Insolubility classification with accurate prediction probabilities using a MetaClassifier

Christian Kramer,[1,2] Bernd Beck[,1] and Timothy Clark*[,2,3]*

[1] Department of Lead Discovery, Boehringer-Ingelheim Pharma GmbH & Co. KG, 88397 Biberach (Germany)

[2] Computer-Chemie-Centrum and Interdisciplinary Center for Molecular Materials, Friedrich-Alexander Universität Erlangen-Nürnberg, Nägelsbachstrasse 52, 91052 Erlangen (Germany)

[3] Centre for Molecular Design, University of Portsmouth, Mercantile House, Hampshire Terrace, Portsmouth, PO1 2EG, United Kingdom.

Supporting Information

Confusion matrix and ROC_AUC matrix scheme for a three-class problem

Based on the confusion matrix scheme shownin Table S1, the accuracy of prediction is calculated as $(AA+BB+CC)/N_{total}$, where N_{total} is the total number of samples.

	Predicted		
Measured	AA	AB	AC
	BA	BB	BC
	CA	CB	CC

Table S1: Confusion matrix scheme for classes A, B and C. (AB, for example is the number of samples measured as A and predicted to be B)

This accuracy measure has the disadvantage that it does not scale according to the distribution of the different classes. Whereas 100% is always perfect and 0% is always completely wrong, it is, for example, very easy to obtain 80% accuracy if 80% of the samples are in one class (simply by assigning all samples to this class). Therefore, the ratios of true and false positives are the accuracy measures of choice. They are scaled to the total number of instances in a specific class. For a two-class problem, their calculation is based on the distribution of negatives/positives. The true positive rate (*TPR*) is calculated as $TPR = TP/(TP+FN)$. The false positive rate (*FPR*) is calculated as $FPR = FP/(FP+TN)$. *TP* stands for true positives, *TN* for true negatives, *FP* for false positives and *FN* for false negatives.

TPR and *FPR* do not take into account the probabilities of correct prediction or the ranking of compounds. For example, if the dataset contains compounds that are predicted with high probabilities of being correct and others for which the classifier cannot give a safe prediction, these two classes of predictions would not be visible from the measures of quality described above. The ROC curve and the measure of the ROC area under the curve (ROC-AUC) were therefore used as improved performance metrics. In the current case, we can create ROC curves based on the probabilities of the predictions being correct. The dataset can be sorted in descending order of these probabilities. *TPR* and the *FPR* can be calculated for each probability level. Plotting each pair of *TPR/FPR* gives the ROC curve.

The ROC curve extends from zero and one. If the area under the ROC curve is unity, the predictions are perfect; if it is 0.5, the predictions are random. The ROC curve automatically scales to asymmetric class distributions and implies compound ranking. Thus, it represents a more informative criterion for classification performance than accuracy if scores or probabilities are available.

For two-class problems there is only one ROC-AUC. For multiclass problems however, the ROC-AUC can be calculated for either one-versus-all or one-versus-one comparisons. In a ROC-AUC matrix, the different ROC-AUCs are shown in Table S2. For one-versus-one ROC-AUC calculations, all other classes are omitted and the scores/probabilities are readjusted to one.

Class 1 vs. all	Class 1 vs. Class 2	Class 1 vs. Class 3
Class 1 vs. Class 2	Class 2 vs. all	Class 2 vs. Class 3
Class 1 vs. Class 3	Class 2 vs. Class 3	Class 3 vs. all

Table S2: ROC-AUC matrix for a three-class problem.

Impact of experimental uncertainty in classification scenario

This analysis can only be carried out for the bounded bins as no experimental uncertainty can be added to a solubility value characterized by an upper or lower bound only. Each measured value was replaced with a normalized Gaussian for the probability of true solubility and the integral for each solubility class was calculated. Figure S1 illustrates this approach schematically.

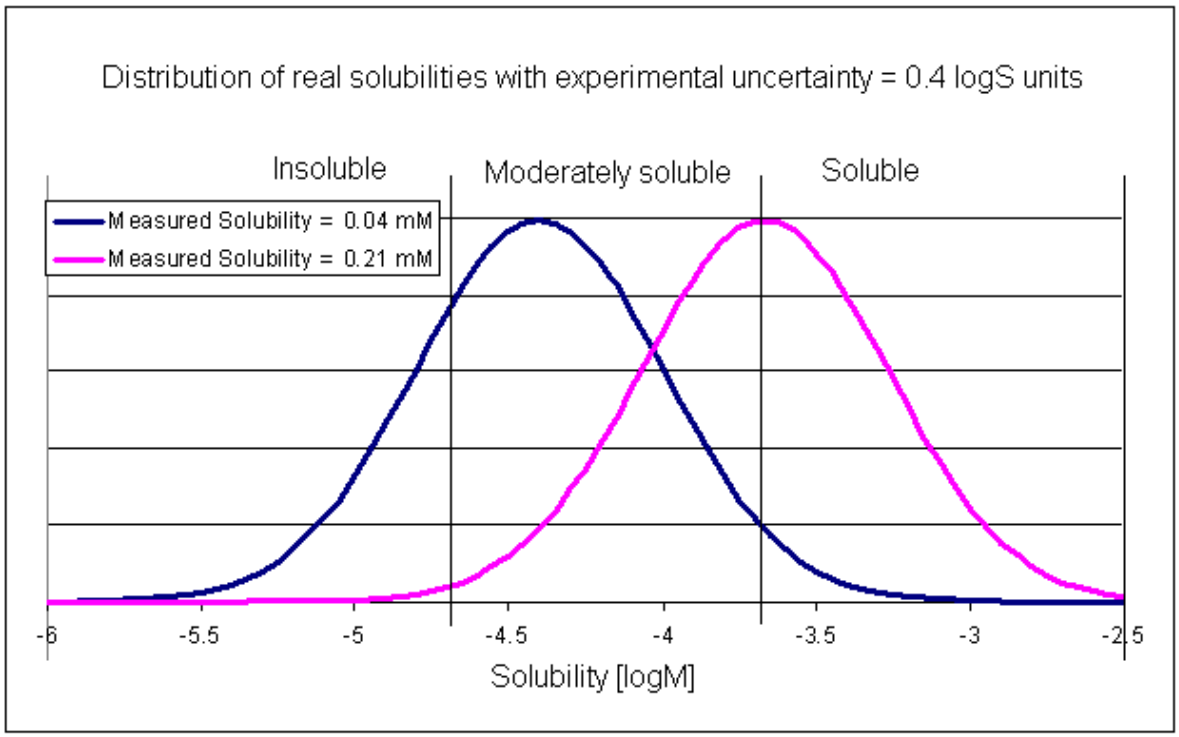

Figure S1: Normal Probability distribution of real solubility with an experimental uncertainty of 0.4 logS units and the classification thresholds used in this study

The probability of actually being insoluble for the compounds with measured solubility of 0.04 mM is 23%. The probability of really being moderately soluble is 73%. For the other compound with measured solubility of 0.21 mM, the probability of really being soluble is 53%, the probability of being moderately soluble is 43%. Larger standard deviations (i.e. less safe measurements) lead to lower maximum achievable accuracies. The real experimental uncertainty of the measured solubilities is probably not normally distributed as a result of the binning approach used. However, considering the large number of compounds is analyzed, the overall distribution is likely to converge towards a normal distribution. Further, this approximation should

suffice to illustrate the extent to which the measurements reproduce the real kinetic solubilities.

Predicted → **Measured ↓**	**Insoluble**	**Moderately soluble**	**Soluble**
90% probability threshold (*N* = 132)			
Insoluble	97 (27%)	0	0
Moderately soluble	7 (2%)	0	4 (2%)
Soluble	0	0	24 (13%)
80% probability threshold (*N* = 277)			
Insoluble	158 (44%)	3 (1%)	0
Moderately soluble	16 (4%)	16 (5%)	10 (5%)
Soluble	0	4 (1%)	70 (37%)
70% probability threshold (*N* = 425)			
Insoluble	198 (55%)	13 (4%)	0
Moderately soluble	29 (8%)	65 (22%)	19 (10%)
Soluble	0	13 (4%)	88 (46%)
60% probability threshold (*N* = 627)			
Insoluble	236 (66%)	28 (10%)	2 (1%)
Moderately soluble	46 (13%)	149 (51%)	25 (13%)
Soluble	1	29 (10%)	113 (59%)

Table S3: Cumulative confusion matrices obtained using probability thresholds between 60% and 90% for solubility assignment. The numbers in brackets give the fraction of the compounds from the total number of insoluble/ moderately soluble/ soluble compounds.

Class assigned via probability → Measured ↓	**Insoluble**	**Moderately soluble**	**Soluble**
90% - 100% probability (*N* = 132), Overall accuracy = 92 %			
Insoluble	97	0	0
Moderately soluble	7	0	4
Soluble	0	0	24
80% - 90% probability (*N* = 145), Overall accuracy = 85 %			
Insoluble	61	3	0
Moderately soluble	9	16	6
Soluble	0	4	46
70% - 80% probability (*N* = 148), Overall accuracy = 72 %			
Insoluble	40	10	0
Moderately soluble	13	49	9
Soluble	0	9	18
60% - 70% probability (*N* = 204), Overall accuracy = 72 %			
Insoluble	38	15	2
Moderately soluble	17	84	6
Soluble	1	16	25
50% - 60% probability (*N* = 179), Overall accuracy = 55 %			
Insoluble	31	24	1
Moderately soluble	21	48	17
Soluble	0	17	20
40% - 50% probability (*N* = 165), Overall accuracy = 47 %			
Insoluble	21	39	1
Moderately soluble	14	39	12
Soluble	0	22	17
30% - 40% probability (*N* = 185), Overall accuracy = 34 %			
Insoluble	16	39	2
Moderately soluble	35	31	19
Soluble	1	27	15

Table S4 continued on next page

Table S4 continued

Class assigned via probability → Measured ↓	**Insoluble**	**Moderately soluble**	**Soluble**
20% - 30% probability (*N* = 173), Overall accuracy = 23 %			
Insoluble	11	45	2
Moderately soluble	32	22	32
Soluble	2	20	7
10% - 20% probability (*N* = 292), Overall accuracy = 14 %			
Insoluble	11	63	9
Moderately soluble	60	17	59
Soluble	12	49	12
0% - 10% probability (*N* = 903), Overall accuracy = 3 %			
Insoluble	9	97	318
Moderately soluble	109	11	153
Soluble	174	26	6

Table S4: Confusion matrices obtained for 10% probability ranges. The overall accuracy for the corresponding bin is given above each confusion matrix for comparison to the predicted accuracy. The total number of compounds is three times the number of compounds in the dataset because each compound occurs once in each predicted class, if the full range of CMPs is analyzed.

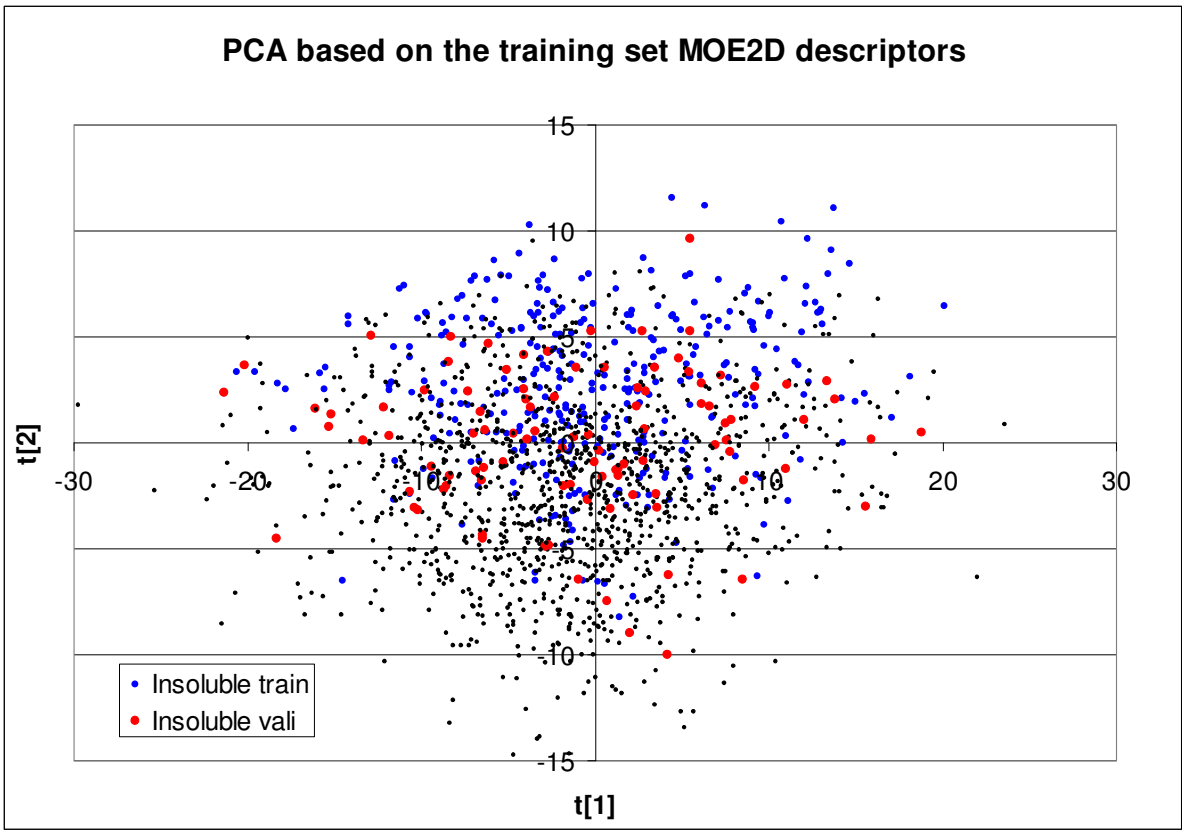

Figure S2: Principal component analysis plot of the training set and the validation set