Supporting information for "Predicting with confidence the efficiency of new dyes in dye sensitized solar cells"

Chung Man Ip[†], Antonio Eleuteri[‡], Alessandro Troisi[†]

[†]*Department of Chemistry and Centre for Scientific Computing, University of Warwick, UK*
[‡]*Department of Medical Physics and Clinical Engineering, Royal Liverpool and Broadgreen University Hospital Trusts, Liverpool, UK*

1. Computation of free energy of oxidation and reorganization energy

Free energy of oxidation and reorganization energy were approximated from the total energy differences computed using the B3LYP functional[1] and the 3-21G* basis set (all structures were optimized at this level). Polarized Continuum Model (PCM)[2] was included to mimic solvent effects in DSSCs, using the solvent parameters appropriate for acetonitrile. The model was built with a solvent excluding surface (SES), where the overlapping index between two interlocking spheres was 0.8 and the minimum radius was 0.5 Å. Construction of SES was based on the GePol method and the set of atomic radii was defined according to the UAKS model.[3] All quantum chemical calculations of this work were performed with Gaussian03.[4] The computational approach to calculate ΔG and λ was identical to that adopted by Maggio et. al. [5]. The largest dye considered (Dye 50) has 699 basis functions and the set of calculation cannot be performed on a semi-automatic fashion with a much larger basis set due to very slow convergence of the computational procedure. For this reason the 7 dyes from the original data set with more than 760 basis function (at the 3-21G* level) have not been included in the analysis. We have also excluded 2 dyes from the original work because they did not have a single carboxylic anchoring group and the single dye with efficiency 0.05% (clearly an outlier).

2. Computation of the surface coverage and the surface dipole density

All dyes have the same (carboxylic) anchoring group and are assumed to adopt the same orientation on the anatase[100] surface as the model compound benzoic acid investigated in ref.[6] (the adsorption mode is non-dissociative molecular mono-dentate). We have therefore rotated the optimized geometry of each dye into the orientation of the benzoic acid on anatase so that the plane of the C-COO group of the reoriented molecule was the same as in the benzoic acid and the C-COO bond was pointing in the same direction as the reference. The anatase[100] plane in the reference structure was perpendicular to the *z* Cartesian axis so that the molecular

electrical dipole moment of the rotated molecule in the *z* direction was used to establish any correlation between surface dipole and efficiency. The area occupied by the molecule in the *xy* plane was estimated by projecting the atomic coordinates in the *xy* plane and assuming that each atom span a circle of radius 1.5 Å.

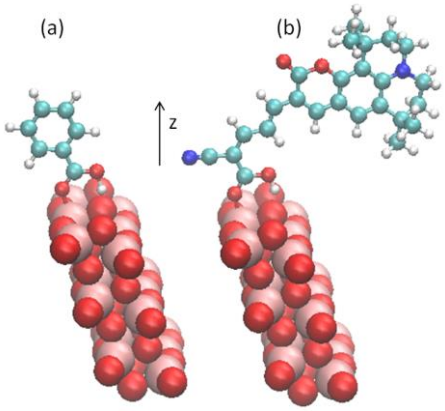

Figure S1. (a) Structure of the optimized model anchoring molecule on anatase(100) and (b) assumed structure of dye 4 on the surface for computing the surface coverage and electrical dipole moment perpendicular to the surface.

3. Computation of the absorption spectrum and its overlap with solar spectrum

The excitation energies and their corresponding oscillation strength to compute adsorption spectrum were obtained by single-point TDDFT calculations with 6-31G* basis set, B3LYP functional and PCM and solvent effects included as in the calculation of reorganization energy. The input geometries were the optimized geometries of neutral dyes from B3LYP/3-21G* calculations (we have performed few tests with full optimization at the 6-31G* level reported below). The number of excited states included in the calculation was set to 11 as the lowest energy of the 11th excited state is typically closer to 4 eV with the lowest energy among all dyes being 3.5 eV. The calculation of the dye absorption above 3.5 eV is however not very important for DSSCs because the optical band gap of TiO2 is just above 3 eV. [7] The solar spectrum employed was the AM 1.5 direct normal plus circumsolar spectrum taken from ASTM G173-03.[8] The simulated absorption spectrum of dye k, $\varepsilon_k(E)$, was computed by:

$$\varepsilon(E) = \left(2\pi\sigma^2\right)^{-1/2} \sum_i f_i \exp(-(E_i - E)^2 / 2\sigma^2)$$

where E_i and f_i were the excitation energy and the dimensionless oscillator strength for i-th transition respectively, σ was a broadening parameter of 0.2 eV. The integral for computing $\tilde{S}_k$ was evaluated numerically in the range between 0 and 10.305 eV.

4. Computation of the orbital asymmetry

OA was the Log of the ratio of the orbital density of LUMO (OD_{LUMO}) to the orbital density of HOMO (OD_{HOMO}) on the anchoring group. The OD of the molecular orbitals was computed by:

$$OD_{MO} = \sum_{i(anchor),\, j} c_{i,MO} c_{j,MO} S_{ij}$$

where S_{ij} is the overlap between basis set function *i* and *j*, $c_{i,MO}$ are the molecular orbital coefficients (*MO* is either LUMO or HOMO of the dye *k*,). *i* ranges over all basis functions on the anchoring group of the molecule and *j* ranges over all basis functions of the molecule. The calculation on the neutral dye was used to obtain the molecular orbital coefficients.

5. On the choice of alternative predictors

With 52 data points it is not advisable to increase the number of predictors as this would generate only an apparently better fit (overfitting). There are several alternatives to the parameters considered here and we excluded those that correlate very strongly with the chosen predictors. For example (i) the HOMO-LUMO gap correlates strongly with the absorption spectrum, (ii) the volume of the dye correlates strongly with the reorganization energy, (iii) the HOMO energy correlates strongly with the oxidation ΔG. Figure S2 illustrates these correlations and give the Person's and Spearman's correlation coefficients to illustrate this point quantitatively. We excluded properties that may affect charge injection rates (like electronic properties of the excited states) as the charge injection is the fastest process, rarely determining the efficiency. Other property, e.g. single-triplet splitting, may become important for metal containing dyes not considered here. We have deliberately excluded simpler topological descriptors used in drug discovery, e.g. number of aromatic rings, for which we are not aware of any physical basis for their relevance for the solar cell efficiency.

For comparison we give in Table S1 the correlation coefficients for the 5 predictors included in the initial model. The correlation between them is very weak, in particular between the predictors that turned out to be the most important, λ and ΔG. The largest, but still very weak, correlation is found between λ and *S*.

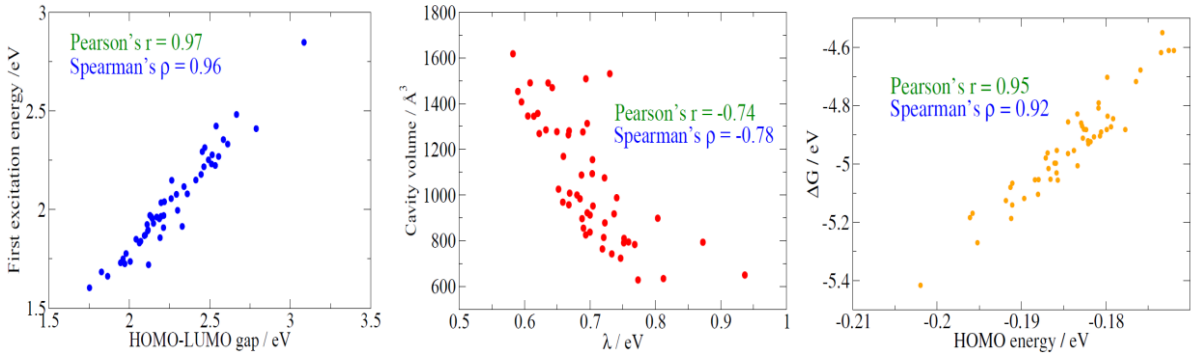

Figure S2. Left: Correlation between HOMO-LUMO gap and the first excitation energies of the dyes. Both quantities are obtained from single-point TDDFT calculations with B3LYP/6-31G* in PCM. Middle: Correlation between cavity volumes (the volumes of the cavities for chromophores generated in PCM) and reorganization energies, λ. The cavity volumes are acquired from the B3LYP/3-21G* optimization of neutral dyes and constructed with the Gepol method. Right: Correlation between ΔG and HOMO energies. The strong correlation between these pairs of properties is reflected by the high Pearson's r and Spearman's ρ statistics.

Table S1. Pearson's r and Spearman's ρ correlation coefficients between the five predictors used in the main statistical analysis. The Pearson's (Spearman's) coefficients are in higher (lower) triangle of the square table.

	λ / eV	ΔG / eV	S	NDD / D/Å²	OA
λ / eV	-	–0.20	–0.60	0.25	0.05
ΔG / eV	–0.10	-	0.36	–0.04	0.03
S	–0.50	0.25	-	–0.17	–0.36
NDD / D/Å²	0.36	0.04	–0.24	-	0.22
OA	–0.23	0.16	–0.01	0.13	-

6. Raw data for the analysis and their distribution

To allow alternative analysis of the data to the interested reader we report the raw data used for our analysis in Table S2. Distribution of the experimental efficiencies and the main predictors λ and ΔG are given in Figure S3.

Table S2. The parameters used for the fitting. The first column reports the dye index as given in Table 1 of ref.7 in the main text and the second column the experimental efficiency. The columns 3-7 contain the computational data used as predictors for the model.

Dye	η / %	ΔG / eV	λ / eV	S	NDD / D/Å^2	OA
1	6.8	-4.86	0.73	2.23	0.26	0.13
2	6.6	-4.83	0.72	2.30	0.22	0.50
3	5.9	-4.70	0.76	1.55	0.18	0.44
4	6	-5.18	0.75	2.02	0.07	0.22
5	7.2	-4.96	0.69	2.25	0.34	0.50
6	7.7	-4.95	0.70	2.45	0.36	0.68
7	8.2	-4.87	0.68	2.72	0.15	-0.35
8	6.5	-5.08	0.67	2.97	0.37	0.74
10	4.4	-5.00	0.69	2.85	0.12	0.69
11	4.5	-4.81	0.75	1.95	0.32	0.68
12	2.9	-4.86	0.75	2.02	0.28	0.70
13	2.3	-4.62	0.70	2.64	0.34	0.79
14a	8	-4.90	0.61	2.63	0.09	1.06
14b	4.8	-4.87	0.60	3.15	0.10	1.07
15a	7.7	-4.91	0.62	2.40	0.08	0.76
15b	5.4	-4.89	0.61	2.40	0.07	0.70
17	6.7	-4.88	0.69	0.71	0.05	4.61
18a	7.4	-4.93	0.63	2.18	0.07	0.93
18b	5.5	-4.88	0.61	2.70	0.07	0.94
19a	5.2	-5.27	0.70	2.12	0.18	0.55
19b	3.8	-5.14	0.70	2.53	0.17	0.50
20	3	-4.68	0.72	2.00	0.24	0.61
21	4.5	-5.13	0.70	2.02	0.05	0.99
22	8	-4.92	0.62	2.65	0.08	1.06
23a	5.2	-5.01	0.66	1.90	0.10	1.22
24	3.4	-5.12	0.74	2.77	0.07	0.26
26	2.5	-5.42	0.81	0.96	0.19	0.34
27	5.2	-5.19	0.72	1.43	0.18	0.63
28	7.3	-5.05	0.72	1.55	0.18	0.46
29	9.1	-5.06	0.70	1.99	0.18	0.97
30	5.9	-5.05	0.69	1.94	0.17	0.60
31	6.9	-4.79	0.74	2.01	0.17	0.68
32	6.2	-5.03	0.69	2.11	0.18	0.91
33	7	-4.85	0.65	2.66	0.19	0.90

34	6.6	-4.92	0.66	2.77	0.23	0.75
35	2.3	-4.88	0.77	1.89	0.33	0.43
37	3.8	-5.10	0.67	1.73	0.18	0.61
38	1.2	-4.88	0.70	1.82	0.21	0.28
39	6.2	-4.61	0.80	2.52	-0.04	1.49
40	3.9	-4.61	0.68	3.26	-0.09	-1.07
41	5.5	-5.17	0.94	0.61	0.18	0.60
42	1.9	-5.07	0.87	1.17	0.02	-0.24
43	6.1	-5.00	0.69	2.34	0.01	-0.69
44	9	-4.96	0.65	2.67	0.00	-0.44
45	5.5	-5.02	0.67	2.71	-0.01	-0.49
46	9.5	-4.91	0.64	2.78	0.01	-0.36
47	5.1	-4.98	0.77	1.57	0.15	-0.90
48	5.4	-4.95	0.67	1.65	0.07	0.56
49	7.2	-4.72	0.59	1.71	0.10	0.80
50	6	-4.84	0.58	3.27	-0.03	-3.63
51	2.9	-4.55	0.73	2.58	0.08	0.95
52	6	-5.05	0.64	2.12	-0.07	-0.13

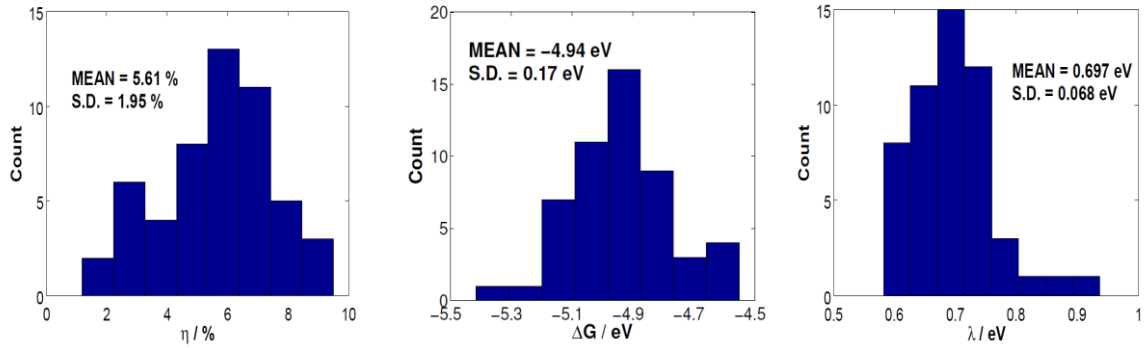

Figure S3. Data distribution of η (%), ΔG (eV) and λ (eV) in the data set.

7. Results of the simple fitting

The results of fitting the data with the expression $\eta_{\text{exp}} = a + b\Delta G + c\Delta G^2 + d\lambda + e\lambda^2$ are given in Table S3, including the 95% confidence interval in the fitting parameters. The interval is reported for completeness but it is not very meaningful considering the nature of the data, i.e. it is known that additional effects beyond ΔG and λ contribute to the coupling. A better way to

evaluate the quality of the fitting is to consider the distribution of the difference between actual efficiency and efficiency computed by the fitting above (given in figure 2b of the main manuscript, the standard deviation was 1.67 %).

Table S3. Results of polynomial fitting.

Parameter	Fitted value	95% confidence interval	Units
A	-290.1	-585.6 : 5.32	%
B	-125.9	-243.0 : -8.77	% eV^{-1}
C	-12.67	-24.54 : -0.804	% eV^{-1}
D	-37.22	-128.28 : 53.8	% eV^{-2}
E	19.32	-42.64 : 81.38	% eV^{-2}

8. A general linear regression model

We have preliminarily constructed a quantile plot (with respect to the Gaussian distribution) of the response variable η (Figure S4) showing that its distribution can be reasonably approximated by a Gaussian distribution, as implied in the simplified model.

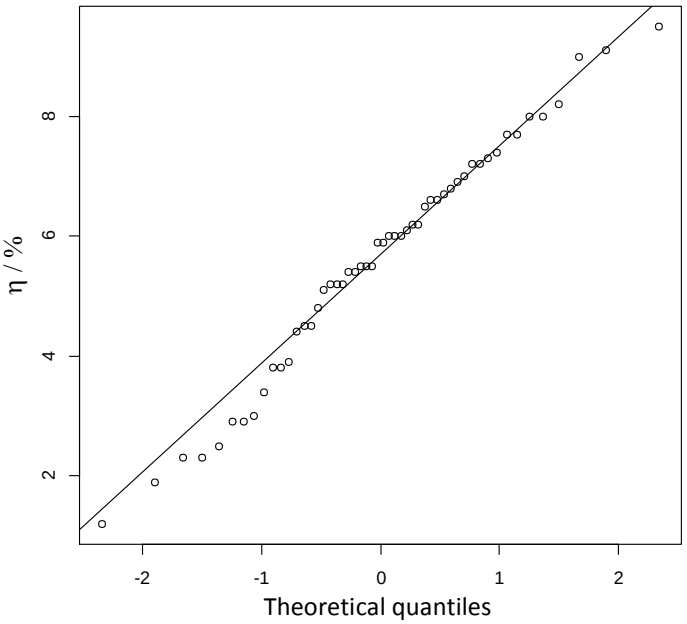

Figure S4. Q-Q plot of the variable η with respect to the Gaussian distribution.

Based on this graph, we will consider satisfying the approximation of the distribution of η by a Gaussian distribution. In a generalized linear model we indicate with **X** the vector of the predictors and possible non-linear functions of them (in our case for example **X**= $(\Delta G, \lambda, S, DD, OA, g_1(\Delta G), g_2(\lambda), \ldots)$). The expectation value of the efficiency conditional on a particular set of predictors is expressed formally as

$$E[\eta \mid \mathbf{X}] = f(\mathbf{X}; \boldsymbol{\beta}) \qquad \text{(S1)}$$

where $\boldsymbol{\beta}$ is a parameter vector to be estimated and f is a nonlinear additive function of the predictors, i.e. $f = g_1(X_1; \beta_1) + g_2(X_2; \beta_2) + \ldots$. However, each of the g_k is a linear function of β_k , so overall *f* is a linear function of $\boldsymbol{\beta}$.

A simple linear regression model (i.e. when g is the identity function and **X** the vector of the five linear predictors) gives rather poor results, as the calibration graph (obtained by bootstrap resampling[9]) illustrated in Figure S5 shows.

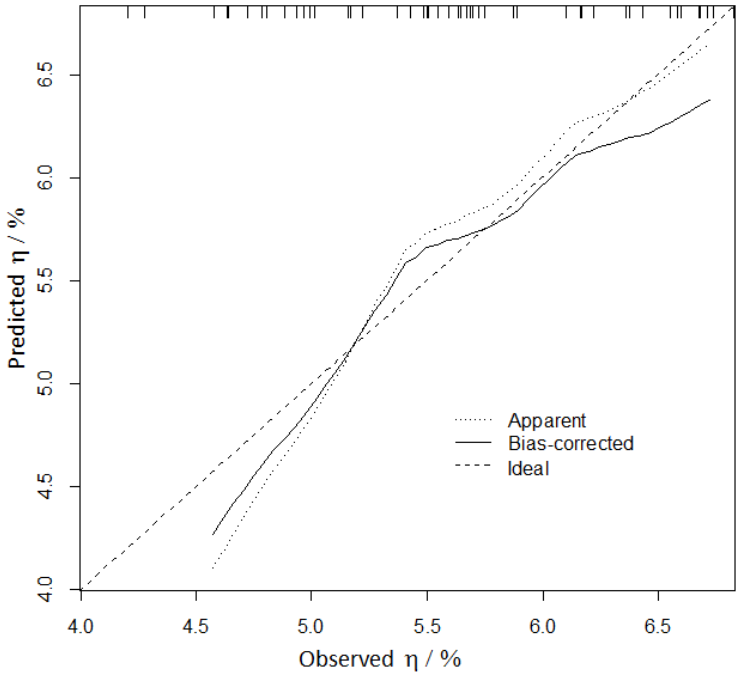

Figure S5. Predicted vs. observed η for a simple linear regression. One hundred bootstrapped predictions of the model were averaged. Apparent and bias-corrected estimates are reported. The mean absolute error for the bias-corrected estimates is 0.149. A perfect fit would lie on the dashed line. The tags on the

upper part of the graph denote the coordinates of the predicted points (the graphs are obtained by smoothing).

The bootstrapped shrinkage estimate is 0.75,[9] which denotes rather poor validation performance (i.e. about 25% lack of fitting).

The poor fit might be explained with effective lack of complexity (as the above model is linear in the data). We will therefore consider a more complex model; given the information provided by the Spearman ρ^2 statistic (figure 1(e) in main manuscript) and the limited sample size, we will introduce nonlinearities for ΔG and λ, but we will also consider penalisation. In the intuitive model we have simply introduced higher powers of ΔG and λ in the fitting function but this is far from ideal when there is no physical reason for such an expansion; furthermore, polynomials have some undesirable properties, most notably non-locality (the fit in one region can be greatly affected by data in other regions;) finally, polynomials tend to infinity quite rapidly when predicting out of the range of the data used for the fit.

A more general approach is to construct restricted (or *natural*) cubic regression splines where the nonlinear function takes the form (in this example it is a third order spline)

$$g(X;\boldsymbol{\beta}) = \beta_0 + \beta_{\text{lin}} X_1 + \beta_{\text{nonlin}} X_2 \qquad \text{(S2)}$$

where $X_1 = X$,

$$X_2 = (X - t_1)_+^3 - \frac{t_3 - t_1}{t_3 - t_2}(X - t_2)_+^3 + \frac{t_2 - t_1}{t_3 - t_2}(X - t_3)_+^3 \qquad \text{(S3)}$$

and $(z)_+$ is equal to z if $z>0$ and zero otherwise. Note $\boldsymbol{\beta} = (\beta_0, \beta_{\text{lin}}, \beta_{\text{nonlin}})$ is a parameter vector.

Equation S2 is known as a third order restricted cubic spline (higher order can be defined but are not used here). Note that constant and linear terms are included in the expansion, and that for $X \geq t_3$ the function is linear. The parameters t_1, t_2, t_3 are known as the *knots* of the spline

and are determined uniquely for the data set, being located respectively at the 0.1, 0.5 and 0.9 quantile of the data for the predictor X. For the predictor ΔG, the function $g_{\Delta G}(\Delta G)$ has nodes $t_{1,\Delta G} = $ -5.1685, $t_{2,\Delta G} = $ -4.9277, and $t_{3,\Delta G} = $ -4.6784; for the predictor λ the function $g_{\lambda}(\lambda)$ has nodes $t_{1,\lambda} = 0.6086$, $t_{2,\lambda} = 0.6936$ and $t_{3,\lambda} = 0.7734$ (the units are eV).

A general model containing all five predictors and the non-linearity in ΔG and λ will have the following form:

$$E[\eta \mid \Delta G, \lambda, S, DD, OA]$$

$$= \beta_0 + g_{\Delta G}(\Delta G; \boldsymbol{\beta}_1) + g_{\lambda}(\lambda; \boldsymbol{\beta}_2) + \beta_3 S + \beta_4 DD + \beta_5 OA \qquad \text{(S4)}$$

The number of parameters in the above model is 8 (two each for the spline expansions; note that the intercept β_0 is common to both and is made explicit, so it is not counted twice to avoid identifiability issues). A widely accepted heuristic is to consider 1/20th to 1/10th of the sample size (52 in this case) as the upper bound for the number of parameters in the model. Violation of this limit can produce biased results and overfitting. However, it is possible to go beyond this limit by penalising model fitting criteria for complexity, as we describe below.

The common approach to fitting a model by maximisation of the likelihood function is equivalent to maximising the likelihood ratio (LR) χ^2 statistic of the model with respect to the "null" model (i.e. the model without any predictors; only β_0 is trivially "fit" to the average of the response.)

Akaike's information criterion (AIC) provides a method for penalising the LR achieved by a given model for its complexity to obtain a more unbiased assessment of the model's worth [9]. The AIC has the form

$$\mathrm{AIC} = \mathrm{LR}\ \chi^2 - 2p \qquad \text{(S5)}$$

where *p* is the number of parameters in the model. As can be seen from the formula, a model with large *p* will reduce the effective LR, so the optimal model will result from a trade-off between maximising LR and reducing *p*. It turns out the above criterion can still be biased when the sample size is small, so the following corrected AIC is used, which also takes into account the sample size *n*

$$\mathrm{AIC}_C = \mathrm{LR}\ \chi^2 - 2p\left(1+\frac{p+1}{n-p-1}\right) \qquad \text{(S6)}$$

From the formula we can see that a small sample size incurs a greater penalty than a large sample size; as the sample size tends to infinity $\mathrm{AIC}_C \underset{n\to\infty}{\to} \mathrm{AIC}$. We have used AIC_C as a model selection criterion.

A look at the ANOVA table of the penalised model (Table S4) shows that there is some predictive power in ΔG and λ, since they have the highest *F* statistics. Notice that the relation with λ effect seems to be mostly linear.

Table S4. ANOVA table of full model.

Factor	***degrees of freedom***	***F test statistic***	***p-value***
ΔG	2	1.39	0.26
nonlinear	1	2.57	0.12
λ	2	3.43	0.04
nonlinear	1	0.37	0.55
NDD	1	0.07	0.79
S	1	0.44	0.51
OA	1	0.24	0.62

Although it would be tempting to simplify the model and remove the remaining variables, it would also bias the results and produce optimistic estimates of the parameters' covariance

matrix (since we have "cheated" looking at a more complex hypothesis, and have already expended degrees of freedom.) Instead, we proceed with the following approach: produce a new set of "estimated efficiency" as predicted by the model, and then regress this "new" response variable (let's call it Z) vs. the chosen subset of predictors. It can be shown [9] that this approach of "simplification by approximation" produces a final model for which confidence limits and statistical tests are unbiased and include the effects of model selection. The simplified model can be written as

$$E[Z \mid \Delta G, \lambda] = \beta_0 + g_{\Delta G}(\Delta G; \boldsymbol{\beta}_1) + g_{\lambda}(\lambda; \boldsymbol{\beta}_2) \qquad \text{(S7)}$$

The calibration graph in figure S6 shows the good performance of the reduced model. The bootstrapped shrinkage estimate is 0.99 and denotes good validation performance (i.e. about 1% lack of fitting). The ANOVA table of the reduced model and its final parameters are given Tables S5 and S6.

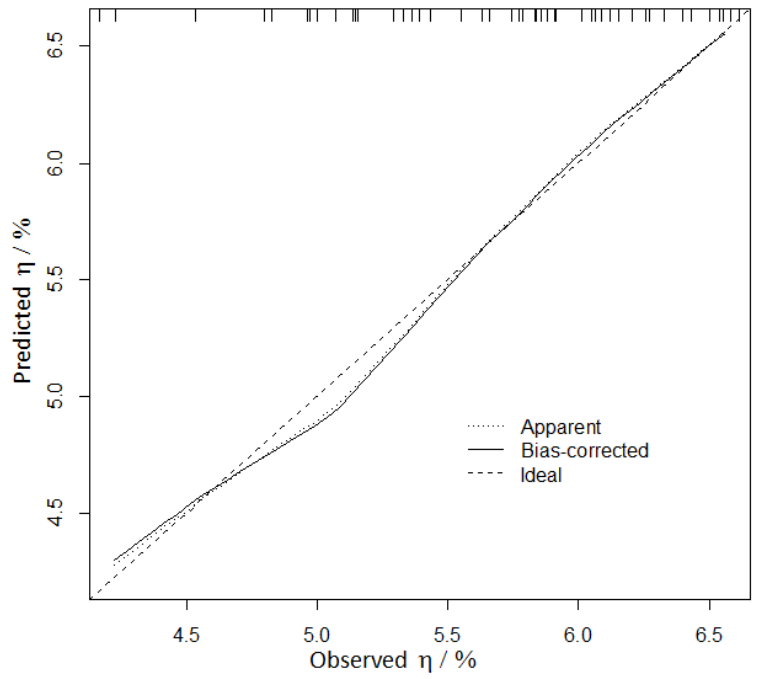

Figure S6. Predicted vs. observed estimated efficiency *Z*. One hundred bootstrapped predictions of the model were averaged. Apparent and bias-corrected estimates are reported. The mean absolute error for the bias-corrected estimates is 0.044. A perfect fit would lie on the dashed line. The tags on the upper part of the graph denote the coordinates of the predicted points (the graphs are obtained by smoothing.)

Table S5. ANOVA table of reduced model.

Factor	***degrees of freedom***	***F test statistic***	***p-value***
ΔG	2	42.48	<0.0001
nonlinear	1	48.86	<0.0001
λ	2	141.04	<0.0001
nonlinear	1	4.32	<0.0001

Table S6. Estimated coefficients of general linear regression model fit (reduced model).

Parameter	Fitted value	95% confidence interval	Units
β_0	20.26	15.54:24.99	%
$\beta_{1,\mathrm{lin}}$	1.531	0.576:2.486	% eV^{-1}
$\beta_{1,\mathrm{nonlin}}$	-3.699	-4.860:-2.537	% eV^{-3}
$\beta_{2,\mathrm{lin}}$	-9.613	-12.373:-6.853	% eV^{-1}
$\beta_{2,\mathrm{nonlin}}$	-0.036	-2.556:2.483	% eV^{-3}

In conclusion, this more rigorous procedure for the regression of η vs. a set of predictors shows that the ones with the greatest predictive capability are ΔG (with a strong nonlinear component) and λ (mostly linear).

References

[1] a) A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648-5652; b) P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, *J. Phys. Chem.* **1994**, *98*, 11623-11627.

[2] J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev.* **2005**, *105*, 2999-3093.

[3] a) J. L. Pascualahuir, E. Silla, *J. Comput. Chem.* **1990**, *11*, 1047-1060; b) E. Silla, I. Tunon, J. L. Pascualahuir, *J. Comput. Chem.* **1991**, *12*, 1077-1088.

[4] Gaussian 03, Revision C.02, M. J. Frisch, e. al.

[5] E. Maggio, N. Martsinovich, A. Troisi, *Angew. Chem.-Int. Edit.* **2013**, *52*, 973-975.

[6] N. Martsinovich, A. Troisi, *J. Phys. Chem. C* **2011**, *115*, 11781-11792.

[7] H. Tang, K. Prasad, R. Sanjinès, P. E. Schmid, F. Lévy, *J. Appl. Phys.* **1994**, *75*, 2042-2047.

[8] A. International.

[9] F. E. Harrel Jr., *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer, New York, **2006**.