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 Physics, University of Liverpool and Department of Medical Physics and Clinical Engineering, Royal Liverpool and Broadgreen University Hospital Trusts, Liverpool, UK*

Abstract

We ask whether it is possible to predict the efficiency of a new dye in dye sensitizes solar cells (DSSC) on the basis of the known performance of existing dyes in the same type of device. We evaluate a number of computable predictors of the efficiency for a large set of dyes whose experimental efficiency is known. We have then used statistical regression methods to establish the relation between the predictors and the efficiency. Our predictions are associated to a rigorously determined confidence level. For a new dye of the same family we are able to predict the probability that its efficiency in a DSSC is larger than a certain threshold. This method is useful for accelerating the discovery of new dyes and establishing more rigorously the existence of specific correlations between structure and property. Within the properties considered we find that the dye efficiency correlates more strongly with its oxidation potential and reorganization energy.

In the development of new, more efficient dye sensitized solar cells (DSSC)[1-3] an important fraction of the research effort is devoted to the synthesis and testing of new dyes. The very few design rules[4-7] emerged over the past years have guided the exploration of a large set of dyes that, when tested under standardized conditions and fabrication methods, should inform the development of new and better dyes. In this paper we ask whether it is possible to predict the efficiency of a new dye on the basis of the known performance of existing dyes. In particular we want to establish the degree of confidence of such predictions.

In material science and physics it is very common to build models of a system under investigation starting from physical principles, and this type of predictive modelling has been part of the development of DSSC since the early days.[8-10] In other fields, like drug discovery, such modelling from first principle is often accompanied by statistical modelling where one looks for correlation between measurable or computable properties (the *predictors*) of a given molecule and a target property, like its pharmacological activity. The development of such quantitative structure activity relations (QSAR) is one of the main approaches currently used to rationalize large medicinal chemistry data set.[11-13] The identification of correlation or lack of correlation between properties can contribute to the understanding of the underlying physical principles for a given problem and, in any case, it can be used to narrow down the exploration of new drugs or materials, when synthesis and testing constitute the slower step.

To build a structure-property relation for dyes in DSSC, we need a sufficiently large database of dyes tested under similar conditions (e.g. same electrolyte, similar fabrication methods). In any convincing statistical analysis the data cannot be handpicked and it is also desirable that they are derived from a relatively uniform set, in this case, for example, a set of dyes with related chemical characteristics. To address both issues we considered 52 dyes listed in table 1 of the review by Mishra et. al., all being synthetic organic dyes tested in similar devices.[7] In this first application we did not include new dyes appeared after the review was published, to avoid the risk of involuntary bias. We have excluded the dyes from the review which did not have the common carboxylic anchoring group, or needed more than 760 basis functions for the electronic structure calculation (for them, a manual optimization outside our automatic procedure was needed). The dyes considered are a fairly representative of the chemical structures used for organic dyes and the experimental efficiencies are very broadly distributed (average 5.61% and standard deviation 1.95%, see also SI), suggesting that the data set is not biased toward high performing dyes.

We aim to find some correlations between the properties of the dyes that can be accessed very easily via routine quantum chemistry calculations and experimental solar cell efficiency η. We can then compute these properties for a new dye and predict the probability that its efficiency in a DSSC is larger than a given threshold. The calculations should be relatively inexpensive so that all the calculations can

be performed semi-automatically for all dyes considered, terminating successfully without user intervention. More importantly, such procedure is useful only if many new potential dyes can be screened rapidly after the statistical regression. Importantly, the systematic and random errors in these computed properties will be fully accounted for in a statistical analysis.

As the extremely broad range of available QSAR demonstrates, there is no best or conclusive way to select predictors to be included in such statistical analysis, and we expect that other improved selections of predictors may be suggested in the future. In this initial report we decided to include computable predictors that are sufficiently independent from one another, easy to evaluate, and expected to influence the efficiency of the device from physical considerations.[14] Importantly, it is not possible to increase the number of predictors for a given data set without risking an overfitting of the data and we follow the common rule-of-thumb of not having more than 1 fitting parameter per 10 data points. A list of the predictors considered in this analysis with a motivation and a brief description of the computational methods is given below (in the SI we provide a full description of the methodology, its justification, the complete data set and further explanation with additional analysis for preferring these predictors over others):

(1) Free energy of dye oxidation in solution, ΔG. This is clearly one of the most relevant parameters for the energetics of a DSSC[14, 15] and it was computed at the B3LYP/3-21G* level in the presence of a continuum model of acetonitrile solvent. [16]

(2) Reorganization energy for oxidation, λ. This parameter enters into the theory of interfacial electron transfer and it is essential in determining the rate of charge recombination to the oxidized dye.[14, 15] It was computed at the same level used for ΔG following the procedure given in ref. [15].

(3) Absorption of solar radiation, *S*. It is expected that a higher efficiency is associated with greater ability to absorb solar radiation. We have computed the absorption spectrum for each dye at the TDDFT/6-31G* level with the inclusion of solvent[16] and evaluated the overlap between the computed absorption spectrum for dye *k*, $\varepsilon_k(E)$, and the solar spectrum, $P(E)$, as $\tilde{S}_k = \int P(E)\varepsilon_k(E)dE$. To have convenient data we have normalized $\tilde{S}_k$ to the value of one of the dyes ($S_k = \tilde{S}_k / S_0$).

(4) Surface dipole density, *NDD*. The dipoles of the ground-state dyes are thought to affect the conduction band level of the semiconductor[17] (if the dyes are in a similar orientation with respect to the surface). We have assumed that the orientation of the dye is guided by the carboxylic anchoring group oriented on the surface as in a calculation of a benzoic acid on TiO_2.[18] We have therefore evaluated (i) the component of the dipole of the dye *k* perpendicular to the surface $\mu_{k,z}$ and (ii) the area of the same dye on the TiO_2 surface A_k. The normalized dipole density (*NDD*) for dye *k* is $NDD_k = \mu_{k,z} / A_k$.

(5) *Orbital asymmetry*, *OA*. A good fraction of dyes, often referred to as donor-pi-acceptor dyes,[4] are synthesized to have a large orbital density of the LUMO on the anchoring group and a small orbital density of the HOMO on the anchoring group (a carboxylic acid) so that charge injection is favoured and charge recombination is prevented. We have defined the quantity *OA* as the Log of the ratio between the orbital density of the LUMO and HOMO on carboxylic acid,[19] where a high *OA* is expected to be beneficial for the cell.

As there is expectedly little correlation between dye and open circuit voltage (determined mostly by the electrolyte and TiO_2 electronic structure) we try to establish a relationship between the five computed properties and the expected efficiency in the form of a function $\eta_{\exp}(\Delta G, \lambda, S, NDD, OA)$. A diagram of the measured efficiency against the computed parameters (figure 1(a-e)) immediately gives some useful indication. It seems that there is an important correlation between reorganization energy and dye efficiency, with higher efficiencies associated with smaller reorganization energy as suggested by phenomenological models.[20] A correlation is also evident between the computed ΔG and the efficiency, i.e. it seems that higher efficiencies are found in a *range* of ΔG as expected from microscopic theories (and also suggesting that ΔG will affect the efficiency non-linearly).[21] Maybe surprisingly, no correlation is evident in the plots of measured efficiency against *S*, *NDD* and *OA*. More quantitatively, Figure 1f shows the Spearman ρ^2 statistics[22] for each predictor and suggests a higher degree of correlation (potentially nonlinear and non-monotonic) between η and the predictors ΔG and λ.

We build a relationship between the expected efficiency and the predictors initially by using an intuitive approach, and then by considering a more rigorous statistical procedure. For the intuitive approach we simply ignore the role of *S*, *NDD* and *OA*, on the basis of the visual inspection of Fig.1(c-e), and construct the simplest 5-parameter non-linear function of ΔG and λ:

$$\eta_{\exp} = a + b\Delta G + c\Delta G^2 + d\lambda + e\lambda^2 \qquad (1)$$

where the parameters *a*, *b*, *c*, *d*, *e* can be uniquely determined to minimize the squared difference between expected and measured efficiency (the resulting function is illustrated in figure 2a – the fitting parameters are in the SI – together with the data points used for the regression). The residuals (difference between predicted and actual values) are normally distributed with standard deviation σ_n = 1.67 % (Fig. 2b), and therefore it is possible to predict the probability that, for a dye with a computed (ΔG, λ) pair, the efficiency is higher than a given threshold η':

$$P(\eta > \eta') = \left(2\pi\sigma_\eta^2\right)^{-1/2} \int_{\eta'}^{\infty} \exp\left(-\left(\eta - \eta_{\exp}\right)^2 \Big/ 2\sigma_\eta^2\right) d\eta \qquad (2)$$

Figure 2c shows the probability that the efficiency is higher than 7% as a function of computed ΔG and λ. Interestingly, there are "accessible" regions of the map with probability higher than 60% and lower than 20%, i.e. the map provides a very good tool for planning the synthesis of new dyes considering it takes only a few minutes to set up the calculation of ΔG and λ (and few hours for their execution on a standard desktop computer). To further characterize the significance of this prediction we can compare it with the prediction of the "null model" based only on the average and standard deviation of the experimental efficiencies, which would predict 24% probability of efficiency larger than 7% for any value of the predictors.

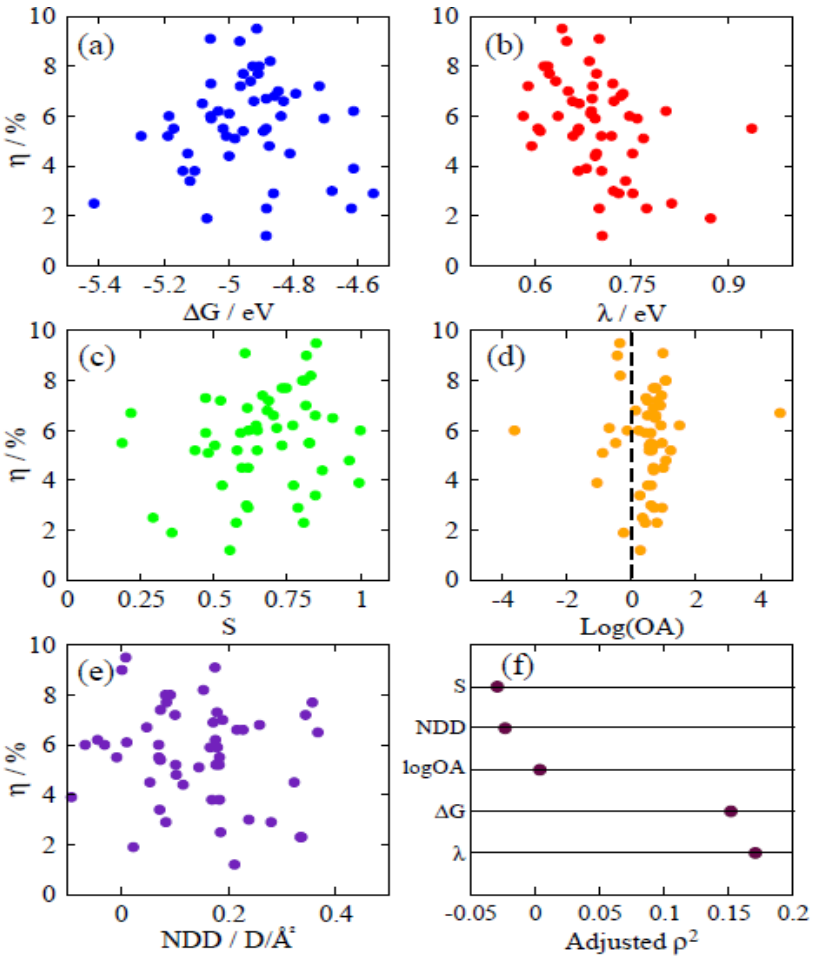

Figure 1. (a)-(e) experimental efficiency of the dyes in reference [7] against five computable parameters ΔG, λ, *S*, *NDD*, *OA*, defined in the text. (f) Strength of marginal relationship between predictors and efficiency using the Spearman ρ^2 statistic.

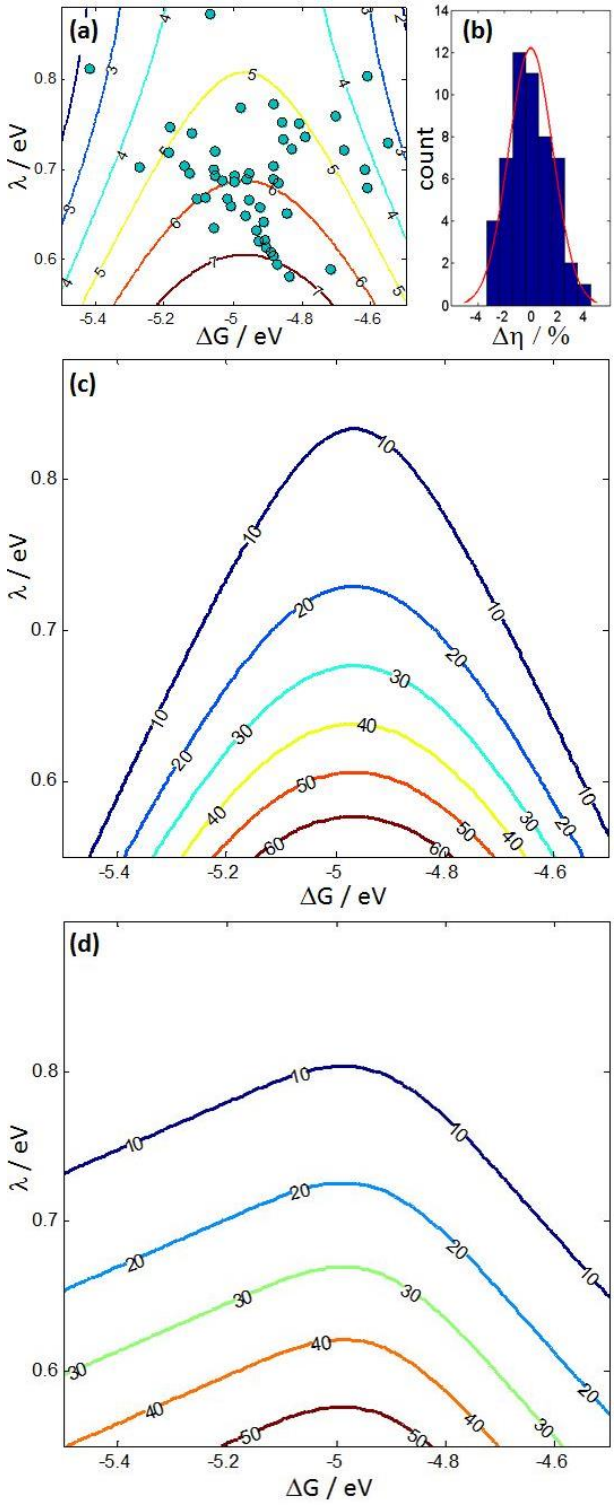

Figure 2. (a) Predicted efficiency (percent) from the fitting in eq. 2 with an indication of the data points included in the fitting procedure. (b) Distribution of the difference between "predicted" and actual efficiency values. (c) Map of the probability (percent) that the efficiency exceeds 7% as a function of the computed parameters ΔG, λ following the polynomial fit in eq. (1). (d) Same as (c) but using the more complex fitting function given in eq. (4).

A more rigorous procedure (described in detail in the SI) is based on the construction of a generalized linear model where the expected value of the efficiency is initially expressed as

$$\eta_{\exp} = \beta_0 + g_{\Delta G}(\Delta G; \beta_1) + g_{\lambda}(\lambda; \beta_2) + \beta_3 S + \beta_4 DD + \beta_5 OA \quad (3)$$

The expression above is linear in *S*, *DD*, and *OA* and contains linear and non-linear components in ΔG and λ (although the overall function will still be linear in all the parameters). In particular, the functions $g_{\Delta G}(\Delta G; \beta_1)$ and $g_{\lambda}(\lambda; \beta_2)$ expand ΔG and λ into restricted cubic splines with parameter vectors β_1, β_2 respectively.[22] The spline expansions are defined uniquely from the data for ΔG and λ. This is a well-established methodology to include non-linear terms in regression procedures where the analytical form of the non-linearity cannot be derived from a physical basis. Eq. 3 contains too many fitting parameters with respect to the 52 data points available and the initial fitting was therefore performed using a statistical penalization procedure.[22] The analysis of variance of the fitting confirms that there is no evidence of correlation between the predictors *S*, *NDD*, *OA* and the efficiency. A reduced model can be built from the total model (3) by using a procedure known as "simplification by approximation", [22] which produces the fitting as

$$\eta_{\exp} = \beta_0 + g_{\Delta G}(\Delta G; \beta_1) + g_{\lambda}(\lambda; \beta_2) \quad (4)$$

The standard deviation of the residuals for this more advanced model is 1.71% and, as before, it is possible to predict the probability that the efficiency is higher than a given threshold for any values of computed ΔG and λ. Figure 2(d) shows a map with the probability of efficiency higher than 7% with this more accurate model. The differences between the intuitive and the rigorous procedures are not large but the rigorous procedure guarantees that we have not neglected the effect of potentially more complex nonlinearities. Moreover the functional form in (4) gives more conservative estimates outside the region where data points are present (where the prediction cannot be trusted), while the polynomial fit of eq. (1) gives unphysical estimates in these regions. In the SI we report additionally the calibration graph of the model in (4) obtained by bootstrap resampling and the rationale for selecting the method based on Akaike's information criterion.[22]

The proposed map can be used to either direct the synthesis of new dyes where the maximum efficiency is predicted, or prepare dyes in the region of the map where there are few or no data points, to learn more about the system in these conditions. Considering that new families of DSSC are now being used, e.g. with different electrolytes,[23-25] we believe that the construction of a similar map should constitute a priority in the rational exploration of the chemical space. The experimental efficiencies in the data set considered here were broadly distributed but, as the field develops in time, there will be a tendency to report only high performing dyes (a problem noted in other contexts[26, 27]) making the statistical

analysis of literature data more complicated, because it should consider the selection bias in the reported data.[28] Sharing the data also on low performing devices would be of course the most desirable alternative.

It is also important to stress the difference between our approach (that looks for correlation between computable properties and a target experimental property) and alternative computational tools for material discovery that generate a large set of "theoretical" materials and directly compute the property of interest (e.g. the band gap[29, 30] or other electronic properties[31, 32]). The latter approach is particularly suitable when the underlying physics is relatively well understood[33] and the direct computation of the property of interest is possible. For DSSC it is currently not possible to compute the efficiency from first principles and a closer alliance between theory and experiment is therefore necessary.

Finally, such analysis in larger and unbiased data sets offers the best opportunity to validate some hypotheses put forward to describe the physics of DSSC. After considering the results, we are not too surprised that the overlap with the solar radiation does not correlate with the efficiency, possibly because cells with small absorptance are not even reported and beyond a threshold of absorptance the efficiency does not change. On the other hand, it is quite surprising to see that there is no effect in having HOMO and LUMO localized in different regions of the dye, considering the enormous effort put into the preparation of large families of donor-pi-acceptor dyes. We cannot exclude that designing donor-pi-acceptor dye is useful but we suggest that the actual benefits of this synthetic strategy can be properly assessed only by a thorough statistical analysis.

In conclusion, we have proposed a general method to predict DSSC efficiency for new dyes from easily computable quantities, including, for the first time, the degree of confidence of such predictions. We have considered carboxylated organic dyes studied with iodide/tridiodide electrolyte but the method can be applied to a different family of DSSCs and the accuracy of its prediction can be improved over time by expanding the set of data and/or the set of predictors.

Acknowledgment. This research was funded by EPRSC. Rupert Ting is thanked for testing the procedure for the calculation of the reorganization energy.

References.

1. A. Hagfeldt, G. Boschloo, L. Sun, L. Kloo and H. Pettersson, *Chem. Rev.*, 2010, **110**, 6595-6663.
2. B. E. Hardin, H. J. Snaith and M. D. McGehee, *Nat. Photonics*, 2012, **6**, 162-169.
3. Z. Ning, Y. Fu and H. Tian, *Energy Environ. Sci*, 2010, **3**, 1170-1181.
4. J. N. Clifford, E. Martinez-Ferrero, A. Viterisi and E. Palomares, *Chem. Soc. Rev.*, 2011, **40**, 1635-1646.
5. Y. Wu and W. Zhu, *Chem. Soc. Rev.*, 2013, **42**, 2039-2058.
6. W. Zeng, Y. Cao, Y. Bai, Y. Wang, Y. Shi, M. Zhang, F. Wang, C. Pan and P. Wang, *Chem. Mat.*, 2010, **22**, 1915-1925.
7. A. Mishra, M. K. R. Fischer and P. Baeuerle, *Angew. Chem.-Int. Edit.*, 2009, **48**, 2474-2499.
8. A. V. Akimov, A. J. Neukirch and O. V. Prezhdo, *Chem. Rev.*, 2013, **113**, 4496-4565.
9. M. Pastore, S. Fantacci and F. De Angelis, *J. Phys. Chem. C*, 2013, **117**, 3685-3700.
10. L. G. C. Rego and V. S. Batista, *J. Am. Chem. Soc.*, 2003, **125**, 7989-7997.
11. H. Gao, J. A. Katzenellenbogen, R. Garg and C. Hansch, *Chem. Rev.*, 1999, **99**, 723-744.
12. M. Karelson, V. S. Lobanov and A. R. Katritzky, *Chem. Rev.*, 1996, **96**, 1027-1043.
13. P. Gramatica, *QSAR Comb Sci*, 2007, **26**, 694-701.
14. J. Bisquert and A. Marcus, *Top. Curr. Chem.*, 2013, **DOI: 10.1007/128_2013_471**.
15. E. Maggio, N. Martsinovich and A. Troisi, *J. Phys. Chem. C*, 2012, **116**, 7638-7649.
16. J. Tomasi, B. Mennucci and R. Cammi, *Chem. Rev.*, 2005, **105**, 2999-3093.
17. S. Ruehle, M. Greenshtein, S.-G. Chen, A. Merson, H. Pizem, C. S. Sukenik, D. Cahen and A. Zaban, *J. Phys. Chem. B*, 2005, **109**, 18907-18913.
18. N. Martsinovich and A. Troisi, *J. Phys. Chem. C*, 2011, **115**, 11781-11792.
19. E. Maggio, N. Martsinovich and A. Troisi, *Angew. Chem.-Int. Edit.*, 2013, **52**, 973-975.
20. J. Bisquert and R. A. Marcus, *Top. Curr. Chem.*, 2014, DOI: 10.1007/1128_2013_1471.
21. E. Maggio and A. Troisi, *J. Phys. Chem. C*, 2013, **117**, 24196-24205.
22. F. E. Harrel Jr., *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer, New York, 2006.
23. W. Xiang, W. Huang, U. Bach and L. Spiccia, *Chem. Comm.*, 2013, **49**, 8997-8999.
24. W. Xiang, F. Huang, Y.-B. Cheng, U. Bach and L. Spiccia, *Energy Environ. Sci*, 2013, **6**, 121-127.
25. A. Yella, H.-W. Lee, H. N. Tsao, C. Yi, A. K. Chandiran, M. K. Nazeeruddin, E. W.-G. Diau, C.-Y. Yeh, S. M. Zakeeruddin and M. Graetzel, *Science*, 2011, **334**, 629-634.
26. K. Dwan, D. G. Altman, J. A. Arnaiz, J. Bloom, A.-W. Chan, E. Cronin, E. Decullier, P. J. Easterbrook, E. Von Elm, C. Gamble, D. Ghersi, J. P. A. Ioannidis, J. Simes and P. R. Williamson, *Plos One*, 2008, **3**.
27. J. P. T. Higgins, D. G. Altman, P. C. Gotzsche, P. Jueni, D. Moher, A. D. Oxman, J. Savovic, K. F. Schulz, L. Weeks, J. A. C. Sterne, G. Cochrane Bias Methods and G. Cochrane Stat Methods, *British Medical Journal*, 2011, **343**.
28. P. A. Puhani, *Journal of Economic Surveys*, 2000, **14**, 53-68.
29. R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sanchez-Carrera, L. Vogt and A. Aspuru-Guzik, *Energy Environ. Sci*, 2011, **4**, 4849-4861.
30. P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon and K. Rajan, *Comput. Mater. Sci.*, 2014, **83**, 185-195.
31. N. M. O'Boyle, C. M. Campbell and G. R. Hutchison, *J. Phys. Chem. C*, 2011, **115**, 16200-16210.
32. K. Yang, W. Setyawan, S. Wang, M. B. Nardelli and S. Curtarolo, *Nat. Mater.*, 2012, **11**, 614-619.
33. S. Curtarolo, D. Morgan, K. Persson, J. Rodgers and G. Ceder, *Phys. Rev. Lett.*, 2003, **91**.

Table of content graphics