



LJMU Research Online

Enoch, SJ

Mechanism-based QSAR modeling of skin sensitization

<http://researchonline.ljmu.ac.uk/2054/>

Article

Citation (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

Enoch, SJ (2015) Mechanism-based QSAR modeling of skin sensitization. Chemical Research in Toxicology, 28 (10). pp. 1975-1986. ISSN 1520-5010

LJMU has developed [LJMU Research Online](#) for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact researchonline@ljmu.ac.uk

<http://researchonline.ljmu.ac.uk/>

Mechanism-based QSAR modeling of skin sensitization

John C Dearden, Mark Hewitt, David W. Roberts, Steven Enoch, Philip Rowe, Katarzyna Przybylak, Daniel Vaughan-Williams, Megan Smith, Girinath G. Pillai, and Alan R. Katritzky

Chem. Res. Toxicol., **Just Accepted Manuscript** • DOI: 10.1021/acs.chemrestox.5b00197 • Publication Date (Web): 18 Sep 2015

Downloaded from <http://pubs.acs.org> on September 21, 2015

Just Accepted

"Just Accepted" manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides "Just Accepted" as a free service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. "Just Accepted" manuscripts appear in full in PDF format accompanied by an HTML abstract. "Just Accepted" manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are accessible to all readers and citable by the Digital Object Identifier (DOI®). "Just Accepted" is an optional service offered to authors. Therefore, the "Just Accepted" Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the "Just Accepted" Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these "Just Accepted" manuscripts.

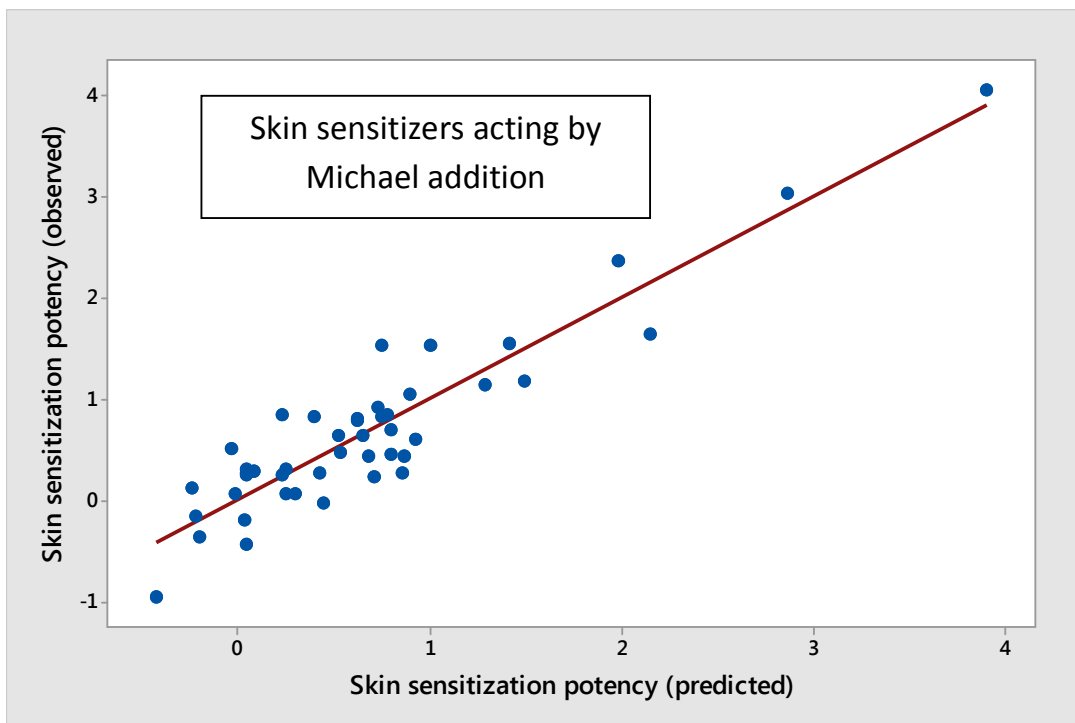


Mechanism-based QSAR modeling of skin sensitization**Byline: QSAR study of skin sensitization****Authors: John C. Dearden,^{*,†} Mark Hewitt,[‡] David W. Roberts,[†] Steven J. Enoch,[†]
Philip H. Rowe,[†] Katarzyna R. Przybylak,[†] G. Daniel Vaughan-Williams,[†] Megan L.
Smith,[†] Girinath G. Pillai,^{§,⊥} and Alan R. Katritzky[§]**[†]**School of Pharmacy & Biomolecular Sciences, Liverpool John Moores University,
Byrom Street, Liverpool L3 3AF, UK**[‡]**School of Pharmacy, University of Wolverhampton, Wulfruna Street, Wolverhampton
WV1 1LY, UK**[§]**Department of Chemistry, University of Florida, Gainesville, FL 32611-7200, USA**[⊥]**Institute of Chemistry, University of Tartu, 50411 Tartu, Estonia****Contact details****John Dearden:** Tel. +44-151-231-2066; Fax +44-151-231-2071; E-mail:
j.c.dearden@ljmu.ac.uk**Mark Hewitt:** E-mail: M.Hewitt@wlv.ac.uk**David Roberts:** E-mail: d.w.roberts@ljmu.ac.uk**Steven Enoch:** E-mail: s.j.enoch@ljmu.ac.uk**Philip Rowe:** E-mail: p.h.rowe@ljmu.ac.uk**Katarzyna Przybylak:** E-mail: k.r.przybylak@ljmu.ac.uk**Daniel Vaughan-Williams:** danielvaughanwilliams@gmail.com**Megan Smith:** E-mail: megan.smith@student.reading.ac.uk**Girinath Pillai:** giribio@mail.com**Alan Katritzky:** deceased

1

2

Table of Contents Graphic



3

Mechanism-based QSAR modeling of skin sensitization

J.C. Dearden¹, M. Hewitt², D.W. Roberts¹, S.J. Enoch¹, P.H. Rowe¹, K.R. Przybylak¹,
G.D. Vaughan-Williams¹, M.L. Smith¹, G. G. Pillai^{3,4} and A.R. Katritzky³

¹School of Pharmacy & Biomolecular Sciences, Liverpool John Moores University,
Byrom Street, Liverpool L3 3AF, UK; ²School of Pharmacy, University of
Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK; ³Department of
Chemistry, University of Florida, Gainesville, FL 32611-7200, USA; ⁴Institute of
Chemistry, University of Tartu, 50411 Tartu, Estonia

Dedication

Professor Alan Katritzky passed away on 10 February 2014. We dedicate this paper to his memory.

Abstract

Many chemicals can induce skin sensitization, and there is a pressing need for non-animal methods to give a quantitative indication of potency. Using two large published data-sets of skin sensitizers, we have allocated each sensitizing chemical to one of ten mechanistic categories, and then developed good QSAR models for the seven categories with a sufficient number of chemicals to allow modeling. Both internal and external validation checks showed that each model had good predictivity.

1

2 **Introduction**

3 Skin sensitization (allergic contact dermatitis) is a common problem arising from the contact
4 of certain chemicals with the skin. Once sensitized, an individual remains so for life, and it is
5 therefore important to know whether or not a chemical possesses skin sensitization potential
6 before skin contact is made.

7 In order for skin sensitization to be induced, a chemical must first penetrate into the viable
8 epidermis and bind to skin proteins/peptides to form an immunogenic complex.¹ The binding
9 is almost always covalent, with the chemical (hapten) acting as an electrophile and the
10 protein as nucleophile; a few haptens operate *via* a free radical mechanism.² The
11 immunogenic complex is taken up by dendritic cells, which convert the complex into a form
12 that can be recognized by T-cells, causing their stimulation and proliferation, and the
13 formation of so-called memory T-cells; this is the induction process.³ Upon re-exposure, the
14 memory T-cells release cytotoxic mediators that cause local tissue inflammation.

15 A number of methods are available for the determination of skin sensitization potential; the
16 current method of choice, and the one initially required for regulatory purposes⁴ is the
17 LLNA,^{5,6} which yields a quantitative endpoint. Much work has also been done on *in silico*
18 prediction of skin sensitization potential, in order to reduce animal usage and save time; this
19 has become more important with the advent of the recent REACH legislation,^{7,8} which
20 requires assessment of toxicity for all chemicals produced in or imported into the European
21 Union at levels above 1 tonne per annum, but which also requires animal testing to be carried
22 out only as a last resort.⁹

23 Despite the LLNA's having a quantitative endpoint, most *in silico* prediction studies of skin
24 sensitization to date have been categorical (i.e. sensitizer/non-sensitizer),¹⁰ as have most other

1 attempts to use biological assays. A small number have used classical QSAR regression to
2 model the LLNA endpoints of, for example, Schiff base electrophiles (aldehydes and
3 ketones),¹¹ Michael acceptors,¹² S_NAr electrophiles,¹³ and diverse organic chemicals.¹⁴
4 Roberts and Patlewicz¹⁵ have reviewed the subject.

5 In order to develop good QSAR models, all chemicals used in the training set should exert
6 their effect by the same mechanism. Since it is often difficult to determine mechanisms of
7 action, the default position has been to use chemicals of the same class (e.g. benzoic acids,¹⁶
8 nitrobenzenes¹⁷) in the expectation that they have a common mechanism. However, with the
9 emphasis in recent years on mechanistically based QSAR modeling, and with current
10 knowledge of mechanisms involved in skin sensitization,¹⁸ we decided to try to use this
11 approach to model the relatively large data-sets of Gerberick et al.¹⁹ and Kern et al.,²⁰
12 comprising 211 chemicals and 108 chemicals respectively.

13 **Methods**

14 *Skin sensitization data*

15 The Gerberick et al.¹⁹ and Kern et al.²⁰ data-sets contain a total of 85 non-sensitizers, which
16 of course cannot be included in MLR modeling. In addition, two chemicals (cinnamic
17 aldehyde and 2-amino-6-chloro-4-nitrophenol) were duplicated in the data-sets. In the case of
18 cinnamic aldehyde, for one duplicate there was some difference between the EC₃ value of 1.4
19 reported by Gerberick et al.¹⁹ and the value of 2.05 reported in the original publication;²¹ in
20 addition, the original publication²¹ reported that the value of 2.05 was an average, indicating
21 that a range of values had been obtained. Because of the doubt about the true EC₃ value, we
22 selected the other duplicate, with an EC₃ value of 3.0. In the case of 2-amino-6-chloro-4-
23 nitrophenol we rejected one EC₃ value (2.2), as it was obtained from an erratic dose-response
24 curve. One chemical (*bis*-3,4-epoxycyclohexyl-ethyl-phenyl-methylsilane) contained

1 silicon and several were ionic chemicals, which could not be handled by our software.
2 Isopropyl myristate was removed because it was listed as a false positive,¹⁹ and methyl
3 hexadecene sulfonate was deleted because the molecular structure and CAS numbers given in
4 Gerberick et al.¹⁹ are incorrect. These deletions left a total of 204 skin sensitizers for
5 modeling.

6 The LLNA involves the topical exposure of the ear dorsum of CBA female mice to 25 μL of
7 at least three different concentrations of test chemical, daily for three days. After a further
8 two days an injection is given of 250 μL of phosphate-buffered saline containing 20 μCi of
9 tritiated thymidine. Five hours later the animals are sacrificed, the draining auricular lymph
10 nodes are excised, and the incorporation of tritiated thymidine measured. From these results,
11 the EC3 value is calculated.

12 It should be noted that EC3 values are reported as g/100 ml. Four potency ranges are used, as
13 follows: EC3 ≥ 10 to ≤ 100 , weak; EC3 ≥ 1 to < 10 , moderate; EC3 ≥ 0.1 to < 1 , strong; EC3
14 < 0.1 , extreme.¹⁹ Use of weight concentrations can give rise to a classification problem.
15 Strictly, concentrations and dosages should be given in molar units (e.g. mmol.L^{-1} , $\mu\text{mol.kg}^{-1}$),
16 for comparison, because effects are initiated by the number of molecules present, and not by
17 how much they weigh.²² Hence we have used SSP, defined as $\text{SSP} = \log(\text{MW}/10\text{EC3})$, in
18 our modeling. The importance of this is demonstrated by two chemicals from our data-set,
19 formaldehyde (MW 30.03) and 3-methylisoeugenol (MW 178.23). They have almost
20 identical skin sensitization potencies (1.692 and 1.695) based on their molar concentrations,
21 yet their EC3 values are quite different (0.61% and 3.6%), meaning that formaldehyde is
22 classified as a strong sensitizer, whilst 3-methylisoeugenol is classified as a moderate
23 sensitizer.

1
2
3 1 Using our in-house expertise,¹⁸ now incorporated into the Toxtree software,²³ together with
4
5 2 additional expert knowledge (DWR and SJE), we classified the chemicals into their
6
7 3 mechanistic categories. The chemicals are listed in Table 1. We have retained the chemical
8
9 4 names used by Gerberick et al.¹⁹ and Kern et al.²⁰ for ease of cross-reference, and have
10
11 5 included CAS numbers for all of the 204 chemicals save for four chemicals whose CAS
12
13 6 numbers we were unable to find.

17 7 Table 1 here

20 8 *QSAR modeling*

23 9 It is widely acknowledged that for a QSAR model to be predictive, external test chemicals
24
25 10 should be similar to one or more chemicals in the training set used to build the model.²⁴⁻²⁶
26
27 11 There are a number of methods used to achieve this,²⁷ although the topic is still open and has
28
29 12 not been completely solved.²⁸ Perhaps the most widely practised approach is that using a
30
31 13 clustering technique on the whole data set in order to select test set chemicals that are similar
32
33 14 to one or more chemicals in the remaining chemicals (i.e. the training set).
34
35

36 15
37
38 16 It has also been pointed out^{24,29} that external test set chemicals should, strictly speaking, be
39
40 17 completely independent of the training set. However, the clustering technique does not
41
42 18 comply with that requirement,^{22,29} since the selection of test chemicals that are very similar to
43
44 19 chemicals in the training set means that they carry the same structural information.³⁰
45
46
47
48 20

49
50 21 In addition, for relatively small data sets such as ours, removal of even a small number of test
51
52 22 set chemicals results in loss of a significant amount of information.³¹ This is of even more
53
54 23 concern when the data set comprises chemicals of a range of chemical classes, as is the case
55
56
57
58
59
60

1
2
3 1 with our skin sensitizers (see Table 1). It is thus likely that the use of leave-many-out and
4
5 2 bootstrap techniques²⁴ would also be inappropriate.
6
7 3
8
9 4 Using the clustering technique for selection of test chemicals, Gramatica et al.³² found that
10
11 5 the four descriptors used to develop a good 93-chemical training set QSAR for K_{oc} prediction
12
13 6 ($R^2 = 0.82$, $s = 0.539$) also yielded a good QSAR on the whole 643-chemical data set ($R^2 =$
14
15 7 0.79 , $s = 0.547$). However, this was not the case with our small data sets. For example, for the
16
17 8 Michael acceptor chemicals, a 6-descriptor QSAR developed using the 36-chemical training
18
19 9 set had $R^2 = 0.866$, $s = 0.344$. When the same 6 descriptors were used to develop a QSAR
20
21 10 for all 45 Michael acceptor chemicals, the result was poor ($R^2 = 0.636$, $s = 0.570$). This
22
23 11 confirms the view of Roy et al.³¹ that removal of test set chemicals from a small data set
24
25 12 results in loss of information, and thus changes the applicability domain of the model. Partly
26
27 13 for this reason, Hawkins³³ recommended that external validation should not be carried out on
28
29 14 data sets much below 50 chemicals, whilst Tropsha²⁷ recommended a minimum of 30-40
30
31 15 chemicals and Gramatica³⁴ recommended a minimum of 25 chemicals. From Table 1 it can
32
33 16 be seen that our data sets range in size from 11 to 45 chemicals, and thus are at least verging
34
35 17 on the size where external validation may be expected not to perform well. It may be noted
36
37 18 also that because of the diversity of our data sets, a greater number of descriptors are required
38
39 19 to give good models.²⁶
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

20
21 The above paragraph indicates that because of the smallness and chemical diversity of our
22
23 22 data sets, we could not expect to obtain good predictive models based on descriptors selected
24
25 23 during development of the training sets. We therefore decided to use for the training sets the
26
27 24 descriptors selected for the corresponding QSARs developed for the full data sets. We
28
29 25 recognise that this means that the training set QSARs are not fully independent of the test set
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 1 chemicals, but we believe that this is no less valid than the widely used clustering approach
4
5 2 for the selection of test set chemicals, which also involves some loss of independence of test
6
7 3 set chemicals. Our approach also means that the applicability domains of the full data sets are
8
9 4 preserved to some extent at least, and thus overcomes the concerns of Hawkins³³ and
10
11 5 Gramatica³⁴ in that respect. We stress, however, that this approach should be used only for
12
13 6 small, very diverse data sets, but in such cases we believe that it fits with the dictum of Albert
14
15 7 Einstein: *Everything should be made as simple as possible, but not simpler.*
16
17 8

18
19
20 9 There were too few chemicals acting by S_N1, pro-S_N2 and S_NAr mechanisms (2, 2, and 4
21
22 10 chemicals respectively) to allow us to develop QSARs in these categories. Hence 196
23
24 11 chemicals constituted our pool of chemicals used for modeling.
25
26 12

27
28
29 13 Various methods can be employed for the splitting of a data-set into training and test sets,
30
31 14 from random selection to activity sampling, clustering techniques, self-organising maps and
32
33 15 formal statistical experimental design.²⁴ Random selection is intuitively unappealing, and
34
35 16 “could result in a subsequent application of the model out of its applicability domain,
36
37 17 resulting in erroneous conclusions on the model’s performance”.³⁴ In addition it does not
38
39 18 provide any rationale for selection.³⁵ However, it was found to yield similar predictive
40
41 19 power to methods based on clustering.³⁵ Activity sampling (e.g. ordering the chemicals
42
43 20 according to their activity, then taking every *n*-th chemical for the test set) ensures a good
44
45 21 coverage of activity, but does not necessarily take account of chemical diversity, and thus
46
47 22 again risks subsequent application outside the applicability domain. The other techniques can
48
49 23 be complex,²⁷ and can give conflicting results.³⁵ Tropsha et al.²⁴ have stated that “the
50
51 24 underlying goal...is to ensure that both the training and test sets separately span the whole
52
53 25 descriptor space occupied by the entire data set and the chemical domains in the two sets are
54
55
56
57
58
59
60

1
2
3 1 not too dissimilar”. Chirico and Gramatica²⁸ have commented that “the topic (of external
4 validation) is still open, and the problem in QSAR modelling has not yet been completely
5 solved, though many techniques have been proposed to validate models”. The above
6 approaches have been designed for large or relatively large data sets, and we did not have that
7 luxury. In fact, the external validation of small heterogeneous data sets has not been
8 addressed before. Martin et al.³⁵ have pointed out that rational design of test sets should
9 ensure that “the compounds in the training and test sets should be close to each other”.
10 However, as stated earlier, selection of test chemicals that are very similar to chemicals in a
11 training set means that they carry the same structural information,³⁰ which would lead to
12 over-estimation of the predictivity of the model. We therefore used a manual sampling
13 approach that ensured a good range of activities and chemical domains in the test sets, whilst
14 never selecting the chemicals with the highest and lowest activities in the whole data sets³⁶ to
15 avoid the risk of extrapolation of the training set models. Care was taken that the test set
16 chemicals covered approximately the same chemical and biological space as the training set
17 chemicals in each category, and were not too close to or too far from the line of best fit in the
18 relevant whole data set model.

19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44 19 It is likely that with small, heterogeneous data sets there is no entirely satisfactory way to
45 demonstrate true prediction capability using QSAR modeling. We believe that the simple
46 method that we have adopted, whilst not perfect, is acceptable, and that the alternatives are
47 open to at least as much criticism as the one that we have used. We recognize that our
48 approach could be controversial, but we believe that it is a useful and pragmatic method for
49 QSAR prediction using small, diverse data sets. We do not recommend it for use with large
50 and/or homogeneous data sets. A reviewer has commented that the Q^2 (leave-one-out) value

1
2
3 1 of each training set could be more valuable than the test set values. In fact, as can be seen
4
5 2 from Table 2, all of our training set Q^2 values are above the recommended lower limit of
6
7 3 0.5,³⁷ and are no more than the recommended³⁸ 0.3 below the corresponding R^2 value, with
8
9
10 4 the exception of the Schiff base model, instead of which we recommend the combined Schiff
11
12 5 base and pro-Schiff base model, which has good statistics ($R^2 = 0.836$, $Q^2 = 0.736$).

13
14 6
15
16 7 A total of about 1600 descriptors were generated from CODESSA,³⁹ MOE⁴⁰ and
17
18 8 winMolconn⁴¹ software. These were pruned, by removal of descriptors with the same values
19
20 9 for all chemicals and by removal of descriptors with high pair-wise collinearity, to about 880
21
22 10 descriptors. Statistical analysis was carried out using the simple wrapper method of step-wise
23
24 11 MLR⁴² in Minitab v17 software⁴³ on the chemicals in each mechanistic category. Modeling
25
26 12 was first performed on the total number of chemicals in each category. Then approximately
27
28 13 20% of the chemicals in each category were removed to serve as a test set, and each model
29
30 14 was re-developed on the remaining (training set) chemicals, using the same descriptors as
31
32 15 were obtained for the model developed with the total number of chemicals in the category.
33
34 16 The predicted skin sensitization potencies of test set chemicals were calculated from the
35
36 17 QSARs developed for the corresponding training set chemicals.

37
38
39
40
41 18 The number in brackets after each coefficient in a QSAR is the standard error on the
42
43 19 coefficient. For a descriptor to be valid, the standard error on its coefficient should be
44
45 20 significantly lower than the value of the coefficient itself. This is also reflected in the p value
46
47 21 for each descriptor, a measure of the probability that the descriptor is there by chance; for a
48
49 22 descriptor to be valid in a QSAR, its p value should generally be < 0.05 (that is, less than a
50
51 23 5% risk that it is present by chance).

52
53
54
55 24

1
2
3 1 The statistics given with each QSAR are: R^2 (indicating the proportion of the variation of
4
5 2 skin sensitization potency (SSP) modeled by the QSAR); R^2_{adj} , which allows comparison
6
7 3 between QSARs with different numbers of descriptors; Q^2 , an internal measure of
8
9 4 predictivity, obtained using the leave-one-out procedure in Minitab; s ; and F (the Fisher
10
11 5 statistic, an indication of the fit of the regression equation to the training set data).
12
13
14 6

15
16 7 We also carried out 20 Y-randomizations of the SSP values within each mechanism in order
17
18 8 to check the robustness of the QSARs generated. For each mechanism, all R^2 values obtained
19
20 9 using randomized SSP values were significantly lower than the values obtained with non-
21
22 10 randomized SSP values.
23
24
25 11

26
27 12 For the test set results, the correlation between observed and predicted SSP values should
28
29 13 have an intercept close to zero and a slope close to unity. However, it has been pointed out
30
31 14 that correlation alone is not an adequate criterion for agreement between predicted and
32
33 15 observed values of biological endpoints.²⁴ To establish agreement it is necessary to exclude
34
35 16 three potential problems: (i) random disagreement, (ii) biased disagreement with one set of
36
37 17 values being systematically greater than (or less than) the other, and (iii) gradient problems
38
39 18 (the points on a graph of predicted versus observed values adhering to a line with a gradient
40
41 19 other than +1.0). Tropsha et al.²⁴ have recommended a multi-step procedure for assessing
42
43 20 how well those criteria are met.
44
45
46 21

47
48
49 22 However, there is a simpler alternative, the ICC, that serves just as well and has been
50
51 23 available for many years.⁴⁴ There are various ways in which the ICC can be calculated but in
52
53 24 some of its forms it will produce a value close to +1.0 only if the data adhere tightly to all
54
55 25 three of the criteria set out above. It can therefore act as a single unified indicator of
56
57
58
59
60

1
2
3 1 agreement between predicted and observed values. In the event that the ICC value was low,
4
5 2 the exact nature of the problem could be diagnosed by plotting the discrepancies between the
6
7 3 values against the average of the two (Bland-Altman plot) as advised by Machin, Campbell
8
9 4 and Walters.⁴⁵ We have used the ICC to assess how well our test set data meet the above
10
11 5 criteria. Weir⁴⁶ has pointed out that the ICC is conceptually akin to R^2 from regression, so it
12
13 6 is reasonable to assume that a value that is considered good for R^2 (say, 0.9), can also be
14
15 7 considered good for the ICC.
16
17
18
19

20
21 9 ICC values were calculated using the Reliability Analysis procedure in SPSS v20.⁴⁷ The
22
23 10 statistical model was set to Two-Way Mixed and the ICC type was set to Absolute
24
25 11 Agreement. The ICC values reported are for those for Single Measures.
26
27
28
29

30
31 13 It is also important that there should be no high pair-wise correlations between the various
32
33 14 descriptors incorporated into a QSAR, otherwise the statistics could be flawed.²³ Using a cut-
34
35 15 off point of $r = 0.9$,⁴⁸ we found no such high correlations between any of the descriptors used
36
37 16 in each QSAR.
38
39
40

41 18 **Results and discussion**

42
43
44 19
45
46 20 The QSARs that we developed for each mechanistic category, as well as that for all 204
47
48 21 chemicals together, are given in Table 2.
49
50
51

52
53 23 Table 2 here
54
55
56
57
58
59
60

1 Explanations of the descriptors are given in Table 3. We recognize that in some cases the
2 explanations are sparse, but descriptor software is frequently short on detail. Table 3 also
3 includes the ranges of SSPs and descriptor values in each mechanistic category, as an
4 indication of the applicability domains of each category. The SSPs cover a very wide range of
5 potency ranging from weak to strong or extreme, save for the oxidation potential category, in
6 which the range is from weak to moderate (EC3 values from 89% to 5%).

7
8 Table 3 here
9

10 For each category with adequate numbers of chemicals, with two exceptions, we were able to
11 formulate good QSARs with good internal and external validation. The first exception is the
12 Schiff base category, for which we could obtain a QSAR that, whilst acceptable, was not
13 good enough for our purposes, namely to provide QSAR models that can offer good
14 prediction. However, by combining the Schiff base chemicals with the five in the pro-Schiff
15 base category we were able to develop a QSAR with good internal and external predictive
16 ability. The second exception is the acyl transfer category, for which a good model could not
17 be developed using all 23 acyl transfer chemicals, owing to one chemical, C11 azlactone,
18 being a pronounced outlier. Several azlactones, with alkyl chains ranging from C4 to C19,
19 have been tested in the LLNA (see Table 1), and they appear to fall into two groups,
20 separated by an activity cliff.⁴⁹ Shorter chain-length azlactones (C4 to C9) are quite potent,
21 with EC3 values between 1% and 3%, whereas longer-chain homologs (C15 to C19) are
22 much weaker, with EC3 values of about 20%. This presumably reflects a change in the rate-
23 determining step (possibly mass transfer) becoming rate-limiting for azlactones with high
24 hydrophobicity.⁵⁰ Our model is able to make this distinction, but it appears that the C11
25 homolog, structurally between these two sub-sets, and which should belong to the low-

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1 potency sub-set, is treated by our model as belonging to the high-potency sub-set. When the
2 C11 azlactone was removed, a good QSAR model was obtained (Table 2, equations 17 and
3 18). The statistical quality of all the models can be seen from Table 4.

4 Table 4 here

5
6 It would, of course, have been possible to increase R^2 and s values for most of the models by
7 increasing the number of descriptors incorporated. However, as we have pointed out
8 elsewhere,²² “the principle of Occam’s razor (principle of parsimony) applies here: ‘One
9 should not increase beyond what is necessary the number of entities required to explain
10 anything’. We suggest that five or six descriptors are generally the maximum that one should
11 generally use in a QSAR/QSPR, partly because it is difficult to comprehend the mechanistic
12 significance of large numbers of descriptors”. We were surprised but very pleased that the
13 two categories with the smallest number of chemicals (acyl transfer and oxidation potential)
14 could nevertheless allow the development of good QSARs. In fact the latter category yielded
15 the best QSAR of all.

16
17 The observed SSPs for all 195 skin sensitizers used in our modeling were correlated with the
18 cumulative SSP values calculated from each appropriate local mechanistic domain QSAR,
19 and as expected a very good correlation was found:

$$21 \text{ SSP (observed)} = 0.000 + 1.000 \text{ SSP (predicted)} \quad (22)$$

$$22 n = 195 \quad R^2 = 0.884 \quad Q^2 = 0.882 \quad \text{ICC} = 0.939 \quad s = 0.296 \quad F = 1471$$

23 A graphical representation of these results is shown in Figure 1.

24 Figure 1 here

1
2
3 1
4
5 2 All test sets yielded very good predictions, fortuitously with all R^2 values higher than those of
6
7 3 the full and training set QSARs.

8
9
10 4 The correlation between observed and predicted SSP values for all 37 test set chemicals was
11
12 5 found to be:

13
14 6
15
16 7
$$\text{SSP (obsd)} = -0.070 + 1.002 \text{ SSP (pred)} \quad (23)$$

17
18 8
19
20 9 $n = 37 \quad R^2 = 0.947 \quad Q^2 = 0.940 \quad \text{ICC} = 0.971 \quad s = 0.209 \quad F = 627.3$

21
22 10
23
24
25 11 The overall ICC of 0.971 for all test set results indicates that the test set results for all
26
27 12 mechanisms were valid. This can also be seen from Figure 2.

28
29
30
31
32 13
33 14 **Figure 2 here**

34 15 The QSAR derived for the complete dataset of 204 active chemicals, covering all the reaction
35
36 16 mechanistic categories, is very much inferior to any of the QSARs for the individual
37
38 17 mechanistic categories (Table 2), and the descriptors found to model the potency best are
39
40 18 different for each mechanistic category, as can be seen from Table 3. These findings
41
42 19 reinforce the argument that for skin sensitization, modeling reaction mechanistic
43
44 20 domains/categories has more realistic prospects of success than attempting a global model.

45
46
47
48 21 The model obtained for Schiff base chemicals was not very good ($n = 35$, $R^2 = 0.837$, $Q^2 =$
49
50 22 0.644 , $s = 0.259$, $F = 19.9$). However, inclusion of the five pro-Schiff base chemicals
51
52 23 improved the model considerably ($n = 40$, $R^2 = 0.850$, $Q^2 = 0.781$, $s = 0.233$, $F = 25.9$).

53
54
55
56 24 It has been found that depending on the reaction mechanism of the protein-binding step, there
57
58 25 are different relationships between model reactivity parameters and potency.⁵⁰⁻⁵² This is

1
2
3 1 argued to be because, depending on the reaction mechanism, relative reactivities towards the
4
5 2 several nucleophilic protein sites will differ. Thus, for example, the Schiff category chemicals
6
7 3 probably sensitize via reaction with amino groups of proteins, whereas the Michael acceptor
8
9
10 4 category chemicals probably sensitize via reaction with protein thiol groups. Even where
11
12 5 compounds from two different mechanistic categories sensitize via reaction with the same
13
14 6 type of protein nucleophile, the proportionality between the *in cutaneo* reactivity and
15
16 7 reactivity determined in a model cannot be assumed to be the same. This should apply
17
18 8 irrespective of whether the model reactivity is based on experimental data with model
19
20 9 nucleophiles, on classical linear free energy relationship indices based on Hammett and Taft
21
22
23 10 substituent constants, on quantum mechanical indices such as activation energy,⁵³ or on
24
25 11 combinations of less transparent descriptors such as those used here. Furthermore, for some
26
27 12 reaction mechanistic categories (Schiff base,^{11,50} S_N2 and acyl transfer⁵⁰), potency has been
28
29 13 found to be dependent not only on reactivity but also on hydrophobicity, whilst for others
30
31 14 (Michael acceptors,¹² S_NAr electrophiles¹³) reactivity parameters alone can give good models
32
33
34 15 for potency. as been argued that depending on the reaction mechanism of the protein-
35
36 16 binding step, there are different relationships between model reactivity parameters and
37
38 17 potency.⁵⁰⁻⁵² This is argued to be because, depending on the reaction mechanism, relative
39
40 18 reactivities towards the several nucleophilic protein sites will differ. Thus for example, the
41
42 19 Schiff base category chemicals probably sensitize via reaction with amino groups of proteins,
43
44 20 whereas the Michael acceptor category chemicals probably sensitize via reaction with protein
45
46 21 thiol groups. Even where compounds from two different mechanistic categories sensitize via
47
48 22 reaction with the same type of protein nucleophile, the proportionality between the *in cutaneo*
49
50 23 reactivity and reactivity determined in a model cannot be assumed to be the same. This
51
52 24 should apply irrespective of whether the model reactivity is based on experimental data with
53
54 25 model nucleophiles, on classical linear free energy relationship indices based on Hammett
55
56
57
58
59
60

1
2
3 1 and Taft substituent constants, on quantum mechanical indices such as activation energy,⁵³ or
4
5 2 on combinations of less transparent descriptors such as those used here. Furthermore, for
6
7 3 some reaction mechanistic categories (Schiff base,^{11,50} S_N2 and acyl transfer⁵⁰), potency has
8
9 4 been found to be dependent not only on reactivity but also on hydrophobicity, while for
10
11 5 others (Michael acceptors,¹² S_NAr electrophiles¹³) reactivity parameters alone give good
12
13 6 models for potency. It has already been mentioned that many descriptors are difficult to
14
15 7 interpret. Those selected for the Michael addition category suggest that reactivity and surface
16
17 8 area, and perhaps especially hydrophobic surface area, enhance skin sensitization potency.
18
19 9 For pro-Michael addition several descriptors represent hydrogen bonding, although there
20
21 10 does not appear to be a consistent pattern; for example, S_{ss}NH has a positive coefficient,
22
23 11 whereas that for v_{surf}_HB7 is negative.
24
25
26
27
28
29
30

31 12
32
33 13 From equation 8 it can be seen that for Schiff base chemicals, polarity and molecular
34
35 14 flexibility increase potency. There are also some specific atom effects (S7 and S10), although,
36
37 15 as the nature of those atoms is not known, no interpretation of those effects can be made. The
38
39 16 situation is somewhat clearer for the combined Schiff base and pro-Schiff base model
40
41 17 (equation 11), with hydrogen bonding (represented by HS6, E_{sol} and possibly DPSA1)
42
43 18 being important for potency, together with molecular shape (dx2 and Kier FI).
44
45
46
47
48
49
50

51 19
52
53 20 S_N2 chemicals appear to require hydrophobicity (S_sCH₃, eaC2C3a) for potency, although
54
55 21 descriptors representing both negative and positive surface area also have positive
56
57 22 coefficients. Electron-donating ability (MNDO_HOMO) decreases potency, which is to be
58
59 23 expected since Michael reactivity is dependent on the electron deficiency of the double or
60
24 triple bond.

1
2
3 1 Acyl transfer appears to be highly dependent on hydrogen bonding, as all four descriptors are
4
5 2 E-state values for different hydrogen atoms. Finally, oxidation potential appears possibly to
6
7 3 be dependent on molecular shape as well as the location of interacting atoms or groups, as
8
9
10 4 contact distances are important (vsurf_DD12, vsurf_DD23).

11
12 5 It should be noted that whilst hydrophobicity (represented in many QSAR studies as log P,
13
14 6 the logarithm of the octanol-water partition coefficient) is not specifically selected as a
15
16
17 7 descriptor in any of our models, it is a composite descriptor with components of polarity,
18
19 8 polarizability, hydrogen bonding and molecular size,⁵⁴ so our models are not incompatible
20
21 9 with previous studies^{11, 50} that found hydrophobicity to be important.
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

10
11 Based on the above perspective, we have shown that quantitative predictive models for
12
13 sensitization potency can be derived by: (i) assigning chemicals to reaction mechanistic
14
15 domains; (ii) determining appropriate reactivity parameters and (if necessary) hydrophobicity
16
17 within a mechanistic domain; (iii) deriving regression-based quantitative mechanistic models
18
19 and using these to estimate the potency for untested chemicals. This chemistry-based
20
21 approach can already enable potency to be predicted for many chemicals.⁵¹ The findings
22
23 presented here strongly reinforce the argument that assignment of chemicals to their reaction
24
25 mechanistic domains (categories) is an essential step before attempting to predict potency by
26
27 *in chemico* or *in silico* approaches.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

21 All the QSARs reported here satisfy all or almost all of the OECD Principles for the
22
23 Validation of (Q)SARs.⁵⁵ The work described here offers one solution to the vital need,
24
25 emphasized by Basketter et al.,⁵⁶ for information on the potency of identified skin sensitizers
26
27 in order to permit risk assessment.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

2 **Conclusions**

3
4 Using in-house expertise, we have allocated 204 skin-sensitizing chemicals to their respective
5 mechanistic categories, and then developed good QSAR models, with good predictive ability,
6 for chemicals in seven out of ten categories. Only one chemical had to be omitted as an
7 outlier, and an explanation is provided for that omission. Data on too few chemicals were
8 available to allow QSAR modeling for three categories, namely S_{N1} , pro- S_{N2} and S_{NAr} . The
9 QSARs reported here can be used, either on their own or as part of a weight-of-evidence
10 approach, in risk assessments of skin sensitization.

12 **Notes**

13 The authors declare no competing financial interests.

14 **Funding**

15 GGP is grateful to the graduate school “Functional Materials and Technologies”, University
16 of Tartu for funding from the European Social Fund under project 1.2.0401.09-0079

18 **Abbreviations**

19 Ac, acyl transfer; CAS, Chemical Abstracts Service; EC₃, the concentration (g/100 ml) that
20 induces a threefold increase in local lymph node proliferative activity relative to controls; *F*,
21 coefficient of variance (Fisher statistic); ICC, intraclass correlation coefficient; LLNA,
22 murine local lymph node assay; MA, Michael addition; MLR, multiple linear regression;
23 MW, molecular weight (relative molecular mass); OxPot, oxidation potential; p-MA, pro-
24 Michael addition; OECD, Organisation for Economic Cooperation and Development; p value,

1
2
3 1 probability that a descriptor is there by chance; p-SB, pro-Schiff base; p-S_N2, pro-bimolecular
4
5 2 aliphatic nucleophilic substitution; Q²; cross-validated coefficient of variation (leave-one-out
6
7 3 procedure); QSAR, quantitative structure-activity relationship; r, correlation coefficient; R²,
8
9 4 coefficient of variation; R²_{adj}, coefficient of variation adjusted for degrees of freedom;
10
11 5 REACH, Registration, Evaluation, Authorisation and restriction of Chemicals; s, standard
12
13 6 error of estimate; SB, Schiff base; S_N1, unimolecular aliphatic nucleophilic substitution; S_N2,
14
15 7 bimolecular aliphatic nucleophilic substitution; S_NAr, bimolecular aromatic nucleophilic
16
17 8 substitution; SSP, skin sensitization potency.
18
19
20
21
22
23
24

25 **References**

- 26
27 1. Patlewicz, G., Roberts, D.W., and Uriarte, E. (2008) A comparison of reactivity
28
29 2 schemes for the prediction (of) skin sensitization potential. *Chem. Res. Toxicol.* *21*,
30
31 521-541.
32
33 14 2. Nilsson, A.-M., Bergström, M.A., Luthman, K., Nilsson, J.L.G., and Karlberg, A.-T.
34
35 15 (2005) A conjugated diene identified as a prohaptens: contact allergenic activity and
36
37 16 chemical reactivity of proposed epoxide metabolites. *Chem. Res. Toxicol.* *18*, 308-316.
38
39 40 3. Roberts, D.W., and Aptula, A.O. (2008) Determinants of skin sensitisation potential. *J.*
41
42 18 *Appl. Toxicol.* *28*, 377-387.
43
44 45 4. Basketter, D.A., McFadden, J.F., Gerberick, F., Cockshott, A., and Kimber, I. (2009)
46
47 20 Nothing is perfect, not even the local lymph node assay: a commentary and the
48
49 21 implications for REACH. *Contact Derm.* *60*, 65-69.
50
51 22 5. Dearman, R.J., Basketter, D.A., and Kimber, I. (1999) Local lymph node assay: use in
52
53 23 hazard and risk assessment. *J. Appl. Toxicol.* *19*, 299-306.
54
55 56 6. Bergström, M.A., Luthman, K., Nilsson, J.L.G., and Karlberg, A.-T. (2006)
57
58 25 Conjugated dienes as prohaptens in contact allergy: in vivo and in vitro studies of
59
60

- 1
2
3 1 structure-activity relationships, sensitizing capacity, and metabolic activation. *Chem.*
4
5 2 *Res. Toxicol.* 19, 760-769.
6
7 3 7. Chaudhry, Q., Piclin, N., Cotterill, J., Pintore, M., Price, N.R., Chrétien, J.R., and
8
9 4 Roncaglioni, A. (2010) Global QSAR models of skin sensitizers for regulatory
10
11 5 purposes. *Chem. Central J.* 4 (Suppl. 1), Article S5.
12
13 6 8. Schaafsma, G., Hertsenberg, A.J., and Marquart, J. (2011) Risk assessment of local
14
15 7 dermal effects and skin sensitisation under the EU chemicals regulation REACH: a
16
17 8 proposal for a qualitative, exposure scenario specific, approach. *Regul. Toxicol.*
18
19 9 *Pharmacol.* 60, 308-317.
20
21 10 9. REACH Regulation: European Parliament: Regulation (EC) N° 1907/2006 of the
22
23 11 European Parliament and of the Council of 18 December 2006 concerning the
24
25 12 Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH);
26
27 13 available at [http://eur-](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=oj:l:2006:396:0001:0849:en:pdf)
28
29 14 [lex.europa.eu/LexUriServ/LexUriServ.do?uri=oj:l:2006:396:0001:0849:en:pdf](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=oj:l:2006:396:0001:0849:en:pdf)
30
31
32 15 10. Patlewicz, G., Aptula, A.O., Roberts, D.W., and Uriarte, E. (2008) A minireview of
33
34 16 available skin sensitization (Q)SARs/expert systems. *QSAR Comb. Sci.* 27, 60-76.
35
36 17 11. Roberts, D.W, Aptula, A.O, and Patlewicz, G. (2006) Mechanistic applicability
37
38 18 domains for non-animal based toxicological endpoints. QSAR analysis of the Schiff
39
40 19 base applicability domain for skin sensitization. *Chem. Res. Toxicol.* 19, 1228-1233.
41
42 20 12. Roberts, D.W., and Natsch, A. (2009) High throughput kinetic profiling approach for
43
44 21 covalent binding to peptides: application to skin sensitization potency of Michael
45
46 22 acceptor electrophiles. *Chem. Res. Toxicol.* 22, 592-603.
47
48 23 13. Roberts, D.W., and Aptula, A.O. (2014) Electrophilic reactivity and skin sensitization
49
50 24 potency of S_NAr electrophiles. *Chem. Res. Toxicol.* 27, 240-246.
51
52
53
54
55
56
57
58
59
60

- 1
2
3 1 14. Nandy, A., Kar, S., and Roy, K. (2013) Development and validation of regression-
4
5 2 based QSAR models for quantification of contributions of molecular fragments to
6
7 3 skin sensitization potency of diverse organic chemicals. *SAR QSAR Environ. Res.* 24,
8
9 4 1009-1013.
10
11 5 15. Roberts, D.W., and Patlewicz, G. (2009) Chemistry based nonanimal predictive
12
13 6 modeling for skin sensitization. In: *Ecotoxicology Modeling* (ed. Devillers, J.),
14
15 7 Springer, Dordrecht, pp. 61-83.
16
17 8 16. Zhao, Y.H., Ji, G.D., Cronin, M.T.D., and Dearden, J.C. (1998) QSAR study of the
18
19 9 toxicity of benzoic acids to *Vibrio fischeri*, *Daphnia magna* and carp. *Sci. Tot.*
20
21 10 *Environ.* 216, 205-215.
22
23 11 17. Dearden, J.C., Cronin, M.T.D., Schultz, T.W., and Lin, D.T. (1995) QSAR study of
24
25 12 the toxicity of nitrobenzenes to *Tetrahymena pyriformis*. *Quant. Struct.-Act. Relat.* 14,
26
27 13 427-432.
28
29 14 18. Enoch, S.J., Madden, J.C., and Cronin, M.T.D. (2008) Identification of mechanisms
30
31 15 of toxic action for skin sensitisation using a SMARTS pattern based approach. *SAR*
32
33 16 *QSAR Environ. Res.* 19, 555-578.
34
35 17 19. Gerberick, G.F., Ryan, C.A., Kern, P.S., Schlatter, H., Dearman, R.J., Kimber, I.,
36
37 18 Patlewicz, G.Y., and Basketter, D.A. (2005) Compilation of historical local lymph
38
39 19 node data for evaluation of skin sensitization alternative methods. *Dermatitis* 16, 157-
40
41 20 202.
42
43 21 20. Kern, P.S., Gerberick, G.F., Ryan, C.A., Kimber, I., Aptula, A., and Basketter, D.A.
44
45 22 (2010) Local lymph node data for the evaluation of skin sensitization alternatives: a
46
47 23 second compilation. *Dermatitis* 21, 8-32.
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 1 21. Patlewicz, G.Y., Wright, Z.M., Basketter, D.A., Pease, C.K., Lepoittevin, J.-P., and
4
5 2 Giménez Arnau, E. (2002) Structure-activity relationships for selected fragrance
6
7 3 allergens. *Contact Derm.* 47, 219-226.
8
9
10 4 22. Dearden, J.C., Cronin, M.T.D., and Kaiser, K.L.E. (2009) How not to develop a
11
12 5 quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR*
13
14 6 *QSAR Environ. Res.* 20, 241-266.
15
16 7 23. Toxtree software; available at <http://toxtree.sourceforge.net>.
17
18 8 24. Tropsha, A., Gramatica, P., and Gombar, V.K. (2003) The importance of being
19
20 9 earnest: validation is the absolute essential for successful application and
21
22 10 interpretation of QSPR models. *QSAR Comb. Sci.* 22, 69-77.
23
24 11 25. He, L., and Jurs, P.C. (2005) Assessing the reliability of a QSAR model's predictions.
25
26 12 *J. Mol. Graph. Model.* 23, 503-523.
27
28 13 26. Weaver, S., and Gleeson, M.P. (2008) The importance of the domain of applicability
29
30 14 in QSAR modeling. *J. Mol. Graph. Model.* 26, 1315-1326.
31
32 15 27. Tropsha, A. (2010) Best practices for QSAR model development, validation, and
33
34 16 exploitation. *Mol. Inf.* 29, 476-488.
35
36 17 28. Chirico, N., and Gramatica, P. (2012) Real external predictivity of QSAR models.
37
38 18 Part 2. New intercomparable thresholds for different validation criteria and the need
39
40 19 for scatter plot inspection. *J. Chem. Inf. Model.* 52, 2044-2058.
41
42 20 29. Consonni, V., Ballabio, D., and Todeschini, R. (2010) Evaluation of model predictive
43
44 21 ability by external validation techniques. *J. Chemometrics* 24, 194-201.
45
46 22 30. Gramatica, P., Cassani, S., Roy, P.P., Kovarich, S., Yap, C.W., and Papa, E. (2012)
47
48 23 QSAR modeling is not "push a button and find a correlation": a case study of toxicity
49
50 24 of (benzo-)triazoles on algae. *Mol. Inf.* 31, 817-835.
51
52
53
54
55
56
57
58
59
60

- 1
2
3 1 31. Roy, K., Mitra, I., Kar, S., Ojha, P.K., Das, R.N., and Kabir, H. (2012). Comparative
4
5 2 studies on some metrics for external validation of QSPR models. *J. Chem. Inf. Model.*
6
7 3 52, 396-408.
8
9
10 4 32. Gramatica, P., Giani, E., and Papa, E. (2007) Statistical external validation and
11
12 5 consensus modeling: a QSPR case study for K_{oc} prediction. *J. Mol. Graph. Model.* 25,
13
14 6 755-766.
15
16 7 33. Hawkins, D.M. (2004) The problem of overfitting. *J. Chem. Inf. Comput. Sci.* 44, 1-
17
18 8 12.
19
20 9 34. Gramatica, P. (2007) Principles of QSAR models validation: internal and external.
21
22 10 *QSAR Comb. Sci.* 26, 694-701.
23
24 11 35. Martin, T.M., Harten, P., Young, D.M., Muratove, E.N., Golbraikh, A., Zhu, H., and
25
26 12 Tropsha, A. (2012) Does rational selection of training and test sets improve the
27
28 13 outcome of QSAR modeling? *J. Chem. Inf. Model.* 52, 2570-2578.
29
30 14 36. Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H., and Tropsha, A. (2003)
31
32 15 Rational selection of training and test sets for the development of validated QSAR
33
34 16 models. *J. Comput.-Aided Molec. Des.* 17, 241-253.
35
36 17 37. Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M., and
37
38 18 Gramatica, P. (2003) Methods for reliability and uncertainty assessment and for
39
40 19 applicability evaluations of classification- and regression-based QSARs. *Environ.*
41
42 20 *Health Perspect.* 111, 1361-1375.
43
44 21 38. Walker, J.D., Dearden, J.C., Schultz, T.W., Jaworska, J., and Comber, M.H.I. (2003)
45
46 22 QSARs for new practitioners. In: Pollution Prevention, Toxicity Screening, Risk
47
48 23 Assessment, and Web Applications (ed. Walker, J.D.), SETAC Press, Pensacola, FL,
49
50 24 pp. 3-18.
51
52 25 39. CODESSA software ; available at <http://www.semichem.com>.
53
54
55
56
57
58
59
60

- 1
2
3 1 40. MOE software; available at www.chemcomp.com.
4
5 2 41. winMolconn software; available at <http://www.molconn.com>.
6
7 3 42. Goodarzi, M., Dejaegher, B., and Vander Heyden, Y. (2012) Feature selection
8 methods in QSAR studies. *J. AOAC Int.* 95, 636-650.
9
10 4 43. Minitab software; available at www.minitab.com.
11
12 5 44. Fisher, R.A. (1963) *Statistical Methods for Research Workers*, Oliver and Boyd,
13
14 6 44. Fisher, R.A. (1963) *Statistical Methods for Research Workers*, Oliver and Boyd,
15
16 7 Edinburgh.
17
18 8 45. Machin, D., Campbell, M.J., and Walters, S.J. (2007) *Medical Statistics*, Wiley,
19
20 9 Chichester.
21
22
23 10 46. Weir, J.P. (2005) Quantifying test-retest reliability using the intraclass correlation
24
25 11 coefficient and the SEM. *J. Strength Cond. Res.* 19, 231-240.
26
27 12 47. SPSS software; available at <http://www.ibm.com>.
28
29 13 48. Chauhan, J.S., Dhanda, S.K., Singla, D., Open Source Drug Discovery Consortium,
30
31 14 Agarwal, S.M., and Raghava, G.P.S. (2014) QSAR-based models for designing
32
33 15 quinazoline/imidazothiazoles/pyrazolopyrimidines based inhibitors against wild and
34
35 16 mutant EGFR. *PLoS One* 9 (7): e101079.
36
37 17 49. Stumpfe D., and Bajorath, J. Exploring activity cliffs in medicinal chemistry. *J. Med.*
38
39 18 *Chem.* 55, 2932-2942.
40
41 19 50. Roberts, D.W, Aptula, A.O, and Patlewicz, G. (2007) Electrophilic chemistry related
42
43 20 to skin sensitization. Reaction mechanistic applicability domain classification for a
44
45 21 published data set of 106 chemicals tested in the mouse local lymph node assay.
46
47 22 *Chem. Res. Toxicol.* 20, 44-60.
48
49 23 51. Roberts, D.W., and Patlewicz, G.Y. (2014) Integrated testing and assessment
50
51 24 approaches for skin sensitization: a commentary. *J. Appl. Toxicol.* 34, 436-440.
52
53 25 52. Roberts D.W. (2013) Allergic contact dermatitis: is the reactive chemistry of skin
54
55
56
57
58
59
60

- 1
2
3 1 sensitizers the whole story? A response. *Contact Derm.* 68, 244–249.
4
5 2 53. Enoch, S.J., and Roberts, D.W. (2013) Predicting skin sensitization potency for
6
7 3 Michael acceptors in the LLNA using quantum mechanics calculations. *Chem. Res.*
8
9
10 4 *Toxicol.* 26, 767-774.
11
12 5 54. Abraham, M.H., Chadha, H.S., Whitney, G.S., and Mitchell, R.C. (1994) Hydrogen
13
14 6 bonding. 32. An analysis of water-octanol and water-alkane partitioning and the $\Delta\log$
15
16 7 P parameter of Seiler. *J. Pharm. Sci.*, 83, 1085-1100.
17
18 8 55. OECD Principles for the Validation of (Q)SARs; available at
19
20
21 9 www.oecd.org/dataoecd/33/37/37849783.pdf.
22
23 10 56. Basketter, D., Alépée, N., Casati, S., Crozier, J., Eigler, D., Griem, P., Hubesch, B.,
24
25 11 de Knecht, J., Landsiedel, R., Louekari, K., Manou, I., Maxwell, G., Mehling, A.,
26
27 12 Netzeva, N., Petry, T., and Rossi, L.H. (2013) Skin sensitisation – Moving forward
28
29 13 with non-animal testing strategies for regulatory purposes in the EU. *Regul. Toxicol.*
30
31
32 14 *Pharmacol.* 67, 531-535.
33
34
35 15
36
37
38 16
39
40
41 17
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1

2 Table 1. Chemicals used in this study, their potencies and mechanisms of action

Name	CAS No.	MW	EC3	Class	SSP	Mechanism
4'-Hydroxychalcone	2657-25-2	224.26	0.002	Extreme	4.050	MA
<i>p</i> -Benzoquinone ^a	106-51-4	108.10	0.0099	Extreme	3.038	MA
2',3',4'-Trihydroxychalcone	1482-74-2	256.25	0.11	Strong	2.367	MA
Methyl 2-octynoate	111-12-6	154.21	0.45	Strong	1.535	MA
2',4'-Dihydroxychalcone	1776-30-3	240.26	0.56	Strong	1.632	MA
Isopropyl isoeugenol	2953-00-7	206.29	0.6	Strong	1.536	MA
β -Phenylcinnamaldehyde	1210-39-5	208.26	0.6	Strong	1.540	MA
Isoeugenol ^a	97-54-1	164.20	1.2	Moderate	1.136	MA
2-Hydroxyethyl acrylate ^a	818-61-1	116.12	1.4	Moderate	0.919	MA
3-Methyl-4-phenyl-1,2,5-thiadiazole-1,1-dioxide (MPT)	3775-21-1	208.24	1.4	Moderate	1.172	MA
6-Methylisoeugenol	13041-12-8	178.23	1.6	Moderate	1.047	MA
Vinyl pyridine	100-43-6	105.14	1.6	Moderate	0.818	MA
5,5-Dimethyl-3-methylene-dihydro-2(3H)-furanone	29043-97-8	126.16	1.8	Moderate	0.846	MA
<i>trans</i> -Anethol ^a	104-46-1	148.21	2.3	Moderate	0.809	MA
<i>trans</i> -2-Decenal	3913-71-1	154.25	2.5	Moderate	0.790	MA
Methyl 2-nonynoate	111-80-8	168.24	2.5	Moderate	0.828	MA
3,4-Dinitrophenol	577-71-9	184.10	2.6	Moderate	0.850	MA
Cinnamic aldehyde	104-55-2	132.16	3	Moderate	0.644	MA
2,4-Hexadienal	142-83-6	96.13	3.5	Moderate	0.439	MA
3-Methylisoeugenol ^a	186743-29-3	178.23	3.6	Moderate	0.695	MA
Benzylidene acetone (4-phenyl-3-buten-2-one)	122-57-6	146.19	3.7	Moderate	0.597	MA
2,4-Heptadienal ^a	5910-85-0	110.16	4	Moderate	0.440	MA
Tropolone	533-75-5	122.12	4.3	Moderate	0.453	MA
5-Methyl-2-phenyl-2-hexenal	21834-92-4	188.27	4.4	Moderate	0.631	MA

1							
2							
3							
4							
5	α -Methylcinnamaldehyde	101-39-3	146.19	4.5	Moderate	0.512	MA
6	<i>trans</i> -2-Hexenal	6728-26-3	98.15	5.5	Moderate	0.252	MA
7	Diethyl maleate	141-05-9	172.18	5.8	Moderate	0.473	MA
8	1,1,3-Trimethyl-2-formylcyclohexa-2,1-diene (safranal)	116-26-7	150.22	7.5	Moderate	0.302	MA
9	Perillaldehyde	2111-75-3	150.22	8.1	Moderate	0.268	MA
10	1-(<i>p</i> -Methoxyphenol)-1-penten-3-one ^a	104-27-8	190.24	9.3	Moderate	0.311	MA
11	Linalool aldehyde	Not known ^b	168.24	9.5	Moderate	0.248	MA
12	2-Ethylhexyl acrylate	103-11-7	184.28	10	Weak	0.265	MA
13	α -Amylcinnamaldehyde	122-40-7	202.30	11	Weak	0.265	MA
14	α -Butylcinnamaldehyde	7492-44-6	188.27	11	Weak	0.233	MA
15	Hexyl cinnamaldehyde	101-86-0	216.32	11	Weak	0.294	MA
16	Butyl acrylate	141-32-2	128.17	11	Weak	0.066	MA
17	R-Carvone ^a	6485-40-1	150.22	12.9	Weak	0.066	MA
18	Benzyl cinnamate	103-41-3	238.29	18.4	Weak	0.112	MA
19	Methyl acrylate ^a	96-33-3	86.09	20	Weak	-0.366	MA
20	Cinnamic alcohol	104-54-1	134.18	21	Weak	-0.195	MA
21	α -iso-Methylionone	127-51-5	206.33	21.8	Weak	-0.024	MA
22	Ethyl acrylate	140-88-5	100.12	28	Weak	-0.447	MA
23	Ethylene glycol dimethacrylate	97-90-5	198.22	28	Weak	-0.150	MA
24	2,2-bis-[4-(2-Hydroxy-3-methacryloxypropoxy)phenyl]-propane	1565-94-2	512.65	45	Weak	0.057	MA
25	Methyl methacrylate	80-62-6	100.12	90	Weak	-0.954	MA
26	Bandrowski's base	20048-27-5	318.38	0.04	Extreme	2.901	p-MA
27	3,4-Diaminonitrobenzene	99-56-9	153.14	0.05	Extreme	2.486	p-MA
28	4-((2-Hydroxyethyl)amino)-3-nitrophenol	65235-31-6	198.18	0.07	Extreme	2.452	p-MA
29	1,4-Dihydroquinone	123-31-9	110.11	0.11	Strong	2.000	p-MA
30	1,4-Phenylenediamine	106-50-3	108.14	0.16	Strong	1.830	p-MA
31	2,5-Diaminotoluene	95-70-5	122.08	0.2	Strong	1.786	p-MA
32	4-Amino-3-nitrophenol	610-81-1	154.12	0.2	Strong	1.887	p-MA
33	Lauryl gallate (dodecyl gallate) ^a	1166-52-5	338.44	0.3	Strong	2.052	p-MA
34	2-Aminophenol	95-55-6	109.13	0.4	Strong	1.436	p-MA
35							
36							
37							
38							
39							
40							
41							
42							
43							
44							
45							
46							
47							
48							
49							

1							
2							
3							
4							
5	2-Methyl-5-hydroxyethylaminophenol	55302-96-0	167.21	0.4	Strong	1.621	p-MA
6	2-Nitro- <i>p</i> -phenylenediamine ^a	5307-14-2	153.14	0.4	Strong	1.583	p-MA
7	1,3-Phenylenediamine ^a	108-45-2	108.14	0.49	Strong	1.344	p-MA
8	R-Carvoxime	55658-55-4	165.23	0.6	Strong	1.440	p-MA
9	Hydroxytyrosol	10897-60-1	154.16	0.6	Strong	1.410	p-MA
10	1,2-Dibromo-2,4-dicyanobutane	35691-65-7	265.94	0.9	Strong	1.471	p-MA
11	1-Naphthol	90-15-3	144.17	1.3	Moderate	1.045	p-MA
12	4-Amino-3-methylphenol	2835-99-6	123.15	1.45	Moderate	0.929	p-MA
13	2-(4-Amino-2-nitrophenylamino)-ethanol	2871-01-4	197.19	2.2	Moderate	0.952	p-MA
14	3-Aminophenol	591-27-5	109.13	3.2	Moderate	0.533	p-MA
15	5-Amino-2-methylphenol ^a	2835-95-2	123.15	3.4	Moderate	0.559	p-MA
16	3-Bromomethyl-5,5-dimethyl-dihydro-2(3H)-furanone	154750-20-6	207.07	3.6	Moderate	0.760	p-MA
17	2-Methoxy-4-methyl-phenol	93-51-6	138.17	5.8	Moderate	0.377	p-MA
18	Anisyl alcohol	105-13-5	138.17	5.9	Moderate	0.370	p-MA
19	Dihydroeugenol	2785-87-7	166.22	6.8	Moderate	0.388	p-MA
20	2-Amino-6-chloro-4-nitrophenol ^a	6358-09-4	188.57	6.85	Moderate	0.440	p-MA
21	1-Amino-2-nitro-4-bis(2-hydroxyethyl)-amino-benzene	29705-39-3	241.24	8.2	Moderate	0.469	p-MA
22	Eugenol	97-53-0	164.20	13	Weak	0.101	p-MA
23	5-Methyleugenol	186743-25-9	178.23	13	Weak	0.137	p-MA
24	6-Methyleugenol	186743-24-8	178.23	17	Weak	0.021	p-MA
25	4-Allylanisole	140-67-0	148.21	18	Weak	-0.084	p-MA
26	2,2'-Azobisphenol ^a	2050-14-8	214.20	27.9	Weak	-0.115	p-MA
27	3-Methyleugenol	186743-26-0	178.23	32	Weak	-0.254	p-MA
28	Glutaraldehyde	111-30-8	100.12	0.1	Strong	2.001	SB
29	Chloroatranol	57074-21-2	186.59	0.4	Strong	1.669	SB
30	Atranol ^a	526-37-4	152.15	0.6	Strong	1.404	SB
31	Formaldehyde	50-00-0	30.03	0.61	Strong	0.692	SB
32	1-Phenyl-1,2-propanedione	579-07-7	148.16	1.3	Moderate	1.057	SB
33	Glyoxal	107-22-2	58.04	1.4	Moderate	0.618	SB
34	Methyl pyruvate ^a	600-22-6	102.09	2.4	Moderate	0.629	SB
35							
36							
37							
38							
39							
40							
41							
42							
43							
44							
45							
46							
47							
48							
49							

1							
2							
3							
4							
5	Phenylacetaldehyde	122-78-1	120.15	3.0	Moderate	0.603	SB
6	α -Methylphenylacetaldehyde	93-53-8	134.18	6.3	Moderate	0.328	SB
7	Undec-10-enal	112-45-8	168.28	6.8	Moderate	0.394	SB
8	1-(2',3',4',5'-Tetramethylphenyl)butane-1,3-dione	167998-73-4	218.30	8.3	Moderate	0.420	SB
9	1-(2',5'-Diethylphenyl)butane-1,3-dione	167998-76-7	218.30	9.6	Moderate	0.357	SB
10	Camphorquinone	465-29-2	166.22	10	Weak	0.221	SB
11	2-Methylundecanal	110-41-8	184.32	10	Weak	0.266	SB
12	2,3-Butanedione ^a	431-03-8	86.09	11	Weak	-0.106	SB
13	1-Phenyloctane-1,3-dione	55846-68-1	218.30	11	Weak	0.298	SB
14	Farnesal	502-67-0	220.36	12	Weak	0.264	SB
15	Citral	5392-40-5	152.44	13	Weak	0.069	SB
16	1-(2',5'-Dimethylphenyl)butane-1,3-dione	56290-55-2	190.24	13	Weak	0.165	SB
17	4-Methylhydrocinnamic aldehyde	5406-12-2	148.21	14	Weak	0.025	SB
18	α -Methyl-1,3-benzodioxole-5-propionaldehyde ^a	1205-17-0	192.21	16.4	Weak	0.069	SB
19	3 and 4-(4-Hydroxy-4-methylpentyl)-3-cyclohexene-1-						
20	carboxaldehyde	31906-04-4	210.32	17	Weak	0.092	SB
21	4- <i>tert</i> -Butyl- α -ethylhydrocinnamal	80-54-6	204.31	19	Weak	0.032	SB
22	N,N-Dibutylaniline ^{ac}	613-29-6	205.30	19.6	Weak	0.020	SB
23	4,4,4-Trifluoro-1-phenylbutane-1,3-dione	326-06-7	216.16	20	Weak	0.034	SB
24	4,4'-Dibromobenzil ^{ac}	35578-47-3	368.02	20.5	Weak	0.254	SB
25	Cyclamen aldehyde ^{ad}	103-95-7	190.29	22	Weak	-0.063	SB
26	<i>cis</i> -6-Nonenal	2277-19-2	140.23	23	Weak	-0.215	SB
27	5-Methyl-2,3-hexanedione	13706-86-0	128.17	26	Weak	-0.307	SB
28	2,2,6,6-Tetramethyl-heptane-3,5-dione	1118-71-4	184.28	27	Weak	-0.166	SB
29	1-Phenyl-2-methylbutane-1,3-dione	6668-24-2	176.22	29	Weak	-0.216	SB
30	3-Ethoxy-1-(2',3',4',5'-tetramethylphenyl)propane-1,3-dione	170928-69-5	248.32	33	Weak	-0.124	SB
31	Hydroxycitronellal	107-75-5	172.27	33	Weak	-0.282	SB
32	2-(4- <i>tert</i> -Amylcyclohexyl)acetaldehyde ^a	620159-84-4	196.33	37	Weak	-0.275	SB
33	Diethyl acetaldehyde	97-96-1	100.16	76	Weak	-0.880	SB
34	3-Dimethylaminopropylamine	109-55-7	102.18	2.2	Moderate	0.667	p-SB
35							
36							
37							
38							
39							
40							
41							
42							
43							
44							
45							
46							
47							
48							
49							

1							
2							
3							
4							
5	Ethylenediamine	107-15-3	60.10	2.2	Moderate	0.436	p-SB
6	Diethylenetriamine ^{ad}	111-40-0	103.17	5.8	Moderate	0.250	p-SB
7	3-Methyl-1-phenylpyrazolone	89-25-8	174.20	8.5	Moderate	0.312	p-SB
8	Geraniol	106-24-1	154.25	26	Weak	-0.227	p-SB
9	1-Chloromethylpyrene	1086-00-6	250.73	0.005	Extreme	3.700	S _N 2
10	5-Chloro 2 methyl 4 isothiazolin-3-one	26172-55-4	149.60	0.009	Extreme	3.221	S _N 2
11	1-Methyl-3-nitro-1-nitrosoguanidine	70-25-7	147.09	0.03	Extreme	2.690	S _N 2
12	<i>N</i> -Methyl- <i>N</i> -nitrosourea	684-93-5	103.08	0.05	Extreme	2.314	S _N 2
13	4-Nitrobenzyl bromide ^a	100-11-8	216.03	0.05	Extreme	2.636	S _N 2
14	β-Propiolactone	57-57-8	72.06	0.15	Strong	1.682	S _N 2
15	Dimethyl sulfate ^a	77-78-1	126.13	0.19	Strong	1.822	S _N 2
16	Benzyl bromide	100-39-0	171.04	0.2	Strong	1.932	S _N 2
17	Methyl dodecane sulfonate	2374-65-4	264.42	0.39	Strong	1.831	S _N 2
18	Iodopropynyl butylcarbamate	55406-53-6	281.09	0.9	Strong	1.495	S _N 2
19	<i>N</i> -ethyl- <i>N</i> -nitrosourea	759-73-9	117.11	1.1	Moderate	1.027	S _N 2
20	Bisphenol A-diglycidyl ether	1675-54-3	340.42	1.5	Moderate	1.356	S _N 2
21	2-Methyl-2H-isothiazol-3-one ^a	2682-20-4	115.15	1.9	Moderate	0.783	S _N 2
22	1,2-Benzisothiazolin-3-one	2634-33-5	151.18	2.3	Moderate	0.818	S _N 2
23	1-Bromohexadecane	112-82-3	305.34	2.3	Moderate	1.123	S _N 2
24	Benzyl salicylate	118-58-1	228.25	2.9	Moderate	0.896	S _N 2
25	Diethyl sulfate	64-67-5	154.18	3.3	Moderate	0.670	S _N 2
26	2-Bromotetradecanoic acid ^a	10520-81-7	307.27	3.4	Moderate	0.956	S _N 2
27	1-Bromoheptadecane	3508-00-7	319.37	4.8	Moderate	0.823	S _N 2
28	1-Bromopentadecane	629-72-1	291.32	5.1	Moderate	0.757	S _N 2
29	Tetramethylthiuram disulfide	137-26-8	240.42	5.2	Moderate	0.665	S _N 2
30	1-Bromoeicosane	4276-49-7	361.45	6.1	Moderate	0.773	S _N 2
31	2-Bromoethylbenzene	103-63-9	185.10	6.2	Moderate	0.475	S _N 2
32	12-Bromo-1-dodecanol ^a	3344-77-2	265.24	6.9	Moderate	0.585	S _N 2
33	Methyl methanesulfonate	66-27-3	110.13	8.1	Moderate	0.133	S _N 2
34	1-Bromodocosane	6938-66-5	389.51	8.3	Moderate	0.671	S _N 2
35							
36							
37							
38							
39							
40							
41							
42							
43							
44							
45							
46							
47							
48							
49							

1							
2							
3							
4							
5	Dodecyl methane sulfonate	51323-71-8	264.42	8.8	Moderate	0.478	S _N 2
6	1-Chlorohexadecane	4860-03-1	260.89	9.1	Moderate	0.457	S _N 2
7	1-Bromotetradecane	112-71-0	277.29	9.2	Moderate	0.479	S _N 2
8	1-Bromohexane	111-25-1	165.07	10	Weak	0.218	S _N 2
9	1-Bromotridecane	765-09-3	263.26	10	Weak	0.420	S _N 2
10	1-Iodododecane	4292-19-7	296.24	13	Weak	0.358	S _N 2
11	1-Iodotetradecane ^a	19218-94-1	324.29	14	Weak	0.365	S _N 2
12	1-Bromooctadecane ^a	112-89-0	333.40	15	Weak	0.347	S _N 2
13	1-Chlorooctadecane	3386-33-2	288.95	16	Weak	0.257	S _N 2
14	Benzyl benzoate	120-51-4	212.25	17	Weak	0.096	S _N 2
15	1-Bromododecane ^a	143-15-7	249.24	18	Weak	0.141	S _N 2
16	12-Bromododecanoic acid	73367-80-3	279.22	18	Weak	0.191	S _N 2
17	1-Iodoheptadecane	544-77-4	352.35	19	Weak	0.268	S _N 2
18	1-Bromoundecane	693-67-4	235.21	20	Weak	0.070	S _N 2
19	1-Chlorotetradecane	2425-54-9	232.84	20	Weak	0.066	S _N 2
20	7-Bromotetradecane	74036-97-8	277.29	21	Weak	0.121	S _N 2
21	1-Iodononane ^a	4282-42-2	254.16	24	Weak	0.025	S _N 2
22	Oleyl methane sulfonate	35709-09-2	346.57	25	Weak	0.142	S _N 2
23	Butyl glycidyl ether	2426-08-6	130.19	31	Weak	-0.377	S _N 2
24	Benzo[a]pyrene	50-32-8	252.32	0.0009	Extreme	4.448	p-S _N 2
25	7,12-Dimethylbenz[α]anthracene	57-97-6	256.35	0.006	Extreme	3.631	p-S _N 2
26	4-Ethoxymethylene-2-phenyl-2-oxazolin-5-one	15646-46-5	217.22	0.003	Extreme	3.860	Ac
27	Tetrachlorosalicylanilide ^a	1154-59-2	351.02	0.04	Extreme	2.943	Ac
28	Fluorescein-5-isothiocyanate	3326-32-7	389.38	0.14	Strong	2.444	Ac
29	2-Methyl -4H,3,1-benzoxazin-4-one	525-76-8	161.16	0.7	Strong	1.362	Ac
30	C6 Azlactone	176665-02-4	197.28	1.3	Moderate	1.181	Ac
31	2-Mercaptobenzothiazole	149-30-4	167.24	1.7	Moderate	0.993	Ac
32	C4 Azlactone	176664-99-6	169.22	1.8	Moderate	0.973	Ac
33	Nonanoyl chloride	764-85-2	176.69	1.8	Moderate	0.992	Ac
34	Methyl 2-sulfophenyl octadecanoate	Not known ^b	454.67	2	Moderate	1.357	Ac
35							
36							
37							
38							
39							
40							
41							
42							
43							
44							
45							
46							
47							
48							
49							

1							
2							
3							
4							
5	Isononanoyl chloride ^a	57077-36-8	176.69	2.7	Moderate	0.816	Ac
6	3,5,5-Trimethylhexanoyl chloride	36727-29-4	176.69	2.7	Moderate	0.816	Ac
7	C9 Azlactone	176665-04-6	239.36	2.8	Moderate	0.932	Ac
8	3-Propylideneophthalide	17369-59-4	174.20	3.7	Moderate	0.673	Ac
9	3,4-Dihydrocoumarin	119-84-6	148.16	5.6	Moderate	0.423	Ac
10	Palmitoyl chloride ^a	112-67-4	274.88	8.8	Moderate	0.495	Ac
11	1,2,4-Benzenetricarboxylic anhydride	552-30-7	192.13	9.2	Moderate	0.320	Ac
12	C11 Azlactone	176665-06-8	267.41	16	Weak	0.223	Ac
13	C15 Azlactone	176665-09-1	323.52	18	Weak	0.255	Ac
14	C17 Azlactone	176665-11-5	351.58	19	Weak	0.267	Ac
15	Phenyl benzoate	93-99-2	198.22	20	Weak	-0.004	Ac
16	Imidazolidinylurea	39236-46-9	388.30	24	Weak	0.209	Ac
17	C19 Azlactone ^a	Not known ^b	379.63	26	Weak	0.164	Ac
18	Penicillin G	61-33-6	334.39	30	Weak	0.047	Ac
19	5-Chlorosalicylanilide	4638-48-6	247.68	5	Moderate	0.695	OxPot
20	α -Phellandrene	99-83-2	136.23	5.4	Moderate	0.402	OxPot
21	β -Phellandrene ^a	555-10-2	136.23	5.6	Moderate	0.386	OxPot
22	(5R)-5-Isopropenyl-2-methyl-1-methylene-2-cyclohexene	Not known ^b	148.25	7.3	Moderate	0.308	OxPot
23	2-(Hexadecyloxy)ethanol	2136-71-2	286.50	8.8	Moderate	0.513	OxPot
24	α -Terpinene	99-86-5	136.24	8.9	Moderate	0.185	OxPot
25	Acetyl cedrene	32388-55-9	246.39	13.9	Weak	0.249	OxPot
26	Abietic acid	514-10-3	302.46	15	Weak	0.305	OxPot
27	Linalool	78-70-6	154.25	30	Weak	-0.289	OxPot
28	R(+) Limonene	5989-27-5	136.24	69	Weak	-0.705	OxPot
29	Aniline ^a	62-53-3	93.13	89	Weak	-0.980	OxPot
30	Chlorothalonil	1897-45-6	265.91	0.004	Extreme	3.823	S _N Ar
31	1-Chloro-2,4-dinitrobenzene	97-00-7	202.55	0.05	Extreme	2.608	S _N Ar
32	2,4,6-Trichloro-1,3,5-triazine	108-77-0	184.41	0.09	Extreme	2.312	S _N Ar
33	Pentachlorophenol	87-86-5	266.34	20	Weak	0.124	S _N Ar
34	Clotrimazole	23593-75-1	344.85	4.8	Moderate	0.856	S _N 1

1
2
3
4
5 *d,l*-Citronellol

106-22-9

156.27

43.5

Weak

-0.445 S_N1

6 1 ^aThese chemicals were used as test set chemicals. Those marked ^{ac} were used only in the SB test set, and those marked ^{ad} were used only in the
7 2 SB + p-SB test set.

8 3 ^bFor compounds with unknown CAS numbers, the SMILES strings are: linalool aldehyde, C=CC(C)(O)CCC=C(C)C=O ; methyl 2-sulfophenyl
9 4 octadecanoate, CCCCCCCCCCCCCCCCCC(C)C(=O)Oc1ccccc1S(O)(=O)=O ; C19 azlactone,
10 5 CCCCCCCCCCCCCCCCCCCC1=NC(C)(C)C(=O)O1 ; (5R)-5-isopropenyl-2-methyl-1-methylene-2-cyclohexene,
11 6 CC(=C)[C@@H]1CC=C(C)C(=C)C1
12 7
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

1

2 **Table 2. Models developed in this work for skin sensitization**

3

4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
Mech.	Model	Eqn.	No. of chemicals	Equation	R^2 (R^2_{adj})	Q^2	s	F	p values																																				
All	Full	1	204	SSP = - 1.164(0.282) + 1.759(0.450) FASA- + 0.174(0.028) eaC2C3a + 0.807(0.155) vsurf_CW2 + 0.012(0.0026) vsurf_D8 - 0.767 (0.202) Hmin - 0.190(0.057) SHCsatu	0.496 (0.480)	0.459	0.689	32.4	<0.001																																				
MA	Full	2	45	SSP = 16.7(2.52) - 0.101(0.020) S4 - 0.760(0.174) HS17 + 0.112(0.015) SlogP_VSA4 + 0.775(0.195) vsurf_CW2 - 8.39(1.14) Max. BC1 - 43.4(7.37) Rel. PMI	0.856 (0.834)	0.793	0.358	37.8	<0.001																																				
MA	Train	3	36	SSP = 16.6(3.77) - 0.094(0.029) S4 - 0.743(0.201) HS17 + 0.113(0.017) SlogP_VSA4 + 0.673(0.257) vsurf_CW2 - 8.26(1.78) Max. BC1 - 42.2(9.9) Rel. PMI	0.825 (0.789)	0.692	0.398	22.9	≤0.015																																				
MA	Test	4	9	SSP (obsd) = - 0.113 + 1.12 SSP (pred) (ICC = 0.977)	0.965	0.937	0.191	195.9																																					
p-MA	Full	5	32	SSP = - 0.360(0.369) + 1.400(0.194) S24 - 0.319(0.046) e1C3O2a + 0.279(0.085) SssNH - 0.337(0.051) vsurf_HB7 + 0.467(0.108) Av. IC2	0.858 (0.831)	0.790	0.349	31.4	≤0.003																																				
p-MA	Train	6	26	SSP = - 0.139(0.454) + 1.348(0.249) S24 + 0.254(0.097) SssNH - 0.318(0.057) e1C3O2a - 0.359(0.098) vsurf_HB7 + 0.401(0.131) Av. IC2	0.848 (0.810)	0.768	0.380	22.3	≤0.01																																				
p-MA	Test	7	6	SSP (obsd) = 0.039 + 0.958 SSP (pred) (ICC = 0.951)	0.887	0.758	0.305	31.5																																					

1														
2														
3														
4														
5	1													
6	2													
7	3													
8	4	SB	Full	8	35	SSP = - 6.99(1.47) + 0.090(0.020) S7 + 0.035(0.014) S10				0.837	0.644	0.259	19.9	≤0.02
9	5					- 3.107(0.717) GCUT_PEOE_1 + 1.880(0.496) vsurf_Wp7 (0.795)								
10	6					+ 2.657(0.702) Av. SI2 + 3.101(1.084) Av. BO								
11	7					+ 0.177(0.026) Kier FI								
12	8													
13	9	SB	Train	9	28	SSP = - 7.54(1.75) + 0.0853(0.0236) S7 + 0.042(0.016) S10				0.838	0.524	0.272	14.8	≤0.15
14	10					- 2.704(0.869) GCUT_PEOE_1 + 1.294(0.852) vsurf_Wp7 (0.781)								
15	11					+ 2.798(0.829) Av. SI2 + 3.573(1.250) Av. BO								
16	12					+ 0.193(0.031) Kier FI								
17	13													
18	14	SB	Test	10	7	SSP (obsd) = 0.060 + 1.02 SSP (pred)				0.904	0.857	0.194	47.0	
19	15													
20	16	SB +	Full	11	40	SSP = 19.22(2.95) + 0.380(0.086) HS6 - 0.238(0.058) dx2				0.850	0.781	0.233	25.9	≤0.005
21	17	p-SB				- 0.0813(0.0107) E_sol + 0.0958(0.0173) Kier FI				(0.817)				
22	18					- 0.00153(0.00047) DPSA1 - 4.542(0.670) Av. valency								
23	19					- 5.885(1.066) relative no. O atoms								
24	20													
25	21	SB +	Train	12	33	SSP = 19.09(3.36) + 0.344(0.107) HS6 - 0.226(0.069) dx2				0.836	0.736	0.251	18.2	≤0.005
26	22	p-SB				- 0.070(0.016) E_sol + 0.103(0.021) Kier FI				(0.790)				
27	23					- 0.00163(0.00053) DPSA1 - 4.490(0.760) Av. valency								
28	24					- 5.960(1.230) relative no. O atoms								
29	25													
30	26	SB +	Test	13	7	SSP (obsd) = - 0.143 + 1.27 SSP (pred) (ICC = 0.936)				0.935	0.838	0.162	71.4	
31	27	p-SB												
32	28													
33	29	S _N 2	Full	14	45	SSP = - 9.468(1.304) + 0.109(0.034) S14 + 0.151(0.050) SsCH ₃				0.852	0.796	0.381	30.3	≤0.005
34	30					+ 4.004(0.717) xvp9 + 0.150(0.037) eaC2C3a				(0.823)				
35	31					+ 8.780(0.864) FASA- + 3.496(0.589) PEOE_VSA_FPOS								
36	32					- 0.473(0.094) MNDO_HOMO								
37														
38														
39														
40														
41														
42														
43														
44														
45														
46														
47														
48														
49														

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

1														
2														
3														
4														
5	1													
6	2													
7	3													
8	4	S _N 2	Train	15	36	SSP = - 9.689 + 0.109(0.039) S14 + 0.149(0.058) SsCH ₃	0.837	0.773	0.419	20.6	≤0.02			
9	5					+ 4.233(0.854) xvp9 + 0.142(0.042) eaC2C3a	(0.797)							
10	6					+ 9.084(1.155) FASA- + 3.699(0.694) PEOE_VSA_FPOS								
11	7					- 0.477(0.123) MNDO_HOMO								
12	8													
13	9	S _N 2	Test	16	9	SSP (obsd) = - 0.023 + 0.889 SSP (pred)	0.951	0.927	0.204	134.7	<0.001			
14	10													
15	11	Ac	Full	17	22	SSP = 0.873(0.088) - 0.616 (0.152) HS14 + 2.644(0.225) HS16	0.921	0.886	0.304	49.5	<0.001			
16	12					- 3.059(0.289) HS17 + 0.633 (0.122) HS29								
17	13	Ac	Train	18	18	SSP = 0.879(0.110) - 0.578(0.210) HS14 + 2.645(0.262) HS16	0.899	0.863	0.342	28.8	≤0.015			
18	14					- 3.079(0.371) HS17 + 0.629(0.142) HS29	(0.867)							
19	15													
20	16	Ac	Test	19	4	SSP (obsd) = - 0.079 + 0.966 SSP (pred) (ICC = 0.995)	0.999	0.992	0.042	2672.7				
21	17													
22	18	OxPot	Full	20	11	SSP = 0.365(0.072) - 0.179(0.017) vsurf_DD12	0.930	0.856	0.156	52.8	<0.001			
23	19					+ 0.0957(0.0200) vsurf_DD23	(0.912)							
24	20													
25	21													
26	22	OxPot	Train	21	9	SSP = 0.363(0.066) - 0.156(0.017) vsurf_DD12	0.931	0.865	0.130	40.4	<0.001			
27	23					+ 0.081(0.018) vsurf_DD23	(0.908)							
28	24													
29	25	OxPot	Test		2	No QSAR with only 2 test chemicals (ICC = 0.945)								
30	26													
31														
32														
33														
34														
35														
36														
37														
38														
39														
40														
41														
42														
43														
44														
45														
46														
47														
48														
49														

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

1

Table 3. Descriptors and SSPs used in the QSAR models, and their ranges**All 204 active sensitizers**

SSP (-0.980 to 4.050)
FASA-: MOE; Fractional accessible surface area of all atoms with negative partial charge (0.067 to 0.703)
eaC2C3a: winMolconn; Bond-type electrotopological state index for single bond between unsubstituted carbon and carbon with three aromatic neighbours (0 to 18.723)
vsurf_CW2: MOE; Capacity factor (Shape, volume, surface area descriptor) (1.160 to 3.211)
vsurf_D8: MOE; Hydrophobic volume (0 to 112.88)
Hmin: CODESSA; Minimum number of hydrogen bond donors and acceptors (0 to 1.514)
SHCsatu: winMolconn: Number of hydrogen atoms on sp³ carbons bonded to sp² carbons (0 to 4.407)

Michael addition

SSP (-0.954 to 4.050)
S4: winMolconn; Atom level E-State for atom 4 (-3.617 to 10.190)
HS17: winMolconn; Hydrogen atom level HE-state for hydrogen atom 17 (0 to 2.690)
SlogP_VSA4: MOE; Sum of van der Waals surface areas such that contribution to log P is in range 0.1-0.15 (0 to 30.233)
vsurf_CW2: MOE; Capacity factor (Shape, volume, surface area descriptor) (1.352 to 2.836)
Max. BC1: CODESSA; Maximum bonding contribution of one (1.84 to 2.14)
Rel. PMI: CODESSA; Relative principal moment of inertia (0 to 0.05)

Pro-Michael addition

SSP (-0.115 to 2.901)
S24: winMolconn; Atom level E-state index for atom 24 (0 to 1.817)
e1C3O2a: winMolconn; Bond-type E-state for single bond between ether oxygen and substituted aromatic carbon (0 to 3.311)
SssNH: winMolconn; Atom type E-state index for >NH nitrogen (0 to 2.952)
vsurf_HB7: MOE; H-bond donor capacity (-3.125 to 3.375)
Av. IC₂: CODESSA; Average information content (₂), a structural descriptor (1.02 to 2.19)

Schiff base

SSP (-0.880 to 2.001)
S7: winMolconn; Atom level E-state for atom 7 (-0.526 to 11.481)
S10: winMolconn; Atom level E-state for atom 10 (-2.017 to 10.595)
GCUT_PEOE_1: MOE; The GCUT descriptors are calculated from the eigenvalues of a modified graph distance adjacency matrix. Each *ij* entry of the adjacency matrix takes the value $1/\sqrt{d_{ij}}$ where d_{ij} is the (modified) graph distance between atoms *i* and *j*. The diagonal takes the value of the PEOE partial charges. The resulting eigenvalues are sorted and the smallest, 1/3-ile, 2/3-ile and largest eigenvalues are reported (-0.468 to -0.187)
vsurf_Wp7: MOE; Polar volume (Shape, volume, surface area descriptors) (0 to 0.50)
Av. SI₂: CODESSA; Average structural information₂, a structural descriptor (0.35 to 0.92)
Av. BO: CODESSA; Average bond order (0.96 to 1.13)
Kier FI: CODESSA; Kier flexibility index (1.25 to 13.94)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
601
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47**Schiff base + pro-Schiff base**

SSP (-0.880 to 2.001)

HS6: winMolconn; Hydrogen atom level HE-state for hydrogen atom 6 (0 to 1.391)

dx2: winMolconn; 2nd Order connectivity index difference between a molecule and its unbranched isomer (0 to 2.588)

E_sol: MOE; Solvation energy (-20.623 to -4.438)

Kier FI : CODESSA ; Kier flexibility index (1.25 to 16.57)

DPSA1 : CODESSA; Difference in positive and negative partial surface areas (-100.41 to 563.06)

Av. valency: CODESSA; Average valency (3.63 to 4.47)

Rel. no. O atoms: CODESSA; Relative number of oxygen atoms (0 to 0.50)

S_N2

SSP (-0.377 to 3.700)

S14: winMolconn; Atom level E-state for atom 14 (-3.234 to 11.013)

SsCH₃: winMolconn; E-state for -CH₃ carbon atoms (0 to 7.701)

xvp9: winMolconn; 9th order valence path molecular connectivity (0 to 0.506)

eaC2C3a: winMolconn; Bond-type E-state for single bond between unsubstituted carbon and carbon with three aromatic neighbours (0 to 12.937)

FASA-: MOE; Fractional accessible surface area of all atoms with negative partial charge (0.103 to 0.673)

PEOE_VSA_FPOS: MOE; Fractional positive van der Waals surface area (0.265 to 0.775)

MNDO_HOMO: MOE; Energy of the highest occupied molecular orbital calculated using the MNDO Hamiltonian [MOPAC] (-12.102 to -8.237)

Acyl transfer

SSP (0.075 to 3.860)

S14: Winmolconn; Hydrogen atom level HE-state for hydrogen atom 14 (0 to 2.749)

HS16: Winmolconn; Hydrogen atom level HE-state for hydrogen atom 16 (0 to 2.711)

HS17: Winmolconn; Hydrogen atom level HE-state for hydrogen atom 17 (0 to 1.514)

HS29: Winmolconn; Hydrogen atom level HE-state for hydrogen atom 29 (0 to 2.898)

Oxidation potential

SSP (-0.980 to 0.695)

vsurf_DD12 : MOE; Contact distances of vsurf_DDmin (3 descriptors) (0.500 to 7.697)

vsurf_DD23 : MOE; Contact distances of vsurf_DDmin (3 descriptors) (0.500 to 6.819)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1

1 **Table 4. Comparison of statistical quality of full data-set QSARs**

2 .

Category	All	MA	pMA	SB	SB+pSB	S _N 2	Acyl	OxPot
Equation	1	2	5	8	11	14	17	20
n	204	45	32	35	40	45	22	11
Descriptors	6	6	5	7	7	7	4	2
R ²	0.496	0.856	0.858	0.837	0.850	0.852	0.921	0.930
R ² _{adj}	0.480	0.834	0.831	0.795	0.817	0.823	0.902	0.912
Q ²	0.459	0.793	0.790	0.644	0.781	0.796	0.886	0.856
s	0.689	0.358	0.349	0.259	0.233	0.381	0.304	0.156
F	32.4	37.8	31.3	19.9	25.9	30.3	49.5	52.8

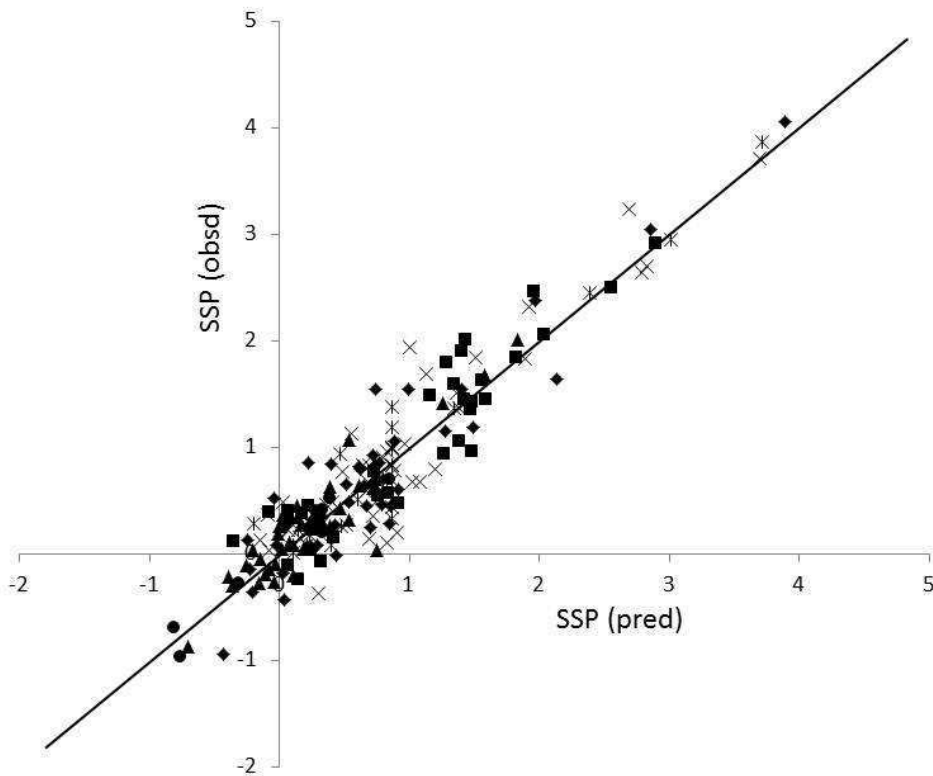
3

4

1

2 **Figure 1.** Observed vs. predicted SSP values for all 196 chemicals. Black diamond = Michael
3 addition; black square = pro-Michael addition; black triangle = Schiff base + pro-Schiff base;
4 cross = S_N2 ; asterisk = acyl transfer; black circle = oxidation potential

5



6

7

8

9

10

11

12

13

14

15

16

17

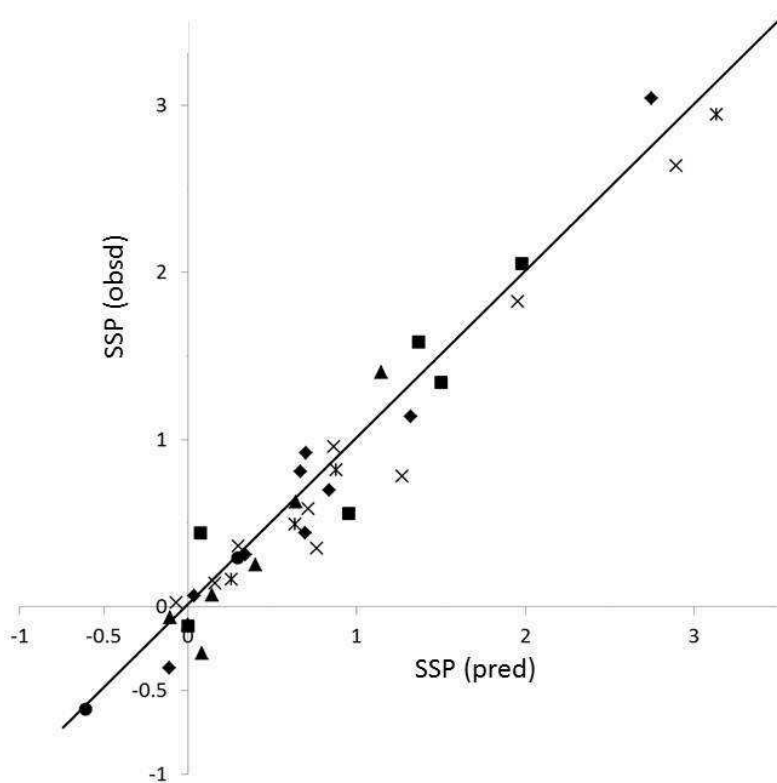
18

19

20

21

Figure 2. Observed vs. predicted SSP values for all 37 test set chemicals. The 45° line on the graph is virtually indistinguishable from that of equation 22. Black diamond = Michael addition; black square = pro-Michael addition; black triangle = Schiff base + pro-Schiff base; cross = S_N2 ; asterisk = acyl transfer; black circle = oxidation potential



8