Table 1: Selected data for Octanol-Water Partition Coefficient, Toxicity (Oral Rat), Aqueous Solubility, and Auto-Ignition Temperature in terms of symbol, number of data-points, internal diversity, data distribution and needs (see Table S1 in supporting material for a full list of properties).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Property | sYMBOL | M | d | dATA dIsTRIBUTION | Needs |
| Octanol-Water Partition Coefficient | LogP | 12193 | 0.86 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\19dist.jpg | Biological activity, environmental risk assessment, transport evaluation |
| Toxicity (Oral Rat) [mol/kg] | LD50 | 4904 | 0.88 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\16dist.jpg | Toxicity evaluation, risk assessment |
| Aqueous Solubility [mol/L] | LogWs | 2565 | 0.87 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\15dist.jpg | Biological activity, environmental risk assessment, transport evaluation |
| Auto-Ignition Temperature [K] | AiT | 571 | 0.87 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\2dist.jpg | Risk assessment, chemical evaluation |

**Table 2.** Comparative performance between the GC-simple-models and the best performing GC-ML-models. The full list of error distribution is provided in Table S2. **Acronyms**: *(GC) Group Contribution; (MARE) Median Absolute Relative Error; (MCL.) Median 95% confidence level; (MSE) Mean Squared Error; (Prc)* *Percentage of data within 1%, 5%, and 10% of absolute relative error; (SWR) Step-Wise Regression..*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | gc-Simple-MODELS | | | | | GC-ML mODELS | | | | | | |
| No | Property | N | Model | Prc (±1%) | Prc (±5%) | Prc (±10%) | Error Distribution | Prc (±1%) | Prc (±5%) | Prc (±10%) | MARE | MSE | MCL |
| 1 | Octanol-Water Partition Coefficient (LogP) | 12193 | SWR | 5.45 | 25.44 | 43.85 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\logP\dist.jpg | 84.75 | 89.09 | 91.97 | 1.06e-7 | 0.12 | 1.5e-4 |
| 2 | Toxicity (Oral Rat) (LD50) [mol/kg] | 4904 | SWR | 1.52 | 6.92 | 13.61 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\LD50\dist.jpg | 89.64 | 90.84 | 92.13 | 0.24 | 0.09 | 6.6e-5 |
| 3 | Aqueous Solubility (LogWs) [mol/L] | 2565 | SWR | 3.12 | 14.36 | 28.63 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\logWs\dist.jpg | 86.24 | 89.43 | 92.09 | 1.14e-07 | 0.08 | 2.6e-4 |
| 4 | Auto Ignition Temperature (AiT) [K] | 571 | SVR | 12.11 | 18.60 | 24.04 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\AiT\dist.jpg | 92.63 | 93.86 | 95.44 | 4.11e-6 | 616.85 | 0.033 |
| 5 | Normal Melting Point (Tm) [K] | 9249 | SWR | 17.60 | 73.28 | 93.90 |  | 80.95 | 84.11 | 87.69 | 7.65e-6 | 0.03 | 6.7e-3 |
| 6 | Normal Boiling Point (Tb) [K] | 5276 | SWR | 56.95 | 99.44 | 100.0 |  | 79.09 | 87.76 | 92.70 | 0.005 | 108.97 | 2.4e-5 |
| 7 | Acid Dissociation Constant (pKa) | 1634 | SVR | 18.37 | 24.80 | 32.15 |  | 85.36 | 87.81 | 89.22 | 5.23e-6 | 0.41 | 5.4e-4 |
| 8 | Hildebrandt Solubility Parameter at 298 K (HSolP) [MPa1/2] | 1384 | SVR | 22.33 | 34.61 | 47.62 |  | 85.33 | 91.47 | 94.58 | 8.34e-6 | 0.39 | 6.5e-4 |
| 9 | Standard Enthalpy Of Formation (Hf) [kJ/mol] | 1059 | SWR | 42.86 | 80.05 | 90.02 |  | 84.14 | 92.63 | 93.20 | 1.19e-8 | 1827.08 | 0.033 |
| 10 | Liquid Molar Volume at 298K (Lmv) [cc/mol] | 1059 | SVR | 22.33 | 34.61 | 47.62 |  | 91.95 | 98.96 | 99.34 | 1.24e-4 | 1.36e-6 | 1.47e-5 |
| 11 | Critical Temperature (Tc) [K] | 776 | SWR | 66.03 | 98.87 | 100.0 |  | 84.41 | 91.37 | 94.20 | 1.06e-6 | 121.89 | 0.023 |
| 12 | Critical Pressure (Pc) [bar] | 774 | SWR | 31.87 | 80.18 | 92.34 |  | 87.47 | 90.05 | 92.76 | 8.58e-8 | 2.03 | 0.037 |
| 13 | Critical Volume (Vc) [cc/mol] | 773 | SWR | 50.31 | 93.22 | 99.00 |  | 86.42 | 90.94 | 94.57 | 7.01e-7 | 1232 | 0.045 |
| 14 | Standard Gibbs Energy Of Formation (Gf) [kJ/mol] | 756 | SVR | 57.14 | 82.54 | 90.21 |  | 85.98 | 91.27 | 91.93 | 1.31e-8 | 1061 | 0.045 |
| 15 | Normal Enthalpy Of Fusion (Hfus) [kJ/mol] | 749 | SVR | 28.44 | 37.25 | 47.4 |  | 83.31 | 87.72 | 91.19 | 1.08e-6 | 2.65 | 3e-3 |
| 16 | Fathead Minnow 96-H LC50 (LC50(FM)) [mol/L] | 705 | SWR | 8.53 | 31.52 | 54.02 |  | 88.23 | 90.50 | 91.21 | 5.91e-6 | 0.06 | 3.4e-4 |
| 17 | Photochemical Oxidation Potential (PCO) | 606 | SWR | 6.42 | 16.90 | 26.30 |  | 84.75 | 89.09 | 91.97 | 3.77e-5 | 0.01 | 1.4e-4 |
| 18 | Bioconcentration Factor (BCF) | 589 | SWR | 8.91 | 19.49 | 30.82 |  | 84.38 | 85.74 | 88.79 | 7.65e-6 | 0.03 | 3.5e-4 |
| 19 | Enthalpy Of Vaporization At 298K (Hv) [kJ/mol] | 425 | SWR | 40.99 | 90.07 | 96.88 |  | 83.06 | 90.12 | 93.14 | 1.42e-6 | 6.09 | 0.004 |
| 20 | Permissible Exposure Limit (OSHA-TWA) [mol/m3] | 422 | SVR | 32.46 | 41.23 | 48.82 |  | 90.76 | 91.47 | 91.94 | 3.58e-6 | 0.15 | 4.6e-4 |

Table 3: Background information on the steps of computing property values from the devised models.

|  |  |  |
| --- | --- | --- |
| Stage | GC-Simple | GC-ML |
| Molecular Representation | Group counts | Group counts |
| Property Calculation | Simple computation | Complex Computation |
| Model Parameters | See SI? | Provided on GitHub |
| Prediction | Property value with acceptable accuracy | Property value with high accuracy and an uncertainty estimate |

Table 4: Estimation of the standard enthalpy of formation and normal boiling point of Isobutylamine using stepwise regression GC models.

|  |  |  |  |
| --- | --- | --- | --- |
| Isobutylamine  CAS No. 78-81-9  SMILES: CC(C)CN | | 2D Molecular structure  ChemSpider 2D Image | Isobutylamine | C4H11N | |
| Molecular formula: C4H11N | Occurrences | Contributions for Hf | Contributions for Tb |
| 1st-order groups |  |  |  |
| CH3 | 2 | -57.2760 | 0.8853 |
| CH | 1 | 7.3487 | 0 |
| CH2NH2 | 1 | -25.6857 | 1.2616 |
| 2nd-order groups |  |  |  |
| (CH3)2CH | 1 | -1.1487 | 0.0071 |
| 3rd-order groups |  |  |  |
| - | - | - | - |
| Universal Const. |  | 35.17 [kJ/mol] | 244.51K |
| Prediction |  | -98.85 [kJ/mol] | 271.81K |
| Experimental Value |  | -98.8 [kJ/mol] | 340.85 K |
| Relative Error |  | 0.06% | 20.25% |

Notation

|  |  |
| --- | --- |
| Notation | |
| NF | First-order groups |
| NS | Second-order groups |
| NT | Third-order groups |
| M | Number of molecules in the dataset |
| *b* | Bias term |
| *c* | Model contribution parameters |
| *KMM* | Covariance function between molecules in the dataset |
| *n* | Group count |
| *n* | Vector of 424 group counts |
| *npred* | Vector of 424 group counts for a to-be-predicted molecule |
| *U* | GP Uncertainty matrix |
| *W* | GP Weights matrix |
| *w* | Tunable parameter/variable in sequential and simultaneous regression |
| *y* | Property values |
| *ypred* | The predicted value of properties |
| *κ* | The kernel (GP covariance) function |
| *Σ* | The covariance property function (uncertainty estimate) |
| *μ* | Mean function of the GP |
| *σ* | The signal variance |
| *σy* | The standard deviation of the property values in the dataset |
| *σε* | The noise variance |
| *𝓁* | The length-scale of the kernel function |
| *ξ, ξ\** | Slack variables |
| *I* | Identity matrix |
| *yexp* | The experimental values of molecules in the dataset of a given property |

SUPPORTING INFORMATION

**A Platform of Machine Learning-Based Next-Generation Property Estimation Methods for CAMD**

Abdulelah S. Alshehri1,2, Anjan K.Tula3, Lei Zhang4, Rafiqul Gani5[[1]](#footnote-1)\*, Fengqi You1[[2]](#footnote-2)\*

1 Robert Frederick Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853, USA

2 Department of Chemical Engineering, College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia

3 College of Control Science and Engineering, Zhejiang University, Hangzhou-310027, China

4 Institute of Chemical Process Systems Engineering, School of Chemical Engineering, Dalian University of Technology, 116024 Dalian, China

5 PSE for SPEED, Skyttemosen 6, DK-3450 Allerød, Denmark

The following tables and figures are included as supporting information

* **Table S1**. Detailed list of physiochemical properties in the dataset, providing information
* **Table of data analysis for all properties (Anjan)**
* about the count, distribution, and needs
* **Table S2:** A list of error distribution for GC-GP models.
* **Table S3**: Property model functions
* **Table S4:** Correlation results

Table S1. Detailed list of physiochemical properties in the dataset, providing information about the count, distribution, and needs.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Property | sYMBOL | m | d | dATA dIsTRIBUTION | Needs |
| Octanol-Water Partition Coefficient | LogP | 12193 | 0.86 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\19dist.jpg | Biological activity, environmental risk assessment, transport evaluation |
| Normal Melting Point [K] | Tm | 9249 | 0.88 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\18dist.jpg | Phase identity, SLE analysis, chemical evaluation |
| Normal Boiling Point [K] | Tb | 5276 | 0.89 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\17dist.jpg | Phase identity, VLE analysis, volatility |
| Toxicity (Oral Rat) [mol/kg] | LD50 | 4904 | 0.88 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\16dist.jpg | Toxicity evaluation, risk assessment |
| Aqueous Solubility [mol/L] | LogWs | 2565 | 0.87 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\15dist.jpg | Biological activity, environmental risk assessment, transport evaluation |
| Acid Dissociation Constant | pKa | 1634 | 0.86 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\14dist.jpg | Biological activity, reactivity evaluation, risk assessment, transport evaluation |
| Hildebrandt Solubility Parameter at 298 K [MPa1/2] | HSolP | 1384 | 0.89 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\13dist.jpg | Equilibrium analysis, solute-solvent interactions, biological activity |
| Standard Enthalpy Of Formation [kJ/mol] | Hf | 1059 | 0.87 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\12dist.jpg | Equilibrium analysis, reactivity and kinetics evaluation |
| Liquid Molar Volume at 298K [cc/mol] | Lmv | 1059 | 0.89 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\11dist.jpg | Equilibrium analysis, chemical evaluation |
| Critical Temperature [K] | Tc | 776 | 0.87 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\10dist.jpg | Phase identity, VLE analysis, volatility |
| Critical Pressure [bar] | Pc | 774 | 0.87 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\9dist.jpg | Phase identity, VLE analysis, volatility |
| Critical Volume [cc/mol] | Vc | 773 | 0.87 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\8dist.jpg | Phase identity, VLE analysis, volatility |
| Standard Gibbs Energy Of Formation [kJ/mol] | Gf | 756 | 0.87 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\7dist.jpg | Equilibrium analysis, reactivity and kinetics evaluation |
| Normal Enthalpy Of Fusion [kJ/mol] | Hfus | 749 | 0.88 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\6dist.jpg | Phase identity, equilibrium analysis, chemical evaluation |
| Fathead Minnow 96-H LC50 [mol/L] | LC50(FM) | 705 | 0.89 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\5dist.jpg | Toxicity evaluation, risk assessment |
| Photochemical Oxidation Potential | PCO | 606 | 0.85 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\4dist.jpg | Environmental risk assessment, reactivity, toxicity evaluation |
| Bioconcentration Factor | BCF | 589 | 0.86 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\3dist.jpg | Environmental risk assessment, toxicity evaluation |
| Auto Ignition Temperature [K] | AiT | 571 | 0.87 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\2dist.jpg | Risk assessment, chemical evaluation |
| Enthalpy Of Vaporization At 298K [kJ/mol] | Hv | 425 | 0.86 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\1dist.jpg | Phase identity, VLE analysis |
| Permissible Exposure Limit [mol/m3] | OSHA-TWA | 422 | 0.91 | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Overall Results\0dist.jpg | Risk assessment, toxicity evaluation |







**Table S3:** A list of error distribution for GC-GP models.

|  |  |  |
| --- | --- | --- |
| No | Property | Error Distribution |
| 1 | Octanol-Water Partition Coefficient (LogP) | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\logP\dist.jpg |
| 2 | Normal Melting Point (Tm) [K] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\Tm\dist.jpg |
| 3 | Normal Boiling Point (Tb) [K] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\Tb\dist.jpg |
| 4 | Toxicity (Oral Rat) (LD50) [mol/kg] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\LD50\dist.jpg |
| 5 | Aqueous Solubility (LogWs) [mol/L] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\logWs\dist.jpg |
| 6 | Acid Dissociation Constant (pKa) | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\pKa\dist.jpg |
| 7 | Hildebrandt Solubility Parameter at 298 K (HSolP) [MPa1/2] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\HsolP\dist.jpg |
| 8 | Standard Enthalpy Of Formation (Hf) [kJ/mol] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\hf\dist.jpg |
| 9 | Liquid Molar Volume at 298K (Lmv) [cc/mol] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\Lmv\dist.jpg |
| 10 | Critical Temperature (Tc) [K] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\Tc\dist.jpg |
| 11 | Critical Pressure (Pc) [bar] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\pc\dist.jpg |
| 12 | Critical Volume (Vc) [cc/mol] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\vc\dist.jpg |
| 13 | Standard Gibbs Energy Of Formation (Gf) [kJ/mol] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\gf\dist.jpg |
| 14 | Normal Enthalpy Of Fusion (Hfus) [kJ/mol] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\hfus\dist.jpg |
| 15 | Fathead Minnow 96-H LC50 (LC50(FM)) [mol/L] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\LC50_FM\dist.jpg |
| 16 | Photochemical Oxidation Potential (PCO) | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\PCO\dist.jpg |
| 17 | Bioconcentration Factor (BCF) | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\BCF\dist.jpg |
| 18 | Auto Ignition Temperature (AiT) [K] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\AiT\dist.jpg |
| 19 | Enthalpy Of Vaporization At 298K (Hv) [kJ/mol] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\hv\dist.jpg |
| 20 | Permissible Exposure Limit (OSHA-TWA) [mol/m3] | C:\Users\asa279\Desktop\MyGitHub\ML-PropertyEstimation\Runbest\OSHA_TWA\dist.jpg |

**Table S3.** A list of GC/SVR model functions for all properties.

|  |  |  |
| --- | --- | --- |
| Property | LEFT-hAND SIDE | rIGHT-hAND SIDE OF gc/SVR MODELS |
| Octanol-Water Partition Coefficient |  |  |
| Normal Melting Point [K] |  |  |
| Normal Boiling Point [K] |  |  |
| Oral Rat [mol/kg] |  |  |
| Aqueous Solubility [mol/L] |  |  |
| Acid Dissociation Constant |  |  |
| Hildebrandt Solubility Parameter at 298 K [MPa1/2] |  |  |
| Standard Enthalpy Of Formation [kJ/mol] |  |  |
| Liquid Molar Volume at 298K [cc/mol] |  |  |
| Critical Temperature [K] |  |  |
| Critical Pressure [bar] |  |  |
| Critical Volume [cc/mol] |  |  |
| Standard Gibbs Energy Of Formation [kJ/mol] |  |  |
| Normal Enthalpy Of Fusion [kJ/mol] |  |  |
| Fathead Minnow 96-H LC50 [mol/L] |  |  |
| Photochemical Oxidation Potential |  |  |
| Bioconcentration Factor |  |  |
| Auto Ignition Temperature [K] |  |  |
| Enthalpy Of Vaporization At 298K [kJ/mol] |  |  |
| Permissible Exposure Limit [mol/m3] |  |  |

1. \*Corresponding author. E-mail: rgani2018@gmail.com

   \*Corresponding author. Phone: (607) 255-1162; Fax: (607) 255-9166; E-mail: fengqi.you@cornell.edu [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)