed scale

		Observed yield of				Predicted yield of				Observed yield of (transformed)				Predicted yield of (transformed)			
Metal	**T**	**2**	**3**	**5**	**6**	**2**	**3**	**5**	**6**	**2**	**3**	**5**	**6**	**2**	**3**	**5**	**6**
Au	80	11.0	7.2	8.4	1.5	24.2	4.7	8.1	2.1	1.08	0.91	0.97	0.40	1.40	0.76	0.96	0.49
Au	100	9.0	13.8	10.1	2.7	11.6	16.5	13.8	3.4	1.00	1.17	1.04	0.56	1.10	1.24	1.17	0.65
Au	120	5.7	40.2	18.0	4.1	5.3	52.5	23.0	5.4	0.83	1.61	1.28	0.71	0.80	1.73	1.38	0.81
Cu	80	7.2	1.1	0.4	0.0	6.6	1.4	0.2	-0.2	0.91	0.32	0.13	0.00	0.88	0.38	0.09	-0.07
Cu	100	6.2	6.2	2.2	0.0	5.5	5.3	2.6	0.1	0.85	0.85	0.51	0.00	0.81	0.80	0.55	0.03
Cu	120	4.6	16.5	10.3	0.6	4.6	15.6	9.3	0.4	0.75	1.24	1.05	0.20	0.75	1.22	1.01	0.13
Ir	80	67.5	3.8	11.4	1.0	48.6	3.4	10.0	0.3	1.84	0.68	1.09	0.29	1.70	0.64	1.04	0.12
Ir	100	18.0	17.3	13.0	1.4	18.3	13.6	12.9	1.0	1.28	1.26	1.14	0.39	1.29	1.16	1.14	0.31
Ir	120	6.3	41.9	17.4	3.2	6.5	47.3	16.6	2.1	0.86	1.63	1.26	0.62	0.88	1.68	1.24	0.49
Ni	80	9.7	2.7	0.5	0.0	11.3	2.5	0.5	0.0	1.03	0.57	0.19	0.00	1.09	0.54	0.18	-0.01
Ni	100	8.5	8.9	2.4	0.0	8.7	8.5	3.0	0.3	0.98	1.00	0.53	0.00	0.99	0.98	0.60	0.10
Ni	120	7.0	21.1	10.3	0.5	6.7	24.9	9.5	0.6	0.90	1.34	1.05	0.19	0.88	1.41	1.02	0.21
Pt	80	60.0	2.6	11.3	1.7	29.7	4.7	9.7	1.9	1.79	0.56	1.09	0.44	1.49	0.76	1.03	0.46
Pt	100	21.8	20.3	15.6	2.3	13.2	16.8	15.1	3.3	1.36	1.33	1.22	0.52	1.15	1.25	1.21	0.63
Pt	120	2.9	64.7	27.4	5.9	5.6	54.9	23.0	5.3	0.60	1.82	1.45	0.84	0.82	1.75	1.38	0.80
Rh	80	51.0	2.0	8.2	2.0	30.2	2.7	7.9	1.1	1.72	0.48	0.96	0.48	1.49	0.57	0.95	0.32
Rh	100	18.1	13.4	12.1	4.0	12.5	10.8	11.4	2.1	1.28	1.16	1.12	0.70	1.13	1.07	1.09	0.50
Rh	120	3.8	50.8	22.6	9.4	4.8	36.4	16.4	3.7	0.68	1.71	1.37	1.02	0.76	1.57	1.24	0.67
Ru	80	15.0	2.2	7.9	0.0	39.1	2.1	8.3	0.4	1.20	0.51	0.95	0.00	1.60	0.49	0.97	0.13
Ru	100	11.3	8.0	7.5	0.0	14.5	9.2	10.7	1.1	1.09	0.95	0.93	0.00	1.19	1.01	1.07	0.32
Ru	120	8.2	25.2	12.9	0.7	5.0	32.3	13.7	2.2	0.96	1.42	1.14	0.23	0.78	1.52	1.17	0.50

Additional references

1. S. Wold, H. Antti, F. Lindgren, and J. Öhman, *Chemometr. Intell. Lab.*, 1998, **44**, 175-185.

2. K. Pöllänen, A. Häkkinen, S.P. Reinikainen, M. Louhi-Kultanen, and L. Nyström, *Chemometr. Intell. Lab.*, 2005, **76**, 25-35.

3. J. Gottfries, E. Johansson, and J. Trygg, *J. Chemometrics*, 2008, **22**, 565-570.

4. J. Gabrielsson, H. Jonsson, C. Airiau, B. Schmidt, R. Escott, and J. Trygg, *Chemometr. Intell. Lab.*, 2006, **84**, 153-158.

5. M. Bylesjö, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, and J. Trygg, *J. Chemometrics*, 2006, **20**, 341-351.

6. J. Trygg, and S. Wold, *J. Chemometrics*, 2002, **16**, 119-128.

7. S. Karlin, and W.J. Studden, *Annals of Mathematical Statistics*, 1966, **37**, 783-815.