


## Article

# DMSO Solubility Assessment for Fragment-Based Screening

Shamkhal Baybekov <sup>1</sup>, Gilles Marcou <sup>1</sup>, Pascal Ramos <sup>2</sup>, Olivier Saurel <sup>2</sup>, Jean-Luc Galzi <sup>3,4</sup>  
and Alexandre Varnek <sup>1,\*</sup>

- <sup>1</sup> Laboratoire de Chémoinformatique UMR 7140 CNRS, Institut Le Bel, University of Strasbourg, 4 Rue Blaise Pascal, 67081 Strasbourg, France; sbaybekov@unistra.fr (S.B.); g.marcou@unistra.fr (G.M.)  
<sup>2</sup> Institut de Pharmacologie et de Biologie Structurale, Université de Toulouse CNRS, UPS, 205 Route de Narbonne, 31077 Toulouse, France; pascal.ramos@ipbs.fr (P.R.); olivier.saurel@ipbs.fr (O.S.)  
<sup>3</sup> Biotechnologie et Signalisation Cellulaire UMR 7242 CNRS, École Supérieure de Biotechnologie de Strasbourg, University of Strasbourg, 300 Boulevard Sébastien Brant, 67412 Illkirch, France; galzi@unistra.fr  
<sup>4</sup> ChemBioFrance—Chimiothèque Nationale UAR3035, 8 Rue de L'école Normale, CEDEX 05, 34296 Montpellier, France  
\* Correspondence: varnek@unistra.fr

**Abstract:** In this paper, we report comprehensive experimental and cheminformatics analyses of the solubility of small organic molecules (“fragments”) in dimethyl sulfoxide (DMSO) in the context of their ability to be tested in screening experiments. Here, DMSO solubility of 939 fragments has been measured experimentally using an NMR technique. A Support Vector Classification model was built on the obtained data using the ISIDA fragment descriptors. The analysis revealed 34 outliers: experimental issues were retrospectively identified for 28 of them. The updated model performs well in 5-fold cross-validation (balanced accuracy = 0.78). The datasets are available on the Zenodo platform (DOI:10.5281/zenodo.4767511) and the model is available on the website of the Laboratory of Cheminformatics.



**Citation:** Baybekov, S.; Marcou, G.; Ramos, P.; Saurel, O.; Galzi, J.-L.; Varnek, A. DMSO Solubility Assessment for Fragment-Based Screening. *Molecules* **2021**, *26*, 3950. <https://doi.org/10.3390/molecules26133950>

Academic Editor: Martin Vogt

Received: 4 June 2021  
Accepted: 23 June 2021  
Published: 28 June 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** DMSO solubility; QSPR; fragment-based screening; outlier detection; NMR

## 1. Introduction

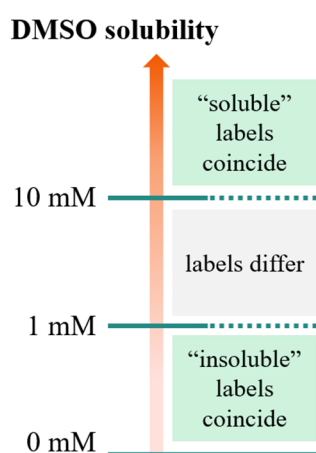
Screening methods have become indisputably an integral part of the drug discovery process [1,2], from hit identification to the evaluation of pharmacological properties. Over the past decades fragment-based screening (FBS) has gained a broad acceptance as an efficient alternative to the conventional high-throughput screening (HTS) [2,3]. This is related to the core idea of FBS, which involves analysis of relatively small libraries containing simple yet diverse organic scaffolds, or fragments, and the identification of hit fragments, that will be developed into more potent lead compounds. Among the basic requirements for fragment-like compounds, well covered by the “rule of three” guidelines [4,5], solubility issues require serious attention [6,7].

Low solubility directly affects the availability of a compound in solution, which may potentially lead to masking of its actual activity. This is notably important for compounds in FBS libraries since the typical concentration of samples is around 1 mM [8–10]. Such a relatively high concentration is related to the low binding affinity of fragments, usually found in the range of  $\mu\text{M}$ – $\text{mM}$  [11]. The assessment of weak ligand–target interactions, requires highly sensitive techniques such as NMR spectroscopy, etc. One of the solvents commonly used in screening methods is dimethyl sulfoxide (DMSO), a well-established standard [12].

Due to the significance of this physicochemical property, the topic of solubility prediction has been and still remains relevant. The challenge of this subject is related to the complexity of the dissolution phenomenon, which is dictated by structural features, solid state, and other physicochemical properties [13]. Very few statistical models designed to predict DMSO solubility have been reported in the literature [12,14], with only one

being publicly available [15]. Thus, Tetko et al. [15] reported a consensus model combining random forest, decision tree and Associative Neural Network individual models, trained on a large and structurally diverse dataset. However, the threshold used for categorizing compounds into “soluble” or “insoluble” classes was set to 10 mM, which is a common concentration of stock solutions.

As illustrated in Figure 1, compounds having a solubility in the range 1–10 mM, are considered soluble according to the FBS definition, but insoluble according to the stock solution definition. This means that the application of the “stock solutions” model by Tetko et al. [15] may lead to discarding compounds predicted as insoluble, but potentially suitable for FBS.



**Figure 1.** Solubility domains defined by the thresholds defined for stock solutions (10 mM) and FBS (1 mM). For these two threshold definitions, the “soluble”/“insoluble” labels coincide for solubility values larger than 10 mM and smaller than 1 mM, respectively. However, in the range 1–10 mM, molecules are considered soluble according to the FBS definition, but insoluble according to the stock solution definition.

This motivated us to develop a classification model predicting fragment solubility in DMSO with a categorical threshold of 1 mM. The model was built on the experimental data provided by the “Plateforme Intégrée de Criblage de Toulouse” (PICT) screening platform. During the training stage, a set of erroneous measurements were identified and removed from the PICT set. The clean dataset was then used for building SVM models. With the help of a Generative Topographic Mapping (GTM) method, the PICT dataset was compared with fragment-like compounds from the Enamine database used for the preparation of the Tetko et al. [15] “stock solutions” model. This analysis revealed some structural motifs present uniquely in PICT. The datasets collected in this work are publicly available on the Zenodo platform (DOI:10.5281/zenodo.4767511). The consensus model is freely accessible on the website of the Laboratory of Chemoinformatics (<http://infochim.u-strasbg.fr/cgi-bin/predictor2.cgi> (accessed on 16 May 2021)).

## 2. Data

### 2.1. Experimental Protocol

In order to design a fragment library for NMR-based FBS, the stock solutions of 939 fragments were prepared at a final concentration of 100 mM in DMSO-d<sub>6</sub>, as described hereafter. The compounds, provided as powder, were dissolved at room temperature in DMSO-d<sub>6</sub> under vigorous shaking until solubilization. Solutions were kept overnight at room temperature, then stored at −20 °C for months. The former solutions were then used for the preparation of a set of diluted solutions with a targeted concentration of 1 mM in DMSO-d<sub>6</sub>, to check by <sup>1</sup>H NMR for each fragment the chemical structure conformity and the solubility. Stock solutions at 100 mM were thawed and kept overnight at room temperature before dilution and running the NMR analysis. NMR experiments were performed

on a Bruker Avance III HD 600 MHz spectrometer ( $^1\text{H}$  Larmor frequency) equipped with a cryoprobe. NMR experiments were performed with a  $30^\circ$  flip angle  $^1\text{H}$  pulse and 1.36 s of acquisition time (with a 20 ppm spectral width and a time domain 32 K complex of data points), and for each sample 32 scans were recorded with a repetition time delay of 5 s. NMR experiments were performed at 298 K and at atmospheric pressure. Quantification was performed with TopSpin, v. 3.5; Bruker Biospin software, by integration of the NMR peaks using the ERETIC2 [16] (Electronic Reference to access in vivo Concentrations) software based on the PULCON method [17]; an internal standard method which correlates the absolute intensities of spectra of compounds to be quantified with a reference spectrum. The reference spectrum was acquired as described above from a 1 mM isoleucine solution in DMSO- $d_6$ . The experimental error of solubility determination was estimated as 50  $\mu\text{M}$ .

## 2.2. Data Description

The PICT dataset contained structures of 939 compounds with their corresponding DMSO concentration values ranging from 0 to 1000  $\mu\text{M}$ . Since the expected concentration for DMSO samples was 1 mM, a threshold for making a division between soluble and insoluble categories was set to 1000  $\mu\text{M}$ . Therefore, if concentration values were equal to 1000  $\mu\text{M}$  it would be classified as soluble, and insoluble if the value was below the given threshold. Experimental error on the concentration was estimated at 50  $\mu\text{M}$ ; therefore, it was decided to remove a segment of the dataset in the range 900–999  $\mu\text{M}$ , as in this range the soluble/insoluble label is ambiguous. After the removal of data points with missing solubility values and the aforementioned “gray area” zone, the number of compounds in the training set was reduced to 822, where 686 and 136 compounds belonged to “soluble” and “insoluble” classes, respectively. The key physicochemical parameters varied across the PICT set in the following ranges: calculated logP  $-3.8 - +3.94$ , molecular weight 150–302 Da, the number of hydrogen bond acceptors 0–6, and the number hydrogen bond donors 0–3.

## 2.3. Data Curation

The chemical structures were standardized using a ChemAxon Standardizer [18]. Applied rules included the removal of solvents, ions, explicitly indicated hydrogen atoms, neutralization, and aromatization. All stereo labels were skipped. A detailed description of the standardization protocol is provided in Supporting Information (“Standardization protocol” section). Erroneous measurements were then detected with the help of the outlier identification procedure (see below).

## 2.4. Filtered Enamine Data

A subset of the fragment-like compounds was extracted from the Enamine dataset used for training of the Tetko et al. model [15] with the help of a filter, matching the same ranges of variation as the PICT dataset for ClogP, molecular weight, number of H-donors and H-acceptors. The filtering resulted in the selection of 8314 fragment-like compounds out of the initial set of 50,620 compounds.

# 3. Method

## 3.1. Molecular Descriptors

ISIDA substructural molecular fragments (SMF) [19] were used in this study. SMF descriptors are derived solely from hydrogen suppressed 2D chemical graphs. They represent fragments of different topologies (sequences of atoms and bonds, sequences of atoms only, atom-centered fragments, triplets) and size (see Table S1 in Supporting Information). The minimal length of fragments varied between 2 and 3, whereas the maximal length varied between 2 and 8. Encoding of a given sequence by its terminal atoms (“atom pairs”) was also considered. A fragment occurrence is a descriptor value. Variation of the descriptors topology; type of sequence (explicit atoms or atom pairs and

size) led to the generation of the pool of 182 subsets of descriptors. ISIDA descriptors were used in numerous QSAR studies [20–22].

### 3.2. Machine Learning Method

Classification models were built using the Support Vector Machine (SVM) machine learning (ML) algorithm. It was used for the selection of optimal descriptor sets, outlier identification and the generation of predictive models. The Libsvm 3.24 package [23] was used for the generation of linear SVM models. The Golden section search method was used in order to find the optimal cost parameter ranging from 0.01 to 1000 with a stopping criterion of 0.1. Optimization was performed to maximize 5-fold cross-validation (5-CV) balanced accuracy (BA).

### 3.3. Modeling Workflow

The modeling workflow consisted of three main stages: (1) detection of erroneous measurements, (2) selection of relevant descriptor spaces and (3) model building and implementation (Figure 2). Detection of erroneous measurements was performed following a protocol from Ruggiu et al. [24] adapted in this study to classification tasks. This approach suggests the preparation of several individual models and the identification of the common badly predicted instances. For the curated PICT dataset, 26 various fragment descriptor spaces were generated. Each subset of descriptors was used for the modeling. Five models providing the best performance in 5-fold cross-validation were selected. At the next step, common false positives and false negatives (“outliers”) detected by all selected models at the training stage were identified and inspected by the experimental team. A vast majority of them were associated with technical problems and discarded from the dataset (see “Results and discussion” section). The resulting “clean” dataset was used in a new round of model building and validation.

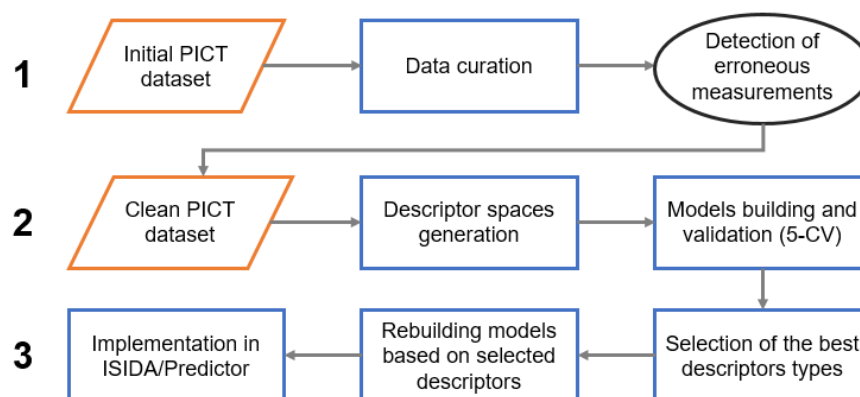


Figure 2. The modeling workflow.

At the next stage, 182 descriptor spaces were generated and used for the building and validation of SVM models. Models performing with  $BA \geq 0.75$  in 5-CV were selected; the highest  $BA = 0.80$  was achieved for the model based on the atom centered fragments connecting atoms pairs derived for the sequences of atoms and bonds of 3–4 atoms length (type “IIAB(3-4)\_R-P”, see Table S1 in Supporting Information). Descriptors involved in the selected models were then used to develop classification models on the entire “clean” PICT dataset. Obtained in such a way, 45 individual models formed a consensus model integrated into the ISIDA Predictor tool [25]. For any new molecule, the tool assigns a solubility label according to the majority of votes for the individual models. The predictive performance of the consensus model is reasonable ( $BA = 0.78$  in 5-CV). Notice that the ISIDA Predictor accounts for the fragment control [26] applicability domain (AD) of each individual model. If a new molecule is outside of the AD, the model is not applied. Along

with the predicted label, the tool provides a confidence estimation based on the ratio of the percentage models and prediction consistency.

The consensus model is freely available on the website of the Laboratory of Chemoinformatics (<http://infochim.u-strasbg.fr/cgi-bin/predictor2.cgi> (accessed on 16 May 2021)). In order to access the model, select the “PhysProp” option in the “general kind of property” section and then choose “Solubility DMSO” option from the “property to model” drop-down list. A user is invited to draw a molecule of interest or upload an SD file containing several compounds. Some screenshots illustrating the functioning of the ISIDA Predictor are given in the Supporting Information (Figure S5).

### 3.4. Generative Topographic Mapping

Generative Topographic Mapping (GTM) [27–30] is a dimensionality reduction method, which transforms a high-dimensional molecular descriptor space into a 2D latent space (“map”). This is achieved by introducing a 2D manifold into the high-dimensional space and adjusting a normal probability density, centered on the nodes of a rectangular grid superposed with the manifold, to the observed data distribution. Once the manifold is fitted, the compounds are projected on this 2D surface. GTM is widely used for the chemical space visualization, analysis, and compounds’ profiling [31].

Two maps were constructed: (i) for the PICT dataset and (ii) for the merged PICT and Enamine datasets. The method hyperparameters and type of fragment descriptors were optimized by maximizing the classes separation (“soluble/insoluble” for the PICT dataset and “PICT/Enamine” for the merged dataset). The compounds were encoded by atom centered fragments, including a given atom and atoms and bonds of its either 3 or 5 coordination spheres for the merged dataset and the PICT dataset, respectively. The data distribution was visualized using “class landscapes” [30], highlighting areas populated by soluble and insoluble compounds.

## 4. Results and Discussion

### 4.1. Data Visualization and Analysis

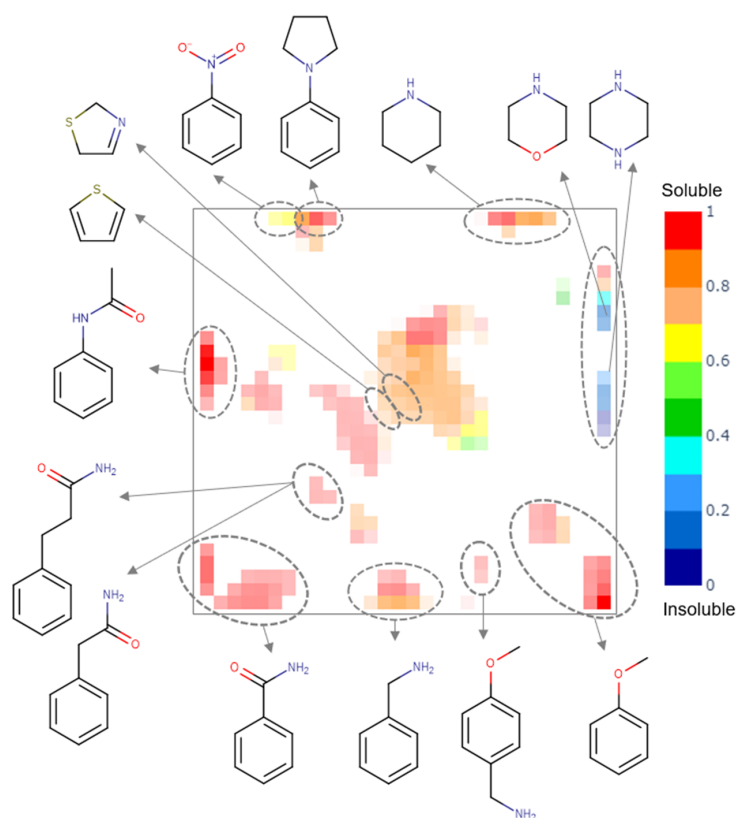
A generative topographic map built for the PICT dataset shows several clusters populated by compounds of a particular chemotype (see Figure 3). Insoluble compounds bear piperazine and morpholine fragments, soluble compounds are mostly aromatic amines, amides, piperidines and ethers, whereas compounds bearing nitro-benzene, thiophene and dihydro-thiazole fragments can be either soluble or insoluble.

A comparative analysis of the PICT and filtered Enamine datasets was performed using a generative topographic map combining both datasets. Figure 4 shows a class landscape in which the color code characterizes the presence of Enamine or PICT compounds in a particular zone of the chemical space. The map well separates blue and red zones populated by Enamine and PICT compounds, respectively, which confirms the structural diversity of the two datasets. Detailed analysis of the red zones, reveal some particular structural motifs present in the PICT and absent in the Enamine dataset (Figure 4).

### 4.2. Erroneous Measurements Detection

As explained above, the outliers are compounds in which the predicted labels systematically do not match the experimental ones for none of the initially developed models. There are 34 outliers which belong to three categories: experimental errors, chemical instability, and unexplained discrepancies. The list includes 31 insoluble compounds predicted as soluble and three soluble molecules predicted as insoluble (see Table S3 in Supporting Information). These modeling results were reported to the PICT team for the reassessment of experimental values. The analysis showed that 15 out of 34 potential outliers resulted from a human error during the sample preparation. Overall, during the revision of the NMR spectra, nine compounds were found to have degradation signs, whereas the values of 19 samples were likely affected by experimental errors. These 28 confirmed outliers were discarded. The remaining six compounds were claimed to have no experimental issues.

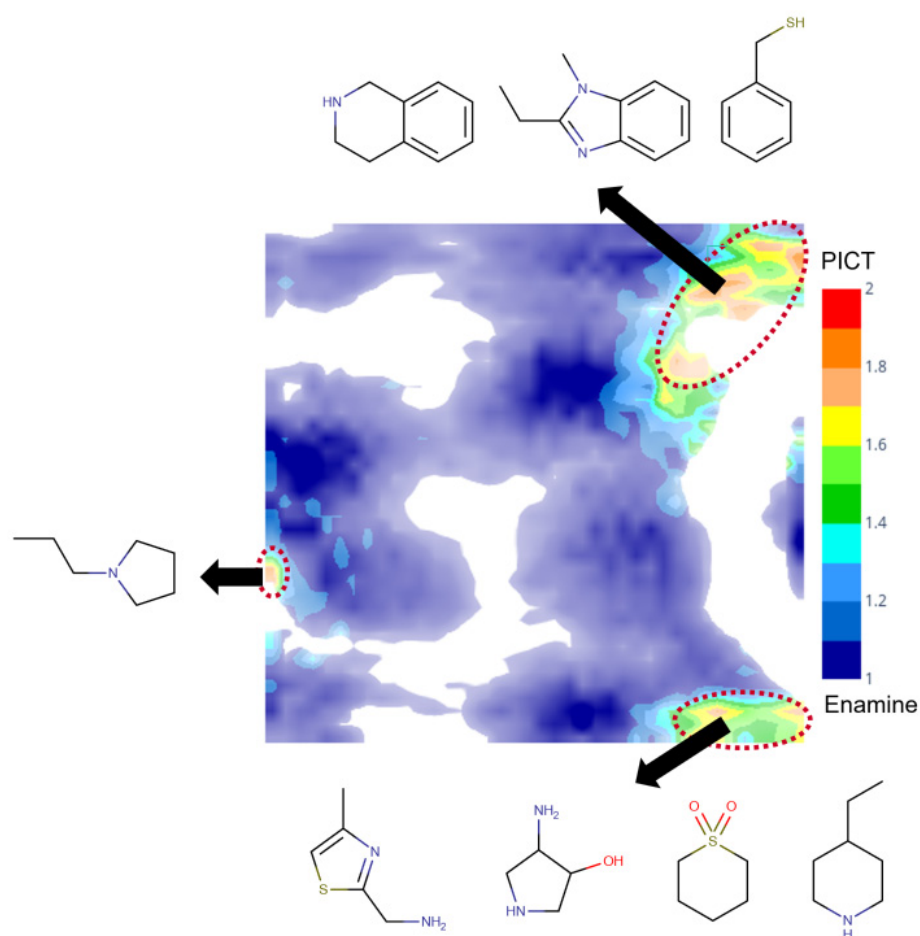
Some incorrectly predicted compounds and their correctly predicted close analogues form some sort of “solubility cliffs” (Table 1). Thus, compounds **1a** and **1b** differ by a methylene bridge between two cyclic fragments; the difference between compounds **2a** and **2b** results from the type of substituent (OH or CH<sub>2</sub>-OH) and its position in the piperidine ring, whereas compound **3b** has two methyl groups more than the compound **3a**. These cliffs are intriguing and require further structure-activity relationship (SAR) exploration, which is beyond the scope of this work.



**Figure 3.** The class landscape for the PICT dataset. Blue and red zones are populated by insoluble and soluble molecules, respectively. Green and yellow zones contain a mixture of soluble and insoluble compounds.

**Table 1.** Example of incorrectly predicted compounds and their correctly predicted close analogues.

Incorrectly Predicted Compounds				Correctly Predicted Similar Compounds			
#	Compound structure	Exp	Pred	#	Compound structure	Exp	Pred
<b>1a</b>		Soluble	Insoluble	<b>1b</b>		Insoluble	Insoluble
<b>2a</b>		Soluble	Insoluble	<b>2b</b>		Insoluble	Insoluble
<b>3a</b>		Insoluble	Soluble	<b>3b</b>		Soluble	Soluble



**Figure 4.** The class landscape depicting the coverage of a fragment-like chemical space by PICT and Enamine datasets. Blue and red zones are populated, respectively, by Enamine and PICT molecules. Green and yellow zones contain a mixture of compounds from the two datasets.

#### 4.3. “Stock Solutions” vs. FBS Models

For the sake of comparison, the “stock solutions” model by Tetko et al. [15] was applied to the PICT dataset and, vice versa, the FBS model was applied to the Enamine dataset. Only 87.4% of the Enamine data were found inside the applicability domain of at least one FBS individual model. On the other hand, 98.6% of the PICT dataset was covered by the AD of the “stock solution” model.

Results given in Table 2 show that both models predicted soluble compounds with a high accuracy, but failed to predict insoluble ones. The latter is not surprising when the FBS model is applied to the Enamine dataset: since solubility assignment thresholds of FBS and stock solution models differ, the compounds with a solubility in the range 1–10 mM are considered soluble according to FBS and insoluble according to stock solution models. On the other hand, the compounds in which the solubility value is smaller than 1 mM are considered insoluble according to both models. This could be explained by the fact that the PICT dataset contains some unique structural motifs, e.g., thiazole, benzimidazole or tetrahydroisoquinoline (see Figure 4). It also looks like these models (at least, the “stock solution” one) are biased toward the training set composition containing mostly soluble compounds.

**Table 2.** Predictive performance of the FBS model on the filtered Enamine data, and of the “stock solution” model on the PICT data. The number of correctly predicted compounds with respect to the total number of compounds is given between the parentheses.

	FBS Model on Enamine Dataset	«Stock Solution» Model on PICT Dataset
Recall (soluble)	0.954 (6828/7156)	1 (676/676)
Recall (insoluble)	0.052 (6/115)	0.01 (1/101)

## 5. Conclusions

This work combines experimental and chemoinformatics studies of the solubility of small molecules (“fragments”) in DMSO in the context of their application in fragment-based screening. Experimentally measured data (PICT dataset) were used for the development of the first classification model for DMSO solubility fragments (FBS model). Unlike the earlier reported “stock solution” model with the categorical threshold “soluble/insoluble” of 10 mM, our model uses a more suitable threshold for fragments of 1 mM. The model displays a reasonable predictive performance in 5-fold cross-validation (BA = 0.78). Both the experimentally measured data and developed model are freely available for users.

We have demonstrated that the developed model can efficiently be used to detect erroneously measured data. Among the 28 picked compounds pointed to by the model, nine compounds were found to have degradation signs, whereas the values of 19 samples were likely affected by experimental errors.

The comparison of the PICT and Enamine datasets performed with the help of a Generative Topographic Mapping approach showed that the PICT dataset contains some unique structural motifs absent in the Enamine collection.

The results reported here demonstrate a synergism between experimental and chemoinformatics teams for obtaining, analyzing and modeling of the DMSO solubility of small molecules (“fragments”) in the context of their application in fragment-based screening.

**Supplementary Materials:** The following are available online: description of standardization rules, description of ISIDA fragment descriptors, description of statistical metrics, a list of models constituting the FBS consensus model, description of GTM parameters of class landscapes, a summary of predictions made on the “gray area” compounds, the outlier detection and removal workflow, a list of outliers, a list of reported classification models for the prediction of DMSO solubility and the screenshots showing the usage of the “Predictor” web-application containing our model, the PICT dataset containing experimental solubility values and class labels and the filtered Enamine dataset.

**Author Contributions:** Conceptualization, S.B., G.M. and J.-L.G.; methodology, S.B. and G.M.; software, S.B. and G.M.; validation, P.R. and O.S.; formal analysis, P.R. and O.S.; investigation, P.R. and O.S.; resources, P.R. and O.S.; data curation, S.B. and G.M.; writing—original draft preparation, S.B.; writing—review and editing, S.B., G.M., P.R., O.S., J.-L.G. and A.V.; supervision, G.M. and A.V.; project administration, G.M. and A.V.; funding acquisition, J.-L.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** The fragment library and the Bruker Avance III HD 600 MHz NMR spectrometer of the Integrated Screening Platform of Toulouse (PICT) were funded by CNRS, Université Paul Sabatier, Infrastructures en Biologie Santé et Agronomie European Structural Funds, and the Midi-Pyrénées Region. This work was supported by ChemBioFrance and the Interdisciplinary Thematic Institute ITI-CSC via the IdEx Unistra (ANR-10-IDEX-0002) within the program Investissement d’Avenir. S.B. thanks the CSC Graduate School funded by the French National Research Agency (CSC-IGS ANR-17-EURE-0016).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on the Zenodo platform (DOI:10.5281/zenodo.4767511) and in Supplementary Materials.



**Conflicts of Interest:** The authors declare no conflict of interest.

**Sample Availability:** Samples of the compounds are not available.

## References

1. Proudfoot, J.R. High-Throughput Screening and Drug Discovery. In *The Practice of Medicinal Chemistry*; Wermuth, C.G., Ed.; Elsevier: New York, NY, USA, 2008; pp. 144–158, ISBN 978-0-12-374194-3. [CrossRef]
2. Farmer, B.T.; Reitz, A.B. Fragment-Based Drug Discovery. In *The Practice of Medicinal Chemistry*; Wermuth, C.G., Ed.; Elsevier: New York, NY, USA, 2008; pp. 228–243, ISBN 978-0-12-374194-3. [CrossRef]
3. Kirsch, P.; Hartman, A.M.; Hirsch, A.K.H.; Empting, M. Concepts and Core Principles of Fragment-Based Drug Design. *Molecules* **2019**, *24*, 4309. [CrossRef]
4. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A ‘Rule of Three’ for fragment-based lead discovery? *Drug Discov. Today* **2003**, *8*, 876–877. [CrossRef]
5. Jhoti, H.; Williams, G.; Rees, D.C.; Murray, C.W. The “rule of three” for fragment-based drug discovery: Where are we now? *Nat. Rev. Drug Discov.* **2013**, *12*, 644. [CrossRef]
6. Siegal, G.; Eiso, A.B.; Schultz, J. Integration of fragment screening and library design. *Drug Discov. Today* **2007**, *12*, 1032–1039. [CrossRef]
7. Lepre, C.A. Library design for NMR-based screening. *Drug Discov. Today* **2001**, *6*, 133–140. [CrossRef]
8. Leach, A.R.; Hann, M.M.; Burrows, J.N.; Griffen, E.J. Fragment screening: An introduction. *Mol. Biosyst.* **2006**, *2*, 429. [CrossRef] [PubMed]
9. Barker, J.; Hestekamp, T.; Whittaker, M. Integrating HTS and fragment-based drug discovery. *Drug Discov. World* **2008**, *9*, 69–75.
10. Lau, W.F.; Withka, J.M.; Hepworth, D.; Magee, T.V.; Du, Y.J.; Bakken, G.A.; Miller, M.D.; Hendsch, Z.S.; Thanabal, V.; Kolodziej, S.A.; et al. Design of a multi-purpose fragment screening library using molecular complexity and orthogonal diversity metrics. *J. Comput. Aided. Mol. Des.* **2011**, *25*, 621–636. [CrossRef] [PubMed]
11. Murray, C.W.; Rees, D.C. The rise of fragment-based drug discovery. *Nat. Chem.* **2009**, *1*, 187–192. [CrossRef]
12. Balakin, K.V.; Ivanenkov, Y.A.; Skorenko, A.V.; Nikolsky, Y.V.; Savchuk, N.P.; Ivashchenko, A.A. In Silico Estimation of DMSO Solubility of Organic Compounds for Bioscreening. *J. Biomol. Screen.* **2004**, *9*, 22–31. [CrossRef]
13. Alsenz, J.; Kansy, M. High throughput solubility measurement in drug discovery and development. *Adv. Drug Deliv. Rev.* **2007**, *59*, 546–567. [CrossRef] [PubMed]
14. Balakin, K.; Savchuk, N.; Tetko, I. In Silico Approaches to Prediction of Aqueous and DMSO Solubility of Drug-Like Compounds: Trends, Problems and Solutions. *Curr. Med. Chem.* **2006**, *13*, 223–241. [CrossRef] [PubMed]
15. Tetko, I.V.; Novotarskyi, S.; Sushko, I.; Ivanov, V.; Petrenko, A.E.; Dieden, R.; Lebon, F.; Mathieu, B. Development of Dimethyl Sulfoxide Solubility Models Using 163 000 Molecules: Using a Domain Applicability Metric to Select More Reliable Predictions. *J. Chem. Inf. Model.* **2013**, *53*, 1990–2000. [CrossRef] [PubMed]
16. Bharti, S.K.; Roy, R. Quantitative <sup>1</sup>H NMR spectroscopy. *Trends Anal. Chem.* **2012**, *35*, 5–26. [CrossRef]
17. Wider, G.; Dreier, L. Measuring protein concentrations by NMR spectroscopy. *J. Am. Chem. Soc.* **2006**, *128*, 2571–2576. [CrossRef] [PubMed]
18. ChemAxon Standardizer. Available online: <http://www.chemaxon.com/> (accessed on 16 May 2021).
19. Varnek, A.; Fourches, D.; Hoonakker, F.; Solov’ev, V.P. Substructural fragments: An universal language to encode reactions, molecular and supramolecular structures. *J. Comput. Aided Mol. Des.* **2005**, *19*, 693–703. [CrossRef]
20. Ruggiu, F.; Solov’ev, V.; Marcou, G.; Horvath, D.; Graton, J.; Le Questel, J.Y.; Varnek, A. Individual hydrogen-bond strength QSPR modelling with ISIDA local descriptors: A step towards polyfunctional molecules. *Mol. Inform.* **2014**, *33*, 477–487. [CrossRef]
21. Glavatskikh, M.; Madzhidov, T.; Solov’ev, V.; Marcou, G.; Horvath, D.; Graton, J.; Le Questel, J.Y.; Varnek, A. Predictive Models for Halogen-bond Basicity of Binding Sites of Polyfunctional Molecules. *Mol. Inform.* **2016**, *35*, 70–80. [CrossRef]
22. Varnek, A.; Fourches, D.; Solov’ev, V.P.; Baulin, V.E.; Turanov, A.N.; Karandashev, V.K.; Fara, D.; Katritzky, A.R. “In silico” design of new uranyl extractants based on phosphoryl-containing podands: QSPR studies, generation and screening of virtual combinatorial library, and experimental tests. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1365–1382. [CrossRef]
23. Chang, C.-C.; Lin, C.-J. {LIBSVM}: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
24. Ruggiu, F.; Gizzi, P.; Galzi, J.L.; Hibert, M.; Haiech, J.; Baskin, I.; Horvath, D.; Marcou, G.; Varnek, A. Quantitative structure-property relationship modeling: A valuable support in high-throughput screening quality control. *Anal. Chem.* **2014**, *86*, 2510–2520. [CrossRef]
25. HOME—Chemoinformatics Laboratory. Available online: <http://infochim.u-strasbg.fr/> (accessed on 16 May 2021).
26. Horvath, D.; Marcou, G.; Varnek, A. A unified approach to the applicability domain problem of QSAR models. *J. Cheminform.* **2010**, *2*, O6. [CrossRef]
27. Kireeva, N.; Baskin, I.I.; Gaspar, H.A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal tool for data visualization, structure-activity modeling and dataset comparison. *Mol. Inform.* **2012**, *31*, 301–312. [CrossRef]
28. Bishop, C.M.; Svensén, M.; Williams, C.K.I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10*, 215–234. [CrossRef]
29. Gaspar, H.A.; Baskin, I.I.; Marcou, G.; Horvath, D.; Varnek, A. Chemical data visualization and analysis with incremental generative topographic mapping: Big data challenge. *J. Chem. Inf. Model.* **2015**, *55*, 84–94. [CrossRef]

- 
30. Horvath, D.; Baskin, I.; Marcou, G.; Varnek, A. Generative Topographic Mapping of Conformational Space. *Mol. Inform.* **2017**, *36*, 1700036. [[CrossRef](#)] [[PubMed](#)]
  31. Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D. Mappability of drug-like space: Towards a polypharmacologically competent map of drug-relevant compounds. *J. Comput. Aided. Mol. Des.* **2015**, *29*, 1087–1108. [[CrossRef](#)] [[PubMed](#)]