Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS
ISSN 2454-6348

Original Research Article DOI: 10.26479/2019.0506.03

MATLAB MACHINE LEARNING & CURVE-FITTING TOOLBOX: PREDICTION OF DRUG AQUEOUS SOLUBILITY

Kamal I.M. Al-Malah*

Chemical Engineering, Higher Colleges of Technology, Ruwais, Abu-Dhabi, UAE.

ABSTRACT: The prediction of aqueous solubility of a set of 246 drug molecules with a broad range, varying from 120 up to 8,330 mg/L, as a function of pertinent molecular properties was examined. MATLAB® Machine Learning (ML) and Optimization Toolbox were used in the data analysis. Both the supervised and supervised learning techniques were used to analyze such highly scattered date, like aqueous solubility of organic drug molecules. The exotic features of machine learning algorithms were shown in the form of figures, pertinent to the selection process of predictor variables. It was found that the drug aqueous solubility data could be best described by the first three important molecular properties: the non-polar molecular mass, *MWNPOL*, the non-polar molar volume, *NPolVol*, and the polar fraction of a molecule, *PolFrac*, as the third refining or tuning-up factor (weight parameter in curve fitting). The polarity index was evaluated based on the atomic mole-fraction of polar atoms, namely, F, O, N, Cl, and Br because such atoms have relatively higher electronegativity values than those of C, H, I, P, and S atoms. The percent relative error (PRE) was also calculated for each individual drug molecule using models based on *MWNPOl, NPolVol*, and *PolFrac*, while assuming that the true value of solubility is the experimentally measured and reported value. It was found that the three models overestimated the aqueous solubility of less soluble materials; specifically, below 200 mg/L. Finally, the entropically driven hydrophobic interactions, manifested via *MWNPOL*, were found to act as anti-solvation factor.

Keywords: MATLAB; Machine Learning; Optimization; Drug Aqueous Solubility.

Article History: Received: September 03, 2019; Revised: October 20, 2019; Accepted: November 11, 2019.

Corresponding Author: Dr. Kamal I. Al-Malah* Ph.D.

Chemical Engineering, Higher Colleges of Technology, Ruwais, Abu-Dhabi, UAE.

E-mail Address: kalmalah@hct.ac.ae, almalak61@hotmail.com.

1. INTRODUCTION

As pointed out earlier [1], “Organic solvents play a critical role in many industrial applications, while they pose a direct impact on health, safety, environmental, feasibility, and economic aspects of a chemical, biochemical, food, and pharmaceutical industries. Of an immense concern, the solubilization of pharmaceutical active ingredients ranks a top priority for injected and oral drug administration. With an increasing pressure to identify high-quality drug candidates, it is critical to assess the Absorption, Distribution, Metabolism, Excretion (ADME) attributes of compounds early during drug discovery stage. This may include properties such as aqueous and non-aqueous solubility, permeability, metabolic stability, and in vivo pharmacokinetics”. Eric *et al.* [2] worked on an approach for the development of a model for prediction of aqueous solubility, based on the implementation of an algorithm for the automatic adjustment of descriptor's relative importance (AARI) in counter-propagation artificial neural networks (CPANN). Using their approach, the interpretability of the model based on artificial neural networks, traditionally considered as “black box” models, was significantly improved. For the development of the model, a data set consisting of 374 diverse drug-like molecules, divided into training (n=280) and test (n=94) sets using self-organizing maps, was used. Heuristic method was applied in preselecting a small number of the most significant descriptors to serve as inputs for CPANN training. The performances of the final model based on 7 descriptors for prediction of solubility were satisfactory for both training and test set. The model was found to be a highly interpretable in terms of solubility, as well as rationalizing structural features that could have an impact on the solubility of the compounds investigated. Their proposed approach can significantly enhance model usability by giving guidance for structural modifications of compounds with the aim of improving solubility in the early phase of drug discovery. Sun *et al.* [3] emphasized the aqueous solubility as one of the most important properties in drug discovery, as it has profound impact on various drug properties, including biological activity, pharmacokinetics (PK), toxicity, and in vivo efficacy. They developed predictive models for kinetic solubility with in-house data generated from 11,780 compounds collected from over 200 NCATS intramural research projects. Based on the customized atom type descriptors, the support vector classification (SVC) models were trained on 80% of the whole dataset, and exhibited high predictive performance for estimating the solubility of the remaining 20% compounds within the test set. Their predictive models of aqueous solubility could be even used to identify insoluble compounds in drug discovery pipeline; to provide design ideas for improving solubility by analyzing the atom types associated with poor solubility; and to prioritize compound libraries to be purchased or synthesized. In previous works, the aqueous solubility of simple inorganic [4] and that of simple (single-carbon) organic [5] molecules were examined and expressed in terms of important molecular properties. Moreover, the aqueous solubility of some organic solvents was examined as a function of some selected molecular descriptors which are thought to affect the solvation process. It was found that to have an organic solvent with a high aqueous solubility, it has to have a low value of both the log partition coefficient between octanol and water, LogKow, and the molecular rugosity, R=V/S, accompanied by a high value of polar to hydrophobic surface area ratio, PHSAR [6]. In addition, paracetamol, also known as acetaminophen, N-(4-hydroxyphenyl)ethanamide, N-(4-hydroxyphenyl)acetamide, or N-acetyl-p-aminophenol, APAP, a medicine used to treat pain and fever, was used as a solid solute to demonstrate how the Aspen Plus simulator could be used as a powerful tool to optimize its solubility, using different organic solvents and water, as well. The minimization of molar Gibbs free energy of a mixture and the maximization of paracetamol solubility were both used as objective functions, from an optimization standpoint [1]. In this work, a set of 246 selected drug molecules were analyzed in light of deciphering the relationship between their aqueous solubilities on the one hand and some of their molecular and physical properties on the other hand. MATLAB Machine Learning (ML) and Optimization Toolbox were both used to analyze the degree of relationship for each player and later quantify such relationships, by expressing the aqueous solubility as a function of the most pertinent and important variables.

2. MATERIALS AND METHODS

MATLAB® has versatile built-in powerful functions which are meant to analyze data, characterized by a significant degree of scatter or bias, and later to decipher the unknown relationship between the dependent variable (or, response variable) and the list of independent variables (or, predictor variables). Each function will be explained on the spot once it is introduced to the reader. Let us briefly introduce what machine learning is all about.

2.1 What is Machine Learning?

As quoted by MATLAB R2019a built-in help, machine learning teaches computers to do what comes naturally to humans: Learn from experience. Machine learning algorithms utilize computational methods to directly learn (or extract) information from data without relying on a deterministic model. The set of algorithms adaptively improve their performance as the number of samples available for learning increases. Machine learning uses two types of techniques: Supervised learning, which trains a model on known input (predictor) and output (response) data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data. The aim of supervised machine learning is to build a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. Supervised learning uses classification and regression techniques to develop predictive models. On the one hand, classification techniques predict categorical responses; for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, image and speech recognition, and credit scoring. On the other hand, regression techniques predict continuous responses, for example, changes in temperature or fluctuations in power demand. Typical applications include electricity load forecasting and algorithmic trading. Machine learning algorithms enable the data analyst to prioritize the input variables based on their impact on or contribution to the response (output) variable. In other words, the investigator can prioritize the list of variables as far as their importance or contribution to the overall portray of the aqueous solubility is concerned. Optimization techniques can then be implemented on the mini set of input variables. In brief, machine learning will facilitate the process of ending up with a deterministic model, in the form of $y = f(X) = f(x_1, x_2)$ only rather than having $y = f(X) = f(x_1, x_2, x_3, \dots, x_n)$, at large.

2.2 Molecular Properties

Table 1 shows 246 drug components and their physical properties. From left, we have the name of the component, its molecular formula, its molecular mass (g/mol), its aqueous solubility, expressed in mg/L, its solid density (g/cm^3), its boiling point, expressed in °C, and finally its melting point, expressed in °C. Molecular data were substantially borrowed from Cao et al. [7] and Yalksowsky [8]. If any molecular data were missing, then, web databanks would be sought, like: www.chemicalbook.com, www.ncbi.nlm.nih.gov, www.chemspider.com, and www.epa.gov.

It is worth mentioning here that there is a pronounced discrepancy in terms of reporting the aqueous solubility value between one source and another. Moreover, even quoting from one source, there is still more than one reported value, depending on the original source of the data. For instance, the handbook [8] itself is a compilation of data, quoted from different sources. In addition, physical and toxicological properties of a drug-like compound may be affected by carrier solvents in commercial formulations. In addition, the boiling point for more than one case is quoted as predicted, but not experimentally measured. An average value was taken should there be more than a value of aqueous solubility. The advantage of having highly biased/scattered data will make ML algorithms more demanding and challenging in terms of deciphering the sacred relationship between the response on one side and predators (predictors) on another side.

Table 1: Drug compounds and their molecular properties.

NAME	Molecular Formula	MW (g/mol)	Aq. Sol (mg/L)	Density (g/cm^3)	BP (°C)	MP (°C)
1,6-Cleve's acid	$C_{10}H_9NO_3S$	223.2	3000	1.502	434.2	173
1_naphthol	$C_{10}H_8O$	144.2	3176	1.100	279	95
2,4,5-trichlorophenol	$C_6H_3Cl_3O$	197.4	3079	1.600	248	68
2,4-DB	$C_{10}H_{10}Cl_2O_3$	249.1	1663	1.400	410	119
2,6-Dibromoquinone-4-chlorimide	$C_6H_2Br_2ClNO$	299.3	1770	2.200	296	80
2-Amino-5-bromobenzoic acid	$C_7H_6BrNO_2$	216.0	2260	1.800	342	213
2-Cyclohexyl-4,6-dinitrophenol	$C_{12}H_{14}N_2O_5$	266.2	1168	1.400	321	107
2-Ethyl-1-hexanol	$C_8H_{18}O$	130.2	2944	0.830	185	-76
2-Naphthol	$C_{10}H_8O$	144.2	740	1.280	285	122
3,4-Dinitrobenzoic acid	$C_7H_4N_2O_6$	212.1	3826	1.674	459	165
4-Amino-2-sulfobenzoic acid	$C_7H_7NO_5S$	217.2	3477	1.709	445.7	187
4-iodophenol	C_6H_5IO	220.0	3628	1.857	329.2	93
5-Aminosalicylic acid	$C_7H_7NO_3$	153.1	840	1.570	403.9	280
5-Bromo-2,4-dihydroxybenzoic acid	$C_7H_5BrO_4$	233.0	2747	2.026	436.7	209
Acetaminophen	$C_8H_9NO_2$	151.2	4114	1.293	420	170
Acetamiprid	$C_{10}H_{11}ClN_4$	222.7	3623	1.330	352	99
Acetanilide	C_8H_9NO	135.2	3806	1.121	304	115
Acetazolamide	$C_4H_6N_4O_3S_2$	222.2	2991	1.744	514	258
Acetochlor	$C_{14}H_{20}ClNO_2$	269.8	2348	1.107	172	1
Acetylacetone	$C_5H_8O_2$	100.1	5221	0.975	140	-23
Acibenzolar-S-methyl	$C_8H_6N_2OS_2$	210.3	887	1.500	267	133
Acrylamide	C_3H_5NO	71.1	5806	1.120	125	84
Acylonitrile	C_3H_3N	53.1	4872	0.801	77	-83
Adenine	$C_5H_5N_5$	135.1	3013	1.612	553.5	360
Adenosine	$C_{10}H_{13}N_5O_4$	267.2	5100	2.080	676.3	234
Adipic acid	$C_6H_{10}O_4$	146.1	4414	1.360	337.5	152
Aldicarb	$C_7H_{14}N_2O_2S$	190.3	3780	1.195	225	100
Allobarbital	$C_{10}H_{12}N_2O_3$	208.2	3258	1.100	468.8	172
Allopurinol	$C_5H_4N_4O$	136.1	569	1.890	423.3	350
Alochlor	$C_{14}H_{20}ClNO_2$	269.8	2380	1.133	100	40
Alpha-acetylbutyrolactone	$C_6H_8O_3$	128.1	5301	1.190	253	-12
Alprenolol	$C_{15}H_{23}NO_2$	249.3	2763	1.000	383.4	108
Amantadine	$C_{10}H_{17}N$	151.2	3326	1.100	373	206
Amitriptyline	$C_{20}H_{23}N$	277.4	892	1.100	398.2	196

NAME	Molecular Formula	MW (g/mol)	Aq. Sol (mg/L)	Density (g/cm^3)	BP (°C)	MP (°C)
Amobarbital	$C_{11}H_{18}N_2O_3$	226.3	2780	1.138	367.9	157
Ancymidol	$C_{15}H_{16}N_2O_2$	256.3	2813	1.300	442.2	110
Aniline	C_6H_7N	93.1	4556	1.020	184.1	-6
Antipyrine	$C_{11}H_{12}N_2O$	188.2	5665	1.190	319	114
ANTU(α-Naphthylthiourea)	$C_{11}H_{10}N_2S$	202.3	2778	1.333	377.6	188
Arabinose	$C_5H_{10}O_5$	150.1	5698	1.585	333.2	158
Ascorbic acid	$C_6H_8O_6$	176.1	5522	1.694	553	191
Aspartic acid	$C_4H_7NO_4$	133.1	3912	1.700	324	270
Aspirin	$C_9H_8O_4$	180.2	3663	1.400	321	135
Asulam	$C_8H_{10}N_2O_4S$	230.2	3699	1.460	382.3	144
Atropine	$C_{17}H_{23}NO_3$	289.4	3459	1.200	429.8	115
Azathioprine	$C_9H_7N_7O_2S$	277.3	2235	1.900	685.7	243
Azintamide	$C_{10}H_{14}ClN_3OS$	259.8	3699	1.270	435.3	97
Baclofen	$C_{10}H_{12}ClNO_2$	213.7	4549	1.300	364.3	207
Badische acid	$C_{10}H_9NO_3S$	223.2	2775	1.500	434.2	173
Barban	$C_{11}H_9Cl_2NO_2$	258.1	1042	1.403	224	75
Barbital	$C_8H_{12}N_2O_3$	184.2	3873	1.100	507.8	190
Bendiocarb	$C_{11}H_{13}NO_4$	223.2	2415	1.250	299	130
Benzidine	$C_{12}H_{12}N_2$	184.2	2505	1.250	401	127
Benzocaine	$C_9H_{11}NO_2$	165.2	2898	1.170	310	89
Benzoic acid	$C_7H_6O_2$	122.1	3350	1.266	249.2	122
Benzylimidazole	$C_{10}H_{10}N_2$	158.2	2942	1.220	310	70
Bromogramine	$C_{11}H_{13}BrN_2$	253.1	1348	1.500	346.9	160
Bronidox	$C_4H_6BrNO_4$	212.0	5737	1.830	280	60
Bupivacaine	$C_{18}H_{28}N_2O$	288.4	2236	1.000	423.4	107
Butamben	$C_{11}H_{15}NO_2$	193.2	182	1.078	303.6	58
Butylparaben	$C_{11}H_{14}O_3$	194.2	198	1.280	369.2	68
Capric acid	$C_{10}H_{20}O_2$	172.3	1791	0.900	269	31
Caproic acid	$C_6H_{12}O_2$	116.2	4012	0.930	203	-3
Carbamazepine	$C_{15}H_{12}N_2O$	236.3	150	1.296	411	191
Carbofuran	$C_{12}H_{15}NO_3$	221.2	2505	1.180	200	148
Carfentrazone-ethyl	$C_{15}H_{14}Cl_2F_3N_3O_3$	412.2	1343	1.457	352.5	-22
Carisoprodol	$C_{12}H_{24}N_2O_4$	260.2	2477	1.100	423	92
Carmustine	$C_5H_9Cl_2N_3O_2$	214.0	3602	1.500	309.5	31
Carnosine	$C_9H_{14}N_4O_3$	226.2	4914	1.400	656	253
Carprofen	$C_{15}H_{12}ClNO_2$	273.7	740	1.400	509	197
Carvedilol	$C_{24}H_{26}N_2O_4$	406.5	1354	1.300	655	114
Cephalothin	$C_{16}H_{16}N_2O_6S_2$	396.4	2660	1.600	757.2	160
Chloramphenicol	$C_{11}H_{12}Cl_2N_2O_5$	323.1	3186	1.547	644.9	151
Chlorpheniramine	$C_{16}H_{19}ClN_2$	274.8	2771	1.100	142	132
Chlorpromazine	$C_{17}H_{19}ClN_2S$	318.9	431	1.200	450	57
Chlorthalidone	$C_{14}H_{11}ClN_2O_4S$	338.8	120	1.600	559.8	225
Chlorzoxazone	$C_7H_4ClNO_2$	169.6	1000	1.486	336	192
Cimetidine	$C_{10}H_{16}N_6S$	252.3	3710	1.300	488	142

NAME	Molecular Formula	MW (g/mol)	Aq. Sol (mg/L)	Density (g/cm^3)	BP (°C)	MP (°C)
Ciprofloxacin	$C_{17}H_{18}FN_3O_3$	331.3	1924	1.500	582	255
Corticosterone	$C_{21}H_{30}O_4$	346.5	199	1.200	500	145
Cortisone	$C_{21}H_{28}O_5$	360.4	255	1.300	534	222
Crotonic Acid	$C_4H_6O_2$	86.1	4934	1.027	185	71
Cumic Acid	$C_{10}H_{12}O_2$	164.2	2179	1.100	271.8	118
Cyanazine	$C_9H_{13}ClN_6$	240.7	2233	1.300	442.4	167
Cyanuric Acid	$C_3H_3N_3O_3$	129.1	3301	2.000	793.4	320
Cyclizine	$C_{18}H_{22}N_2$	266.4	3000	1.100	363.7	105
Cyclobarbital	$C_{12}H_{16}N_2O_3$	236.3	3204	1.200	549	172
Cycloleucine	$C_6H_{11}NO_2$	129.2	4698	1.200	420	328
Cyproconazole	$C_{15}H_{18}ClN_3O$	291.8	2146	1.300	375	106
Cyprodinil	$C_{14}H_{15}N_3$	225.3	1114	1.200	406	76
Cystine	$C_6H_{12}N_2O_4S_2$	240.3	2049	1.600	387	246
Cytosine	$C_4H_5N_3O$	111.1	7543	1.600	283.2	91
Danofloxacin	$C_{19}H_{20}FN_3O_3$	357.4	2654	1.500	569	317
Dapsone	$C_{12}H_{12}N_2O_2S$	248.3	150	1.400	475	175
Dehydroacetic Acid	$C_8H_8O_4$	168.1	2839	1.300	270	111
Deoxycorticosterone	$C_{21}H_{30}O_3$	330.5	145	1.200	456	141
Deprenyl	$C_{13}H_{17}N$	187.3	2760	1.000	272.5	141
Desipramine	$C_{18}H_{22}N_2$	266.4	1799	1.000	407	216
Dexamethasone	$C_{22}H_{29}FO_5$	392.5	1949	1.300	538	263
Diazepam	$C_{16}H_{13}ClN_2O$	284.7	1699	1.300	497.4	128
Diazoxide	$C_8H_7ClN_2O_2S$	230.7	2000	1.600	415	330
Dicamba	$C_8H_6Cl_2O_3$	221.0	2920	1.500	326	115
Dichlobenil	$C_7H_3Cl_2N$	172.0	1327	1.309	270	145
Difenoconazole	$C_{19}H_{17}Cl_2N_3O_3$	406.3	1177	1.400	220	76
Difloxacin	$C_{21}H_{19}F_2N_3O_3$	399.4	2000	1.400	595	322
Digallic Acid	$C_{14}H_{10}O_9$	322.2	2699	1.800	565	268
Diltiazem	$C_{22}H_{26}N_2O_4S$	414.5	2458	1.300	594	210
Dimethenamid	$C_{12}H_{18}ClNO_2S$	275.8	3079	1.187	383	139
Dimethirimol	$C_{11}H_{19}N_3O$	209.3	3079	1.100	350	102
Diphenydramine	$C_{17}H_{21}NO$	255.4	2461	1.000	343.7	168
Diphenylhydantoin (Phenytoin)	$C_{15}H_{12}N_2O_2$	252.3	1544	1.300	464	295
DL-Camphor	$C_{10}H_{16}O$	152.2	1600	0.992	204	180
Enrofloxacin (Baytril)	$C_{19}H_{22}FN_3O_3$	359.4	2375	1.400	560	221
EPTC	$C_9H_{19}NOS$	189.3	2574	0.955	232	60
Equilin	$C_{18}H_{20}O_2$	268.3	150	1.200	459	239
Ethinamate	$C_9H_{13}NO_2$	167.2	3398	1.100	237.3	96
Ethirimol	$C_{11}H_{19}N_3O$	209.3	2301	1.100	365.7	160
Ethofumesate	$C_{13}H_{18}O_5S$	286.3	1699	1.300	409.1	71
Ethohexadiol	$C_8H_{18}O_2$	146.2	4623	0.900	244	-40
Ethoprop	$C_8H_{19}O_2PS_2$	242.3	2875	1.100	310.2	-13
Ethylparaben	$C_9H_{10}O_3$	166.2	885	1.171	297.5	117

NAME	Molecular Formula	MW (g/mol)	Aq. Sol (mg/L)	Density (g/cm^3)	BP (°C)	MP (°C)
Famotidine(Pepcid)	$C_8H_{15}N_7O_2S_3$	337.5	2881	1.800	662.4	163
Fenbufen	$C_{16}H_{14}O_3$	254.3	344	1.157	470.2	186
Fenoprofen	$C_{15}H_{14}O_3$	242.3	1681	1.200	381.3	169
Fenpiclonil	$C_{11}H_6Cl_2N_2$	237.1	682	1.500	437.5	150
Fludrocortisone	$C_{21}H_{29}FO_5$	380.4	2146	1.300	564.7	261
Flufenacet	$C_{14}H_{13}F_4N_3O_2S$	363.3	1748	1.312	401.5	79
Flumequine	$C_{14}H_{12}FNO_3$	261.2	1681	1.500	439.7	254
Flumioxazin	$C_{19}H_{15}FN_2O_4$	354.3	253	1.500	644.4	204
Flurbiprofen	$C_{15}H_{13}FO_2$	244.3	1235	1.200	162.4	110
Fluspirilene	$C_{29}H_{31}F_2N_3O$	475.6	1000	1.300	668.9	189
Fumaric acid	$C_4H_4O_4$	116.1	3845	1.500	355	287
Furazolidone	$C_8H_7N_3O_5$	225.2	1603	1.700	353.4	255
Ganciclovir	$C_9H_{13}N_5O_4$	255.2	3633	1.360	398.5	250
Glipizide	$C_{21}H_{27}N_5O_4S$	445.5	161	1.300	689	208
Gluconolactone	$C_6H_{10}O_6$	178.1	5770	1.700	446	155
Glutamic acid	$C_5H_9NO_4$	147.1	3933	1.538	333.8	205
Glycine	$C_2H_5NO_2$	75.1	5396	1.600	240.9	240
Glyphosate	$C_3H_8NO_5P$	169.1	4079	1.704	465.8	215
Guaifenesin	$C_{10}H_{14}O_4$	198.2	4698	1.200	215	80
Guanine	$C_5H_5N_5O$	151.1	748	2.200	493.8	300
Haloperidol	$C_{21}H_{23}ClFNO_2$	375.9	1147	1.200	529	149
Heptabarbital	$C_{13}H_{18}N_2O_3$	250.3	2398	1.300	427.4	174
Hexazinone	$C_{12}H_{20}N_4O_2$	252.3	4519	1.300	332.8	116
Hexobarbital	$C_{12}H_{16}N_2O_3$	236.3	2699	1.200	530.7	146
Histidine	$C_6H_9N_3O_2$	155.2	4658	1.400	458.9	282
Hydrochlorothiazide	$C_7H_8ClN_3O_4S_2$	297.7	722	1.700	577	274
Hydrocortisone	$C_{21}H_{30}O_5$	362.5	2505	1.081	414.1	218
Hydroflumethiazide	$C_8H_8F_3N_3O_4S_2$	331.3	2516	1.700	531.6	272
Hydroquinone	$C_6H_6O_2$	110.1	4857	1.300	286	172
Hydroxyphenamate	$C_{11}H_{15}NO_3$	209.2	4397	1.200	415.4	55
Hydroxyproline	$C_5H_9NO_3$	131.1	5557	1.400	355.2	273
Hymexazol	$C_4H_5NO_2$	99.1	4929	1.200	363.6	86
Hyoscyamine	$C_{17}H_{23}NO_3$	289.4	3560	1.200	429.8	108
Ibuprofen	$C_{13}H_{18}O_2$	206.3	1716	1.030	364.8	76
Idoxuridine	$C_9H_{11}IN_2O_5$	354.2	3301	2.100	573.0	191
Imazapyr	$C_{13}H_{15}N_3O_3$	261.3	4053	1.300	425.1	171
Imazaquin	$C_{17}H_{17}N_3O_3$	311.3	1955	1.400	609.3	221
Imazethapyr	$C_{15}H_{19}N_3O_3$	289.3	3146	1.300	446.8	171
Indoprofen	$C_{17}H_{15}NO_3$	281.3	128	1.300	511.3	213
Iridomyrmecin	$C_{10}H_{16}O_2$	168.2	3301	1.000	270.5	60
Isoflurophate	$C_6H_{14}FO_3P$	184.1	4187	1.060	183	-82
Isoleucine	$C_6H_{13}NO_2$	131.2	4536	1.000	408.1	268
Isoniazid	$C_6H_7N_3O$	137.1	5146	1.200	329.8	171
Isophorone	$C_9H_{14}O$	138.2	4079	0.922	215	-8
Ketanserin	$C_{22}H_{22}FN_3O_3$	395.4	1000	1.300	607.5	231

NAME	Molecular Formula	MW (g/mol)	Aq. Sol (mg/L)	Density (g/cm^3)	BP (°C)	MP (°C)
Khellin	$C_{14}H_{12}O_5$	260.2	3017	1.300	587	154
Lindane	$C_6H_6Cl_6$	290.8	864	1.600	323.3	112
Linuron	$C_9H_{10}Cl_2N_2O_2$	249.1	1876	1.490	185	93
Lomefloxacin	$C_{17}H_{19}F_2N_3O_3$	351.3	3212	1.300	542.7	240
Malathion	$C_{10}H_{19}O_6PS_2$	330.4	2159	1.300	385.1	3
Maprotiline	$C_{20}H_{23}N$	277.4	748	1.100	399.6	93
Methocarbamol	$C_{11}H_{15}NO_5$	241.2	7200	1.300	472.5	93
Methomyl(Lannate)	$C_5H_{10}N_2O_2S$	162.2	4763	1.200	228	78
Methylparaben (Methyl-p-hydroxybenzoate)	$C_8H_8O_3$	152.1	2500	1.460	275	131
Metoclopramide	$C_{14}H_{22}ClN_3O_2$	299.8	1914	1.200	418.7	147
Metronidazole	$C_6H_9N_3O_3$	171.1	4012	1.399	301.12	159
Miconazole	$C_{18}H_{14}Cl_4N_2O$	416.1	544	1.400	555.1	161
Minoxidil	$C_9H_{15}N_5O$	209.2	2200	1.520	348.6	248
Nadolol	$C_{17}H_{27}NO_4$	309.4	8330	1.190	526.4	125
Nalidixic acid	$C_{12}H_{12}N_2O_3$	232.2	1756	1.224	374.4	229
Naloxone	$C_{19}H_{21}NO_4$	327.4	2617	1.400	532.8	202
Naproxen	$C_{14}H_{14}O_3$	230.3	863	1.200	404	153
Niflumic acid	$C_{13}H_9F_3N_2O_2$	282.2	845	1.400	378	203
Nitrofurantoin	$C_8H_6N_4O_5$	238.2	2067	1.582	380.8	268
Norfloxacin	$C_{16}H_{18}FN_3O_3$	319.3	2800	1.300	695.6	220
Nortriptyline	$C_{19}H_{21}N$	263.4	1398	1.100	403.4	214
Ofloxacin	$C_{18}H_{20}FN_3O_4$	361.4	4292	1.500	571.5	246
Oxytetracycline	$C_{22}H_{24}N_2O_9$	460.4	2580	1.634	727.8	183
p-Aminobenzoic acid	$C_7H_7NO_2$	137.1	5390	1.374	340	188
p-Aminosalicylic acid	$C_7H_7NO_3$	153.1	1690	1.490	347	150
Papaverine	$C_{20}H_{21}NO_4$	339.4	1663	1.200	483.2	147
p-Fluorobenzoic acid	$C_7H_5FO_2$	140.1	3079	1.300	253.7	183
Phenacetin	$C_{10}H_{13}NO_2$	179.2	766	1.000	243	134
Phenantroline	$C_{12}H_8N_2$	180.2	3638	1.300	330	117
Phenazopyridine	$C_{11}H_{11}N_5$	213.2	1137	1.300	277.4	139
Phenobarbital	$C_{12}H_{12}N_2O_3$	232.2	3072	1.200	568.8	175
Phenolphthalein	$C_{20}H_{14}O_4$	318.3	2603	1.300	558	260
Phenylbutazone	$C_6H_6O_4$	308.4	1098	1.200	425	105
Phenytoin	$C_{15}H_{12}N_2O_2$	252.3	1412	1.300	464	295
Phthalazine	$C_8H_6N_2$	130.1	4698	1.200	317	90
Phthalic acid	$C_8H_6O_4$	166.1	3730	1.593	378.3	210
Phthalimide	$C_8H_5NO_2$	147.1	2556	1.210	366	234
p-Hydroxybenzoic Acid	$C_7H_6O_3$	138.1	3699	1.460	336	215
Picloram	$C_6H_3Cl_3N_2O_2$	241.5	2633	1.800	421	209
Picric Acid	$C_6H_3N_3O_7$	229.1	4103	1.850	300	122
Pindolol	$C_{14}H_{20}N_2O_2$	248.3	1602	1.200	457.1	169

NAME	Molecular Formula	MW (g/mol)	Aq. Sol (mg/L)	Density (g/cm^3)	BP (°C)	MP (°C)
Piroxicam	$C_{15}H_{13}N_3O_4S$	331.3	716	1.500	568.5	200
Praziquantel	$C_{19}H_{24}N_2O_2$	312.4	400	1.200	544	137
Prednisolone	$C_{21}H_{28}O_5$	360.4	223	1.300	570	235
Primidone	$C_{12}H_{14}N_2O_2$	218.2	500	1.200	443	281
Procaine	$C_{13}H_{20}N_2O_2$	236.3	3653	1.100	373.6	61
Propranolol	$C_{16}H_{21}NO_2$	259.3	1919	1.100	434.9	96
Propylparaben	$C_{10}H_{12}O_3$	180.2	500	1.100	294	96
Quinidine	$C_{20}H_{24}N_2O_2$	324.4	140	1.200	496	174
Quinine	$C_{20}H_{24}N_2O_2$	324.4	2724	1.200	496	177
Ranitidine	$C_{13}H_{22}N_4O_3S$	314.4	2996	1.200	437.1	70
Salicylamide	$C_7H_7NO_2$	137.1	2060	1.300	348.5	140
Salicylic acid	$C_7H_6O_3$	138.1	2240	1.400	373	158
Sparfloxacin	$C_{19}H_{22}F_2N_4O_3$	392.4	2335	1.400	640.4	265
Strychnine	$C_{21}H_{22}N_2O_2$	334.4	180	1.360	560	280
Sulfacetamide	$C_8H_{10}N_2O_3S$	214.2	8293	1.400	450	183
Sulfamerazine	$C_{11}H_{12}N_4O_2S$	264.3	202	1.400	519	236
Sulfamethazine	$C_{12}H_{14}N_4O_2S$	278.3	2706	1.460	526	176
Sulfamethoxazole	$C_{10}H_{11}N_3O_3S$	253.3	610	1.500	482	171
Sulfanilamide	$C_6H_8N_2O_2S$	172.2	7500	1.400	400	165
Sulfathiazole	$C_9H_9N_3O_2S_2$	255.3	2718	1.600	480	202
Sulindac	$C_{20}H_{17}FO_3S$	356.4	1041	1.400	582	183
Sulpiride	$C_{15}H_{23}N_3O_4S$	341.4	2280	1.200	530	179
Testosterone	$C_{19}H_{28}O_2$	288.4	1390	1.100	432.9	154
Tetracaine	$C_{15}H_{24}N_2O_2$	264.4	2412	1.000	389.4	149
Tetracycline	$C_{22}H_{24}N_2O_8$	444.4	2722	1.700	738.2	170
Theobromine	$C_7H_8N_4O_2$	180.2	330	1.600	483.5	350
Theophylline	$C_7H_8N_4O_2$	180.2	7360	1.500	454	273
Thiamphenicol	$C_{12}H_{15}Cl_2NO_5S$	356.2	3560	1.500	696	165
Thionazin	$C_8H_{13}N_2O_3PS$	248.2	3057	1.300	307	-2
Thymine	$C_5H_6N_2O_2$	126.13	3820	1.200	378	316
Thymol	$C_{10}H_{14}O$	150.2	2991	0.970	233	50
Tolmetin	$C_{15}H_{15}NO_3$	257.3	1322	1.200	483.2	156
Trichloromethiazide	$C_8H_8Cl_3N_3O_4S_2$	380.7	2053	1.700	631.3	250
Trimethoprim	$C_{14}H_{18}N_4O_3$	290.3	2512	1.300	405	201
Trimipramine	$C_{20}H_{26}N_2$	294.4	681	1.000	411.8	45
Tryptamine	$C_{10}H_{12}N_2$	160.2	1903	1.200	378.8	115
Uracil	$C_4H_4N_2O_2$	112.1	3600	1.300	367	330
Verapamil	$C_{27}H_{38}N_2O_4$	454.6	1682	1.100	586.2	228
Warfarin	$C_{19}H_{16}O_4$	308.3	708	1.300	515.2	162

Based on molecular properties listed in Table 1, the following additional molecular properties are defined as input arguments for calculation of subsequent molecular properties, where the latter will serve as input (predictor) variables. To demonstrate how such properties are calculated, Table 2 shows molecular properties of 1,6-Cleve's acid.

Table 2: 1,6-Cleve's acid as an example to demonstrate the definition of additional molecular properties.

NAME	Molecular Formula	MW	Aq. Sol (mg/L)	Density (g/cm^3)	BP (°C)	MP (°C)
1,6-Cleve's acid	$C_{10}H_9NO_3S$	223.3	3000	1.502	434.2	173

The following equations represent the molecular (or mole-) fraction of each atomic species as part of the molecular constituent of a given drug compound.

$$C_{Frac} = \frac{Number\ \ of\ \ C\ \ Atoms}{Total\ \ Number\ \ of\ \ Atoms} = \frac{10}{24} = 0.41666 \tag{1}$$

$$H_{Frac} = \frac{Number\ \ of\ \ H\ \ Atoms}{Total\ \ Number\ \ of\ \ Atoms} = \frac{9}{24} = 0.37500 \tag{2}$$

$$N_{Frac} = \frac{Number\ \ of\ \ N\ \ Atoms}{Total\ \ Number\ \ of\ \ Atoms} = \frac{1}{24} = 0.04167 \tag{3}$$

$$O_{Frac} = \frac{Number\ \ of\ \ O\ \ Atoms}{Total\ \ Number\ \ of\ \ Atoms} = \frac{3}{24} = 0.12500 \tag{4}$$

$$S_{Frac} = \frac{Number\ \ of\ \ S\ \ Atoms}{Total\ \ Number\ \ of\ \ Atoms} = \frac{1}{24} = 0.04167 \tag{5}$$

Notice that $C_{Frac} + H_{Frac} + N_{Frac} + O_{Frac} + S_{Frac} = 1$ (6)

Table 3 shows the electronegativity value [9] for each atom present in the previous drug molecules.

Table 3: Electronegativity (eV/electron)♦ of atoms constituting drug molecules.

Atom	Electronegativity (eV)/electron♦	Atom	Electronegativity (eV)/electron♦
H	13.6	**S**	13.6
C	13.9	**Cl**	16.3
N	16.9	**Br**	15.2
O	18.6	**I**	13.4
F	23.3	**P**	12.8

♦*[9]:* Martin Rahm, Tao Zeng, Roald Hoffmann. "Electronegativity Seen as the Ground-State Average Valence Electron Binding Energy". Journal of the American Chemical Society 2019, 141: 342−351.

The following equations define the electronegativity contribution for each atomic species, based on its mole-fraction times its atomic electronegativity, as shown in Table 2. For example, let us take 1,6-Cleve's acid, $C_{10}H_9NO_3S$, then X_C, represents the electronegativity contribution of carbon atoms.

$$X_C = C_{Frac} \times EN_C = 0.41666 \times 13.9 = 5.7916 \ \ eV \tag{7}$$

Other atomic contributions are shown below:

$$X_H = H_{Frac} \times EN_H = 0.37500 \times 13.6 = 5.1000 \ \ eV \tag{8}$$

$$X_{Halo} = Halo_{Frac} \times EN_{Halogen} = 0 \times EN_{Halo} = 0.0 \ \ eV \tag{9}$$

where halo stands for a halogen atom, like F, Cl, Br, and I.

$$X_N = N_{Frac} \times EN_N = 0.04167 \times 16.9 = 0.7042 \ \ eV \tag{10}$$

$$X_O = O_{Frac} \times EN_O = 0.12500 \times 18.6 = 2.3250 \ \ eV \tag{11}$$

$$X_P = P_{Frac} \times EN_P = 0 \times 12.8 = 0.0 \ \ eV \tag{12}$$

$$X_S = S_{Frac} \times EN_S = 0.04167 \times 13.6 = 0.5671 \ \ eV \tag{13}$$

$$X_{Total} = \sum_{i=1}^{n} X_i = X_C + X_H + X_{Halo} + X_N + X_O + X_P + X_S \tag{14}$$

$$PolFrac = \frac{X_{Polar}}{X_{Total}} = \frac{[X_F+X_O+X_N+X_{Cl}+X_{Br}]}{X_{non-Polar}+X_{Polar}} = \frac{[X_F+X_O+X_N+X_{Cl}+X_{Br}]}{[\{X_C+X_H+X_I+X_P+X_S\}+\{X_F+X_O+X_N+X_{Cl}+X_{Br}\}]} \tag{15}$$

Notice that the electronegativities of F, O, N, Cl, and Br atoms have relatively higher values than those of C, H, I, P, and S atoms.

$$MW_{POL}[\frac{g}{mol}] = (PoleFrac) \times MW \tag{16}$$

$$MW_{NPOL}[\frac{g}{mol}] = (1 - PoleFrac) \times MW \tag{17}$$

$$MolVol \ \ \left[\frac{L}{mol}\right] = \frac{MW \ \left[\frac{g}{mol}\right]}{\left(Density\left[\frac{g}{cm^3}\right]\right) \times \frac{1{,}000 \ \ cm^3}{L}} \tag{18}$$

$$NPolVol \ \ \left[\frac{L}{mol}\right] = (1.0 - PolFrac) \times MolVol \tag{19}$$

$$BPMPR = \left(\frac{\Delta T_{Liquid}}{\Delta T_{Solid}}\right) = \left(\frac{T_{BP}-T_{MP}}{T_{MP}-0K}\right) = \frac{T_{BP}}{T_{MP}} - 1 = \left\{\frac{BP[^\circ C]+273}{MP[^\circ C]+273} - 1\right\} > 0 \tag{20}$$

The following five linearly independent predictors are included in the analysis:

$BPMPR, X_{POL}, PolFrac$, MW_{NPOL}, *and* , $NPolVol$. The aqueous solubility, expressed in mg/L, is the response variable.

3. RESULTS AND DISCUSSION

3.1 Raw Data Acquisition

The reason for the normalization step of raw data (i.e., original predictors' data, X) is simply to make the predictors likely equal in terms of the foothold (weight) and distance (lever) separating each from the response variable. Moreover, if the original predictor has some physical dimensions, then the normalization will transform the predictor into a dimensionless property. For example, molecular properties: $MW_{NPOL}, X_{POL}, and\ \ NPolVol$ are all dimensional whereas PolFrac and BPMPR are both dimensionless. ML algorithms deal with predictors from an abstract (i.e., dummy arguments) point of view. Hence, the normalization step is needed in this regard. Of course, the normalized data will scatter between 0 and 1.

Figure 1 shows MATLAB code, used in all upcoming m-files intended for carrying out subsequent machine learning or optimization step. The code simply fetches molecular properties from Table 1, make data acquisition to define new molecular properties (equations 1 up to 20) and normalize the data for further analysis. Each line of code is preceded by a comment statement to explain what it means.

```
%% RAW MOLECULAR DATA ACQUISITION.
%% The data found in Table 1 will be converted into a numeric 246x5 matrix.
% The matrix represents five molecular predictors of 246 drug molecules.
% Reading # of constituting atoms of a molecule.
Cnum=DrugSol4d.Cnum;
Hnum=DrugSol4d.Hnum;
Nnum=DrugSol4d.Nnum;
Onum=DrugSol4d.Onum;
Snum=DrugSol4d.Snum;
Fnum=DrugSol4d.Fnum;
Clnum=DrugSol4d.Clnum;
Brnum=DrugSol4d.Brnum;
Inum=DrugSol4d.Inum;
Pnum=DrugSol4d.Pnum;
% Reading the response variable; i.e., solubility in mg/L.
resp=DrugSol4d.Sol_PPM;
% Reading total number of atoms for each drug molecule.
TotAtom=DrugSol4d.TotAtom;
% Reading the boiling point/melting point ratio.
BPMPR=DrugSol4d.BPMPR;
% BPMPR=normalize(BPMPR,'range');
% Reading the molar volume of a drug molecule.
MolVol=DrugSol4d.MolVol;
% Reading the molecular mass of a drug molecule.
MW=DrugSol4d.MW;
% Defining the electronegativity contribution of each atom in a drug molecule.
XC=13.9*(Cnum./TotAtom);
XH=13.6*(Hnum./TotAtom);
```

```
XN=16.9*(Nnum./TotAtom);
XO=18.6*(Onum./TotAtom);
XS=13.6*(Snum./TotAtom);
XF=23.3*(Fnum./TotAtom);
XCl=16.3*(Clnum./TotAtom);
XBr=15.2*(Brnum./TotAtom);
XI=13.4*(Inum./TotAtom);
XP=12.8*(Pnum./TotAtom);
% Defining the electronegativity contribution of polar atoms.
XPol=XN+XO+XF+XCl+XBr;
% Defining the electronegativity contribution of non-polar atoms.
XNPol=XC+XH+XS+XI+XP;
% Defining the polar fraction of a drug molecule.
PolFrac=XPol./(XPol+XNPol);
% Defining the non-polar molecular mass of a drug molecule.
MWNPOL=MW.*(1.00-PolFrac);
% Defining the non-polar molar volume a drug molecule.
NPolVol=MolVol.*(1.00-PolFrac);
% Normalizing the five predictor raw data to vary between 0 and 1
Xraw(:,1)=BPMPR;
Xraw(:,2)=XPol;
Xraw(:,3)=PolFrac;
Xraw(:,4)=MWNPOL;
Xraw(:,5)=NPolVol;
X=normalize(Xraw, 'range');
% defining the labels of predictors to be used later.
labels={'BPMPR','XPol','PolFrac','MWNPOL','NPolVol'};
% Normalizing the response variable, Y, to vary between 0 and 1.
Y=normalize(resp,'range');
```

Figure 1: MATLAB code to define and normalize molecular and solubility data for further analysis.

3.2 P-Value Prediction, Using Least Square Boosted Regression Ensemble

Figure 2 shows that MATLAB's ML **fitrensemble** function is used. The function: **tModel** = **fitrensemble(X,Y,'Method','LSBoost', …)** returns optimized hyperparameters of a boosted regression ensemble, using the linear square boost (LSBoost) algorithm and using surrogate splits, based on the predictor, X, and response, Y, data. The additional arguments are meant to further improve the optimization of the resulting model by varying the number of learning cycles, the maximum number of surrogate splits, and the learn rate. Furthermore, the optimization has flexibility to repartition the cross-validation between every iteration. For a better reproducibility, the random seed is set and the expected-improvement-plus acquisition function is used.

```
% For reproducibility, set the random seed.
rng default;
tModel = fitrensemble(X,Y,'Method','LSBoost','Learner',templateTree('Surrogate','on'),...
    'OptimizeHyperparameters',{'NumLearningCycles','MaxNumSplits','LearnRate'},...
    'HyperparameterOptimizationOptions',struct('Repartition',true,...
    'AcquisitionFunctionName','expected-improvement-plus'));
%%predictorImportance function outputs the probability (i.e., how important)for each
predictor.
p = predictorImportance(tModel);
% Sorting the predictors based on their p-values in descending order.
[sortedp,idp]=sort(p,'descend');
figure(3);
% View predictor importance on a bar plot
bar(sortedp)
%Assign labels in light of re-ordering the predictors.
Predictlabel=labels(idp);
% Define the x-axis labels.
xticklabels(Predictlabel);
% Define the y-axis label.
ylabel('Probability');
```

Figure 2: MATLAB code for the ensemble of least square boosted regression, as well as, predicting and presenting the importance of each predictor.

The result of executing both codes shown in figures 1 and 2 is depicted in Figure 3, where it shows, in descending order, the importance of each predictor, expressed in terms of its p-value. Based on the exploited least square algorithm of boosted regression ensemble, MW_{NPOL} (Eq. 17), $NPolVol$ (Eq. 19) and $PolFrac$ (Eq. 15) are the first three important molecular properties which can explain variation in aqueous solubility.

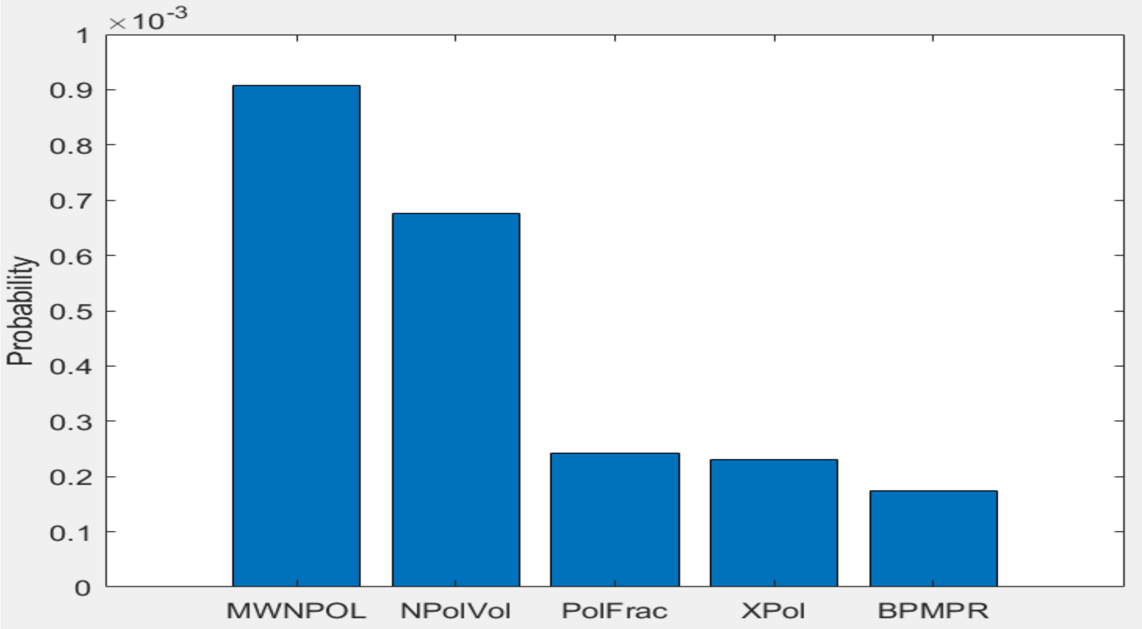

Figure 3: Prediction of importance for the five predictors, using fitrensemble model.

3.3 Principle Component Analysis (PCA)

Figure 4 shows the code for applying MATLAB ML Principal Component Analysis (PCA) of X (predictors) data, X, without incorporation of the response variable, Y. The command is:

[pcs,~,~,~,pexp,~] = pca(X); (21)

In Eq. (21), out of the six potential output terms, left-hand side of Eq. (21), the following two terms are defined:

pcs: The principal component coefficients, also known as loadings, for the $n \times p$ data matrix, X. n is number of data points and p number of predictors (or parameters). The coefficient matrix, pcs, is $p \times p$. Each column of pcs contains coefficients for one principal component, and the columns are in descending order of component variance.

pexp: The percentage of the total variance explained by each principal component. The concept of principal component analysis in ML simply aims at potential transformation of the original set made of the six predictors into a new set of less number of principal components. For example, in our case study, the original five predictors can be reduced to three principal components (with 97.2 % accuracy; see Figure 5) or even down to two principal components (with 88.5 % accuracy; see Figure 5). Obviously, the model accuracy decreases with decreasing the number of chosen principal components in the final list.

It should be noticed that the code present in Figure 1 must precede the code in Figure 4, below. It is omitted here to avoid redundancy in coding.

```
%% Method: Feature Transformation with Principal Component Analysis, PCA.
[pcs,~,~,~,pexp,~] = pca(X);
%[coeff,score,latent,tsquared,explained,mu]=pca(X);
% Prepare a window for the upcoming figure.
figure(5);
% Pareto charts display the values in the vector Y as bars drawn in
% descending order. Values in Y must be nonnegative and not include NaNs (not a number).
% Only the first 95% of the cumulative distribution is displayed.
pareto(pexp);
% Prepare x-axis
xticks([1 2 3]);
xticklabels({'PC#1', 'PC#2', 'PC#3'});
% Sort in descending order the percentage of the total variance
% explained by each principal component.
%[sortedp,idp]=sort(pexp,'descend');
% Pareto charts display the values in the vector Y as bars drawn in
% descending order. Values in Y must be nonnegative and not include NaNs.
% Only the first 95% of the cumulative distribution is displayed.
% Prepare a window for the image screen of predictors.
pcssqrd=pcs.^2;
figure(6);
% Plot a colored image screen showing the contribution of predictors to PC.
% imagesc(abs(pcs(:,1:3)));
imagesc(pcssqrd(:,1:3));
% Populate y-axis with predictor labels
yticks([1 2 3 4 5]);
yticklabels(labels);
% Populate x-axis with PC#1, PC#2, and PC#3.
xticks([1 2 3]);
xticklabels({'PC#1', 'PC#2', 'PC#3'});
colorbar;
```

Figure 4: MATLAB code for Principal Component Analysis (PCA) of X data, using squared principal component coefficients.

The results of executing both codes, shown in figures 1 and 4, are depicted in Figure 5 and 6. Figure 5 shows, in descending order, the first 95% of the cumulative distribution. In fact, the cumulative distribution amounts to 99.2 % of the total distribution of X. Notice that the first two principal components, together, account for 86.7 % of the total distribution of X. It is worth mentioning here that each principal component is a cumulative contribution, emanating from the original five predictors. The contribution of each individual predictor can be seen in Figure 6.

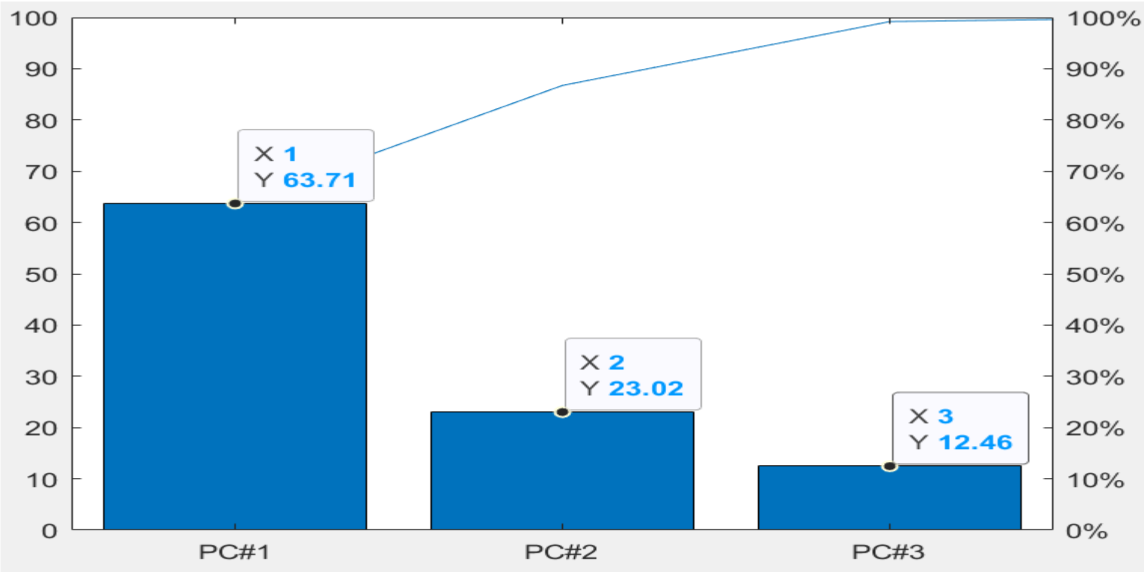

Figure 5: Pareto plot for 95 % cumulative distribution of principal components.

As shown in Figure 6, the contribution of each individual predictor is given in the form of colored area. Notice that I found a better approach; instead of taking the absoultute value of the coefficient, I take the square root of each coefficient where the sum of all coeffients will add up to unity for each column of the three principal component columns. The new squared matrix is named pcssqrd (5×5). One may conclude that NPolVol, MWNPOL, and PolFrac are the first three important predictors. The conclusion is in harmony with the previous finding, based on p-value (Figure 3), using the least square boosted regression ensemble.

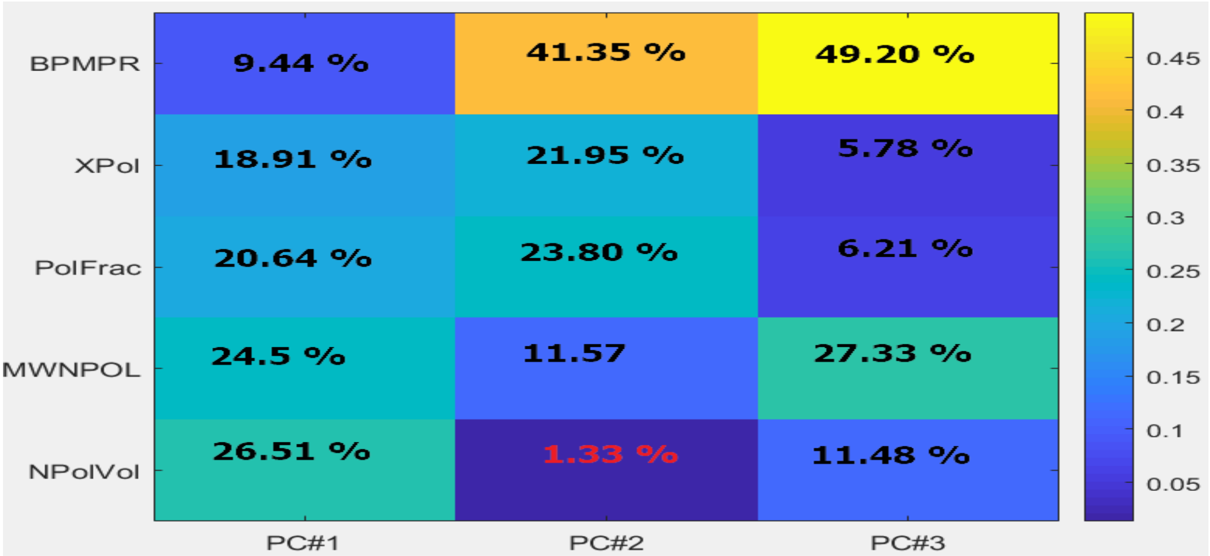

Figure 6: The colored image screen for the individual contribution of each original predictor as part of the first, second, and third principal component, PC, using the square value of each coefficient, where the sum adds up to unity for each PC.

3.4 Sequential Feature Selection

The following MATLAB code:

ferror = @(Xtrain,ytrain,Xtest,ytest) nnz(predict(fmodel(Xtrain,ytrain),Xtest) ~= ytest); (22)

creates an anonymous function named ferror that takes four inputs: Xtrain, ytrain, Xtest, and ytest, and returns the number of inaccurate predictions for ytest.

tokeep = sequentialfs(ferror,X,Y,'cv',part,'options',statset('Display','final')); (23)

selects a subset of features from the data matrix X, which best predicts the data in y by sequentially selecting features until there is no improvement in prediction. Rows of X correspond to observations; columns correspond to variables or features. Y is a column vector of response values or class labels for each observation in X. X and Ymust have the same number of rows. **ferror** is a function handle to a function that defines the criterion used to select features and to determine when to stop. The output **tokeep** is a logical vector indicating which features (or, predictor columns) are finally chosen. Notice here that there are more than one **fmodel** to fit into Eq. (22). Any of the following **fmodel** types can be used:

```
% Fit binary decision tree for multiclass classification.
fmodel = @(X,Y) fitcknn(X,Y,"NumNeighbors",20);                    (24)
```

```
% Fit multiclass models for support vector machines or other classifiers
temp=templateSVM("KernelFunction","polynomial");                    (25a)
fmodel=@(X,Y)fitcecoc(X,Y,"Learners",temp);                         (25b)

temp=templateSVM("KernelFunction","linear");                        (26a)
fmodel=@(X,Y)fitcecoc(X,Y,"Learners",temp);%                        (26b)

temp=templateSVM("KernelFunction","gaussian");                      (27a)
fmodel=@(X,Y)fitcecoc(X,Y,"Learners",temp);%                        (27b)

fmodel = @(X,Y) fitcnb(X,Y,'Distribution','kernel');                (28)
```

Figure 7 shows MATLAB code for sequential feature selection, in addition to the generation of a 3-D plot for aqueous solubility as a function of the first two sequentially selected predictors.

```
%% Perform sequential feature selection
% rng('default') puts the settings of the random number generator used
% by rand, randi, and randn to their default values.
rng('default');
% creates an object part that does not partition the data. Both the training
% set and the test set contain all of the original n observations.
part = cvpartition(Y,'resubstitution');
ti = cputime;
% Fit k-nearest neighbor classifier.%
fmodel = @(X,Y) fitcknn(X,Y,"NumNeighbors",20);
ferror = @(Xtrain,ytrain,Xtest,ytest) nnz(predict(fmodel(Xtrain,ytrain),Xtest) ~= ytest);
% The output tokeep is a logical vector indicating which features
% (or, predictor columns) are finally chosen.
tokeep = sequentialfs(ferror,X,Y,'cv',part,...
    'options',statset('Display','final'));
elapsetime=cputime-ti;
KeptX=X(:,tokeep);
X1=KeptX(:,1);
X2=KeptX(:,2);
figure(8);
mylabel=labels(tokeep);
plot3(X1,X2,Y,'o');
xlabel(mylabel(1,1),'FontSize',14,'FontWeight','bold');
ylabel(mylabel(1,2),'FontSize',14,'FontWeight','bold');
zlabel('Solubility','FontSize',14,'FontWeight','bold');
grid on;
```

Figure 7: Sequential feature selection for the five predictors, using sequential feature selection, based on one of the multi-class classifiers.

It is worth mentioning here that the first sequentially selected predictors vary from one fmodel case to another (equations 24 through 28) and also by varying some additional input parameters found in the selected fmodel equation itself.

Table 4 shows the results of attempting to first sequentially selecting predictors for each examined case. MWNPOL and PolFrac were first sequentially selected among the five predictors.

Table 4: The first two sequentially selected predictors using different fmodel equations

#	fmodel equation number	The first sequentially selected predictors		CPU Time
1	24	MWNPOL	PolFrac	0.4 s
2	25a-25b	No sequentially selected predictors		33.6 minute
3	26a-26b	No sequentially selected predictors		32.8 minute
4	27a-27b	No sequentially selected predictors		36.1 minute
5	28	MWNPOL	PolFrac	23.5 s

3.5 Curve-Fitting: Robust Least Squares

Based on the previously examined ML methods, one can conclude that the following three predictors turn out to be the most important in terms of explaining variation in Y as a function of X: MWNPOL, NPolVol, and PolFrac. Let us consider the solubility as a function of one set at a time and calculate robust least squares regression parameters.

For the sake of simplicity, I will pick up the following pairwise combination of predictors and examine the model goodness of each.

$$Sol\ \ (\frac{mg}{L}) = f(\text{MWNPOL}, \text{NPolVol}) \tag{29}$$

$$Sol\ \ (\frac{mg}{L}) = f(\text{MWNPOL}, \text{PolFrac}) \tag{30}$$

$$Sol\ \ (\frac{mg}{L}) = f(\text{NPolVol}, \text{PolFrac}) \tag{31}$$

Let us examine the three models and see which gives the best fit under robust least squares regression. It is usually assumed that the response errors follow a normal distribution, and that extreme values are rare. Still, extreme values, called outliers, do occur. The main disadvantage of least squares fitting is its sensitivity to outliers. Outliers have a large influence on the fit because squaring the residuals magnifies the effects of these extreme data points. To minimize the influence of outliers, one can fit his/her data using robust least-squares regression. The optimization toolbox provides these two robust regression methods. The Least Absolute Residuals (LAR) method finds a curve that minimizes the absolute difference of the residuals, rather than the squared differences. Therefore, extreme values have a lesser influence on the fit.

The other available robust method is the ‘bi-square’. This method minimizes a weighted sum of squares, where the weight given to each data point depends on how far the point is from the fitted line. Points near the line get full weight. Points farther from the line get reduced weight or even down to zero weight. This process of elimination can be set by comparing the absolute value of the residual of a given data point to the median absolute deviation of the residuals. The weight will be set to zero if the absolute value of the residual is greater than six times the median, for example. Both methods will work much better than the case when the robust option is disabled, if the predictor data is characterized by a large degree of scatter, which is the case of describing the aqueous solubility of drug-like molecules. Table 5 shows the robust linear least square results using the raw data, presented in Table 1. The weight factor, for each of the three cases, is indicated as X_3. To demonstrate the importance of the weight factor, consider the first model, Eq. (29); without the inclusion of the weight factor, X_3, as third “variable” in the regression process, the adjusted R^2 will drop from 0.9710 down to 0.2215. So does the case for the second model given by Eq. (30). For the third and last model, Eq. (31), the incorporation of a weight factor did not improve the regression process. In fact, the inclusion of the weight factor, X_3, in the third regression case kept the adjusted R^2 the same but the root mean square error (RMSE) was drastically blown up from 284.6 up to 3,693.

Table 5: Curve-fitted parameters using the least absolute residual regression with and without a weight factor.

Eq. #	**Selected Predictors** (X_1, X_2) X_3: **Weight Factor**	**Model Parameters:** $Aq_{Sol} = a \times X_1 + b \times X_2 + C$ **(95 % C.I.)**	**Adjusted R-square:**	**Root Mean Square Error (RMSE)**
29	(MWNPOL, NPolVol) X_3: PolFrac	*a* = -15 (-16.38, -13.62) *b = 4,697 (3,026, 6,367)* *c = 4,870 (4,776, 4,964)*	0.9710	130.0
30	(MWNPOL, PolFrac) X_3: NPolVol	*a = -9.533 (-10.01, -9.058)* *b = 1,175 (791.1, 1,559)* *c = 4,283 (4,135, 4,432)*	0.9708	100.2
31	(NPolVol, PolFrac) X_3: None	*a = -13,320 (-14,060, -12,570)* *b = -1,539 (-1,985, -1,093)* *c = 5,025 (4,836, 5,214)*	0.9696	284.6

3.6 Implementation of Curve-Fitted Model to Candidate Drugs

Let us take the 1,6-Cleve’s acid and apply model #1, presented in Section 3.5.

NAME	**Molecular Formula**	**MW**	**Aq. Sol (mg/L)**	**Density (g/cm³)**	**BP (°C)**	**MP (°C)**
1,6-Cleve’s acid	$C_{10}H_9NO_3S$	223.3	3000	1.502	434.2	173.4

$$MW = 10 \times 12 + 9 \times 1 + 1 \times 14 + 3 \times 16 + 1 \times 32 = 223 \quad (32)$$

$$X_C = C_{Frac} \times EN_C = \frac{10}{24} \times 13.9 = 5.792 \ \ eV \quad (33)$$

$$X_H = H_{Frac} \times EN_H = \frac{9}{24} \times 13.6 = 5.100 \ \ eV \quad (34)$$

$$X_N = N_{Frac} \times EN_N = \frac{1}{24} \times 16.9 = 0.7042 \ \ eV \quad (35)$$

$$X_O = O_{Frac} \times EN_O = \frac{3}{24} \times 18.6 = 2.325 \ \ eV \quad (36)$$

$$X_S = S_{Frac} \times EN_S = \frac{1}{24} \times 13.6 = 0.5667 \ \ eV \quad (37)$$

$$PolFrac = \frac{X_{Polar}}{X_{Total}} = \frac{[X_N + X_O]}{[X_C + X_H + X_S + X_N + X_O]} = \frac{3.029}{11.459 + 3.029} = \frac{3.029}{14.488} = 0.2091 \quad (38)$$

$$MW_{NPOL}\left[\frac{g}{mol}\right] = (1 - PoleFrac) \times MW = (1 - 0.2091) \times 223 = 176.4 \frac{g}{mol} \quad (39)$$

$$MolVol \ \ \left[\frac{L}{mol}\right] = \frac{MW \ \left[\frac{g}{mol}\right]}{\left(Density\left[\frac{g}{cm^3}\right]\right) \times \frac{1{,}000 \ \ cm^3}{L}} = \frac{223}{1.502 \times 1000} = 0.14847 \frac{L}{mol} \quad (40)$$

$$NPolVol \ \ \left[\frac{L}{mol}\right] = (1.0 - PolFrac) \times MolVol = (0.7909) \times 0.14847 = 0.11742 \frac{L}{mol} \quad (41)$$

$$Sol_{mg/L} = (-15 \times MWNPOL) + (4{,}697 \times NPolVol) + 4{,}870 \quad (42)$$

$$Sol_{mg/L} = -2{,}646 + 551 + 4{,}870 = 2{,}775 \text{ mg/L} \quad (43)$$

$$Percent \ \ Relative \ \ Error \ \ (PRE) = \frac{|Measured - Predicted|}{Measured} \times 100\% \quad (44)$$

$$PRE = \frac{|Measured - Predicted|}{Measured} \times 100\% = \frac{|3{,}000 - 2{,}775|}{3{,}000} \times 100\% = 7.5\% \quad (45)$$

Table 6 shows PRE for each of the three models, shown in Table 5. Figure 8 shows MATLAB code for calculation of percent relative error, PRE, for each drug compound.

```
y=resp;
Y1=-15.0*MWNPOL+4697*NPolVol+4870;
Y2=-9.533*MWNPOL+1175.0*PolFrac+4283.0;
Y3=-13320.0*NPolVol-1539.0*PolFrac+5025.0;
PRE1=(abs(Y1-y)./y)*100;
PRE2=(abs(Y2-y)./y)*100;
PRE3=(abs(Y3-y)./y)*100;
```

Figure 8: MATLAB Code for calculation of percent relative error, PRE, based on the assumption that the measured solubility is the “true” value compared with that predicted by any of the three molecular models.

Table 6: The calculated percent relative error (PRE) using the three molecular models given in Table 5. The measured solubility is considered as the "true" value versus those given by such models.

NAME	PRE1	PRE2	PRE3	NAME	PRE1	PRE2	PRE3
1,6-Cleve's acid	8	5	5	Fenoprofen	50	43	47
1_naphthol	8	3	4	Fenpiclonil	293	308	346
2,4,5-trichlorophenol	8	13	11	Fludrocortisone	46	37	32
2,4-DB	61	67	72	Flufenacet	17	32	16
2,6-Dibromoquinone-4-chlorimide	50	78	88	Flumequine	44	53	71
2-Amino-5-bromobenzoic acid	30	38	52	Flumioxazin	527	652	782
2-Cyclohexyl-4,6-dinitrophenol	121	133	136	Flurbiprofen	105	97	100
2-Ethyl-1-hexanol	26	7	1	Fluspirilene	88	55	44
2-Naphthol	353	317	376	Fumaric acid	5	7	1
3,4-Dinitrobenzoic acid	7	0	11	Furazolidone	102	118	109
4-Amino-2-sulfobenzoic acid	10	4	3	Ganciclovir	19	15	21
4-iodophenol	34	31	5	Glipizide	376	566	545
5-Aminosalicylic acid	321	322	336	Gluconolactone	41	38	38
5-Bromo-2,4-dihydroxybenzoic acid	8	19	27	Glutamic acid	6	6	7
Acetaminophen	16	19	16	Glycine	20	21	25
Acetamiprid	20	20	19	Glyphosate	11	7	11
Acetanilide	5	13	10	Guaifenesin	34	37	37
Acetazolamide	11	20	14	Guanine	403	430	414
Acetochlor	0	6	15	Haloperidol	10	17	5
Acetylacetone	23	29	31	Heptabarbital	5	5	10
Acibenzolar-S-methyl	222	226	263	Hexazinone	44	43	42
Acrylamide	26	30	31	Hexobarbital	2	1	3
Acylonitrile	8	17	17	Histidine	23	23	24
Adenine	28	31	24	Hydrochlorothiazide	260	310	314
Adenosine	49	42	34	Hydrocortisone	42	45	64

NAME	PRE1	PRE2	PRE3	NAME	PRE1	PRE2	PRE3
Adipic acid	18	20	19	Hydro-flumethiazide	3	21	15
Aldicarb	16	20	20	Hydroquinone	21	25	21
Allobarbital	3	7	16	Hydroxy-phenamate	33	36	36
Allopurinol	579	603	573	Hydroxy-proline	32	33	33
Alochlor	2	7	13	Hymexazol	17	19	23
Alpha-acetyl-butyrolactone	29	32	33	Hyoscyamine	43	44	44
Alprenolol	8	19	32	Ibuprofen	68	50	43
Amantadine	0	11	4	Idoxuridine	48	30	9
Amitriptyline	122	96	92	Imazapyr	37	36	36
Amobarbital	3	1	7	Imazaquin	2	7	19
Ancymidol	16	16	9	Imazethapyr	30	28	26
Aniline	13	22	17	Indoprofen	1557	1578	1723
Antipyrine	46	49	47	Iridomyrmecin	0	11	14
ANTU(α-Naphthylthiourea)	0	4	10	Isoflurophate	20	24	31
Arabinose	36	36	35	Isoleucine	17	24	27
Ascorbic acid	36	34	35	Isoniazid	27	29	32
Aspartic acid	1	1	3	Isophorone	12	24	25
Aspirin	11	11	8	Ketanserin	16	39	42
Asulam	20	17	16	Khellin	17	15	14
Atropine	41	43	42	Lindane	209	245	239
Azathioprine	25	42	45	Linuron	47	55	60
Azintamide	32	31	31	Lomefloxacin	44	38	40
Baclofen	36	37	35	Malathion	13	7	7
Badische acid	0	3	13	Maprotiline	164	133	129
Barban	148	158	168	Methocarbamol	62	61	61
Barbital	13	16	23	Methomyl (Lannate)	26	29	31
Bendiocarb	20	19	17	Methylparaben	38	35	44
Benzidine	19	12	24	Metoclopramide	11	11	5
Benzocaine	16	9	10	Metronidazole	11	10	15
Benzoic acid	11	5	10	Miconazole	92	157	190
Benzylimidazole	12	4	13	Minoxidil	35	38	48
Bromogramine	67	73	113	Nadolol	77	77	78
Bronidox	41	36	39	Nalidixic acid	61	59	54
Bupivacaine	5	18	39	Naloxone	39	32	18
Butamben	1595	1459	1410	Naproxen	206	191	201
Butylparaben	1404	1342	1451	Niflumic acid	211	234	220
Capric acid	86	61	43	Nitrofurantoin	60	73	56
Caproic acid	4	13	16	Norfloxacin	28	23	24

NAME	PRE1	PRE2	PRE3	NAME	PRE1	PRE2	PRE3
Carbamazepine	1576	1536	1704	Nortriptyline	52	35	35
Carbofuran	13	8	7	Ofloxacin	64	57	49
Carfentrazone-ethyl	25	57	51	Oxytetracycline	75	52	30
Carisoprodol	4	1	13	p-Aminobenzoic acid	33	35	32
Carmustine	8	3	9	p-Aminosalicylic acid	110	110	114
Carnosine	39	38	38	Papaverine	2	0	6
Carprofen	196	207	250	p-Fluorobenzoic acid	20	18	16
Carvedilol	36	19	10	Phenacetin	330	291	261
Cephalothin	54	37	19	Phenantroline	16	21	12
Chloramphenicol	28	17	17	Phenazopyridine	161	160	162
Chlorpheniramine	21	27	33	Phenobarbital	7	9	13
Chlorpromazine	283	276	286	Phenolphthalein	33	30	23
Chlorthalidone	1432	1725	2013	Phenylbutazone	129	139	99
Chlorzoxazone	250	255	250	Phenytoin	74	74	85
Cimetidine	30	29	28	Phthalazine	22	28	24
Ciprofloxacin	8	6	24	Phthalic acid	8	7	4
Corticosterone	603	609	604	Phthalimide	41	36	34
Cortisone	396	452	525	p-Hydroxybenzoic Acid	2	4	0
Crotonic Acid	16	21	23	Picloram	27	40	28
Cumic Acid	52	39	42	Picric Acid	10	0	16
Cyanazine	30	33	26	Pindolol	55	49	52
Cyanuric Acid	26	34	15	Piroxicam	162	202	238
Cyclizine	27	35	37	Praziquantel	345	338	341
Cyclobarbital	14	16	19	Prednisolone	467	531	615
Cycloleucine	21	25	24	Primidone	470	448	450
Cyproconazole	3	1	6	Procaine	27	32	36
Cyprodinil	138	125	136	Propranolol	23	13	6
Cystine	36	45	54	Propylparaben	548	503	489
Cytosine	47	47	49	Quinidine	1080	1071	1072
Danofloxacin	45	33	18	Quinine	39	40	40
Dapsone	1549	1582	1769	Ranitidine	31	29	36
Dehydroacetic Acid	21	19	18	Salicylamide	77	71	74
Deoxycorticosterone	919	901	944	Salicylic acid	63	59	63
Deprenyl	9	5	7	Sparfloxacin	42	28	23
Desipramine	27	9	11	Strychnine	690	768	976
Dexamethasone	47	36	30	Sulfacetamide	64	63	62

NAME	PRE1	PRE2	PRE3	NAME	PRE1	PRE2	PRE3
Diazepam	24	26	36	Sulfamerazine	1155	1214	1259
Diazoxide	44	52	61	Sulfamethazine	15	9	0
Dicamba	5	10	9	Sulfamethoxazole	334	360	385
Dichlobenil	158	155	151	Sulfanilamide	55	55	55
Difenoconazole	0	29	42	Sulfathiazole	7	0	12
Difloxacin	38	22	14	Sulindac	22	43	80
Digallic Acid	18	2	8	Sulpiride	21	18	27
Diltiazem	67	58	53	Testosterone	39	24	17
Dimethenamid	27	29	30	Tetracaine	2	9	28
Dimethirimol	4	11	14	Tetracycline	76	54	27
Diphenydramine	2	16	29	Theobromine	939	973	964
Diphenylhydantoin (Phenytoin)	59	59	69	Theophylline	53	52	53
DL-Camphor	112	85	88	Thiamphenicol	50	40	35
Enrofloxacin (Baytril)	38	28	18	Thionazin	13	12	12
EPTC	21	6	4	Thymine	3	1	6
Equilin	1295	1215	1328	Thymol	15	0	0
Ethinamate	1	8	10	Tolmetin	85	80	79
Ethirimol	28	19	14	Trichloromethiazide	4	29	29
Ethofumesate	28	32	37	Trimethoprim	9	5	6
Ethohexadiol	23	32	37	Trimipramine	192	146	80
Ethoprop	14	21	23	Tryptamine	72	59	72
Ethylparaben	280	260	261	Uracil	14	14	3
Famotidine (Pepcid)	32	17	2	Verapamil	69	68	103
Fenbufen	603	565	557	Warfarin	158	168	194

From Table 6, it can be seen that the model overestimates the solubility of the following eleven drug compounds: butamben, butylparaben, carbamazepine, chlorthalidone, dapsone, deoxycorticosterone, equilin, indoprofen, quinidine, sulfamerazine, and theobromine. Scrutinizing the experimental solubility data, one can see that they all fall below 200 mg/L, except for theobromine, which amounts to 330 mg/l; however, the solubility of theobromine is also reported as 610 mg/L [8]. Another source [10] reported the value as: "One gram dissolves in about 200 mL water, 150 mL boiling water". The latter value amounts to 5,000 mg/L. The three models predict a solubility value of 3,429, 3540, and 3512 mg/L, respectively. What I argue here regarding theobromine aqueous solubility will extend to solubility of any other drug molecule, as well. The variation in experimental solubility is quite significant and that it will be very difficult to rely on one reported value of aqueous solubility of a given drug molecule. This opens the door for a future work to consider a more giant set of drug aqueous solubility data and make further classification, based on the reported value as practically insoluble, barely or slightly soluble, relatively soluble, soluble, and highly soluble subsets of drug molecules. The last important point to pinpoint here is simply what drives solvation process of a drug in water. Based on the arrived conclusion that at the top of the examined five predictors, it was found that $MW_{NPOL}\left[\frac{g}{mol}\right] = (1 - PoleFrac) \times MW$ ranks number one among the rest of the list. Let us expatiate a little bit on this predictor. Notice that the value of MW_{NPOL} will grow up by two independent variables: The non-polar fraction given by $(1 - PoleFrac)$ and the size of the molecule itself given by the molecular mass, MW. The multiplication of such two molecular properties should tell us about the influence of the hydrophobic non-polar core of the molecule on the overall solvation process. If we scrutinize this first predictor throughout the examined three models, we will find that the slope is negative for MW_{NPOL} (a term in both equations 29 and 30). Although it will be too early to explain in a more detail the contribution of each molecular predictor, but one can say at this stage that the since the slope is negative it simply implies that the anti-solvation (i.e., phase separation) process is entropically driven, mainly by water molecules surrounding and surmounting the organic solute. The solvation process will accommodate the non-polar organic moiety into a polar medium, like water. This being the case, water molecules surrounding an organic molecule are characterized by a higher degree of order at this polar/non-polar interface, where they assume a locally ordered, quasi-solid structure (a "cage-like" structure, clathrate, or iceberg structure) with some loss of H-bonding capacity. As phase separation between a substantially hydrophobic (high MW_{NPOL}) drug and water is thermodynamically more stable than the monodisperse case (i.e., solution), it turns out that $\Delta S_{solvent}$ is the predominant driving force that underlies the process of phase separation in this case. The effect of $\Delta S_{solvent}$ is usually referred to as a hydrophobic or entropic effect [11].

4. CONCLUSION

The supervised machine learning techniques can be used to decipher the relationship between the response on one side and predictor variables on another side. The unsupervised machine learning techniques, on the other hand, can be used to weigh the importance of predictor variables relative to each other without the influence of the response variable. In general, Using MATLAB supervised and unsupervised machine learning algorithms, the drug aqueous solubility data can be best described by the first three important molecular properties: MW_{NPOL}, $NPolVol$, and $PolFrac$, as the third refining or tuning-up factor (weight parameter in curve-fitting). MW_{NPOL} is thought to represent the entropically driven hydrophobic interactions which favor phase separation (anti-solvation) over making up a solution. The robust, linear regression method was used to quantitatively predict the relationship between aqueous solubility and the above three selected predictors. The robust approach relies on the least absolute residuals (LAR) optimization criterion, which tries to find a curve that minimizes the absolute difference of the residuals, rather than the squared differences. Therefore, extreme values have a lesser influence on the fit. The adjusted R^2 was found to be around 0.97 for any of the three models given by equations 29 through 31 and as shown in detail in Table 5. The percent relative error (PRE) was also calculated for each individual drug molecule using the above three models while assuming that the true value of solubility is the experimentally measured and reported value. It was found that the three models overestimate the aqueous solubility of less soluble materials, i.e., below 200 mg/L.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are base of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The authors confirm that the data supporting the findings of this research are available within the article.

FUNDING

None

CONFLICT OF INTEREST

There is no conflict of interest.

REFERENCES

1. Al-Malah, Kamal I. Optimization of Drug Solubility Using Aspen Plus: Acetaminophen Solubility, a Case Study. Int J Pharmacognosy 2018; 5(11): 724-31. http://ijpjournal.com/bft-article/optimization-of-drug-solubility-using-aspen-plus-acetaminophen-solubility-a-case-study/?view=fulltext.
2. Erić, S., Kalinić, M., Popović, A., Zloh, M., Kuzmanovski, I. Prediction of aqueous solubility of drug-like molecules using a novel algorithm for automatic adjustment of relative importance of descriptors implemented in counter-propagation artificial neural networks. International Journal of Pharmaceutics 2012; 437(1–2): 232-241. ISSN 0378-5173. (http://www.sciencedirect.com/science/article/pii/S0378517312008174).
3. Sun, H., Pranav, Sh., Nguyen, K., Yu, KR, Kerns, E., Kabir, Md, Wang Y., Xu X. Predictive models of aqueous solubility of organic compounds built on A large dataset of high integrity. Bioorganic & Medicinal Chemistry 2019; 27(14): 3110-3114. ISSN 0968-0896, (http://www.sciencedirect.com/science/article/pii/S0968089619303475).
4. Al-Malah, K. Aqueous solubility of a diatomic molecule as a function of its size & electronegativity difference. J Mol Model 2011; 17:325–331.
5. Al-Malah, K. Aqueous solubility of a simple (single-carbon) organic molecule as a function of its size & dipole moment. J Mol Model 2011; 17:1029–1034.
6. Al-Malah, K. Prediction of aqueous solubility of organic solvents as a function of selected molecular properties. Journal of Pharmaceutics & Drug Delivery Research 2012; 1(2): 1-7 https://www.scitechnol.com/prediction-of-aqueous-solubility-of-organic-solvents-as-a-function-of-selected-molecular-properties-l0x8.pdf.
7. Cao, D-S, Xu, Q, Liang, Y-Z, Chen, X, Li, H-D. Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. J. Chemometrics, 2010; 24(9): 584-595. (www.interscience.wiley.com)
8. Yalkowsky, S, He, Y., Jain, P. Handbook of Aqueous Solubility Data. CRC Press, Taylor & Francis Group, Boca Raton, Florida, USA. ISBN-13: 978-1-4398-0246-5 (E-book-PDF), 2010.
9. Rahm, M, Zeng, T, Hoffmann, R. Electronegativity Seen as the Ground-State Average Valence Electron Binding Energy”. Journal of the American Chemical Society 2019; 141: 342−351.
10. O'Neil, M.J. (ed.). The Merck Index - An Encyclopedia of Chemicals, Drugs, and Biologicals. Cambridge, UK: Royal Society of Chemistry. 2013; p. 1719.
11. Al-Malah, K. A Macroscopic Model for Apparent Protein Adsorption Equilibrium at Hydrophobic Solid/Liquid Interfaces, Ph.D. Dissertation, Oregon State University, Corvallis.1993; p. 9-10.