Documentation

Contents

What is ProperSea?

ProperSea is a service that predicts a variety of physico-chemical properties so that scientists are able to get an idea of a compound's properties even if it has never been synthesized. To do this, ProperSea uses a combination of statistical modelling and semi-empirical quantum chemistry.

These properties are provided as guidance only, and come with no guarantee. The statistically predicted properties are not always reliable for compounds that differ markedly from the training data. Most of the training data were organic chemicals, so predictions may not be accurate for inorganic and organometallic compounds. ProperSea keeps the user informed of these sources of error by providing uncertainty intervals and quantifying each model's applicability domain.

Basic Properties

ProperSea uses RDKit to display basic properties of each molecule:

- Number of rotatable bonds
- Number of hydrogen bond donors
- Number of hydrogen bond acceptors
- Polar surface area
- Molar refractivity
- Polarizability

Polar surface area and molar refractivity are predicted with fragment contributions. Polarizability is calculated from molar refractivity.

IUPAC Name

ProperSea predicts IUPAC names with a machine-learning model trained on freely available data from PubChem. Not all predictions are of the same quality, so ProperSea does not provide a prediction where the confidence is low. In particular, predictions for inorganic and organometallic compounds are not reliable as these were not well represented in the training data.

Semi-Empirical Properties

ProperSea uses the semi-empirical PM6 method to predict the following properties:

- Heat of formation
- Ionization energy
- Dipole moment

If the molecule has a hydroxyl groups, ProperSea uses COSMO solvent screening to calculate the acid disassociation constant (pK_a) for each hydrogen. These properties are not predicted for large molecules where the calculation is too computationally expensive.

Statistical Properties

ProperSea uses Bayesian Additive Regression Trees (BART) to predict the following physico-chemical properties:

- Boiling point
- Melting point

- Flash point
- Density
- Solubility
- Viscosity
- Surface tension
- logP (log$K_{octanol/water}$)
- log$K_{octanol/air}$
- Refractive index

The first three properties are predicted at ambient pressure. The rest are predicted at ambient temperature and ambient pressure. Depending on the molecule, not all properties are displayed. For example, if the molecule is inorganic, no flash point is predicted. And if the phase is predicted to be solid, viscosity and surface tension are not predicted.

BART is a tree-based regression method, similar to gradient boosting. BART builds numerous regression trees on a training set, and combines the trees into a predictive model. Because BART is a Bayesian technique, it predicts a distribution rather than a point value, so users have an indication of the confidence of the prediction.

Data Sources

PhysProp Database

PhysProp is a collection of experimental datasets published by the United States Environmental Protection Agency (EPA) and bundled with their EPI Suite software. The models were trained on a curated version of the data. Although the datasets relate principally to biological and environmental properties, some contain physico-chemical properties:

Property	Number of compounds
Boiling Point	5591
Melting Point	9120
logP	14544
logK_{OA}	277

The PhysProp collection includes sources for every single datapoint. Although experimental conditions are not specified, it can be assumed that melting and boiling point were measured at ambient pressure, and logP and logK_{OA} were measured at ambient temperature and pressure.

EPA TEST

The EPA's Toxicity Estimation Software Tool (TEST) includes experimental data from a variety of sources. The following physical properties were used to train models for ProperSea:

Property	Conditions	Number of compounds
Flash Point [?]	-	8362
Surface Tension [?]	25 °C	1416
Viscosity [?]	25 °C	557
Density [?]	-	8909

No conditions are specified for flash point and density, but it can be assumed that these were measured at ambient pressure (and ambient temperature for density). The density dataset only includes solids and liquids, so experimental densities vary negligibly with typical differences in ambient conditions.

Refractive index is calculated from density with the Lorentz-Lorenz equation:

$$\frac{n^2-1}{n^2+2}=\frac{4\pi}{3}N\alpha_m$$

n: refractive index

N: number density

α_m: mean polarizability from fragment contributions

AqSolDB

AqSolDB is a curated dataset of aqueous solubility for 9982 compounds, measured at 25 ± 5 °C. It was compiled from nine openly-accessible solubility datasets.

Molecular Descriptors

Molecules are generally approximated as a graph with variable types of edge and node. To train a regression model, each graph must be converted to a set of numerical descriptors. There are broadly two types of molecular descriptor. Two-dimensional descriptors (e.g. ring count, bond count, weight...) can be calculated directly from the molecular graph, while three-dimensional descriptors (e.g. moment of inertia, charged partial surface area...) require an optimized 3D representation.

Descriptors for ProperSea are generated with Mordred, which outputs 1825 molecular descriptors.

Bayesian Additive Regression Trees

Bayesian Additive Regression Trees (BART) model a relationship as a sum of m regression trees:

$$Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon$$

Y: measured value of property

x: molecular descriptors

$g(x; T_j, M_j)$: output of tree T_j with terminal nodes M_j

ϵ: normally distributed experimental error with standard deviation σ

The model places a soft constraint on the regression trees with a prior distribution over the possible values of terminal nodes, which can be informed by the range of values found in the training set. The model also uses a prior distribution over tree depth that favours shallow regression trees. These priors act as regularizing constraints, preventing overfitting. All models were trained with a fixed ensemble of 300 trees. The training was done with the R package dbarts, using all 1825 2D and 3D descriptors from Mordred.

During training, BART learns a posterior distribution over the trees in the ensemble:

$$p((T_1, M_1), \ldots, (T_m, M_m), \sigma \mid y_{\text{train}})$$

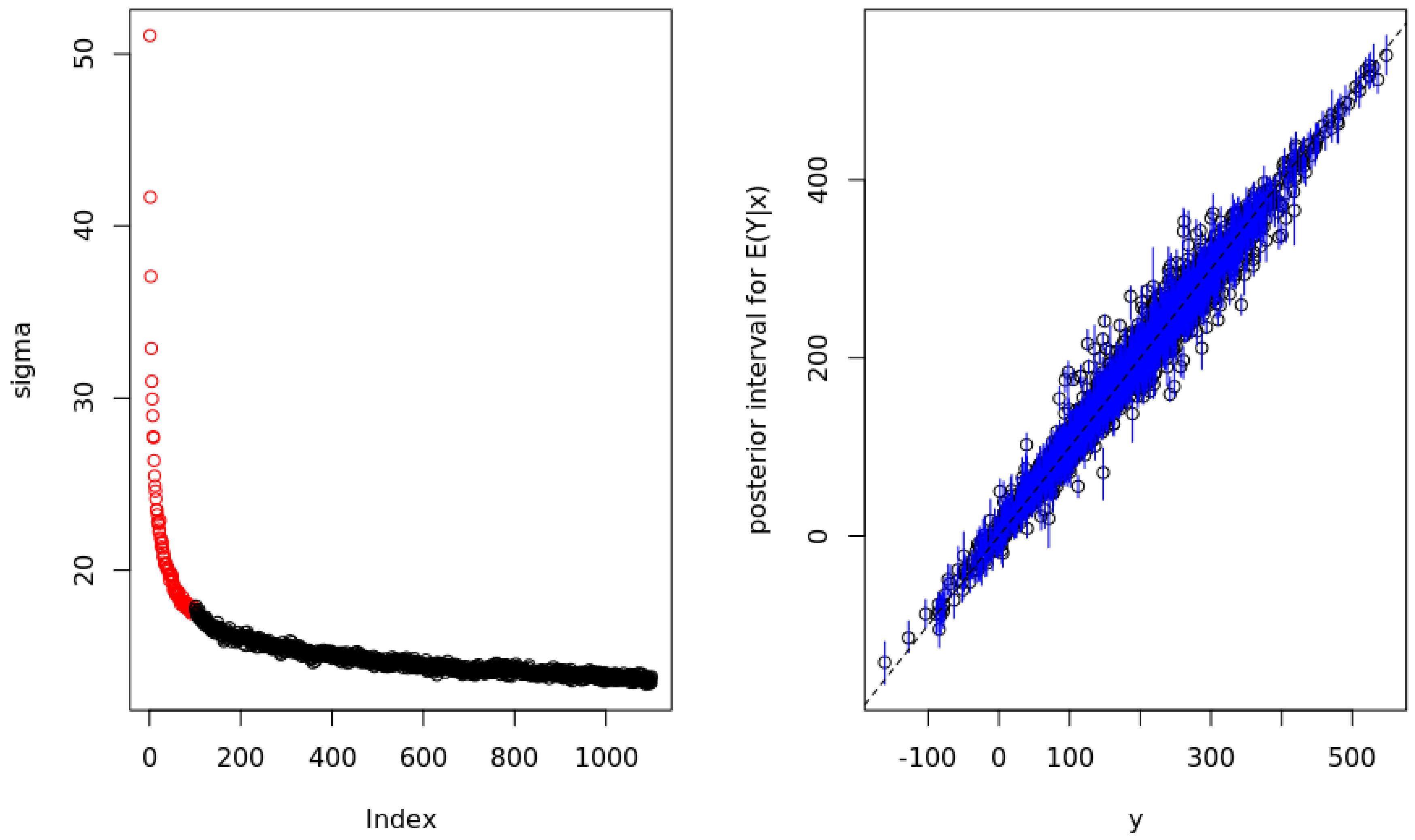
After training, 1000 ensembles of decision trees were drawn from the final distribution with a Gibbs sampler. Because ProperSea uses these pre-sampled ensembles, property prediction is relatively fast. The ensembles give a distribution of values for the predicted property, which ProperSea summarizes with a mean value and a 95% credible interval.

Model Validation

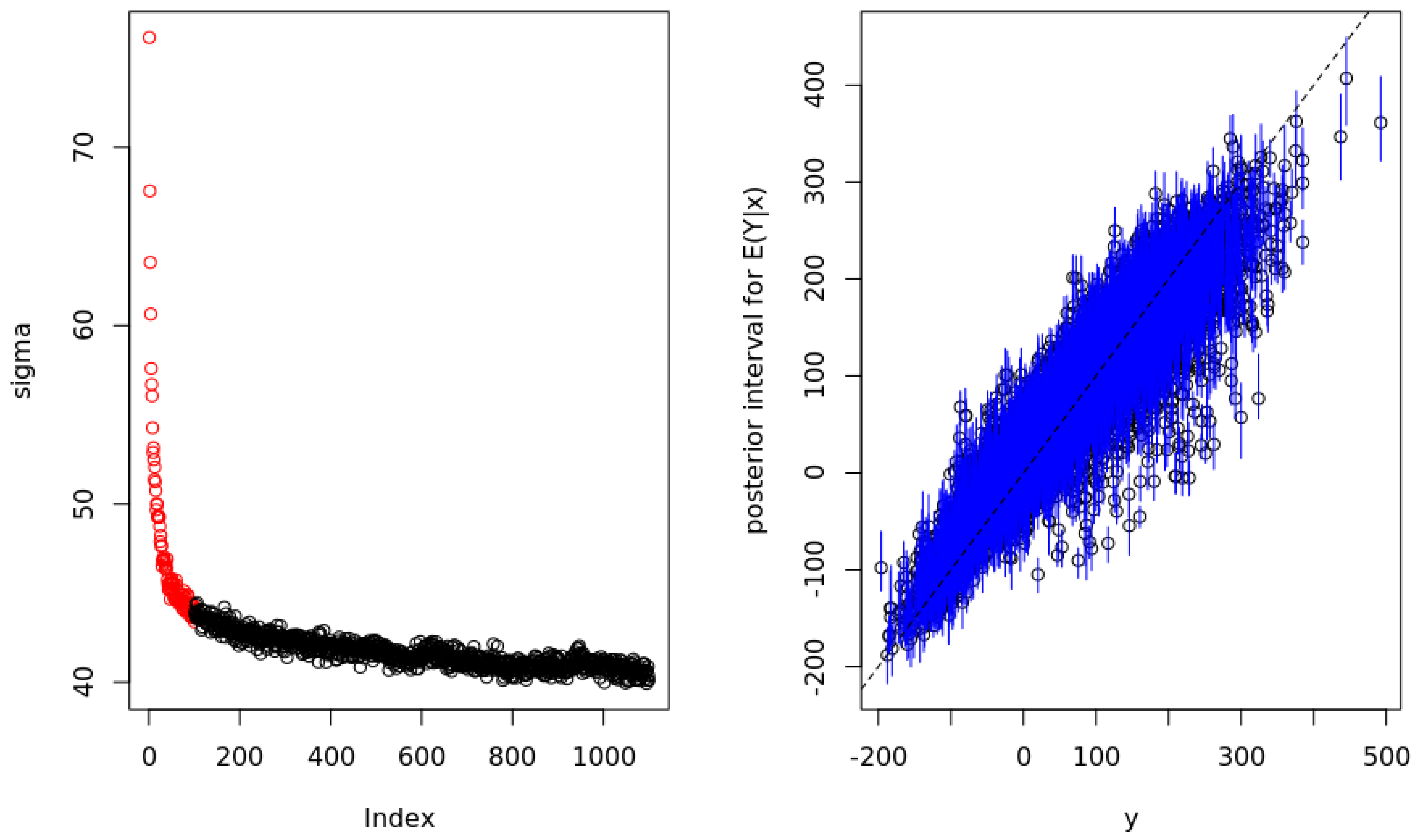
Each model was validated with 5-fold cross validation to ensure that predictions generalized to unseen data. Each final BART model was then trained on the entirety of the data. BART does not tend to overfit the data because the Bayesian priors induce regularization, and the use of an ensemble of 300 trees is effectively model averaging.

The fitting results for each property are summarized below with two plots. The first shows 1000 samples of the experimental error parameter (σ) drawn from the posterior distribution, with 100 discarded burn-in samples shown in red. A smaller σ means the model is a better fit to the data, although even a totally accurate model will have a non-zero σ if the experimental data is noisy.

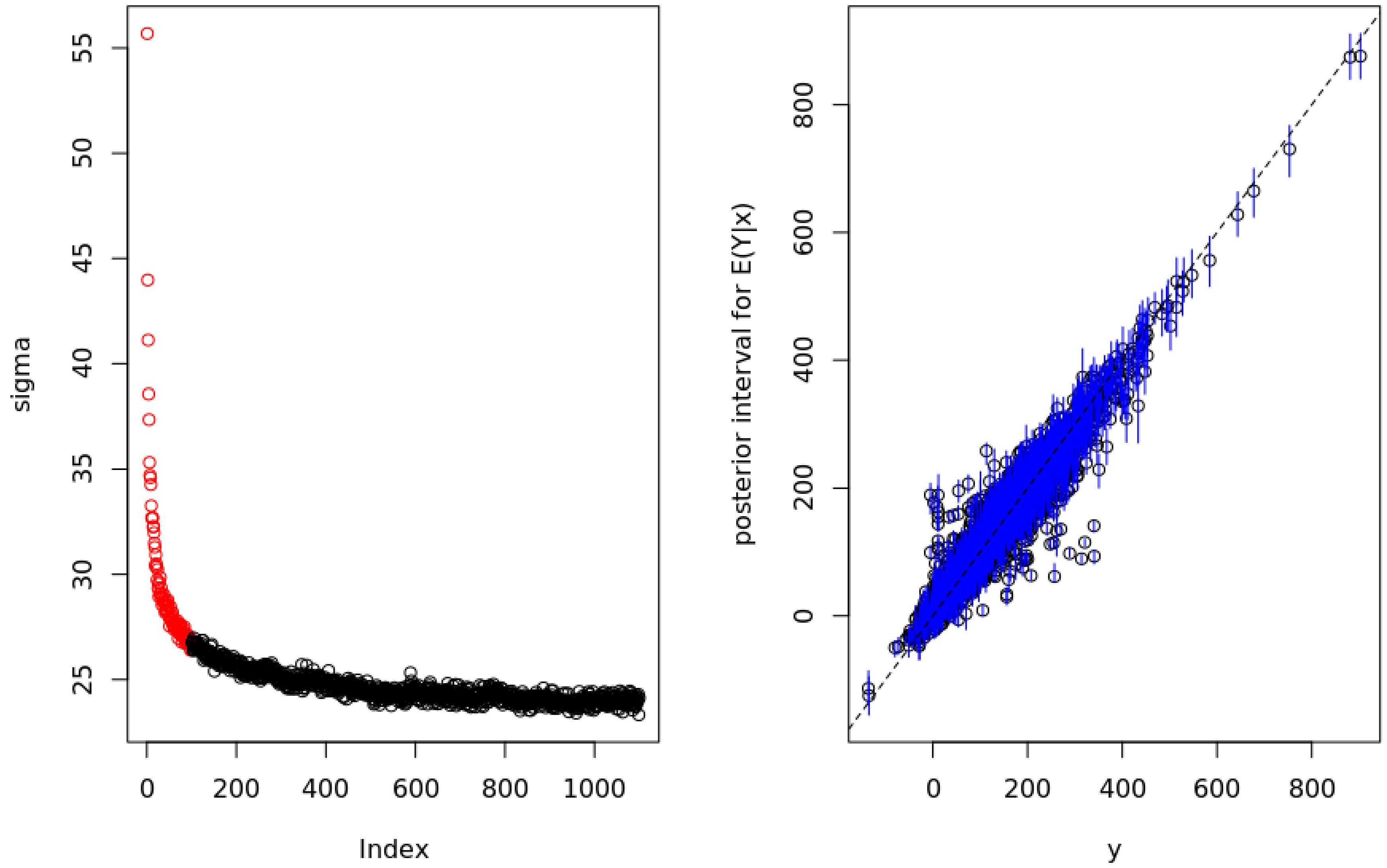
The second plot for each property shows the predicted distribution for each item in the training set (median and 90% confidence interval), plotted against the experimental value.



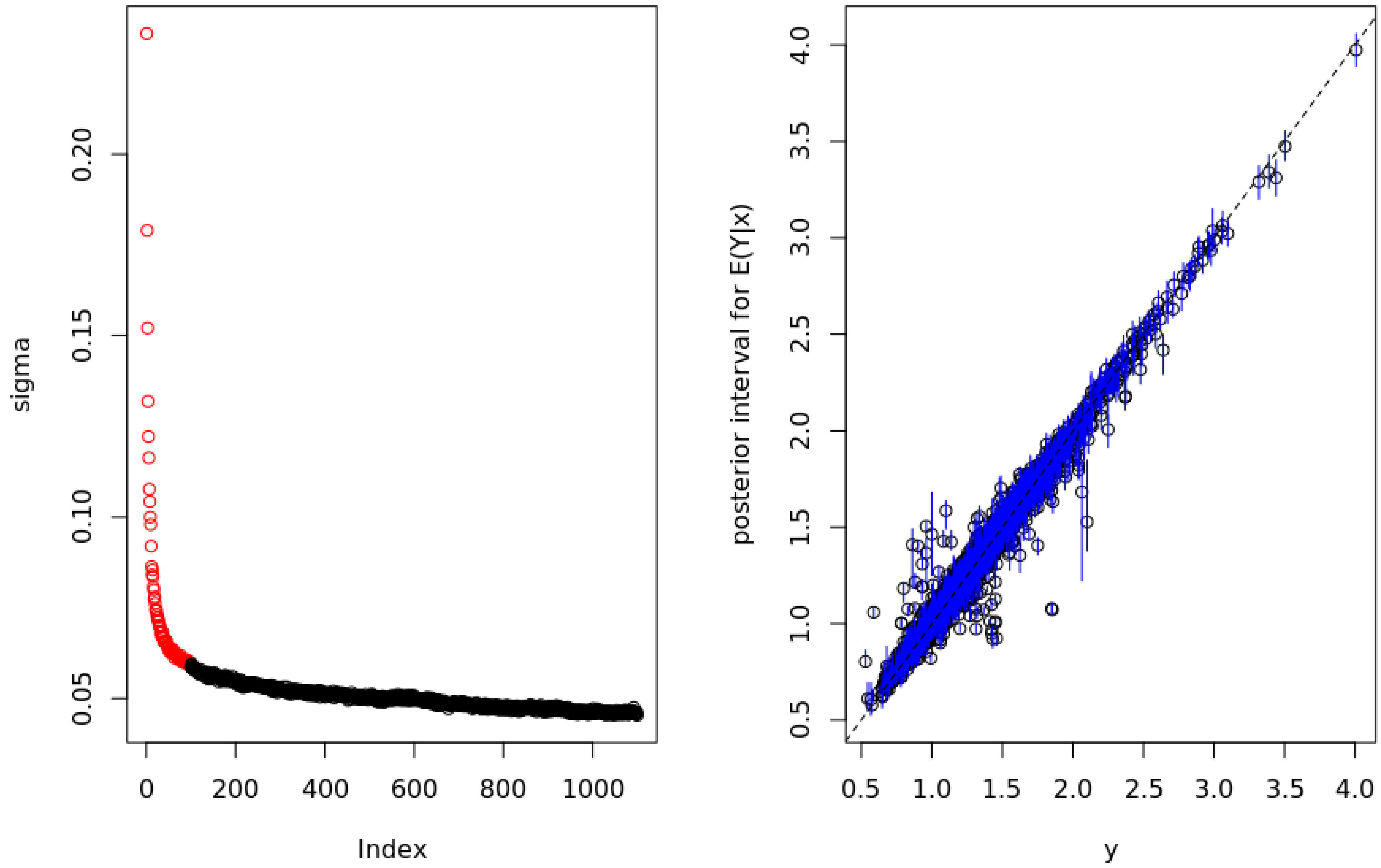
Boiling Point



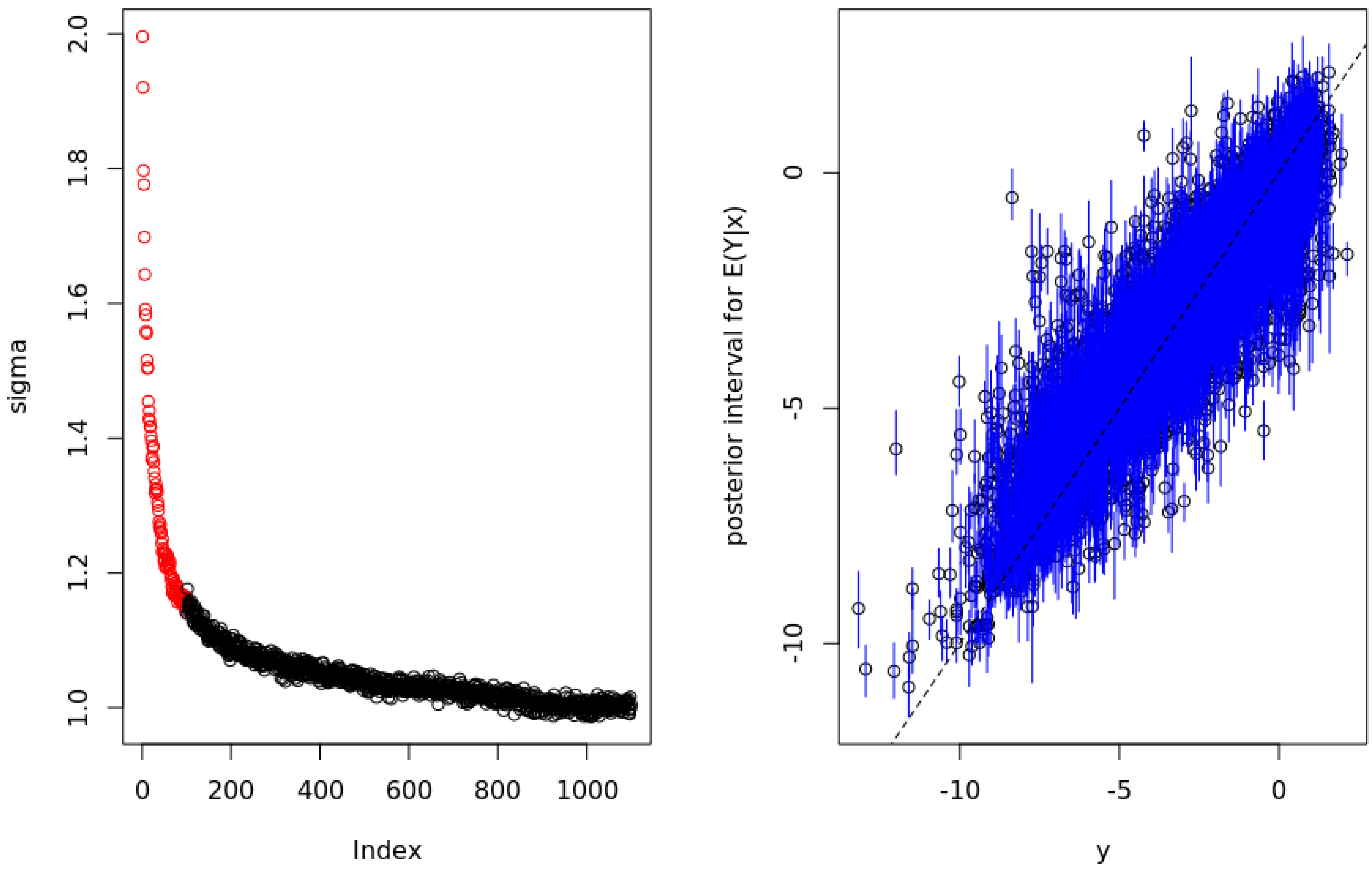
Melting Point



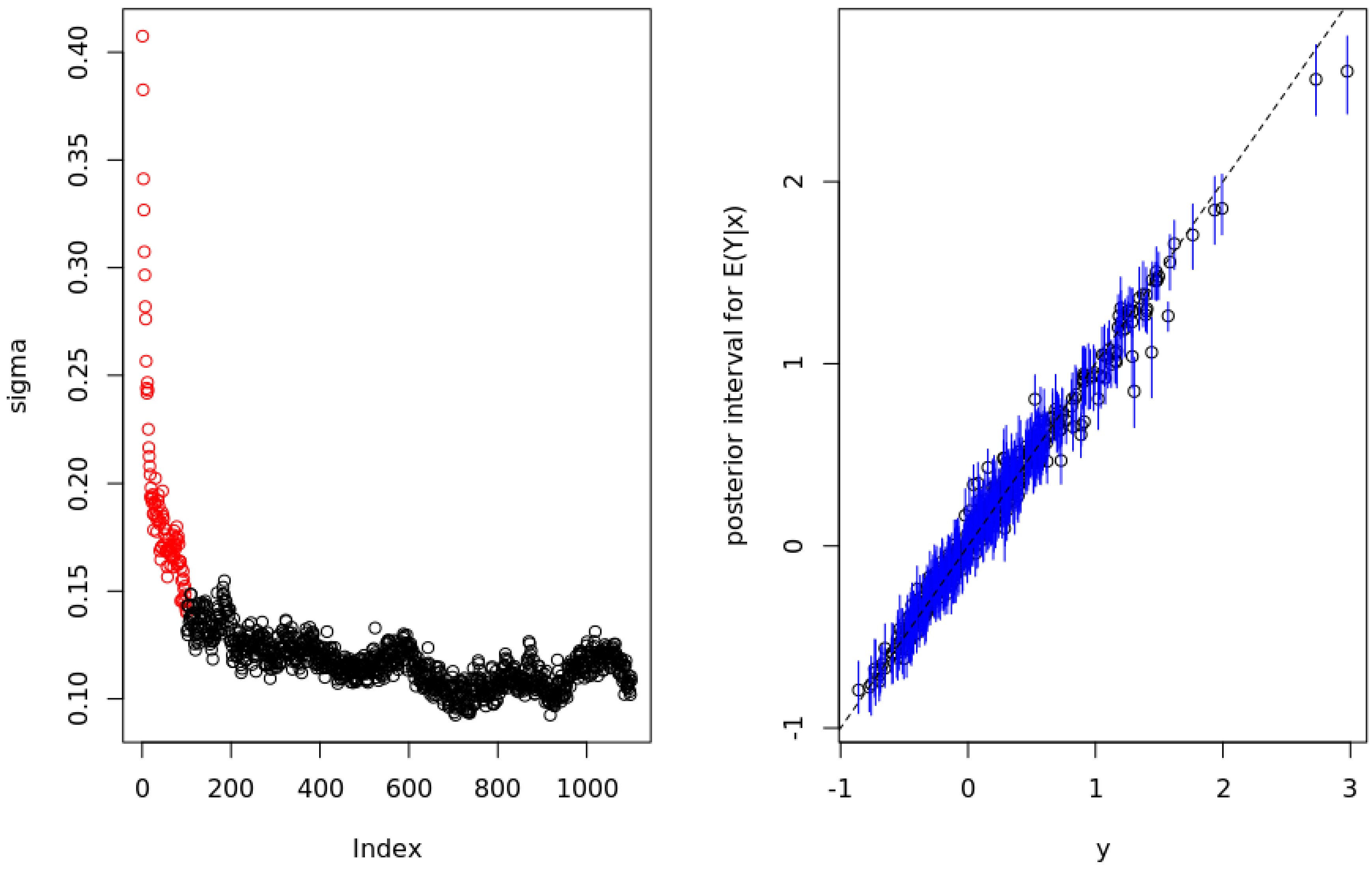
Flash Point



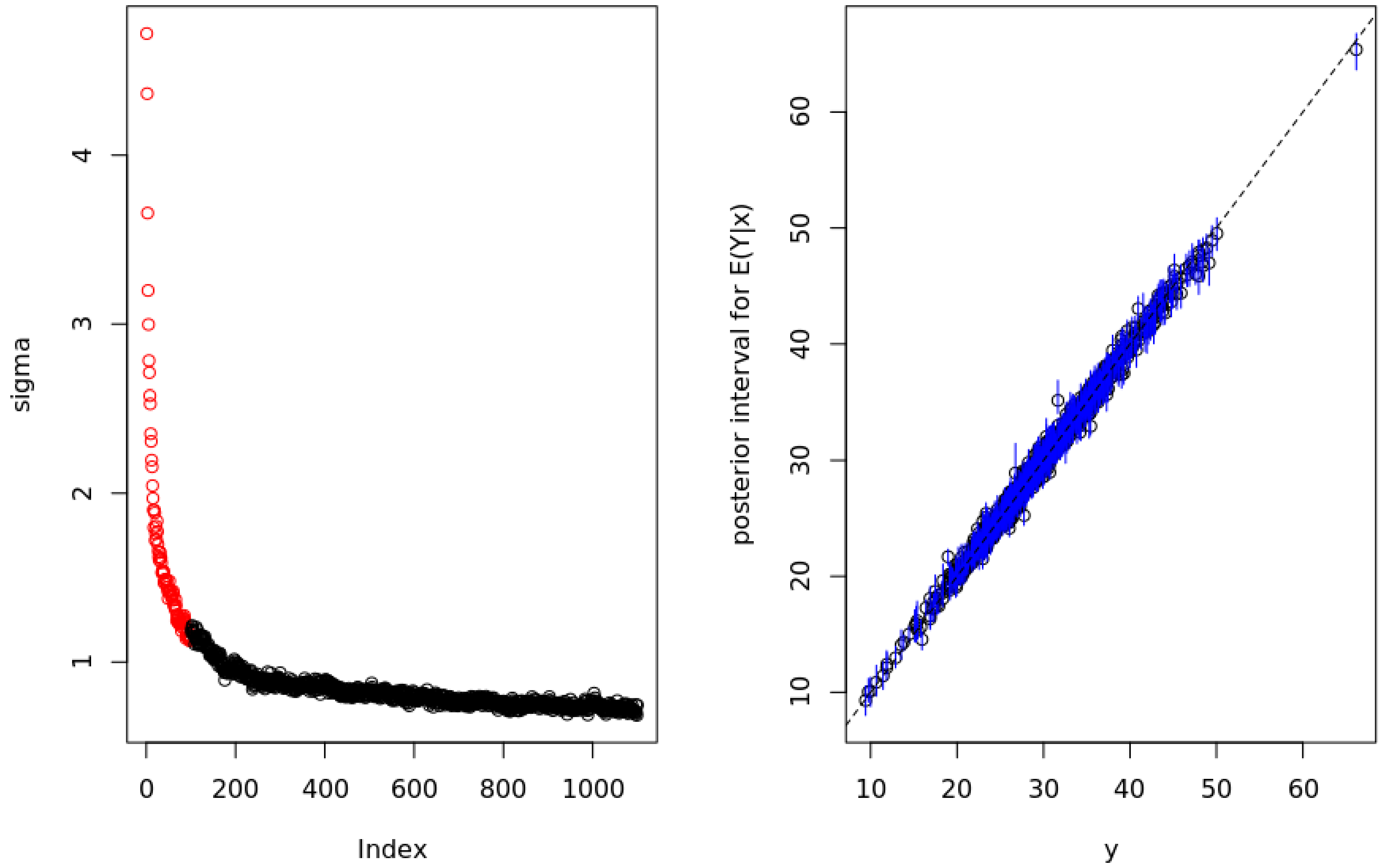
Density



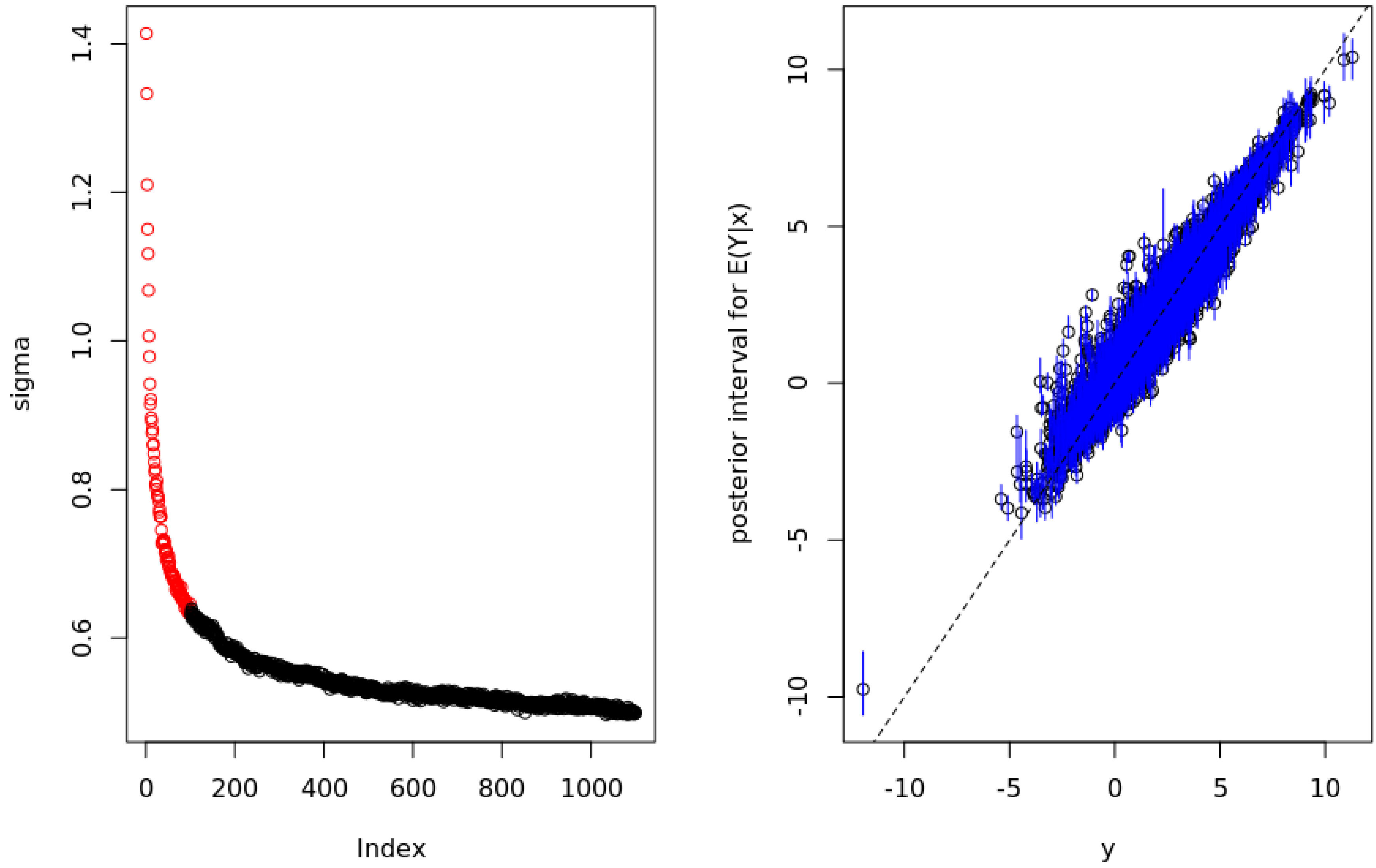
Log Solubility



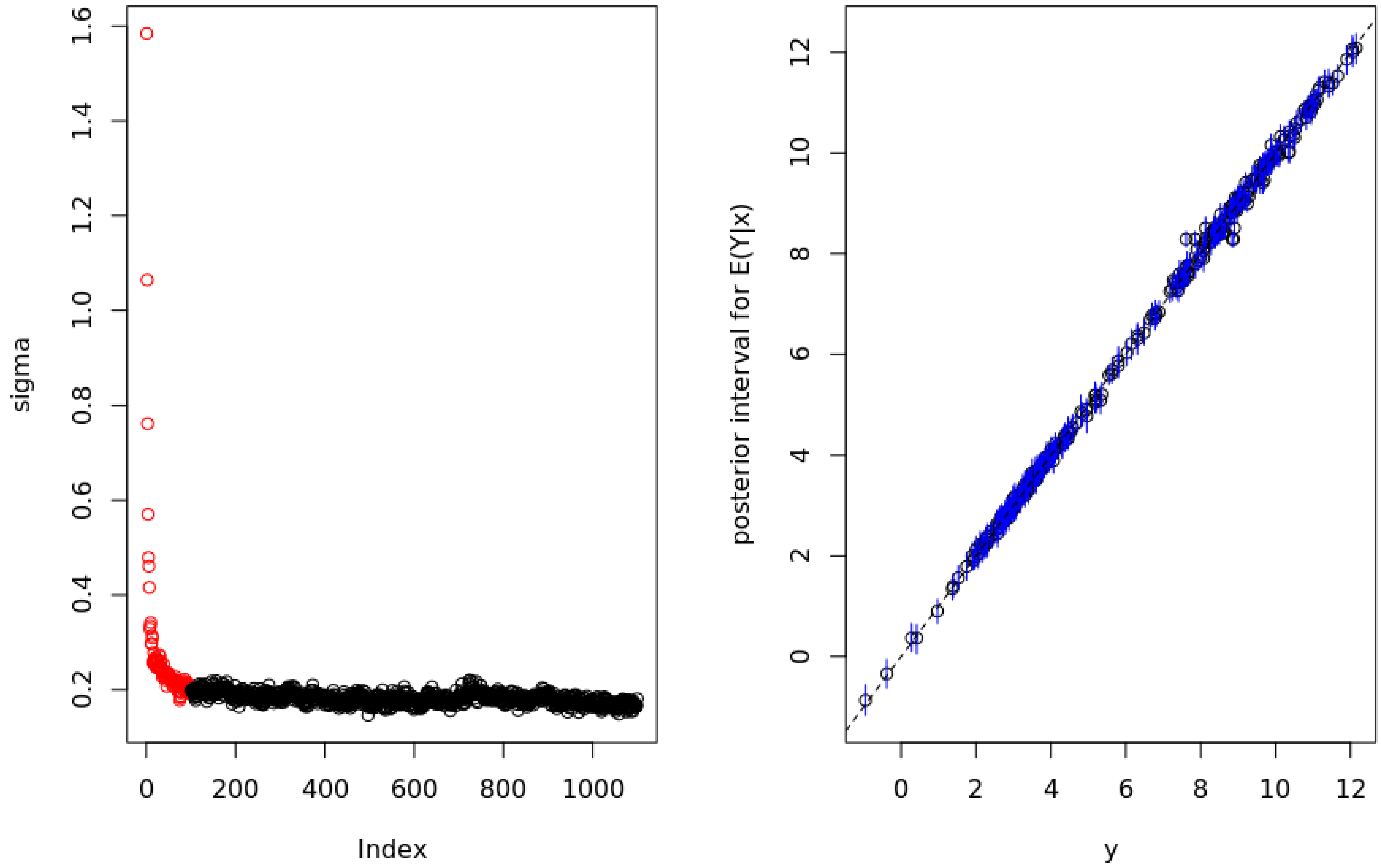
Log Viscosity



Surface Tension



logP



logK$_{octanol/air}$

Applicability Domain

To further quantify the trustworthiness of each prediction, ProperSea compares the molecule being queried to the molecules in each training set. To do this, molecular fingerprints (essentially a binary representation of the molecule) have been pre-computed for each training set. ProperSea computes the corresponding fingerprint for the query molecule, and compares it to the training set with Tanimoto's distance measure.

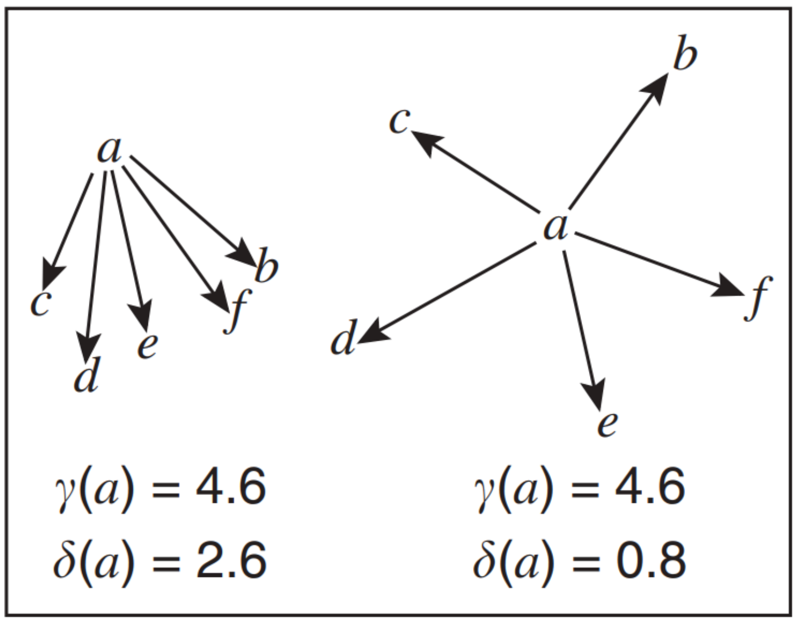
The distances are summarized using indices proposed by Harmeling:

γ: average distance to five closest neighbours

δ: distance to average fingerprint of closest five neighbours

The first index, γ, gives an indication of the sparsity of the region of fingerprint space where the query molecule is located. This is an intuitive measure, but struggles to differentiate between the case where there are no molecules in the training set with similar attributes to the query molecule, and the case where the query molecule is surrounded by training set molecules with a degree of similarity but no exact match. The second index, δ, identifies query molecules that have very little similarity to any item in the training set.

The difference between between γ and δ is best illustrated in the figure below from Harmeling's paper, where point a is compared to five neighbouring point:



The γ index is unable to differentiate between the case where point a is on the edge of the applicability domain, and the case where it is within the applicability domain but distant from all neighbours. The δ index allows ProperSea to identify search queries that are on the edge of the applicability domain, and are therefore least trustworthy.

As the Harmeling indices are not straighforward to interpret by the user, the reliability of the prediction is qualified with the following table:

	$\gamma \leq 0.3$	$0.3 < \gamma \leq 0.6$	$\gamma > 0.6$
$\delta \leq 0.3$	High	Medium	Low
$0.3 < \delta \leq 0.6$	Medium	Medium	Low
$\delta > 0.6$	Very Low	Very Low	Very Low