Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc



New QSPR study for the prediction of aqueous solubility of drug-like compounds

Pablo R. Duchowicz^{a,*}, Alan Talevi^{a,b}, Luis E. Bruno-Blanch^b, Eduardo A. Castro^a

^a Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64, C.C. 16, Suc.4, La Plata 1900, Argentina ^b Medicinal Chemistry, Biological Sciences Department, Faculty of Exact Sciences, La Plata National University (UNLP), 47 y 115, La Plata 1900, Argentina

ARTICLE INFO

Article history: Received 8 July 2008 Revised 22 July 2008 Accepted 23 July 2008 Available online 29 July 2008

Keywords: QSPR theory Aqueous solubility DRAGON molecular descriptors Replacement method variable subset selection Group contribution method

ABSTRACT

Solubility has become one of the key physicochemical screens at early stages of the drug development process. Solubility prediction through Quantitative Structure–Property Relationships (QSPR) modeling is a growing area of modern pharmaceutical research, being compatible with both High Throughput Screening technologies and limited compound availability characteristic of early stages of drug development. We resort to the QSPR theory for analyzing the aqueous solubility exhibited by 145 diverse drug-like organic compounds (0.781 being the average Tanimoto distances between all possible pairs of compounds in the training set). An accurate and generally applicable model is derived, consisting on a linear regression equation that involves three DRAGON molecular descriptors selected from more than a thousand available. Alternatively, we apply the linear QSPR to other 21 commonly employed validation compounds, leading to solubility estimations that compare fairly well with the performance achieved by previously reported Group Contribution Methods.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

1.1. Importance of solubility in the early stages of a drug development program

In the past, traditional drug development scheme focused on biological activity and potency of drugs. As a consequence, many drug development programs usually failed, because of toxicological and pharmaco-kinetical issues, at late stages of the development process, when large investments had already been made. Thus, modern drug development paradigm (usually referred as 'fail early, fail cheap') includes determination and/or estimation of physicochemical properties related to bioavailability at the very first stages of drug development, when a lead compound is being sought.¹ For many reasons, solubility stands out among such properties (along with pK_a , lipophilicity and stability) as one of the key physicochemical screens in early compound screening, which explains why solubility determination and estimation have been subjects of several recent publications and reviews in Medicinal Chemistry and Pharmaceutical specialized journals.²⁻⁷ Among these reasons we may list:

- 1. To elicit their pharmacological activity, orally administered drugs should exhibit certain solubility in physiological intestinal fluids to be present in the dissolved state at the site of absorption. Aqueous solubility is a major indicator for the solubility in the intestinal fluids and its contribution to bioavailability issues. Note that 56 out of 100 product launches between 1995 and 2002 belong either to class II or class IV of the Biopharmaceutical Classification System, which means their oral bioavailability may be improved by enhancing their solubility.²
- 2. Determination of the true concentration of the free drug is critical in the in vitro assays; wrong conclusions regarding efficacy or toxicity may be drawn if unexpected low solubility or precipitation of the drug occurs.^{2,6–8} Achievement of solution state is usually also needed for adequate in vivo testing.
- 3. Low solubility of compounds contributes to extent timelines, since material engineering of the drug or formulation efforts should be used to produce dosage forms that consistently deliver the desired dose of the drug in the site of absorption.^{2–7}
- 4. Compounds with high solubility are more easily metabolized and eliminated from the organism, thus leading to lower probability of adverse effects and bioaccumulation.⁹

1.2. Solubility measurement and prediction

Solubility measurements determine either the thermodynamic or the kinetic solubility of the compounds. Thermodynamic solu-

^{*} Corresponding author. Tel.: +54 221 425 7430/7291; fax: +54 221 425 4642. *E-mail addresses:* pduchowicz@gmail.com, pabloducho@gmail.com (P.R. Duchowicz).

^{0968-0896/\$ -} see front matter \circledcirc 2008 Elsevier Ltd. All rights reserved. doi:10.1016/j.bmc.2008.07.067



Figure 1. Molecular structures for the training set compounds (*N* = 97).

bility measurements are performed by dispensing a purified crystalline solid compound in a liquid, allowing an incubation time (typically, 24–48 h) to ensure equilibrium.^{6,7} The necessary time to measure thermodynamic solubility is not compatible with modern High Throughput Screening (HTS) technologies: standard equilibrium solubility measures are restricted to about 25–50 compounds a week if handled by one specialist. Moreover, they demand 3–10 mg of purified compound, at an early stage when usually only a few milligrams of product are available which should also be used to measure other important absorption, distribution, metabolization, and elimination (ADME) parameters and biological activity.⁶ True HTS solubility assays are only available in a few specialized companies; and they involve complex task such as automatically handling powders with different characteristics, with the consequent cross-contamination potential, power loss during movement of dosing heads and difficulties in equipment cleaning.





Kinetic solubility measurement starts from a pre-dissolved sample of the compound, usually in dimethyl-sulfoxide (DMSO). Small volumes of the stock solution are added incrementally to the aqueous solution of interest until the solubility limit is reached, the resulting precipitation being detected optically. Although faster than thermodynamic solubility measurement, the DMSO might well operate as a co-solvent, dramatically enhancing the solubility of lipophilic compounds.⁶ Because of these reasons and also because the sample is in amorphous state, kinetic solubility tends to overestimate thermodynamic solubility. With this background, the use of QSPR methodologies to predict aqueous solubility appears as an interesting, increasingly popular alternative to solubility measurement: they are compatible with both HTS technologies and limited compound availability typical of early stages of development, since none of the samples of compound is needed for the estimation of solubility and relatively few computational time is needed for the predictions. Balakin et al. have proposed the following classification of in silico approaches for the assessment of aqueous solubility³:



Figure 1. (continued)

 Table 1

 Different linear methods applied on the same 21-test set compounds

Lead author	Method	Type of descriptors	Number of parameters	rms	N/d	Ref.
Klopman	GCM	2D substructures	34	1.213	0.62	18
Yan	MLR	3D descriptors	40	1.286	0.53	50
Hou	GCM	atomic	78	0.664	0.27	51
Huuskonen	MLR	topologicals	30	0.810	0.70	52
Duchowicz	MLR	Dragon	3	1.202	7.00	This study

Table 2

Experimental and predicted values for \log_{10} Sol (mg ml⁻¹)

Table 2 (continued)

Pred. Eq. (3) -3.08 -3.469 -0.87 -0.9510.019 -1.966 1.118 -1.725-0.065-2.385 -0.064-0.26 -1.0440.772 -0.377 -3.456-2.248-2.072 -3.228 -1.35 -1.204-1.213

-1.585 -2.17 -4.587 -2.002-2.5322.145 -0.611 0.07 0.997 1.776 1.717 2.883 1.729 0.936 -2.916 -0.651-0.2681.705 -1.325 -0.956 1.128 0.05 1.917 2.116 0.688 -0.193-1.316 -0.625 -0.272 0.614 1.384 1.64 0.542 -2.602-0.417-0.913 -1.373 0.648 0.586 0.349 0.776 0.228 0.303 0.645 -0.035 -0.426 -0.4240.222

-3.932 0.104 1.316

No. Chernical name Exp. Prof. B; (1) B Deck (prof. = -3.06) 1 2,45.5 Tichfornylenol 0.079 0.941 78 Digalic Add 0.031 2 2,45.0 Biomanguinane-4 chlorinulu -1.231 -1.134 60 Dimentificantial 0.079 3 2,47.0 Informaguinane-4 chlorinulu -1.237 -1.134 60 Dimentificantial 0.079 3 2,47.0 Informaguinane-4 chlorinulu -1.237 -1.134 60 Dimentificantial 0.079 4 Amino - Suttificeronic acid 0.425 -0.301 81 Dimentificantial -2.387 7 4.Amino - Suttificeronic acid 0.335 85 Eptimina -3.280 11 Accelatoric -0.052 0.312 91 Etholematic -3.301 12 Actelatoria -0.052 0.312 91 Etholematic -3.301 13 Actelatoria -0.324 10.31 1.706 96 Featoria -3.301 14 Actelatori <th>лреппи</th> <th>citital and predicted values for log10501 (ing</th> <th>5 m)</th> <th></th> <th>No.</th> <th>Chemical name</th> <th>Exp.</th>	лреппи	citital and predicted values for log10501 (ing	5 m)		No.	Chemical name	Exp.
Troining set	No.	Chemical name	Exp.	Pred. Eq. (3)	76	Diclofop-methyl	-3.096
1 2.45-Trichkorophenol 0.079 0.043 78 Digalic Acid 0.731 2 2.45-Dihomophenol 1.237 1.138 30 Dimetheaming 0.079 3 2.5-Dihomophenol 1.037 1.138 30 Dimetheaming 0.079 4 2.5-Dihomophenol 0.637 82 Dimetheaming 0.238 6 3.4-Dihirbeenoic acid 0.876 0.533 83 Dimetheaming 0.386 7 4.7310 -0.397 91 Ethnamaz 0.386 7 Actinibic -0.036 0.373 84 Ethnamaz 0.386 7 Actionamide 0.036 0.373 84 Ethnamaz 0.386 7 Actionamide 0.0375 89 Ethnamaz 0.386 0.387 81 Ethnamaz 0.386 7 Actionamide 0.237 91 Ethnamaz 0.386 0.387 81 Ethnamaz 0.386 0.387 81 Ethnamaz 0.387	Training :	set			77	Difenoconazole	-1.823
2 2.4-08 -1.337 1-129 79 Dimechamid 0.079 3 2.6-050 morphics -1.323 -1.127 80 Dimechamorphics -1.728 6 3.4-050 morphics -1.323 -1.127 80 Dimechamorphics -1.728 7 4 -Amino-2-unblemonic acid 0.477 0.134 84 EFTC 9436 8 Accentricle 3.332 2.2939 86 Exhanata 0.388 9 Accentricle 0.032 -0.06 87 Exhirmat -0.038 10 Accentricle 0.020 -0.175 89 Exhirmat -0.038 11 Accentricle -0.000 1.755 89 Exhirmat -0.000 12 Accentricle 2.081 1.021 90 Exhirmat -0.000 13 Activitacio -0.010 1.755 89 Exhirmat -0.000 14 Activitacio 2.031 -0.212 90 Exhirmat -0.000 15 Achonica 2.030 1.024 91 Exhirmat -0.000 16 Accontric acid 1.037 0.013 1.026 96 Futocorphica -2.081 <th>1</th> <th>2,4,5-Trichlorophenol</th> <th>0.079</th> <th>-0.943</th> <th>78</th> <th>Digallic Acid</th> <th>-0.301</th>	1	2,4,5-Trichlorophenol	0.079	-0.943	78	Digallic Acid	-0.301
3 2.4-Cibbromoçumen-4-thorimide -1.130 -1.134 80 Dimethinimal 0.077 5 2.4-Dibromoçumen-4-chorimide -0.056 0.221 82 Dimethinimal 2.480 6 3.4-Dibritheencis and -0.056 0.221 82 Dimethinimal 2.480 7 Accamico -0.057 83 Eprilin -2.450 7 Accamico 0.0623 86 Epriline-4-8.00 7 Accamico 0.066 0.033 88 Erbahosadial 1.831 7 Accamiride -0.009 1.765 83 Erbahosadial 1.831 7 Accaszelamide -0.009 1.783 80 Erbahosadial 1.832 12 Accaszelamide -0.009 1.783 80 Erbahosadial 1.832 13 Accaszelamide -0.009 1.001 1.783 80 Erbahosadial 1.833 14 Accaszelamide -0.009 1.002 82 Erbahosadial 1.2334 <th>2</th> <td>2,4-DB</td> <td>-1.337</td> <td>-1.29</td> <th>79</th> <td>Dimethenamid</td> <td>0.079</td>	2	2,4-DB	-1.337	-1.29	79	Dimethenamid	0.079
4 2-2-cyclohecyd-I-dexand -1.273 81 Dimerhomorph -1.275 2 2-Bityd-H-exand -0.056 0.321 82 Dimerhomorph 2.389 6 3-4.Dimerhomorph 0.374 83 Dimerhomorph 2.389 7 4.Amino 3.372 2.292 85 Dimerhomorph -0.499 9 Acctaratike -0.061 7.65 89 Etholewatel -1.031 12 Acctaratike -0.062 -0.612 90 Dimorph om 2.612 -0.613 13 Acctaratinide -0.062 -0.012 90 Etholewatel -0.620 14 Acctaratinide 2.013 1.705 96 Etholewatel -2.366 15 Acctaratinide 2.013 1.706 96 Enduratel -2.366 16 Actaratinide 1.872 2.396 94 Fernagener-thyl -3.046 16 Actaratinide 1.706 96 Enduratel -2.247 16	3	2,6-Dibromoquinone-4-chlorimide	-1.230	-1.194	80	Dimethirimol	0.079
5 2-Ethyl-1-bcanol -0.056 0.231 82 Dimorphalmine 2.2487 7 4-Amino-2-uniforenzoic acid 0.477 0.134 84 PTC -0.4387 7 4-Amino-2-uniforenzoic acid 0.477 0.134 84 PTC -0.4387 7 4-Amino-2-uniforenzoic acid 0.623 -0.6 87 Ethinanate -0.308 10 Acctanolide 0.060 0.633 88 Etholewaldol 1.631 11 Acctazolamide -0.002 1.765 89 Etholewaldol 1.631 12 Acctazolamide -0.002 1.763 90 Etholewaldol 1.632 13 Acctazolamide 2.038 1.021 90 Etholewaldol 1.632 14 Actazolamide 2.038 2.46 94 Fromazoro-ethyl -2.438 15 Actazolamide 0.013 1.056 96 Fromazoro-ethyl -2.438 16 Actazolamide 0.021 0.24 90 <th>4</th> <td>2-Cyclohexyl-4,6-dinitrophenol</td> <td>-1.823</td> <td>-1.275</td> <th>81</th> <td>Dimethomorph</td> <td>-1.728</td>	4	2-Cyclohexyl-4,6-dinitrophenol	-1.823	-1.275	81	Dimethomorph	-1.728
6 3.4. Dintrobersoic acid 0.826 0.830 B3 Dinconazole -2.397 8 Accentinocyl -4.173 -4.308 B4 EPTC -4.436 8 Accentinocyl -4.173 -4.308 B5 Epulin -2.830 9 Actanilide 0.000 1.765 B9 Ethoframe -0.600 11 Actanilide -0.000 1.765 B9 Ethofrexate -0.610 12 Actanilide -0.010 1.765 B9 Ethofrexate -0.600 13 Acterchlor -0.612 -0.1178 B1 Etofreyrox -6.000 14 Acterylactina 1.272 1.378 B1 Etofreyrox -1.202 14 Acterylactina 1.372 2.396 Feabucina -2.369 15 Acheronactina 1.372 2.396 Feabucina -2.394 16 Acheronactina 1.372 2.396 Feabucina -2.394 16 Acher	5	2-Ethyl-1-hexanol	-0.056	0.321	82	Dimorpholamine	2.698
7 4 Amine-2-subberneic and 0.477 0.134 84 PTC 0.436 9 Acctanué 3.33 2.929 86 Eluinanué 0.38 9 Acrianué 3.33 2.929 86 Eluinanué 0.38 11 Actazolamide 0.000 1.765 89 Elubrexdiol 1.83 12 Actazolamide 0.000 1.765 89 Elubreydiol 1.83 13 Actazolar-Sinethyl 2.211 1.071 91 Eluforprox -5.609 14 Acrylycetone 2.268 1.021 92 Fenbactanole -2.649 15 Actexitic acid 1.013 1.706 96 Fenbactanole -2.649 16 Aconitic acid 1.013 1.708 96 Fenbactanole -2.641 16 Addenine 2.046 96 Fenbactanole -2.641 24 Addenine 1.024 -0.015 Fer -2.641 24 Add	6	3,4-Dinitrobenzoic acid	0.826	-0.503	83	Diniconazole	-2.397
8 Accegatinocyl. -4.173 -4.268 85 Epullin -2.880 10 Actsanipird 0.033 -0.6 87 Ethinanate -0.038 10 Actsanipird 0.032 -0.6 87 Ethinanate -0.031 11 Actsachinitie 0.052 -0.812 90 Ethinanate -0.812 13 Actsachinitie -0.612 -0.812 91 Ethinanate -0.600 14 Actsachinitie -2.21 1.078 91 Ethinanate -0.600 15 Actsachinitie -2.21 1.078 91 Ethinanate -2.866 16 Aconitic acid -2.868 1.021 95 Fenderont -2.318 10 Adapta acid 1.14 0.251 95 Fenderont -2.318 11 Adapta acid 0.11 1.705 96 Fludrasscia -2.401 24 Adapta acid 0.258 0.408 99 Flutanania acid -2.40	7	4-Amino-2-sulfobenzoic acid	0.477	0.134	84	EPTC	-0.426
9 Acctanide 3.32 2.929 86 Bihamate 0.388 0 Acctanide 0.060 0.753 80 Bhahamate -1.301 11 Acctanide 0.080 1.755 80 Bhahesoldo 1.633 12 Actanida 0.021 1.755 80 Bhahamate -1.301 13 Acctanida 0.221 0.210 91 Baforapore -0.309 14 Acchiric acid 2.468 1.021 93 Fenbulen -2.316 15 Accylonitrile 1.372 2.396 95 Fenbulen -2.318 16 Actine 0.730 -0.115 For art of -2.277 11 Addicab 0.700 -0.232 101 For arc of -2.777 13 Anite acid -0.200 102 For arc of -2.777 14 Acida -0.201 1.045 101 For arc of -2.777 14 Actionaria -	8	Acequinocyl	-4.173	-4.506	85	Equilin	-2.850
10 Acctamiptid 0.633 60 77 Ethinimal -0.690 12 Acctazalamide -0.000 1.753 88 Ethicinesate -1.301 12 Acctazalamide -0.000 1.753 89 Ethicinesate -1.301 12 Acctazalamide -0.000 1.753 89 Ethicinesate -1.301 13 Acctazalamide -0.000 1.753 89 Ethicinesate -0.000 15 Achenzolar-Smethyl -2.113 -0.212 92 Fenburcen -3.06 16 Acrylamide 2.266 2.46 94 Fenozynop ethyl -3.04 17 Achenic 0.013 1.706 95 Fendyconit -2.318 21 Alacine 0.013 1.706 96 Fluitesate -2.041 24 Adhipa etci 0.239 -0.028 100 Fluitesate -2.041 24 Anlobatiral 0.258 -0.020 0.439 104 Fura	9	Acetamide	3.352	2.929	86	Ethinamate	0.398
11 Acetanilde 0.066 0.063 88 Ehdumesate -1.01 12 Accacahamide -0.052 -0.012 90 Ehdensate -1.021 13 Acceconitor -2.011 -1.012 90 Ehdensate -1.021 14 Activication encluyi -2.013 1.021 90 Endensate -2.005 16 Aconitic acid 1.872 2.396 95 Fenpicioni -2.306 17 Acylanitic 1.872 2.396 96 Fuforonitonic -2.304 18 Acrylonitric 1.872 2.396 96 Fuforonitonic -2.2041 24 Addicarb 0.780 -0.015 Test set vol -2.2041 24 Addicarb 0.780 -0.232 100 Futneazzt -2.041 24 Addicarb 0.780 -0.232 101 Futneazzt -2.041 24 Addicarb 0.230 0.468 96 Fufenanic acid -2.031	10	Acetamiprid	0.623	-0.6	87	Ethirimol	-0.699
12 Accessolation -0.009 1.765 89 Etholescalul 1.622 14 Accessolator -0.212 1.978 91 Etholescalul -0.025 15 Acternalis-smethyl -2.211 1.978 91 Etholescalul -0.006 15 Acternalis-smethyl -2.111 -0.212 92 Feinzonzole -0.008 16 Acrylamidio 2.206 92 Feinzonzole -0.318 17 Acquamidio 1.144 0.051 97 Fufenzet -0.232 21 Adlacab 0.730 -0.115 Text set vol -0.224 23 Allidochlor 2.39 0.015 Fufenzet -2.041 24 Allaba-acetylburyolactone 2.01 1.538 101 Fufenzet -2.042 24 Allaba-acetylburyolactone -0.208 100 Fufenzet -0.043 25 Andrina-acetylburyolactone -0.371 106 Garcelovi -0.383 26 An	11	Acetanilide	0.806	0.363	88	Ethofumesate	-1.301
13 Acerochlor -0.652 -0.812 90 Ethoprop -0.123 15 Acibenzolar-Surethyl -2.113 -0.212 92 Frankuronazole -3.690 16 Acontraciód 2.898 1.021 93 Frankuronazole -3.286 17 Acquantic 2.899 1.021 93 Frankuronazole -3.288 18 Admine 2.914 2.899 96 Futronazole -0.843 20 Adiora 0.714 0.951 97 Futronazole -0.854 21 Alaine 2.214 2.244 1.241 -0.221 98 Futronazole -2.041 22 Adicabre 0.720 -0.628 100 Flusinezaria -2.000 23 Alicharbe 0.700 -0.233 103 Futranito acid -0.241 24 Anilozaria -3.030 -2.334 104 Futranity -0.633 25 Amintaz -3.030 -2.334 1046	12	Acetazolamide	-0.009	1.765	89	Ethohexadiol	1.623
14 Accerylactione 2.21 1.378 91 Explemition 6.000 15 Accionital-S-methyl -2.113 -0.212 92 Febluorinazio -2.636 16 Aconita acid 2.088 1.021 93 Febluorinazio -2.636 17 Acylinide 1.091 94 Febluorinazio -0.384 18 Admine 2.144 2.441 -0.222 Adhine -2.644 21 Adicarbia 0.228 0.022 98 Flufenanic acid -2.441 23 Alidochior -0.620 -0.208 100 Flufenazic acid -2.747 24 Alidozhiari 0.239 -0.323 101 Flufenazic acid 0.300 27 Aninozania -0.620 -0.201 102 Flufenazic acid 0.433 38 Annicopromazine -3.030 -1.333 101 Flufenazic acid 0.433 39 Ansitra -0.220 -0.633 106 Flufenazic acid	13	Acetochlor	-0.652	-0.812	90	Ethoprop	-0.125
15 Achemzalar-S-methyl -2.113 -0.212 92 Fenbulate -3.899 67 Acrylamide 2.866 2.46 94 Fenovapro-relynyl -3.046 77 Acrylamide 1.872 2.396 95 Fenovapro-relynyl -2.318 78 Admine 2.14 2.144 97 Fenovapro-relynyl -2.318 20 Admine 2.214 2.441 - - - - - - 21 Aldiochar 0.780 -0.115 Yess et all -	14	Acetylacetone	2.221	1.978	91	Etofenprox	-6.000
16 Accontric actic 2.806 2.46 93 Fendulen -2.306 18 Acrylonitrile 1.872 2.386 94 Fenculation -2.318 19 Adeine 0.013 1.301 95 Fenculation -2.318 20 Adipic acid 1.214 0.201 97 For set rof -0.824 21 Adiacab 0.780 -0.022 98 Fufemaric acid -2.041 23 Adiabicab 0.780 -0.028 100 Fufemaric acid -2.401 24 Allobachital 0.228 0.468 99 Funitoacin -2.400 25 Adhitochita 0.700 -0.282 102 Folic acid -2.812 26 Ahitoactinbita -0.200 -0.282 102 Folic acid 0.833 30 Anobabrital -0.200 -2.834 104 Furametry -0.648 30 Anobabrital -0.220 0.679 107 Chronolatone	15	Acibenzolar-S-methyl	-2.113	-0.212	92	Fenbuconazole	-3.699
17 Acrylamide 2.806 2.46 94 Henoxapor-Hyly 2.406 18 Acterine 0.013 1.706 95 Feupleoni 2.314 19 Adenine 0.214 2.214 2.441 2.212 21 Adaine 2.214 2.441 2.213 Adaine 2.217 23 Adiacchior 1.253 0.028 98 Futenamic add 2.217 24 Adaine 0.620 0.282 100 Futenamic add 2.377 25 Aninopromazine 2.331 103 Futnacic add 2.375 26 Aninopromazine 2.334 104 Futnacic add 2.373 31 Antypindol 0.187 0.7.19 106 Garcclovir .0.333 32 Antine 1.556 0.979 107 Giaconalactone .1.397 33 ANTU 0.222 0.663 108 Glutamic acid 0.933 34 Arabin	16	Aconitic acid	2.698	1.021	93	Fenbuten	-2.656
18 Actylonitrile 1.872 2.99 95 Perpletomin -2.418 20 Adipic and 1.414 0.951 96 Fludenacet -0.854 21 Alainie 2.214 2.414 - First viol - 21 Alainie 0.241 - First viol - - 23 Allidachlor 0.258 0.468 95 Fluminozarin - - 24 Allochlor -0.620 -0.208 100 Fluikacet-methyl -2.000 26 Apha-acrylhotymolectone -3.230 -1.345 101 Fluikacet -2.895 28 Aninopromazine -3.239 -0.231 104 Fluikacet -0.633 30 Anobarbital -0.220 -0.637 102 Cluconolactone -2.795 31 Anzymidol -0.187 -0.719 106 Garciclori 0.333 Antine -0.528 -0.631 108 Cluconolactone 2.797	17	Acrylamide	2.806	2.46	94	Fenoxaprop-ethyl	-3.046
19 Ademne 0.013 1.706 90 Fuldiscottisolies -0.834 20 Adjine 2.214 2.441 -0.227 21 Adiance 2.214 2.441 -0.227 22 Adiactho 0.780 -0.115 Terr set val -2.2411 23 Adiacthor 1.254 -0.022 98 Fufenzanic acid -2.2411 24 Adiacthor 0.253 -0.023 99 Futanciz acid -2.070 24 Alphane onthytrobactone -0.000 -2.334 101 Futanciz acid -0.070 25 Anninapromazine -2.339 -1.383 103 Futanciz acid -0.397 36 Anosharhital -0.220 0.497 105 Futancitone -1.397 31 Anzy -0.187 -0.166 Guacclovin 0.633 32 Antinapromazine -5.252 2.005 110 Clybnoste 1.093 33 ANTU -0.222 0.063	18	Acrylonitrile	1.872	2.396	95	Fenpicionii	-2.318
20 Adapte acid 1.414 0.951 9' Futeract -1.252 21 Alainine 2.214 2.4411 Ters set val	19	Adenine	0.013	1.706	96	Fludrocortisone	-0.854
21 Alkine 2.214 2.44 22 Alkidochlor 1.294 -0.015 Test set val 23 Alkidochlor 1.294 -0.022 98 Flurnanic acid -2.247 24 Alkobrital 0.288 0.468 99 Flurnioszain -2.247 25 Alkobrital 0.288 0.468 99 Flurnioszain -2.307 26 Alpha-actylibutyrolactone 2.301 1.458 101 Flurniarc acid 0.845 27 Annitraz -3.00 -2.334 104 Furazolidone -1.397 28 Annitraz -3.00 -2.334 105 Furazolidone 2.770 31 Anzymidol -0.187 -0.719 106 Guarcio/ort 0.333 33 Antriaz -3.00 -2.334 109 Glytine 2.396 34 Arabinose 2.688 2.168 109 Glytine 2.396 35 Ascontric acid 0.912 2.095	20	Adipic acid	1.414	0.951	97	Flufenacet	-1.252
22 Aldicarb 0.780 -0.012 Particle -2.241 23 Aldicharbir 1.294 -0.022 98 Flurenamic acid -2.241 24 Aldobarbital 0.288 0.468 99 Flurenamic acid -2.200 25 Alchor -0.620 -0.208 100 Flurpinterne-embryl -3.000 27 Anincarbilde 0.700 -0.233 103 Furamethyr -0.648 28 Aminopromazine -3.230 -1.331 106 Furamethyr -0.648 30 Annobarbital -0.227 0.497 106 Furamethyr -0.648 31 Antroprintal -0.185 -0.633 108 Clutaric acid 0.239 31 Antroprintal -0.122 -0.063 108 Clutaric acid 0.239 33 Antroprintal 0.4089 -0.106 112 Haloperiol -1.698 34 Antroprintal 0.699 -0.762 114 Hexatone	21	Alanine	2.214	2.441	I		
23 Allidochlor 1.294 -0.022 98 Flutchanic add -2.247 24 Allobarital 0.286 0.468 99 Fluminoszán -2.2747 25 Allochlor -0.620 -0.208 100 Fluspireline -2.3070 26 Alpha-actylbutyrolactone 2.301 1.458 101 Fluthacet-methyl -3.070 27 Annicarbalide 0.700 -0.282 102 Folt acid -2.374 28 Aninpromazine -3.239 -1.333 103 Furnarticy acid 0.845 29 Aniniraz -0.020 0.497 105 Furzacildone -1.397 31 Anzymidol -0.187 -0.719 106 Gluconolactone 2.770 33 Ahrine 1.556 0.879 107 Gluconolactone 2.770 34 Arabinose 2.688 2.168 109 Glytins acid -0.282 35 Ascorbic acid 0.9912 2.095 111	22	Aldicarb	0.780	-0.115	Test set val		
24 Allobarthital 0.28 0.468 99 Funding and the second s	23	Allidochlor	1.294	-0.022	98	Flufenamic acid	-2.041
25 Alachiof -0.020 -0.028 100 PHUSPINERE -2.030 26 Alpha-ectylbutyrolactone 2.301 LASS 101 Fluthiacet-methyl -3.070 27 Amicarbalide 0.700 -0.282 102 Flutaicatical -2.373 30 Amobarbital -0.220 0.497 105 Furazolidore -1.397 31 Ancymidol -0.187 -0.719 106 Ganciclovir 0.633 32 Aniline 1.556 0.979 107 Cilciconolactone 2.770 33 Anzinose 2.688 2.168 109 Cilyrine 2.395 34 Arabinose 2.699 -0.106 112 Halpostate 1.698 35 Ascorbic acid 0.922 -0.611 126 Haipostate 1.692 36 Aspartic acid 0.925 -0.817 116 Hydrocortisone -0.492 37 Asianfenicol 1.025 -0.817 124 Haperba	24	Allobarbital	0.258	0.468	99	Flumioxazin	-2./4/
46 Alpha-acetypolutyrolactone 2.301 1.433 101 Public ethertypi -3.000 27 Aminopromazine -3.239 -1.533 103 Funancia acid 0.2434 28 Aminopromazine -3.239 -1.533 103 Funancia acid 0.0488 30 Amobarbial -0.220 0.497 105 Furametpyr -0.648 31 Ancymidol -0.187 -0.719 106 Canciclovir 0.633 32 Amine 1.556 0.979 107 Cluconalcone 2.770 33 ArUU -0.222 0.065 110 Calymbacet 2.396 34 Arabinose 2.698 2.168 109 Clytone 2.393 35 Ascothicacid 0.912 2.005 111 Cualambrial -0.602 36 Asparticacid 0.912 2.0205 113 Heptazinone 1.519 37 Aulan 0.699 -0.762 114 Heazanote	25	Alochlor	-0.620	-0.208	100	Fluspirilene	-2.000
22 AnnicarDailoé 0.700 -0.282 102 FOIC action -2.790 28 Amitoraz -3.000 -2.534 104 Furnaric acid 0.845 29 Amitraz -3.000 -2.534 106 Furnazolidore -0.1397 31 Ancymidol -0.187 -0.719 106 Carciclovir 0.633 32 Ahline 1.556 0.979 107 Ciluconolactone 2.700 33 Anzimine 1.552 0.979 107 Ciluconolactone 2.700 34 Arabinose 2.698 2.168 109 Cilycine 2.396 35 Ascorbic acid 0.522 2.005 110 Cilyboste 1.079 36 Aspartic acid 0.999 -0.106 112 Halperidol -1.853 37 Asalamenicol 1.301 0.258 113 Heptababilal -0.602 38 Aziaménicol -0.200 -3.393 115 Histicline -1.658 41 Badiche acid -0.207 -0.817 116	26	Alpha-acetylbutyrolactone	2.301	1.458	101	Fluthlacet-methyl	-3.070
Aminopromazine -3.239 -1.933 103 Purametry and the action of the	27	Amicarbalide	0.700	-0.282	102	Folic acid	-2.795
29 Affilt Affilt -0.20 0.497 104 Putazolikone -0.204e 30 Amobarbital -0.187 -0.719 106 Ganciclovir 0.633 31 Ancymidol -0.220 -0.633 108 Giluconolactone 2.737 33 ANTU -0.222 -0.663 108 Gilutamic acid 0.933 34 Arabinose 2.698 2.168 109 Glyonatta acid 0.239 36 Asparita acid 0.9212 2.005 111 Gualfenesin 1.698 37 Asulam 0.699 -0.066 112 Hajperiolol -1.853 38 Azianfenicol 1.301 0.258 113 Heptabrital -0.622 38 Azianfenicol -0.200 -3.393 115 Histoline 1.658 41 Badische acid -0.202 -0.317 116 Hydroguinethaizade -0.523 42 Barbai -0.782 1.26 120 Hydrosyphenamate 1.397 43 Barbital 0.877 118 Hydro	28	Aminopromazine	-3.239	-1.933	103		0.845
30 Antrobation -0.220 0.937 103 Probability -1.337 31 Anzymidol -0.187 -0.719 106 Canciclovir 0.233 32 Aniline 1.556 0.979 107 Glucanic acid 0.233 33 ANTU -0.220 -0.663 108 Glutanic acid 0.233 34 Arabinose 2.698 2.168 109 Glutanic acid 0.239 35 Ascotic acid 0.912 2.095 111 Guidenesin 1.698 37 Asulan 0.699 -0.106 112 Haloperiolo -1.853 38 Azidamfenicol 1.301 0.258 113 Hershabital -0.602 40 Azoxystrobin -0.200 -3.393 115 Histidine 1.658 41 Badische acid -0.253 -0.573 118 Hydrooguinone 1.857 42 Barban -0.873 -0.267 118 Hydroxyphenamate 1.	29	Amitraz	-3.000	-2.534	104	Furanetpyr	-0.648
AltCyliniton -0.187 -0.79 100 Callectivity 0.033 Antiline 1.556 0.979 107 Ciluconiactone 2.770 33 ANTU -0.222 -0.663 108 Cilutamic acid 0.933 34 Arabinose 2.592 2.005 110 Cilutamic acid 0.933 35 Ascorbic acid 0.512 2.095 111 Guaidanfeincol 1.698 36 Aspartic acid 0.699 -0.762 114 Hestabrital -0.602 38 Azianmenico 1.510 0.258 113 Heptabrital -0.602 39 Azintamide 0.699 -0.762 114 Hestabrital -0.602 31 Barbina -0.225 -0.817 116 Hydrocortisone -0.495 32 Barbina -0.237 0.857 118 Hydrocortisone -0.248 33 Barbina -0.458 -0.207 121 Hydroxoyphenamate 1.397 <tr< td=""><th>5U 21</th><td>Amodarditai</td><td>-0.220</td><td>0.497</td><th>105</th><td>Furazolidolle</td><td>-1.397</td></tr<>	5U 21	Amodarditai	-0.220	0.497	105	Furazolidolle	-1.397
22 Animite 1.336 0.37.9 107 Chickan barlow 2.774 33 ANTU -0.222 -0.663 108 Chickan barlow 2.393 34 Arabinose 2.698 2.168 109 Cilytanic acid 2.393 35 Ascorbic acid 0.912 2.095 111 Guaifenesin 1.693 36 Aspartic acid 0.912 2.095 111 Halperidol -1.853 37 Asulam 0.699 -0.762 114 Heytaszinone 1.519 40 Azoxystrobin -2.000 -3.393 115 Histidine 1.658 41 Badische acid -0.225 -0.617 116 Hydrocortisone -0.485 42 Barbial 0.873 0.557 118 Hydrocyphenamate 1.397 43 Benzidine -0.495 -1.026 120 Hydroxyphenamate 1.397 44 Bendiocarb -0.237 121 Hydroxyphenamate 1.397)))	Anityiniuoi	-0.167	-0.719	100	Chicopolactopo	0.055
33 ArNo -0.222 -0.063 106 Glutalitic Actu 0.333 34 Arabinose 2.098 2.168 109 Glyphosate 1.079 35 Ascorbic acid 2.522 2.005 110 Glyphosate 1.079 37 Asulam 0.699 -0.106 112 Haloperidol -1.853 38 Azidamfenicol 1.301 0.258 113 Heptababital -0.602 39 Azintamide 0.699 -0.762 114 Hexazinone 1.619 40 Azoxystrobin -2.000 -3.933 115 Histidine -0.602 41 Badische acid -0.225 -0.817 116 Hydroxyphenamate -0.495 42 Barban -1.958 -2.248 117 Hydroxyphenamate 1.357 43 Barbital 0.873 0.857 118 Hydroxyphenamate 1.397 44 Bendiocarb -0.458 -0.207 120 Hydroxyphenamate 1.397 45 Bifenox -3.397 -3.207 121	32	Annine	1.550	0.979	107	Clutamic acid	2.770
24 AddDitUSe 2.080 2.180 100 GlyCine 2.197 35 Ascorbic acid 2.522 2.005 110 GlyDhesate 1.1079 36 Aspartic acid 0.912 2.095 111 Guaifenesin 1.693 37 Asulam 0.699 -0.762 114 Hexazinone 1.519 38 Azidamfenicol -2.000 -3.393 115 Histidine 1.658 40 Azoxystrobin -0.225 -0.817 116 Hydroortisone -0.495 42 Barban -1.958 -2.248 117 Hydrofumente 1.537 43 Barban -0.552 -0.596 119 Hydroxyphenamet 1.537 44 Bendiocarb -0.585 -0.206 120 Hydroxyphenamet 1.537 45 Benzidine -0.495 -1.026 120 Hydroxyphenamet 1.537 46 Bifentrhin -0.695 -0.217 123 Imazapir 1.053 47 Bifentrhin -0.658 -0.217 123 <t< th=""><th>55 74</th><th>Anito</th><th>-0.222</th><th>-0.003</th><th>100</th><th>Chucino</th><th>0.955</th></t<>	55 74	Anito	-0.222	-0.003	100	Chucino	0.955
36 Aspartic acid 0.912 2.003 110 Galynnoate 1.079 37 Asulam 0.699 -0.106 112 Haloperidol -1.853 38 Azidamfenicol 1.301 0.258 114 Headbarbital -0.602 39 Azintamide 0.699 -0.762 114 Hexazinone -0.602 40 Azoxystrobin -2.000 -3.393 115 Histidine 1.658 41 Badische acid -0.225 -0.817 116 Hydrocurisone -0.495 42 Barbial 0.877 0.857 118 Hydroquinone 1.857 43 Barbial 0.695 -1.026 120 Hydroxyproline 2.257 44 Bendiocarb -0.585 -0.596 119 Hydroxyproline 2.257 45 Bitenox -3.397 -3.207 121 Hydroxyproline 2.257 46 Bifenox -3.297 123 Imazapyr 1.053 47 Bitenthin -0.0511 124 Imazapyr 1.053 <th>54 25</th> <td>Arabinose Accorbic acid</td> <td>2.098</td> <td>2.108</td> <th>109</th> <td>Chapter</td> <td>2.590</td>	54 25	Arabinose Accorbic acid	2.098	2.108	109	Chapter	2.590
37 Astulan 0.512 2.093 111 Chaineman 1.1398 37 Astulan 0.699 -0.106 112 Helpperidol -1.853 38 Azidamfenicol 1.301 0.258 113 Heptabrital -0.602 39 Azintamide 0.699 -0.762 114 Hexazinone 1.519 40 Azoxystrobin -0.225 -0.817 116 Hydrocortisone -0.6425 42 Barban -1.958 -2.248 117 Hydroxyptenamate 1.397 43 Barbital 0.873 0.857 118 Hydroxyptenamate 1.397 44 Benzidine -0.495 -1.026 120 Hydroxyptenamate 1.397 45 Benzidine -0.495 -1.026 120 Hydroxyptenamate 1.392 47 Bifentrhin -4.000 -3.5 122 Idoxuridine 0.301 48 Biotin -0.658 -0.217 123 Imazaquin -1.045 50 Caproic acid 1.012 1.128 Isofuronymec)))C	Assorbic acid	2.522	2.005	111	Guaifonosin	1.079
J Astham 0.039 -0.000 112 Halperhold -1.025 38 Azintamide 0.699 -0.762 114 Hexatione 1.519 40 Azoxystrobin -2.000 -3.393 115 Histidine 1.519 41 Badische acid -0.225 -0.817 116 Hydrogurione -0.495 42 Barban -1.958 -2.248 117 Hydrogurione 1.857 43 Bendiocarb -0.585 -0.596 119 Hydroxyphenamate 1.397 44 Bendiocarb -0.495 -1.026 120 Hydroxyphenamate 1.397 45 Bendiocarb -0.495 -1.026 120 Hydroxyphenamate 1.397 46 Bifenox -3.397 -3.207 121 Hymexzol 1.529 47 Bifenthin -4.000 -3.5 122 Idoxuridine 0.301 48 Botin -0.658 -0.217 123 Imazaquin -1.045 50 Caproic acid 1.012 1.128 Imazaquin	90 97	Asulam	0.912	2.095	111	Haloperidol	1.058
36 Azitaminido 1.501 0.2.50 112 Heptazinona -0.005 39 Azitaminide 0.699 -0.762 114 Hexazinone 1.519 40 Azoxystrobin -2.000 -3.393 115 Histidine 1.558 41 Batische acid -0.225 -0.817 116 Hydrocutisone -0.495 42 Barban -1.958 -0.596 119 Hydroxyptenamate 1.357 43 Barbital 0.873 0.857 118 Hydroxyptenamate 1.357 44 Bendiocarb -0.0495 -1.026 120 Hydroxyptenamate 1.357 45 Benzidine -0.495 -1.026 120 Hydroxyptenamate 1.352 46 Bifentxin -0.495 -0.217 123 Imazapyr 1.053 47 Bifentxin -0.029 -0.511 124 Imazaquin -1.045 50 Capric acid 1.012 1.128 125 Imazethapyr 0.146 51 Carbofuran -0.495 -0.336 12	2/ 20	Azidamfonicol	1 201	-0.100	112	Hentabarbital	-1.855
30 According to the set of the	20	Azintamide	0.600	0.250	113	Heyazinone	1 519
Aboxystrobin -2.000 -0.000 -0.000 -0.495 41 Batische acid -0.225 -0.817 116 Hydroorisone -0.495 42 Barban -1.958 -2.248 117 Hydroflumethiazide -0.525 43 Barbial 0.873 0.857 118 Hydroxyphenamate 1.397 44 Benzidine -0.495 -1.026 120 Hydroxyphenamate 1.397 45 Benzidine -0.495 -1.026 120 Hydroxyphenamate 1.529 46 Bifentrhin -4.000 -3.5 122 Idoxuridine 0.301 48 Biotin -0.658 -0.217 123 Imazaquin -1.045 50 Capric acid 1.012 1.128 126 Iridomyrmecin 0.301 52 Carbosuffan -3.522 -2.296 127 Isofluctine 1.187 53 Cariboxin -0.701 -0.274 128 Isoleucine 1.536 5	10	Azovystrobin	2 000	3 303	115	Histidine	1,515
1Data barban -0.522 -0.611 110Hydrofulbric 0.752 42Barbial 0.873 0.857 118Hydroquinone 1.857 43Barbital 0.733 0.857 118Hydroxyphenamate 1.397 44Bendiocarb -0.495 -1.026 120Hydroxyphenamate 1.397 45Benzidine -0.495 -1.026 120Hydroxyphenamate 1.397 46Bifenox -3.397 -3.207 121Hymexazol 1.929 47Bifenthin -0.658 -0.217 123Imazapyr 1.053 48Biotin -0.658 -0.217 124Imazaquin -1.045 50Capric acid 1.012 1.128 125Imazaquin -1.045 51Carbofuran -0.495 -0.836 127Isofurophate 1.187 52Carbosulfan -3.522 -2.266 127Isofurophate 1.167 53Carbosurian -0.701 -0.274 128Isoleucine 1.039 54Carisoprodol -0.523 1.088 130Isophorone 1.079 55Carisoprodol -0.523 1.088 130Isophorone 1.079 56Carmosine 1.914 0.791 132Khellin 0.017 57Carmosine 1.934 1.788 134Linuron -1.124 59Crotonic Acid 0.301 1.614 137 <i>p</i> -Fluorobenzoic acid 0.0	11	Badische acid	-2.000	0.817	116	Hydrocortisone	_0.495
2 Darkin 1.150 2.120 113 Hydroquinone 1.857 33 Barbital 0.873 0.857 118 Hydroxyphenamate 1.397 44 Bendicarb -0.495 -1.026 120 Hydroxyproline 2.557 46 Bifenox -3.397 -3.207 121 Hydroxyproline 2.557 47 Bifentrhin -4.000 -3.5 122 Idoxuridine 0.301 48 Biotin -0.658 -0.217 123 Imazapyr 1.053 49 Capric acid 1.012 1.128 Imazethapyr 0.146 51 Carbofuran -0.495 -0.274 128 Isoleucine 1.536 52 Carboxulfan -3.522 -2.266 127 Isolfurophate 1.187 53 Carisoprodol -0.523 1.088 130 Isophorone 1.079 54 Carinostrane 0.907 131 Ketanserin -2.000 55	17	Barban	-1.958	-0.817	117	Hydroflumethiazide	-0.523
1110101110	13	Barbital	0.873	0.857	118	Hydroquinone	1 857
Senzidine -0.495 -1.026 120 Hydroxyproline 2.557 46 Bifenox -3.397 -3.207 121 Hymexazol 1.929 47 Bifenthin -4.000 -3.5 122 Idoxuridine 0.301 48 Biotin -0.658 -0.217 123 Imazapyr 1.053 49 Capric acid 1.012 1.128 124 Imazapyr 0.146 50 Caproic acid 1.012 1.128 125 Imazethapyr 0.146 51 Carbosufan -0.495 -0.836 126 Iridomyrmecin 0.301 52 Carbosufan -0.701 -0.274 128 Isoleucine 1.187 53 Carboxin -0.701 -0.274 128 Isoleucine 1.536 54 Carfentrazone-ethyl -1.657 -2.226 129 Isoniazid 2.146 55 Carisoprodol -0.523 1.088 134 Linuron -1.224	13 14	Bendiocarb	-0.585	-0.596	119	Hydroxyphenamate	1 397
Bifenox -3.397 -12.07 121 Hymexazol 1.929 47 Bifentrhin -4.000 -3.5 122 Idoxuridine 0.301 48 Biotin -0.658 -0.217 123 Imazapyr 1.053 49 Capric acid -1.012 1.128 Imazetquin -1.045 50 Caproir acid 1.012 1.128 Imazetquin -1.045 51 Carbosulfan -3.522 -2.296 127 Isoflurophate 1.187 52 Carbosulfan -0.701 -0.274 128 Isoleucine 1.036 54 Carfsoprodol -0.523 1.088 130 Isophorone 1.079 56 Carmosine 0.602 0.597 131 Ketanserin -2.000 57 Carosine 0.914 0.791 132 Khellin 0.017 58 Ló-Cleve's acid 0.000 -0.577 133 Lenacil -2.221 59 Crotonic Acid	15	Benzidine	-0.495	-1.026	120	Hydroxyproline	2 557
Differ Differ <thdiffer< th=""> <thdiffer< th=""> <thdiffer< td="" th<=""><th>16</th><td>Bifenox</td><td>-3 397</td><td>-3 207</td><th>121</th><td>Hymexazol</td><td>1 929</td></thdiffer<></thdiffer<></thdiffer<>	16	Bifenox	-3 397	-3 207	121	Hymexazol	1 929
Biotin Doto Doto <thdoto< th=""> Doto Doto <t< td=""><th>17</th><td>Bifentrhin</td><td>-4 000</td><td>-3.5</td><th>122</th><td>Idoxuridine</td><td>0 301</td></t<></thdoto<>	17	Bifentrhin	-4 000	-3.5	122	Idoxuridine	0 301
Description Description <thdescription< th=""> <thdescription< th=""></thdescription<></thdescription<>	18	Biotin	-0.658	-0.217	123	Imazapyr	1 053
Caproic acid 1.012 1.128 125 Image that 10.3 model 50 Carbofuran -0.495 -0.836 126 Iridomyrmecin 0.301 52 Carbosulfan -3.522 -2.296 127 Isoflurophate 1.187 53 Carboxin -0.701 -0.274 128 Isoleucine 1.536 54 Carfentrazone-ethyl -1.657 -2.226 129 Isoniazid 2.146 55 Carisoprodol -0.523 1.088 130 Isophorone 1.079 56 Carmosine 1.914 0.791 132 Khellin 0.017 57 Carnosine 1.914 0.791 132 Lenacil -2.221 59 Crotonic Acid 1.934 1.788 134 Linuron -1.124 60 Cumic Acid -0.821 -0.404 135 Methomyl 1.763 61 Cyanazine -0.767 -0.417 136 PABA 0.769 6	19	Capric acid	-1 209	-0.511	124	Imazaguin	-1.045
51 Carbofuran -0.495 -0.836 126 Internation process 0.301 52 Carbosulfan -3.522 -2.296 127 Isoflurophate 1.187 53 Carboxin -0.701 -0.274 128 Isoleucine 1.536 54 Carfentrazone-ethyl -1.657 -2.226 129 Isoniazid 2.146 55 Carisoprodol -0.523 1.088 130 Isophorone 1.079 56 Carmustine 0.602 0.597 131 Ketanserin -2.000 57 Carnosine 1.914 0.7991 132 Khellin 0.017 58 1.6-Cleve's acid 0.000 -0.577 133 Lenacil -2.221 59 Crotonic Acid 1.934 1.788 134 Linuron -1.124 60 Cumic Acid 0.301 1.614 137 p-Fluorobenzoic acid 0.079 61 Cyanazine -0.767 -0.417 136 PABA 0.769 62 Cyanuric Acid 0.301 1.614 137 <th>50</th> <td>Caproic acid</td> <td>1.203</td> <td>1 128</td> <th>125</th> <td>Imazethapyr</td> <td>0.146</td>	50	Caproic acid	1.203	1 128	125	Imazethapyr	0.146
52 Carbosulfan -3.522 -2.296 127 Isoflurophate 1.187 53 Carbosulfan -0.701 -0.274 128 Isoflurophate 1.536 54 Carfentrazone-ethyl -1.657 -2.226 129 Isoniazid 2.146 55 Carisoprodol -0.523 1.088 130 Isophorone 1.079 56 Carmustine 0.602 0.597 131 Ketanserin -2.000 57 Carnosine 1.914 0.791 132 Khellin 0.017 58 1.6-Cleve's acid 0.000 -0.577 133 Lenacil -2.221 59 Crotonic Acid 1.934 1.788 134 Linuron -1.124 60 Cumic Acid -0.821 -0.404 135 Methomyl 1.763 61 Cyanazine -0.767 -0.417 136 PABA 0.769 62 Cyanuric Acid 0.301 1.614 137 p-Fluorobenzoic acid 0.079 63 Cyclobarbital 0.204 0.241 139	51	Carbofuran	-0.495	-0.836	126	Iridomyrmecin	0 301
53 Carboxin -0.701 -0.274 128 Isoleucine 1.536 54 Carfentrazone-ethyl -1.657 -2.226 129 Isoniazid 2.146 55 Carisoprodol -0.523 1.088 130 Isophorone 1.079 56 Carnustine 0.602 0.597 131 Ketanserin -2.000 57 Carnosine 1.914 0.791 132 Khellin 0.017 58 1.6-Cleve's acid 0.000 -0.577 133 Lenacil -2.221 59 Crotonic Acid 1.934 1.788 134 Linuron -1.124 60 Cumic Acid -0.821 -0.404 135 Methomyl 1.763 61 Cyanazine -0.767 -0.417 136 PABA 0.769 62 Cyanuric Acid 0.301 1.614 137 p-Fluorobenzoic acid 0.079 63 Cyclobarbital 0.204 0.241 139 Phthalizene 1.698 64 Cyclobarbital 0.2051 1.391 141 <	52	Carbosulfan	-3 522	-2.296	127	Isoflurophate	1,187
54 Carfentrazone-ethyl -1.657 -2.226 129 Isoniazid 2.146 55 Carisoprodol -0.523 1.088 130 Isophorone 1.079 56 Carnosine 0.602 0.597 131 Ketanserin -2.000 57 Carnosine 1.914 0.791 132 Khellin 0.017 58 1.6-Cleve's acid 0.000 -0.577 133 Lenacil -2.221 59 Crotonic Acid 1.934 1.788 134 Linuron -1.124 60 Cumic Acid -0.821 -0.404 135 Methomyl 1.763 61 Cyanazine -0.767 -0.417 136 PABA 0.079 62 Cyanuric Acid 0.301 1.614 137 p-Fluorobenzoic acid 0.079 63 Cyclobarbital 0.204 0.241 139 Phthalazine 1.698 64 Cyclobarbital 0.204 0.241 139 Phthaliz Acid 0.846 65 Cyclobarbital 0.204 1.391 141	53	Carboxin	-0.701	-0.274	128	Isoleucine	1.536
55 Carisoprodol -0.523 1.088 130 Isophorone 1.079 56 Carmustine 0.602 0.597 131 Ketanserin -2.000 57 Carnosine 1.914 0.791 132 Khellin 0.017 58 1,6-Cleve's acid 0.000 -0.577 133 Lenacil -2.221 59 Crotonic Acid 1.934 1.788 134 Linuron -1.124 60 Cumic Acid -0.821 -0.404 135 Methomyl 1.763 61 Cyanazine -0.767 -0.417 136 PABA 0.769 62 Cyanuric Acid 0.301 1.614 137 p-Fluorobenzoic acid 0.079 63 Cyclobarbital 0.204 0.241 139 Phthalic Acid 0.846 64 Cyclobarbital 0.204 0.241 139 Phthalic Acid 0.846 65 Cyclobarbital 0.051 1.391 141 p-Hydroxybenzoic Acid 0.669 66 Cymoxanil -0.051 0.781 144	54	Carfentrazone-ethyl	-1.657	-2.226	129	Isoniazid	2.146
56 Carmustine 0.602 0.597 131 Ketanserin -2.000 57 Carnosine 1.914 0.791 132 Khellin 0.017 58 1,6-Cleve's acid 0.000 -0.577 133 Lenacil -2.221 59 Crotonic Acid 1.934 1.788 134 Linuron -1.124 60 Cumic Acid -0.821 -0.404 135 Methomyl 1.763 61 Cyanazine -0.767 -0.417 136 PABA 0.769 62 Cyanuric Acid 0.301 1.614 137 p-Fluorobenzoic acid 0.079 63 Cyclizine 0.000 -1.525 138 Phthaliazine 1.698 64 Cyclobarbital 0.204 0.241 139 Phthalic Acid 0.846 65 Cyclobarbital 0.204 0.241 139 Phthalimide -0.444 66 Cymoxanil -0.051 1.391 140 Phthalimide -0.444 66 Cyprocinazole -0.854 -1.399 142 <t< td=""><th>55</th><td>Carisoprodol</td><td>-0.523</td><td>1.088</td><th>130</th><td>Isophorone</td><td>1.079</td></t<>	55	Carisoprodol	-0.523	1.088	130	Isophorone	1.079
57 Carnosine 1.914 0.791 132 Khellin 0.007 58 1,6-Cleve's acid 0.000 -0.577 133 Lenacil -2.221 59 Crotonic Acid 1.934 1.788 134 Linuron -1.124 60 Cumic Acid -0.821 -0.404 135 Methomyl 1.763 61 Cyanazine -0.767 -0.417 136 PABA 0.769 62 Cyanuric Acid 0.301 1.614 137 p-Fluorobenzoic acid 0.0079 63 Cyclizine 0.000 -1.525 138 Phthalazine 1.698 64 Cyclobarbital 0.204 0.241 139 Phthalizine -0.444 66 Cymoxanil -0.051 1.391 141 p-Hydroxybenzoic Acid 0.699 67 Cyprocinazole -0.854 -1.399 142 Picloram -0.367 68 Cyprodinil -1.886 -1.58 143 Picric Acid 1.103 69 Cystine -0.951 0.781 144 <	56	Carmustine	0.602	0 597	131	Ketanserin	-2.000
58 1,6-Cleve's acid 0,000 -0,577 133 Lenacil -2,221 59 Crotonic Acid 1,934 1,788 134 Linuron -1,124 60 Cumic Acid -0,821 -0,404 135 Methomyl 1,763 61 Cyanazine -0,767 -0,417 136 PABA 0,769 62 Cyanuric Acid 0,301 1,614 137 p-Fluorobenzoic acid 0,079 63 Cyclizine 0,000 -1,525 138 Phthalazine 1,698 64 Cyclobarbital 0,204 0,241 139 Phthalic Acid 0,846 65 Cycloleucine 1,698 1,183 140 Phthaliride -0,444 66 Cymoxanil -0,051 1,391 141 p-Hydroxybenzoic Acid 0,699 67 Cyproconazole -0,854 -1,399 142 Picloram -0,367 68 Cyprodinil -1,886 -1,58 143 Picric Acid 1,103 69 Cystine -0,051 0,781 144	57	Carnosine	1 914	0.791	132	Khellin	0.017
59 Crotonic Acid 1.934 1.788 134 Linuron -1.124 60 Cumic Acid -0.821 -0.404 135 Methomyl 1.763 61 Cyanazine -0.767 -0.417 136 PABA 0.769 62 Cyanuric Acid 0.301 1.614 137 p-Fluorobenzoic acid 0.079 63 Cyclizine 0.000 -1.525 138 Phthalazine 1.698 64 Cyclobarbital 0.204 0.241 139 Phthaliz Acid 0.846 65 Cyclobarbital 0.204 0.241 139 Phthaliz Acid 0.846 66 Cymoxanil -0.051 1.83 140 Phthaliainide -0.444 66 Cyproconazole -0.854 -1.399 142 Picloram -0.367 67 Cyprodinil -1.886 -1.58 143 Picric Acid 1.103 69 Cystine -0.951 0.781 144 Pirimicarb 0.431 70 Dehydroacetic Acid -0.161 0.997 145	58	1.6-Cleve's acid	0.000	-0.577	133	Lenacil	-2.221
60 Cumic Acid -0.821 -0.404 135 Methomyl 1.763 61 Cyanazine -0.767 -0.417 136 PABA 0.769 62 Cyanuric Acid 0.301 1.614 137 p-Fluorobenzoic acid 0.079 63 Cyclizine 0.000 -1.525 138 Phthalazine 1.698 64 Cyclobarbital 0.204 0.241 139 Phthalic Acid 0.846 65 Cyclobarbital 0.204 0.241 139 Phthalic Acid 0.846 66 Cymoxanil -0.051 1.391 141 p-Hydroxybenzoic Acid 0.699 67 Cyproconazole -0.854 -1.399 142 Picloram -0.367 68 Cyprodinil -1.886 -1.58 143 Picric Acid 1.103 69 Cystine -0.951 0.781 144 Pirimicarb 0.431 70 Dehydroacetic Acid -0.161 0.997 145 Thionazin 0.057 71 Dexamethasone -1.051 -0.785 <td< th=""><th>59</th><th>Crotonic Acid</th><th>1 934</th><th>1 788</th><th>134</th><th>Linuron</th><th>-1.124</th></td<>	59	Crotonic Acid	1 934	1 788	134	Linuron	-1.124
61 Cyanazine -0.767 -0.417 136 PABA 0.769 62 Cyanuric Acid 0.301 1.614 137 p-Fluorobenzoic acid 0.079 63 Cyclizine 0.000 -1.525 138 Phthalazine 1.698 64 Cyclobarbital 0.204 0.241 139 Phthalic Acid 0.846 65 Cycloleucine 1.698 1.183 140 Phthalimide -0.444 66 Cymoxanil -0.051 1.391 141 p-Hydroxybenzoic Acid 0.699 67 Cyproconazole -0.854 -1.399 142 Picloram -0.367 68 Cyprodinil -1.886 -1.58 143 Picric Acid 1.103 70 Dehydroacetic Acid -0.161 0.997 145 Thionazin 0.057 71 Dexamethasone -1.051 -0.785 - - - - 72 Diallate -1.853 -1.154 Test set 21 - - - - - - - - - <th>50</th> <th>Cumic Acid</th> <th>-0.821</th> <th>-0.404</th> <th>135</th> <th>Methomyl</th> <th>1.763</th>	50	Cumic Acid	-0.821	-0.404	135	Methomyl	1.763
62 Cyanuric Acid 0.301 1.614 137 p-Fluorobenzoic acid 0.079 63 Cyclizine 0.000 -1.525 138 Phthalazine 1.698 64 Cyclobarbital 0.204 0.241 139 Phthalic Acid 0.846 65 Cycloleucine 1.698 1.183 140 Phthalimide -0.444 66 Cymoxanil -0.051 1.391 141 p-Hydroxybenzoic Acid 0.699 67 Cyproconazole -0.854 -1.399 142 Picloram -0.367 68 Cyprodinil -1.886 -1.58 143 Picric Acid 1.031 70 Dehydroacetic Acid -0.051 0.997 145 Thionazin 0.057 71 Dexamethasone -1.051 -0.785 - - - - 72 Diallate -1.853 -1.154 Test set 21 - - - - - - - - - - - - - - - - - - -	50 61	Cvanazine	-0.767	-0.417	136	PABA	0.769
63 Cyclizine 0.000 -1.525 138 Phthalazine 1.698 64 Cyclobarbital 0.204 0.241 139 Phthalazine 1.698 65 Cycloleucine 1.698 1.183 140 Phthalazine 0.846 66 Cymoxanil -0.051 1.391 141 p-Hydroxybenzoic Acid 0.699 67 Cyproconazole -0.854 -1.399 142 Picloram -0.367 68 Cyprodinil -1.886 -1.58 143 Picric Acid 1.103 69 Cystine -0.951 0.781 144 Pirimicarb 0.431 70 Dehydroacetic Acid -0.061 0.997 145 Thionazin 0.057 71 Dexamethasone -1.051 -0.785 - - - - 72 Diallate -1.853 -1.154 Test set 21 -<	62	Cyanuric Acid	0.301	1.614	137	p-Fluorobenzoic acid	0.079
64 Cyclobarbital 0.224 0.241 139 Phthalic Acid 0.846 65 Cycloleucine 1.698 1.183 140 Phthalic Acid 0.846 66 Cymoxanil -0.051 1.391 141 p-Hydroxybenzoic Acid 0.699 67 Cyproconazole -0.854 -1.399 142 Picloram -0.367 68 Cyprodinil -1.886 -1.58 143 Picric Acid 1.103 69 Cystine -0.951 0.781 144 Pirimicarb 0.431 70 Dehydroacetic Acid -0.051 -0.785 Thionazin 0.057 71 Dexamethasone -1.051 -0.785 Test set 21 Test set 21 73 Dicamba -0.080 -1.003 146 2.2', 4.5,5'-PCB -5.376 74 Dichlobenil -1.673 -0.705 147 Benzocaine -0.02 75 Dichlobenthion 2.6(10 2.466 148 Theenshulling 0.020 <th>63</th> <th>Cyclizine</th> <th>0,000</th> <th>-1.525</th> <th>138</th> <th>Phthalazine</th> <th>1.698</th>	63	Cyclizine	0,000	-1.525	138	Phthalazine	1.698
65 Cycloleucine 1.698 1.183 140 Phthalimide -0.444 66 Cymoxanil -0.051 1.391 141 p-Hydroxybenzoic Acid 0.699 67 Cyproconazole -0.854 -1.399 142 Picloram -0.367 68 Cyprodinil -1.886 -1.58 143 Picric Acid 1.103 69 Cystine -0.951 0.781 144 Pirimicarb 0.431 70 Dehydroacetic Acid -0.161 0.997 145 Thionazin 0.057 71 Dexamethasone -1.051 -0.785 - - - - 72 Diallate -1.853 -1.154 Test set 21 - - - - - - 5.376 74 Dichlobenil -1.673 -0.705 147 Benzocaine -0.102 - 0.102 75 Dichlobenthion 2.610 2.456 149 Theenbulling - 0.002	64	Cyclobarbital	0.204	0.241	139	Phthalic Acid	0.846
66 Cymoxanil -0.051 1.391 141 p-Hydroxybenzoic Acid 0.699 67 Cyproconazole -0.854 -1.399 142 Picloram -0.367 68 Cyprodinil -1.886 -1.58 143 Picric Acid 1.103 69 Cystine -0.951 0.781 144 Pirimicarb 0.431 70 Dehydroacetic Acid -0.161 0.997 145 Thionazin 0.057 71 Dexamethasone -1.051 -0.785 - - - - 72 Diallate -1.853 -1.154 Test set 21 -	65	Cycloleucine	1.698	1.183	140	Phthalimide	-0.444
67 Cyproconazole -0.854 -1.399 142 Picloram -0.367 68 Cyprodinil -1.886 -1.58 143 Picric Acid 1.103 69 Cystine -0.951 0.781 144 Pirimicarb 0.431 70 Dehydroacetic Acid -0.161 0.997 145 Thionazin 0.057 71 Dexamethasone -1.051 -0.785 - - - - 72 Diallate -1.853 -1.154 Test set 21 -	56	Cymoxanil	-0.051	1,391	141	p-Hydroxybenzoic Acid	0.699
68 Cyprodinil -1.886 -1.58 143 Picric Acid 1.103 69 Cystine -0.951 0.781 144 Pirimicarb 0.431 70 Dehydroacetic Acid -0.161 0.997 145 Thionazin 0.057 71 Dexamethasone -1.051 -0.785 - - - - 72 Diallate -1.853 -1.154 Test set 21 - 0.057 7 - - - - - 0.057 - 0.057 7 - - - - - - 0.057 - <	67	Cyproconazole	-0.854	-1.399	142	Picloram	-0.367
G9 Cystine -0.951 0.781 144 Pirimicarb 0.431 70 Dehydroacetic Acid -0.161 0.997 145 Thionazin 0.057 71 Dexamethasone -1.051 -0.785 Test set 21 73 Dicamba -0.080 -1.003 146 2,2',4,5,5'-PCB -5.376 74 Dichlobenil -1.673 -0.705 147 Benzocaine -0.02 75 Dichlobenil -2.650 2.456 148 Theoremulting 0.002	58	Cyprodinil	-1.886	-1.58	143	Picric Acid	1.103
70 Dehydroacetic Acid -0.161 0.997 145 Thionazin 0.057 71 Dexamethasone -1.051 -0.785 -0.785 -0.737 -0.733 -0.733 -0.080 -1.003 146 2,2',4,5,5'-PCB -5.376 73 Dicamba -0.080 -1.003 146 2,2',4,5,5'-PCB -5.376 74 Dichlobenil -1.673 -0.705 147 Benzocaine -0.02 75 Dichlobenil -2.465 149 Theoremultical 0.002	59	Cystine	-0.951	0.781	144	Pirimicarb	0.431
71 Dexamethasone -1.051 -0.785 -0.785 72 Diallate -1.853 -1.154 Test set 21 73 Dicamba -0.080 -1.003 146 2,2',4,5,5'-PCB -5.376 74 Dichlobenil -1.673 -0.705 147 Benzocaine -0.102 75 Dichlobenih 2.610 2.455 149 Theorem. 0.002	70	Dehydroacetic Acid	-0.161	0.997	145	Thionazin	0.057
72 Diallate -1.853 -1.154 Test set 21 73 Dicamba -0.080 -1.003 146 2,2',4,5,5'-PCB -5.376 74 Dichlobenil -1.673 -0.705 147 Benzocaine -0.102 75 Dichlofenthion 2.610 -2.455 149 Theoremultica 0.005	71	Dexamethasone	-1.051	-0.785	. 10		0.057
73 Dicamba -0.080 -1.003 146 2,2',4,5,5'-PCB -5.376 74 Dichlobenil -1.673 -0.705 147 Benzocaine -0.102 75 Dichlobenil -2.610 -2.456 148 Theoremultical -0.002	72	Diallate	-1.853	-1 154	Test set 21		
74 Dichlobenil -1.673 -0.705 147 Benzocaine -0.102 75 Dichlobention 2.610 2.456 148 Theoremultication 0.002	73	Dicamba	-0.080	-1.003	146	2.2'.4.5.5'-PCB	-5.376
75 Dichlefonthion 2 610 2 456 140 Theoremulting 0.000	74	Dichlobenil	-1 673	-0.705	147	Benzocaine	-0.102
- 1000000000000000000000000000000000000	75	Dichlofenthion	_3.610	-2 456	148	Theophylline	0.886

Table 2 (continued)

No.	Chemical name	Exp.	Pred. Eq. (3)
149	Antipyrine	2.665	0
150	Atrazine	-1.216	-0.509
151	Phenobarbital	0.026	-0.208
152	Diuron	-1.392	-0.713
153	Nitrofurantoin	-1.003	0.511
154	Phenytoin	-1.588	-0.581
155	Testosterone	-1.610	-1.955
156	Lindane	-2.136	-2.381
157	Parathion	-1.826	-1.453
158	Phenolphthalein	-0.397	-2.581
159	Malathion	-0.841	-0.523
160	Chlorpyrifos	-3.125	-1.833
161	Prostaglandin E2	0.077	-2.344
162	DDT	-5.530	-4.244
163	Chlordane	-4.247	-4.302
164	Diazepam	-1.301	-2.282
165	Aspirin	0.663	-0.074
166	Diazinon	-1.397	-0.535

- Solubility methods based on other experimental measurements, such as the melting point and the experimental log*P* value. Although they present good accuracy, the greatest drawback of these methodologies is the requirement of experimentally measure one or more physicochemical properties, which in some cases might be difficult or impossible to determine (e.g., compounds with very low or very high log*P* values and compounds with very high melting points that decompose before melting).
- Methods exploring 3D structure, which suppose either low speed of calculation (when ab initio approaches are employed) or previous optimization of molecular structures.
- 3. Methods using low dimensional descriptors (1D-2D). These include the group contribution methods (GCM) and QSAR approaches relying on topological descriptors. They are not computationally demanding, neither they require optimization of the molecular structure. GCM are easy to apply, relying solely on the sum of contributions of each molecular structure fragment to the aqueous solubility.¹⁰⁻¹² The basic assumption of this approach is the transferability concept for a group; if this hypothesis does not hold, then GCM can be corrected with experimental data when available to achieve better predictions. The methods proposed by Nirmalakhandan et al.,¹³ Suzuki et al., ¹⁴ Kuhne et al.,¹⁵ Lee et al.,¹⁶ and Klopman et al.^{17,18} belong to this category. Among all these approaches, only Klopman et al.'s approach is a pure and general group contribution model without using additional experimental parameters. Although GCM have a simple and practical implementation, some common drawbacks of this methodology are the following: (a) they require a large data set to obtain a contribution of each functional group; (b) in its basic form (without corrections) it cannot model isomeric structures; (c) they may contain a 'missing fragment' problem, which means that if a compound contains a missing fragment which cannot be defined by the group contribution model, its aqueous solubility cannot be precisely predicted; (d) measured data are not always available to extend these methods to strange compounds such as molecules containing fused aromatic rings or to organo-metallic compounds.

Table 1 summarizes different linear estimation methods for solubility prediction in terms of type and number of structural descriptors used to derive the model, and the root mean square error (rms) against a common test set of 21 'classic compounds' found in many solubility prediction papers.¹⁸ It has been pointed out that solubility modeling efforts have suffered from some ba-

sic concerns, among them: training sets that are not drug-like, lack of structural diversity, unknown experimental error, incorrect tautomers or structures, neglect of ionization and crystal packing effects, over-sampling of compounds with low molecular weight and range in solubility data that is not pharmaceutically relevant.4,19 In this paper, we present a QSPR that answers to some of these issues, since it was developed from a structural diverse training set composed by drug-like compounds with more than half the data set presenting solubility values below 1 mg ml⁻¹. Note that low solubility compounds are actually the ones one would like to obtain more accurate predictions,^{4,19} since they have higher probability of presenting difficulties in pre-clinic and clinic assays and formulation stages. We were also careful not to over-represent compounds with low molecular weight. In the present analysis, we decided to use Multivariable Linear Regression (MLR)-based methods instead of the GCM approach for analyzing the aqueous solubilities of 166 organic compounds. A great number of theoretical molecular descriptors are simultaneously explored by including definitions of all classes. For this task, we employ the linear variable subset selection approach Replacement Method (RM),^{20–23} and we draw conclusions by contrasting our results with other previously reported linear models of the literature.

2. Methods

2.1. Data set

The experimental aqueous solubilities (Sol) measured at 298 K and expressed in mg ml⁻¹ for 145 structurally diverse drug-like organic compounds were extracted from Merck Index 13th.²⁴ Solubility data were checked at ChemID Plus (National Library of Medicine, National Institute of Health).²⁵ No differences in solubility data were found between Merck Index and ChemID records except for crotonic acid (Δ Sol = 0.053 log units), cyanazine (Δ Sol = 0.003), dexamethasone (Δ Sol = 0.050) and PABA (Δ Sol = 0.016). In those cases ChemID data were considered. None solubility record at 25° was found in ChemID Plus for 4-amino-2-sulfobenzoic acid, acequinocyl, aconitic acid, amicarbalide, aminopromazine, ascorbic acid, axocystrobin, ethirimol and furametpyr. For modeling purposes, these data are converted into logarithm units (log₁₀Sol) and are presented in Table 2; all the molecular structures are drawn in Figures 1 and 2. The molecular set was split into a 97-compound training set (train) and a 48-compounds test set (val), selecting the members of each set in such a way to share similar structural characteristics of the compounds. Additionally, we also used an external molecular set (test set 21) that was not involved during the model design, composed of 21 well-known compounds found in many solubility prediction papers,^{4,18} in order to further examining the model's validation. In a recent work, we have already used this 145 data set (plus aspirin, diazepam, and diazinon, which in the current study are part of the classic 21-compound test) for modeling of aqueous solubility through the RM.²⁶ In that opportunity, however, we conditioned the model to include at least one out of twelve descriptors inspired in Lipinski rules.²⁷ In the present study, we imposed no restriction regarding the descriptors included in the models.

Note that most of the drugs that compose the training and test sets accomplish several drug-likeness criteria. It can be noticed that more than 99% of the data set observes the Lipinski-rule criteria for estimating drug oral bioavailability,²⁸ while more than 93% accomplish Veber et al. rule.²⁹ More than 99% of the data set also accomplishes more general rules for evaluating drug-likeness extracted from several recent publications^{30–32}: 100 \leq molecular weight \leq 800 g mol⁻¹; log*P* \leq 7; number of H bond acceptors \leq 10; number of H bond donors \leq 5; rotatable bonds \leq 15; halogen



Figure 2. Molecular structures for the test set compounds (N = 48).

atoms \leqslant 7; alkyl chains \leqslant (CH₂)₆CH₃; no perfluorinated chains: CF₂CF₂CF₃; no big size ring with more than seven members; no presence of other atoms than C, O, N, S, P, F, Cl, Br, I, Na, K, Mg, Ca or Li and; presence of at least one N or O atom. Moreover, note that low molecular weight compounds are not over-represented in this molecular set. The structural diversity of the training set was assessed through calculation of the average Tanimoto intermolecular distances (based on atom pairs) for all the possible pairs of structures that could be derived from the training set. For this pur-

pose, we used de PowerMV software provided by the National Institute of Statistical Sciences.³³ According to the results, the average Tanimoto intermolecular distance for the training set is 0.781 with a SD of 0.412, which confirms the high structural diversity of the training set. Figure 3 includes a histogram representing the distribution of the 166 aqueous solubilities under study, which suggests that the experimental sample is normally distributed over more than four logarithmic units and can thus be employed in regression analysis.



Figure 2. (continued)



Figure 3. Normal distribution of the experimental $log_{10}Sol$ values under analysis N = 166.

2.2. Molecular descriptors

The structures of the compounds were firstly pre-optimized with the Molecular Mechanics Force Field (MM+) procedure included in the Hyperchem 6.03 package,³⁴ and the resulting geometries were further refined by means of the Semi-Empirical Molecular Orbitals Method PM3 (Parametric Method-3) using the Polak-Ribiere algorithm and a gradient norm limit of 0.01 kcal Å⁻¹.

We computed 1497 molecular descriptors using the software Dragon 5.0,³⁵ including descriptors of all types such as Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecu-

lar Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups, Atom-Centred Fragments, Empirical and Properties.³⁶ Furthermore, four molecular descriptors were derived taking into consideration Lipinski's rule, based on combinations of the detour index *dd* from the Chemical Graph Theory (calculated as the ratio between the half sum of the elements of the Detour Matrix (DD) and molecular features related to solubility such as the number of H donors (D), the number of H acceptors (A), and the number of hetero-atoms (H) present in the molecular structure).^{26,27,37} We also considered the square and cubic roots of these last descriptors. Finally, five quantum-chemical descriptors not provided by the program Dragon were added to the pool: molecular dipole moments, total energies, homo-lumo energies, and homo-lumo gap (Δ homo-lumo) calculated at the PM3 level. The total pool of explored descriptors consisted on D = 1514variables.

2.3. Model search

The computer system Matlab 5.0 was used in all our calculations.³⁸ Our purpose was to search the optimal subset of *d* descriptors from the total number of *D* descriptors which to accomplish the following criterion: $d \ll D$ and *d* with minimum standard deviation *S*:

$$S = \frac{1}{(N-d-1)} \sum_{i=1}^{N} \operatorname{res}_{i}^{2}$$
(1)

where *N* is the number of molecules in the training set, and res_{*i*} the residual for molecule *i* (difference between the experimental and predicted property **p**). More precisely, we want to obtain the global minimum of *S*(**d**) where **d** is a point in a space of D!/[d!(D-d)!]

Table 3

Linear QSPR models established for the training set of aqueous solubilities (N = 97)

d ^a	Descriptors involved	R ^b	S ^c	FIT ^d	R _{loo} ^e	S _{loo} ^f	$R_{\rm val}{}^{\rm g}$	S_{val}^{h}
1	DP03	0.722	1.257	1.053	0.708	1.283	0.794	1.047
2	DP03, MLOGP	0.831	1.016	2.071	0.817	1.054	0.798	0.983
3	X1sol, RDF060u, MLOGP	0.871	0.903	2.747	0.849	0.971	0.848	0.899
4	X1sol, RDF060u, RDF020e, MLOGP	0.889	0.844	3.078	0.870	0.911	0.838	0.986
5	Sp, nR09, H3D, Mor04u, MLOGP	0.895	0.829	2.991	0.878	0.890	0.891	0.758

The best relationship found appears in bold.

^a *d*: number of descriptors in the linear regression.

^b *R*: correlation coefficient of the model.

^c *S*: standard deviation of the model.

^d FIT: Kubinyi function.

^e R_{loo}: R of Leave-One-Out.

 $^{\rm f}$ S_{loo}: S of Leave-One-Out.

^g R_{val} : *R* of validation test set.

^h S_{val} : S of validation test set.

Table 4

Symbols for molecular descriptors involved in different models

Molecular descriptor	Dim ^a	Туре	Description
DP03	3D	Randic molecular profiles	Molecular profile No. 3
MLOGP	1D	Properties	Moriguchi octanol-water partition coefficient
X1sol	2D	Topological	Solvation connectivity index chi-1
RDF060u	3D	Radial Distribution Function	Radial distribution function – 6.0/unweighted
RDF020e	3D	Radial Distribution Function	Radial distribution function - 2.0/weighted by atomic Sanderson electronegativities
Sp	0D	Constitutional	Sum of atomic polarizabilities (scaled on carbon atom)
nR09	0D	Constitutional	Number of nine-membered rings
H3D	3D	Geometrical	3D-Harary index
Mor04u	3D	3D-MoRSE	3D-MoRSE-signal 04/unweighted

^a Dim, dimensionality of the descriptor.

ones. Usually, a full search (FS) of optimal variables is unfeasible because it requires D!/[d!(D - d)!] linear regressions. Some time ago we proposed the Replacement Method (RM) that produces linear QSPR-QSAR models that are quite close the FS ones with much less computational work.^{20–23} This technique approaches the minimum of *S* by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of *d* descriptors $\mathbf{d} = \{X_1, X_2, \dots, X_d\}$. The RM gives models with better statistical parameters than the Forward Stepwise Regression procedure and similar ones to the more elaborated Genetic Algorithms.^{39,40}

The Kubinyi function $(FIT)^{41}$ is a statistical parameter that closely relates to the Fisher ratio (*F*), but avoids the main disadvantage of the latter that is too sensitive to changes in small *d* values and poorly sensitive to changes in large *d* values. The *FIT*(**d**) criterion has a low sensitivity to changes in small *d* values and a substantially increasing sensitivity for large *d* values. The greater the *FIT* value the better the linear equation. It is given by the following equation, where *R*(**d**) is the correlation coefficient for a model with *d* descriptors.

$$FIT = \frac{R^2(N-d-1)}{(N+d^2)(1-R^2)}$$
(2)

2.4. Model internal validation

The theoretical 'internal validation' practiced over each developed linear model is based on the Leave-More-Out Cross-Validation procedure (*l*-*n*%-*o*),⁴² with *n*% representing the percentage of molecules removed from the training set. The number of cases for random data removal analyzed in every *l*-*n*%-*o* is of 5,000,000. The percentage *n*% depends simultaneously upon the number of compounds (*N*), as one cannot remove many molecules from the training set if a small sample is analyzed as the normality condition of the fitted data has to be obeyed, and upon their structural diversity, since if the molecules are structurally very different, more compounds would have to be removed from the set for checking the predictive performance of the model. We choose the value of n% = 10% (10 compounds) in Cross-Validation in order to properly validate the QSAR equations.

In addition, we applied the *y*-randomization technique⁴³ with the purpose of demonstrating that the model established does not result from happenstance but involves a real structure–property relationship. This method consists on scrambling the experimental property of each compound in such a way that it does not correspond to the respective compound. After analyzing 5,000,000 cases of *y*-randomization for each developed QSPR, the smallest *S* value obtained using this procedure turned out to be a poorer value when compared to the one found when considering the true calibration.

2.5. Orthogonalization procedure

We employ the orthogonalization procedure introduced several years ago by Randic^{44,45} as a way of improving the statistical interpretation of the model built by interrelated indices. From our point of view, the co-linearity of the molecular descriptors should be as low as possible, because the interrelatedness among the different descriptors can lead to highly unstable regression coefficients, which makes it impossible to know the relative importance of an index and underestimates the utility of the regression coefficients of the model. The crucial step of the orthogonalization process is the choice of an appropriate order of orthogonalization, which in present analysis is the order that maximizes the correlation between each orthogonal descriptor and the observed aqueous solubilities. From now on, an orthogonalized descriptor will be represented with notation Ω .

3. Results and discussion

The application of the RM method on the training set of 97 heterogeneous drugs leads to the best 1–5 variables linear regression models listed in Table 3, while the specific details for all the molecular descriptors reported in this article are provided in Table 4. A close inspection of Table 3 reveals that the best linear QSPR equation found for modeling the aqueous solubility of the organic compounds includes the following satisfactory three molecular descriptors relationship:

$$\begin{split} log_{10}Sol = & -0.435(\pm 0.03) \cdot \Omega(X1sol) - 0.503(\pm 0.06) \cdot \Omega(MLOGP) \\ & + 0.0767(\pm 0.01) \cdot \Omega(RDF060u) + 2.970(\pm 0.3) \end{split}$$

$$\begin{split} N_{\rm train} &= 97, \, N_{\rm train}/d = 32.333, \, R = 0.871, \, S = 0.903, \, FIT = 2.747 \\ R_{\rm loo} &= 0.849, \quad S_{\rm loo} = 0.971, \quad R_{l-10\%-o} = 0.809, \quad S_{l-10\%-o} = 1.090, \\ p &< 10^{-4} \\ N_{\rm val} &= 48, \, R_{\rm val} = 0.848, \, S_{\rm val} = 0.899 \end{split}$$

Here, the absolute errors of the regression coefficients are provided in parentheses, *p* is the significance of the model, and loo sub-index stands for the Leave-One-Out Cross-Validation technique.⁴² The QSPR derived does not incorporate redundant structural information, as it involves orthogonal descriptors. Furthermore, by means of a proper standardization³⁹ of such orthogonal variables it is feasible to assign a greater importance to those molecular descriptors that exhibit larger absolute standardized coefficients (st.coeff.). The order of appearance of each descriptor within the QSPR of Eq. 3 corresponds to its order of importance in the established relationship, and each variable includes the following standardized coefficients: $\Omega(X1sol)$: 0.71, $\Omega(MLOGP)$: 0.42, and $\Omega(RDF060u)$: 0.28.

Table 2 also includes the predicted residuals as obtained via Eq. (3) for the training and test sets, while the plot of predicted versus experimental aqueous solubilities shown in Figure 4a suggests that the 97 training and 48 test set val compounds follow a straight line. The behavior of the plotted residuals in terms of the predictions in Figure 4b leads to a normal distribution. This figure includes two calibration outliers with a residual exceeding the value 2S = 1.806: compounds 15 (Acibenzolar-S-Methyl, 1.902) and 91(Etofenprox, -2.545), while none of the training compounds exceed the value 3S = 2.709; the presence of these outliers may be attributed exclusively to be a pure consequence of the limited number of structural descriptors participating in Eq. 3, since this model have a high ratio of number of observations to number of parameters (N/d = 32.333). The predictive power of the linear model is satisfactory as revealed by its stability upon the inclusion or exclusion of compounds, as measured by the loo parameters $R_{\rm loo}$ = 0.849 and $S_{\rm loo}$ = 0.971, and by the more severe test of higher percentage of compounds exclusion $R_{l-10\%-o} = 0.809$ and $S_{l-10\% - 0}$ = 1.090. These results are in the range of a validated model: $R_{l-n\%-o}$ must be greater than the value of 0.50, according to the specialized literature.⁴⁶ Furthermore, the predictive capability of the so-established equation is demonstrated by its performance in the test set val, leading to $R_{val} = 0.848$ and $S_{val} = 0.899$. Finally, after analyzing 5,000,000 cases fory-randomization, the smallest S value obtained using this procedure was 1.650, a poorer value when compared to the one found considering the true calibration (S 0.903). In this way, the robustness of the model could be assessed, showing that the calibration was not a fortuitous correlation and therefore results in a structure-activity relationship.

The three structural descriptors mentioned in Eq. 3 quantify different aspects of the molecular geometry and can be classified as follows: (i) a topological 2D-descriptor: X1sol, the solvation connectivity index chi-1, (ii) a Property 1D-descriptor: MLOGP, the



Figure 4. (a) Predicted (Eq. 3) versus experimental log₁₀Sol for the training and test sets. (b) Dispersion plot of the residuals for the training and test sets according to Eq. 3.

Moriguchi octanol-water partition coefficient; and (iii) a Radial Distribution Function 3D-descriptor: RDF060u, the radial distribution function -6.0/unweighted. As can be appreciated, different definitions of descriptors are needed to correctly represent the structures for the drug-like heterogeneous compounds. Figure 5 includes the histograms of the 166 organic compounds for each of the three descriptors appearing in the optimal QSPR equation found.

The most important structural factor of the model, the bidimensional descriptor X1sol, was proposed by Zefirov and Palyulin⁴⁷ in 1991 in order to treat the enthalpies of non-specific solvation. For instance, the solvation enthalpy of propane (CH₃CH₂CH₃) and di-methyl-mercury (CH₃HgCH₃) differs enormously, but both of these molecules are represented by the same hydrogen depleted graph, and, hence, have the identical topological indices which do not take into account atom types. The solvation index was created exactly to differentiate such cases, having the following general formula when calculated for hydrogen- and fluorine-depleted molecular graphs:

$$\mathsf{Xmsol} = (1/2^{m+1}) \sum \frac{Z_i Z_j \dots Z_k}{\left(\delta_i \delta_j \dots \delta_k\right)^{1/2}} \tag{4}$$

where *m* is the order of index; summation is over all sub-graphs of order *m*; $\delta_i \delta_i \dots \delta_k$ are connectivities of vertexes of sub-graph; and



Figure 5. Histograms for the molecular descriptors appearing in the QSPR solubility model (*N* = 166).

 $Z_i Z_j \dots Z_k$ are coefficients characterizing the atom size, which coincide to the number of the period in the Periodic Table. The term $1/2^{m+1}$ just normalizes values of Xmsol to provide their coincidence with the connectivity index Xm for the elements of the second row. The second important descriptor involved in Eq. 3 corresponds to the Moriguchi octanol–water partition coefficient,⁴⁸ revealing that a compound's hydrofobicity plays a crucial role in explaining the aqueous solubility data. Finally, the contribution of a 3D-Radial Distribution Function⁴⁹ helps to improve the predictive power of the QSPR. Such a kind of molecular descriptor defined for an ensemble of atoms may be interpreted as the probability distribution of finding an atom in a spherical volume of certain radius, incorporating

different types of atomic properties in order to differentiate the nature and contribution of atoms to the property being modeled. For the case of RDF060u, the sphere radius is of 6.0 Å and no atomic property is employed, thus characterizing the molecular size.

It is feasible to discuss the numerical effect of the optimal subset of structural descriptors selected in Eq. 3 on the aqueous solubility predictions. Since the orthogonal descriptor $\Omega(X1sol)$ is numerically positive for all the structures under study, its contribution to log₁₀Sol results in a negative quantity, according to the regression coefficient (-0.435). This causes that chemical compounds displaying greater values of $\Omega(X1sol)$ would tend to exhibit lower predicted values of aqueous solubilities. For the case of the orthogonal variable $\Omega(MLOGP)$, drugs manifesting higher positive values of this descriptor would tend to manifest their preference to the octanol lipophilic phase rather than to the water phase. and according to the sign of the regression coefficient in Eq. 3 (-0.503) would lead to a lower prediction of the aqueous solubilities. Finally, the tri-dimensional descriptor $\Omega(RDF060u)$ would tend to lead to higher predictions of log₁₀Sol whenever it presents higher numerical values.

Applying now the designed QSPR model of Eq. 3 to the classical test set 21, whose data are considered 'unknown' and that do not participate during the model development (as is the case of test set val), leads to a square root mean quadratic residual (rms) of 1.202. The statistical quality achieved on this test set is comparable to that obtained by the previously reported models for aqueous solubilities in Table 1, and the main advantage here is that only three molecular descriptors are employed to model the physical property and thus leads to a favorable ratio N/d = 7. This equation results in a superior predictive quality than that obtained by the GCM of Klopman (rms = 1.213) involving 34 parameters,¹⁸ and also outperforms the MLR of Yan (rms = 1.286) using 40 parameters.⁵⁰

4. Conclusions

The chemical information encoded by three theoretical molecular descriptors of the one-, two-, and three-types participating in a linear QSPR model enabled to explain the variation of the experimental aqueous solubilities in a satisfactory extent, and allowed a proper characterization of structurally heterogeneous drug-like organic compounds from both the training and test sets. The QSPR designed involved molecular descriptors that have a quite direct interpretation, and this relationship proved to have general applicability. The statistical parameters of the proposed model compare fairly well with others published previously based on Group Contribution methods. Furthermore, among the different linear regression based-algorithms, the Replacement Method continues demonstrating to be an efficient technique for the search of a reduced set of numerical variables from a huge number of them. This has application for the analysis of any physicochemical, biological, or pharmacological property of interest.

Acknowledgments

P.R.D. and A.T. thank to the National Council of Scientific and Technological Research (CONICET) for supporting this work. L.B.B. thanks ANPCyT and Universidad Nacional de La Plata for supporting this work.

References and notes

- 1. Schuster, D.; Laggner, C.; Langer, T. Curr. Pharm. Des. 2005, 11, 3545.
- 2. Stegemann, S.; Leveiller, F.; Franchi, D.; de Jong, H.; Lindén, H. *Eur. J. Pharm. Sci.* **2007**, *31*, 249.
- 3. Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. Curr. Med. Chem. 2006, 13, 226.
- 4. Delaney, J. S. Drug Discov. Today 2005, 10, 289.
- 5. Goodwin, J. J. Drug Discov. Today Technol. 2006, 3, 67.

- 6. Alsenz, J.; Kansy, M. Adv. Drug Deliv. Rev. 2007, 59, 546.
- 7. Bhattachar, S. N.; Deschenes, L.; Wesley, J. A. Drug Discov. Today 2006, 11, 1012.
- 8. Di, L.; Kerns, E. H. Drug Discov. Today **2006**, *11*, 446. 9. Smith C. L.; Hansch, C. Food Chem. Toxicol. **2000**, 38, 6
- 9. Smith, C. J.; Hansch, C. Food Chem. Toxicol. 2000, 38, 637.
- Artist, http://www.ddbst.de/new/Win_DDBSP/frame_Artist.htm.
 ChemEng Software Design, http://www.cesd.com/chempage.htm.
- Predict, http://www.mwsoftware.com/dragon/desc.html.
- 13. Nirmalakhandan, N. N. P.; Speece, R. E. Environ. Sci. Technol. **1989**, 23, 708.
- 14. Suzuki, T. J. Comput.-Aided Mol. Des. **1991**, 5, 149.
- 15. Kuhne, R.; Ebert, R. U.; Kleint, F.; Schmidt, G.; Schuurmann, G. Chemosphere 1995, 30, 2061.
- 16. Lee, Y.; Myrdal, P. B.; Yalkowsky, S. H. Chemosphere 1996, 33, 2129.
- 17. Klopman, G.; Zhu, H. J. Chem. Inf. Model. 2001, 41, 439.
- 18. Klopman, G.; Wang, S.; Balthasar, D. M. J. Chem. Inf. Model. 1992, 32, 474.
- 19. Johnson, S. R.; Zheng, W. AAPS J. 2006, 8, E27.
- Duchowicz, P. R.; Castro, E. A.; Fernández, F. M.; González, M. P. Chem. Phys. Lett. 2005, 412, 376.
- Duchowicz, P. R.; Castro, E. A.; Fernández, F. M. MATCH Commun. Math. Comput. Chem. 2006, 55, 179.
- Duchowicz, P. R.; Fernández, M.; Caballero, J.; Castro, E. A.; Fernández, F. M. Bioorg. Med. Chem. 2006, 14, 5876.
- Helguera, A. M.; Duchowicz, P. R.; Pérez, M. A. C.; Castro, E. A.; Cordeiro, M. N. D. S.; González, M. P. Chemometr. Intell. Lab. 2006, 81, 180.
- 24. The Merck Index An Encyclopedia of Chemicals, Drugs, and Biologicals; Merck & Co.: NJ, 2001.
- Division of Specialized Information Services, National Institute of Health. ChemID Plus. http://chem.sis.nlm.nih.gov/chemidplus/.
- Duchowicz, P. R.; Talevi, A.; Bellera, C.; Bruno-Blanch, L. E.; Castro, E. A. Bioorg. Med. Chem. 2007, 15, 3711.
- 27. Talevi, A.; Castro, E. A.; Bruno-Blanch, L. E. J. Arg. Chem. Soc. 2006, 44, 129.
- Lipinski, C. A.; Lombardo, F.; Dominy, D. W.; Feeney, P. J. Adv. Drug Deliver. Rev. 2001, 46, 3.

- Veber, D. F.; Johnson, S. R.; Cheng, H.; Smith, B. R.; Ward, K. W.; Kopple, K. D. J. Med. Chem. 2002, 45, 2615.
- 30. Charifson, P. S.; Walters, W. P. J. Comput. Aided Mol. Des. 2002, 16, 311.
- 31. Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. Mol. Divers. 2006, 10,
- 339.
 Walters, W. P.; Murcko, M. A. Adv. Drug Deliv. Rev. 2002, 54, 255.
- Vallers, W. F., Marcko, M. P. Juv. Didg Denv. Rev. 2002, 94, 255.
 Liu, K.; Feng, J.; Young, S. S. J. Chem. Inf. Model. 2005, 45, 515. PowerMV v.0.61.
- http://www.niss.org/PowerMV.
- 34. Hyperchem 6.03 (Hypercube) http://www.hyper.com.
- 35. Dragon 5.0, Evaluation Version, http://www.disat.unimib.it/chm.
- 36. Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors; Wiley VCH: Weinheim, Germany, 2000.
- 37. Harary, F. Graph Theory; Addison-Wesley, 1969.
- 38. Matlab 7.0, The MathWorks Inc.
- Draper, N. R.; Smith, H. Applied Regression Analysis; John Wiley& Sons: New York, 1981.
- 40. So, S. S.; Karplus, M. J. Med. Chem. 1996, 39, 1521.
- 41. Kubinyi, H. Quant.-Struct.-Act. Relat. 1994, 13, 393.
- 42. Hawkins, D. M.; Basak, S. C.; Mills, D. J. Chem. Inf. Model. 2003, 43, 579.
- 43. Wold, S.; Eriksson, L. Chemometrics Methods in Molecular Design; VCH: Weinheim, 1995.
- 44. Randic, M. J. Chem. Inf. Model. 1991, 31, 311.
- 45. Randic, M. New J. Chem. 1991, 15, 517.
- 46. Golbraikh, A.; Tropsha, A. J. Mol. Graphics Model. 2002, 20, 269.
- Antipin, I. S.; Arslanov, N. A.; Palyulin, V. A.; Konovalov, A. I.; Zefirov, N. S. Dokl. Akad. Nauk. SSSR 1991, 316, 925 (Chem. Abstr. 115, 91390).
- Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsuchita, Y. Chem. Pharm. Bull. 1992, 40, 127.
- 49. Consonni, V.; Todeschini, R.; Pavan, M. J. Chem. Inf. Model. 2002, 42, 693.
- 50. Yan, A.; Gasteiger, J. J. Chem. Inf. Model. 2003, 43, 429.
- 51. Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. J. Chem. Inf. Model. 2004, 44, 266.
- 52. Huuskonen, J. J. Chem. Inf. Model. 2000, 40, 773.