Comparison of Test and Training Sets

Product Use

Table S-1 shows the product use distribution for the training set of 72 substances and the test set of 24 substances. All product categories are represented in each set in generally similar proportions.

Table S-1. Distribution of product categories for training and test sets

Product	Percentage of Substances: Training Set[a]	Number of Substances: Training Set	Percentage of Substances: Test Set[b]	Number of Substances: Test Set
Manufacturing	56%	40	33%	8
Food additives	33%	24	46%	11
Pharmaceuticals	28%	20	33%	8
Intermediate in chemical synthesis	25%	18	13%	3
Pesticide (other)	22%	16	8%	2
Personal care products	24%	17	13%	3
Fragrance agent	18%	13	33%	8
Pesticide (antimicrobial)	17%	12	17%	4
Cosmetics	17%	12	25%	6
Solvent	8%	6	4%	1
Household product	4%	3	8%	2
Other[c]	1%	1	4%	1

[a] Percentage of 72 substances. Total of all percentages exceeds 100 because most substances were associated with more than one product category.

[b] Percentage of 24 substances. Total of all percentages exceeds 100 because most substances were associated with more than one product category.

[c] Represents a rubber product for the training set and an antioxidant for the test set.

Structural Diversity

Structural diversity of the test and training sets was assessed using ChemoTyper v1.0, a free software developed under contract with the U.S Food and Drug Administration. ChemoTyper uses 729 chemotypes, which are generic structural fragments that represent chemical structural features, including connected and nonconnected chemical patterns as well as atom, bond, and molecular-based properties (Yang et al. 2015 J. Chem. Inf.

Model. DOI: 10.1021/ci500667v). The 96 substances in the database represented 183 chemotypes. The distribution of the chemotypes for the training and test sets are shown in Fig. S-2.

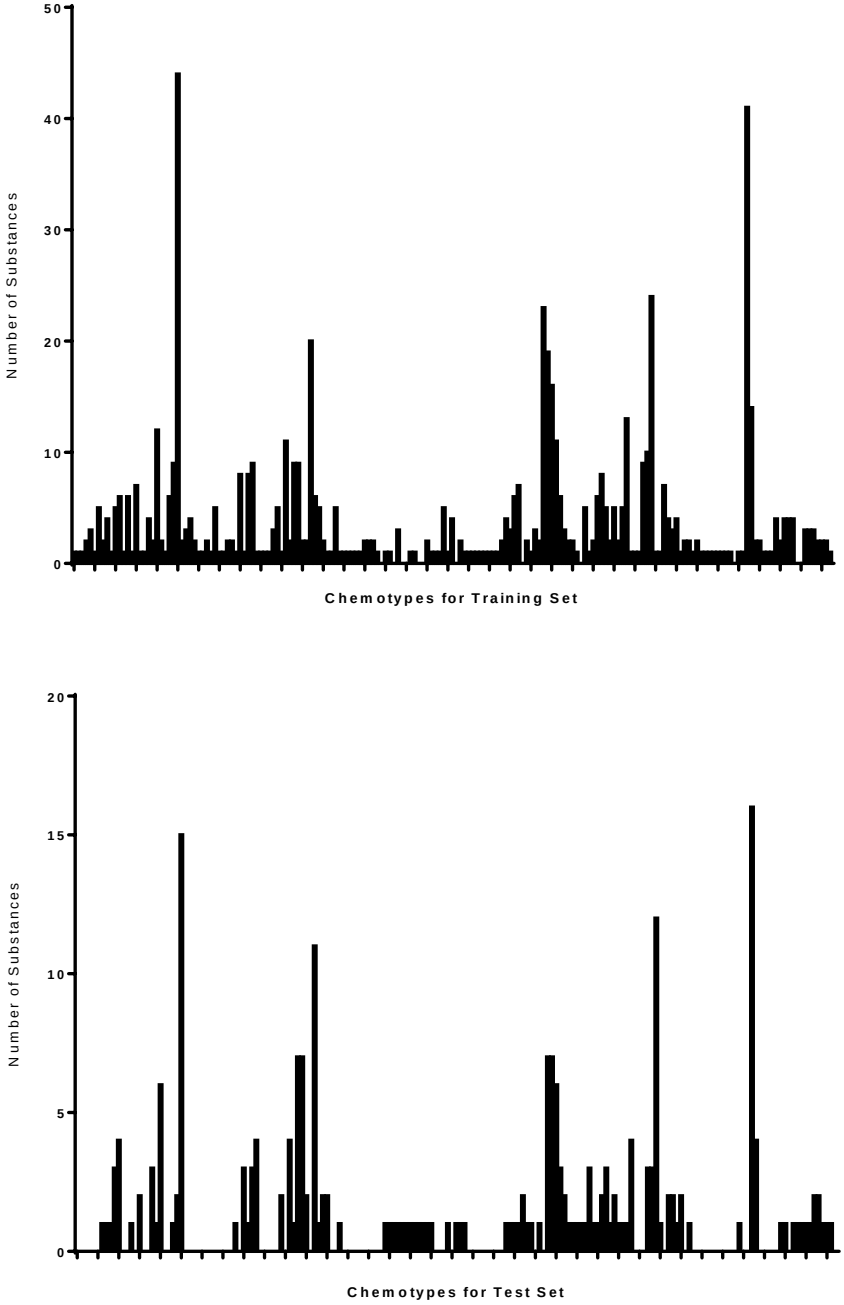

Figure S-2. Frequency of appearance of chemotypes in training and test sets. Height of bars represents the number of substances that included each of 183 chemotypes.

Prehapten and Prohapten Status

Fig. S-3 shows the distribution of haptens, prehaptens, pre/prohaptens, and prohaptens for the sensitizers in the training and test sets, as determined by a review of the literature. The hapten status of four substances in the training set and two substances in the test set could not be identified.

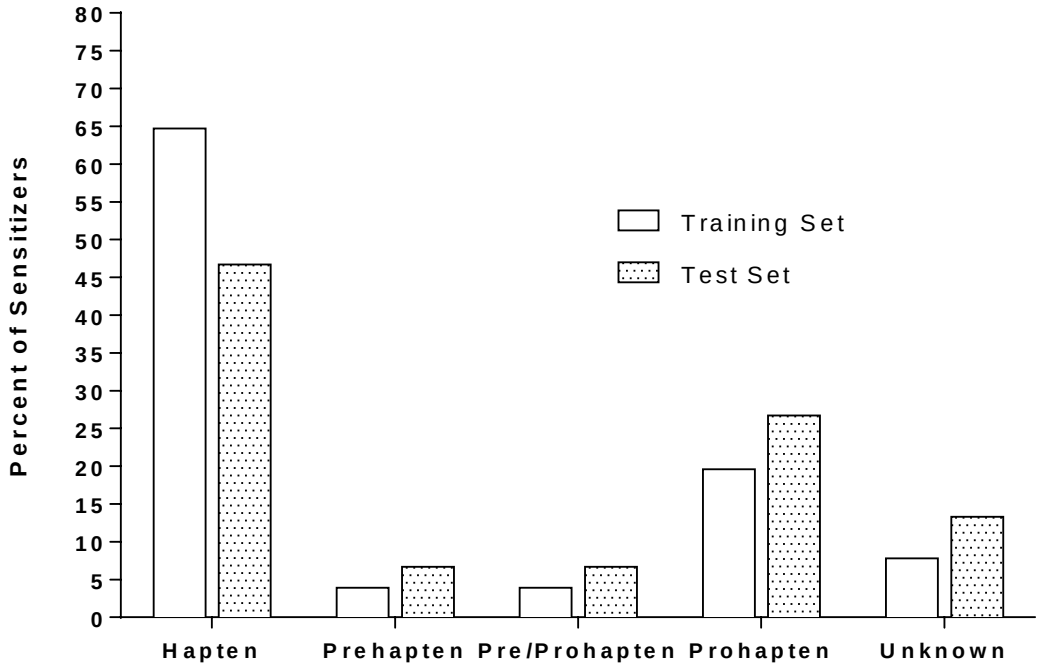

Figure S-3. Distribution of hapten status for training and test sets. Bars show the percentages of various types of sensitizers in the training (51 sensitizers/72 substances) and test (15 sensitizers/24 substances) sets.

Mechanism of Protein Binding

Protein binding alerts for skin sensitization by OASIS v1.2 in QSAR Toolbox v3.2 were used to characterize the protein binding mechanism of substances in the training and test sets. This system identifies structural features (i.e., protein binding alerts) in the test substance molecules responsible for interaction with skin proteins. There are 100 structural alerts that have been categorized into 11 mechanistic domains. Each of the mechanistic domains has been separated into at least two mechanistic alerts. The distributions of the members of the training and test sets among the 11 mechanistic domains for protein binding alerts for skin sensitization are shown in Table S-2**.**

Table S-2. Distribution of mechanistic domains for protein binding for training and test sets

Mechanistic Domain[a]	Percentage of Substances: Training Set[b]	Number of Substances: Training Set	Percentage of Substances: Test Set[c]	Number of Substances: Test Set
No alert	47%	34	67%	16
Acylation	15%	11	8%	2
Michael addition	17%	12	4%	1
Schiff base formation	11%	8	17%	4
SN2	8%	6	4%	1
Nucleophilic addition	0%	0	0%	0
SNAr	1%	1	0%	0
SNVinyl	0%	0	0%	0

[a] Using protein binding alerts for skin sensitization by OASIS v1.2 from the OECD QSAR Toolbox v3.2.

[b] Percentage of 72 substances. One substance was associated with more than one mechanistic domain.

[c] Percentage of 24 substances.

Physicochemical Properties

Figs. S-4 through S-9 present the distribution of physicochemical properties for the substances in the training and test sets**.**

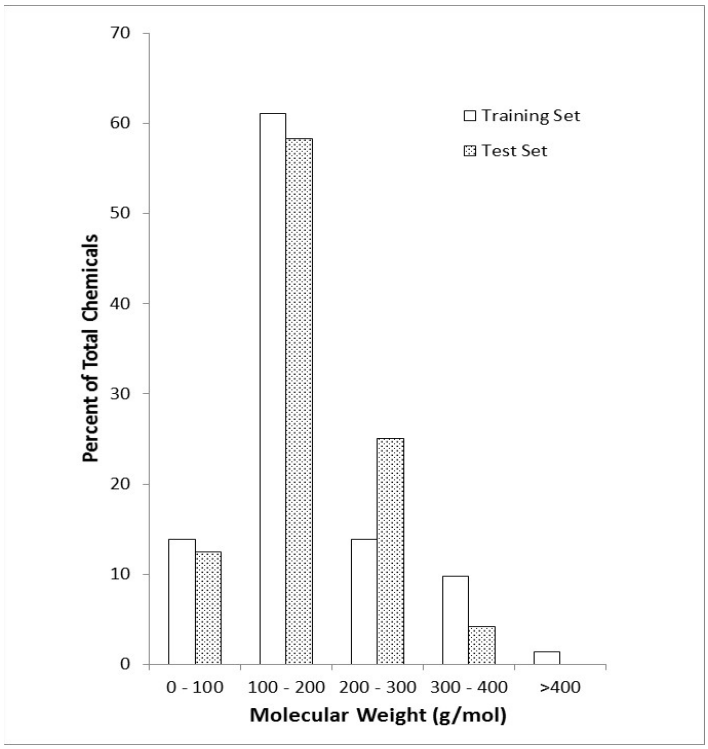

Figure S-4. Distribution of molecular weights in training and test sets

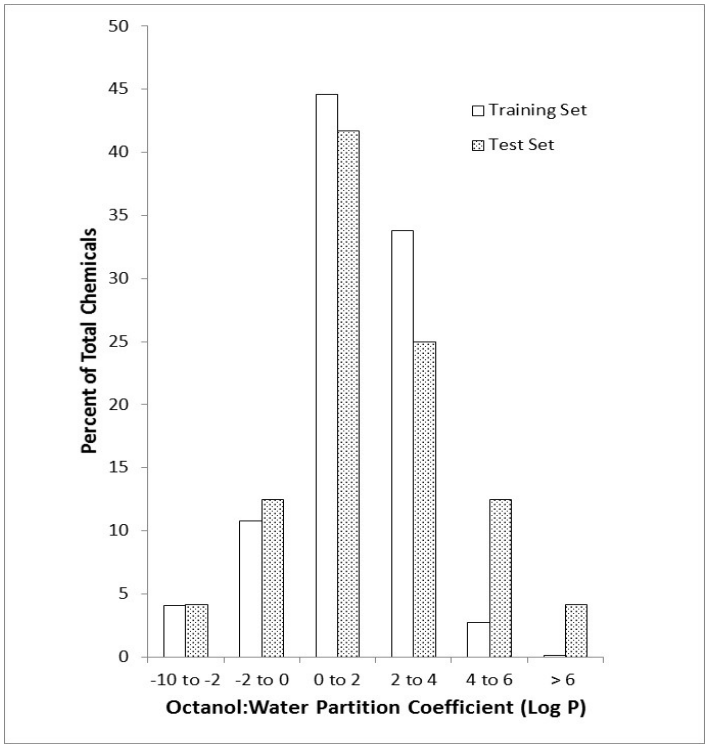

Figure S-5. Distribution of octanol:water partition coefficients in training and test sets

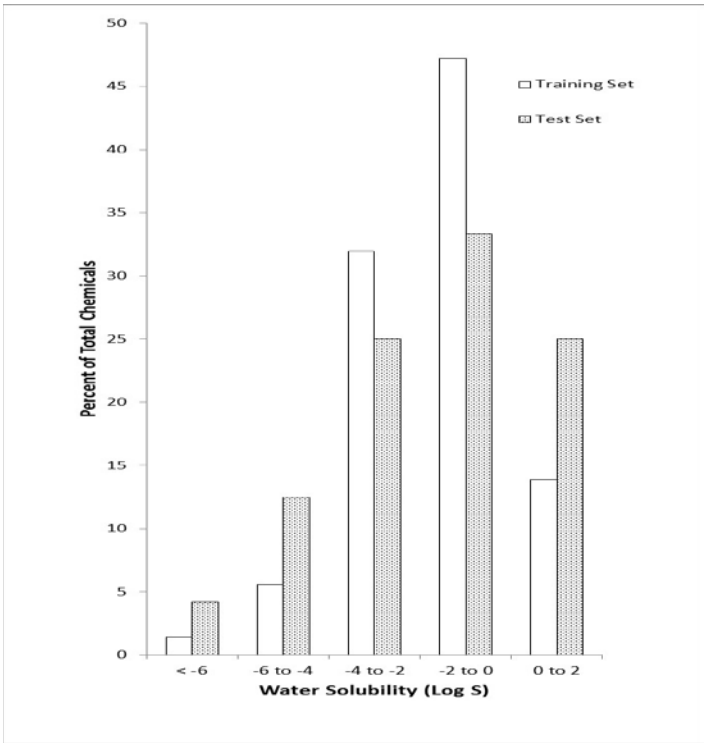

Figure S-6. Distribution of water solubilities in training and test sets

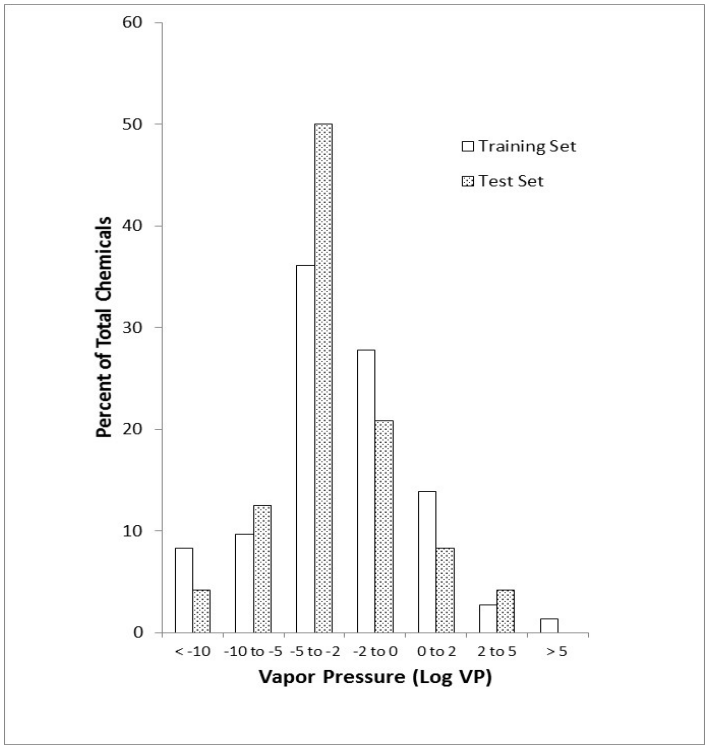

Figure S-7. Distribution of vapor pressures in training and test sets

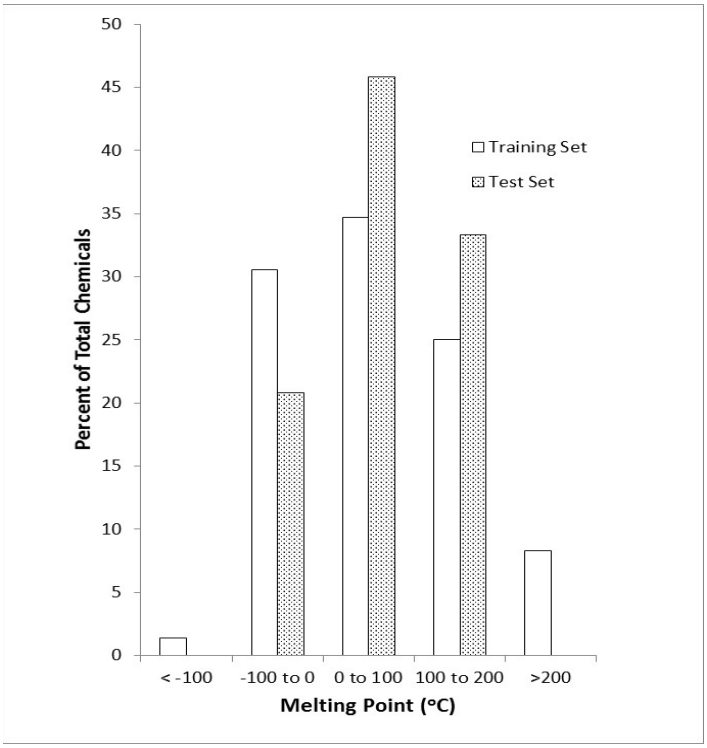

Figure S-8. Distribution of melting points in training and test sets

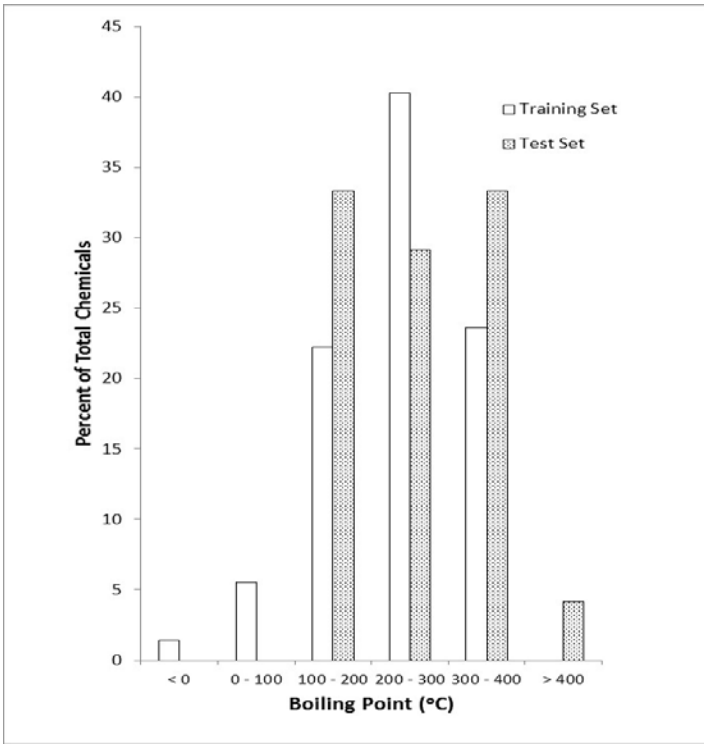

Figure S-9. Distribution of boiling points in training and test sets