Journal of Computer-Aided Molecular Design (2005) 19: 731–738
DOI 10.1007/s10822-005-9026-y

Possibilities for transfer of relevant data without revealing structural information

Omoshile O. Clement[a,b,*] & Osman F. Güner[a,†]
[a]*Accelrys Inc., 10188 Telesis Court, San Diego, CA, 92121, USA;* [b]*Present address: Bio-Rad Laboratories, Informatics Division, 3316 Spring Garden St, Philadelphia, PA, 19104-2596, USA*

Received 25 May 2005; accepted 20 October 2005

Key words: ADME/Tox, *in silico*, QSAR

Summary

In this paper, we discuss how we safely exchanged proprietary data between third parties in the early years of predictive ADME/Tox model development. At that time, industry scientists wanted to evaluate predictive models, but were not willing to share their structures with software vendors. At the same time, model developers were willing to run the scientists' structures through their models, but they were not willing to reveal which descriptors were important for a particular predictive model. We developed a process where scientists could perform calculations on a broad number of commercially available public descriptors and forward results as a property file, instead of their structures. Meanwhile, the model developer could extract descriptors used in the predictive model, run the model, and pass results back to the scientist. On the following pages, we discuss the pros and cons of this approach, and we address questions such as: Can structural information that is proprietary be compromised from descriptors in ADME/Tox models? Can ADME/Tox predictions be made purely from descriptors, without the explicit knowledge of chemical structures, proprietary or otherwise?

Introduction

In an article published in 2003 [1], Stephen A. Hill, CEO of ArQule Inc., based in Cambridge, MA, stated that the productivity of drug discovery is deteriorating. "If you look at the total number of dollars put into drug discovery and look at what comes out at the end, it is not a pretty picture. There is a real problem there for the future of the industry in terms of making drug discovery productive."

To address this problem, a paradigm shift in the drug discovery process is occurring where focus is placed on *early* attrition of poor drug candidates. Driving this paradigm shift is the movement towards *in silico* and high-throughput *in vitro* approaches. Increasingly, computational methods are being used to predict the absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox) properties of structures *before* synthesis, so as to identify favorable compounds that warrant synthesis and screening.

In order to develop and apply *in silico* technologies as filters for weeding out undesirable drug candidates early in the discovery process, there is a need to have access to 'reliable' predictive models.

*To whom correspondence should be addressed. E-mail: omoshile_clement@bio-rad.com

†No longer at this address. Can be reached at osman@san.rr.com

The quality of the predictive models depends on the quality of the data with which they are built, and a vast number of proprietary biological, chemical, pharmacokinetic, and pharmacodynamics data which could help this task, resides within the pharmaceutical industry and are generally unavailable to model developers. This is a difficult concept to bring to fruition because these data constitute the bulk of the intellectual properties inherent within the industry.

Although a large database of chemical and biological data resides within the pharmaceutical industry, and constitute important intellectual property for these companies, these data may not be diverse enough to develop over-arching predictive *in silico* models that produce high confidence predictions across a range of compounds covering diverse structural families. Companies, themselves use their own data to build models, but such models are restricted to the limited chemistry representation in the company's database (i.e., local models). The predictive value of these local models is limited as compounds that are outside the coverage of the model domain are not predicted reliably. In order to ensure broader chemistry coverage (i.e., global models) such data from a number of different companies will be needed. A likely solution was proposed to address this problem – the creation of a consortium of pharmaceutical companies whose pooled data could be used to develop predictive ADME/Tox models that enable the identification and removal of potentially undesirable compounds early in the search for new drug candidates.

In the early part of 2000, Pharmacopeia Labs based in Princeton, NJ, along with its computational chemistry software subsidiary, Accelrys Inc., embarked on the creation of an ADME Consortium to facilitate this data sharing and model building process. Since this required companies to provide proprietary data to this consortium, soliciting members was a difficult challenge to surmount. The critical issue was figuring out how to protect the intellectual property (IP) rights of each company's data from their competitors. This was such a huge concern that this consortium did not take off as a result. One of the biggest reasons for its failure was the reluctance of scientists within the pharmaceutical industry to provide proprietary data to a third party (software vendor), especially where there was no credible means of protecting such data from competitors. Unfortunately, this data is imperative for the development of good, robust, and transferable *in silico* ADME/Tox models for use within the industry.

Two key questions that are pertinent to ask are: (1) Is it possible for pharmaceutical companies to provide their proprietary data to the scientific public domain without revealing or compromising their chemical structures? and (2) What level of safety can guarantee the integrity of such data in the public domain? These questions remained a burning issue so much that a symposium was created to debate them at the 229th American Chemical Society (ACS) Meeting in San Diego, CA, jointly organized by two Divisions of the ACS: Chemical Information, and Computers in Chemistry.

In introductory remarks at this workshop [2], panelists Hugo Kubinyi and Doug Livingston suggested that the large amount of data collected over many years within the pharmaceutical industry can be harnessed by computational tools to sieve relevant information to help the drug discovery efforts. However, such data is rarely, if ever, made available to the scientific public domain. Conversely, the industry demands that better and more robust *in silico* tools be developed to help weed out unwanted compounds early in the discovery process, even while many, if not all, are unwilling to share the data necessary to accomplish such a task.

What is needed, according to Tudor Oprea [3, 4], one of the organizers of this ACS workshop, is a cooperative data-exchange venture between software developers and the pharmaceutical industry. To make such a venture successful, researchers must devise computational methods to mask structures of chemical compounds, while still providing molecular modeling experts with enough information about the compound's properties to develop accurate predictive methods for ADME/Tox studies.

Software vendors such as Accelrys, Inc. are commercializing ADME/Tox predictive models [5] using data sourced from literature publications and databases of content providers. These models are often applied to filter large compound libraries to make predictions about aqueous solubility [6], blood–brain barrier penetration, human intestinal absorption [7], didactic liver toxicity [8], inhibition of CYP2D6 enzyme [9], and protein plasma binding (based on 1DSim algorithm [10]). Clearly such models are potentially very useful in the industry, and hence there is a need to evaluate their robustness and coverage of chemical space.

Such evaluations were undertaken in contractual agreements between Accelrys and the industry. The question arises as to how exchange of data was safely carried out in such projects? Using a generic example, we describe here a scenario occasionally faced by software vendors and their user community: there is a need to evaluate software that could give useful predictions about in-house compounds, but such compounds cannot be sent to the software vendor; conversely, the software vendor is unable to provide the *in silico* model without licensing such tools. One approach to tackle this is as described in the workflow shown in Figure 1.

Company-X wants a quantitative analysis of an SAR data. Built into the dataset are in-house (proprietary) set of descriptors that company-X does not want to disclose to an outside party. How would such data be safely exchanged between both parties to accomplish the goal? Accelrys has a script specifically designed for such protection [11]. The script works on a tab-delimited file by renaming all rows containing molecule names and/or structures as “M_1 ... M_n”, and renaming all columns containing descriptor names as “X_1 .... X_n”. With this script available, a scientist at Company-X computes a wide range of descriptors for the SAR data, exports the calculated data and structure names as a tab-delimited file while preserving the column and row names, runs the script to convert row and column names to Ms and Xs, respectively (see Table 1), and provides the numerical ‘M×X’ matrix dataset to the software vendor. A predictive QSAR model can then be generated from such data without the need to identify the structures or descriptors. A representative example of the outcome of such a scenario is provided in Figure 2 below.

The QSAR model generated will inherently be decipherable only to the individual who provided the data, since that person can deconvolute the model into the actual set of descriptors found to correlate with biological activity.

In the following paragraphs, we present another proof-of-concept example. Company-X wants to evaluate a specific ADME/Tox prediction model before making a purchase decision. This company needs to examine if the predictive *in silico* model covers the company’s chemistry space, or if it will need re-training. Here, Company-X does not want to provide their data to the software vendor, and the software vendor does not have a trial license to provide for evaluation. Company-X and the software vendor enter a contractual agreement to deploy the ADME/Tox model against a test chemical library, with the structures protected by converting the data into numerical format.

In this example, the evaluation is to be performed on a small, focused library from a

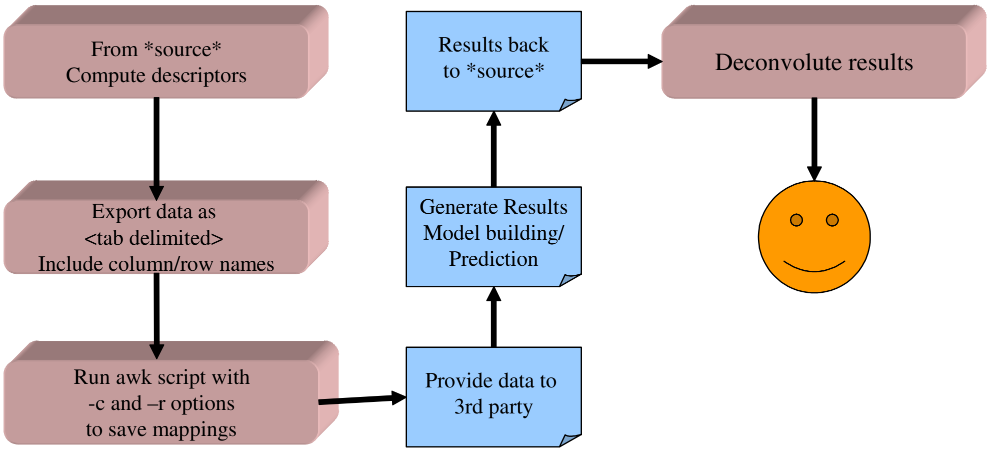
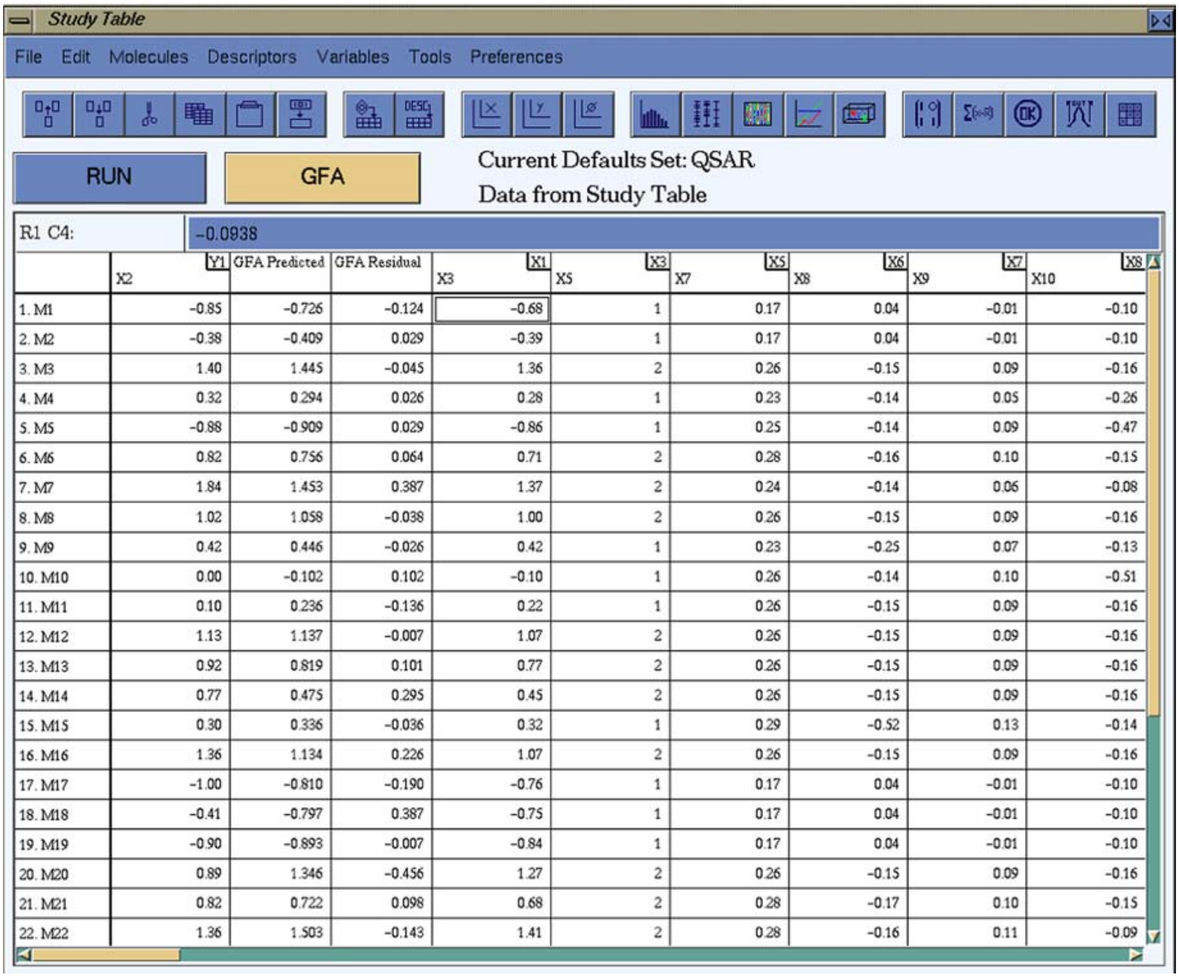


Figure 1. Workflow for a “Proof-of-Concept” scenario where structural and descriptor data are protected during their exchange between two parties.

Table 1. Cerius^2 study table showing rows and columns described as M_1 ... M_n and X_1 ... X_n, respectively.



Current Defaults Set: QSAR
Data from Study Table

R1 C4: -0.0938

	Y1	GFA Predicted	GFA Residual	X1	X3	X5	X6	X7	X8
	X2			X3	X5	X7	X8	X9	X10
1. M1	-0.85	-0.726	-0.124	-0.68	1	0.17	0.04	-0.01	-0.10
2. M2	-0.38	-0.409	0.029	-0.39	1	0.17	0.04	-0.01	-0.10
3. M3	1.40	1.445	-0.045	1.36	2	0.26	-0.15	0.09	-0.16
4. M4	0.32	0.294	0.026	0.28	1	0.23	-0.14	0.05	-0.26
5. M5	-0.88	-0.909	0.029	-0.86	1	0.25	-0.14	0.09	-0.47
6. M6	0.82	0.756	0.064	0.71	2	0.28	-0.16	0.10	-0.15
7. M7	1.84	1.453	0.387	1.37	2	0.24	-0.14	0.06	-0.08
8. M8	1.02	1.058	-0.038	1.00	2	0.26	-0.15	0.09	-0.16
9. M9	0.42	0.446	-0.026	0.42	1	0.23	-0.25	0.07	-0.13
10. M10	0.00	-0.102	0.102	-0.10	1	0.26	-0.14	0.10	-0.51
11. M11	0.10	0.236	-0.136	0.22	1	0.26	-0.15	0.09	-0.16
12. M12	1.13	1.137	-0.007	1.07	2	0.26	-0.15	0.09	-0.16
13. M13	0.92	0.819	0.101	0.77	2	0.26	-0.15	0.09	-0.16
14. M14	0.77	0.475	0.295	0.45	2	0.26	-0.15	0.09	-0.16
15. M15	0.30	0.336	-0.036	0.32	1	0.29	-0.52	0.13	-0.14
16. M16	1.36	1.134	0.226	1.07	2	0.26	-0.15	0.09	-0.16
17. M17	-1.00	-0.810	-0.190	-0.76	1	0.17	0.04	-0.01	-0.10
18. M18	-0.41	-0.797	0.387	-0.75	1	0.17	0.04	-0.01	-0.10
19. M19	-0.90	-0.893	-0.007	-0.84	1	0.17	0.04	-0.01	-0.10
20. M20	0.89	1.346	-0.456	1.27	2	0.26	-0.15	0.09	-0.16
21. M21	0.82	0.722	0.098	0.68	2	0.28	-0.17	0.10	-0.15
22. M22	1.36	1.503	-0.143	1.41	2	0.28	-0.16	0.11	-0.09

hit-to-lead project. To protect the structures, the vendor recommended that Company-X convert the dataset into a numerical set of computed descriptors. Descriptors such as number of rotatable bonds, number of aromatic rings, heteroatom counts, hydrogen bond donor and acceptor counts, and many 3D descriptors that could reveal information on structures should be avoided. The computed descriptors, in the form of numerical data are sent to the software vendor (see Table 2). Using the vendor's ADME/Tox model, a prediction was made (see Figure 3) and sent back to the customer [12].

No attempt to deconvolute the data into structures is required to generate a model prediction. However, one can speculate on how challenging it would be for an individual to deconvolute the numerical data into its structural composition. We propose a workflow that can be attempted (see Figure 4) to determine the structure using a variety of computed descriptor values. This is a brute-force method much like guessing a password by trying all different combinations of alpha-numeric characters. Given an infinite amount of computer time, this may result with a successful guess, even then, only if one is lucky.

The process workflow is as follows: One starts with a random structure and computes the Molecular Weight. Using a smart algorithm, a series of small modifications is made to the structure with the objective to match the target Molecular Weight. This process continues to iterate until

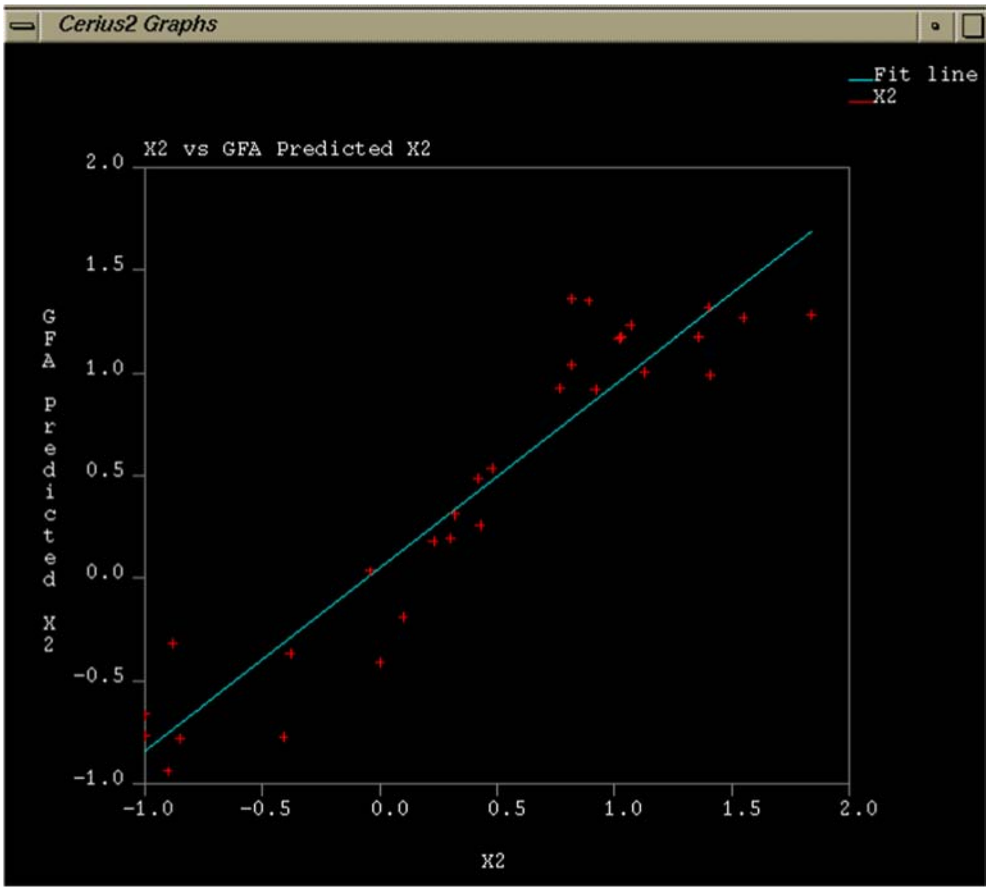
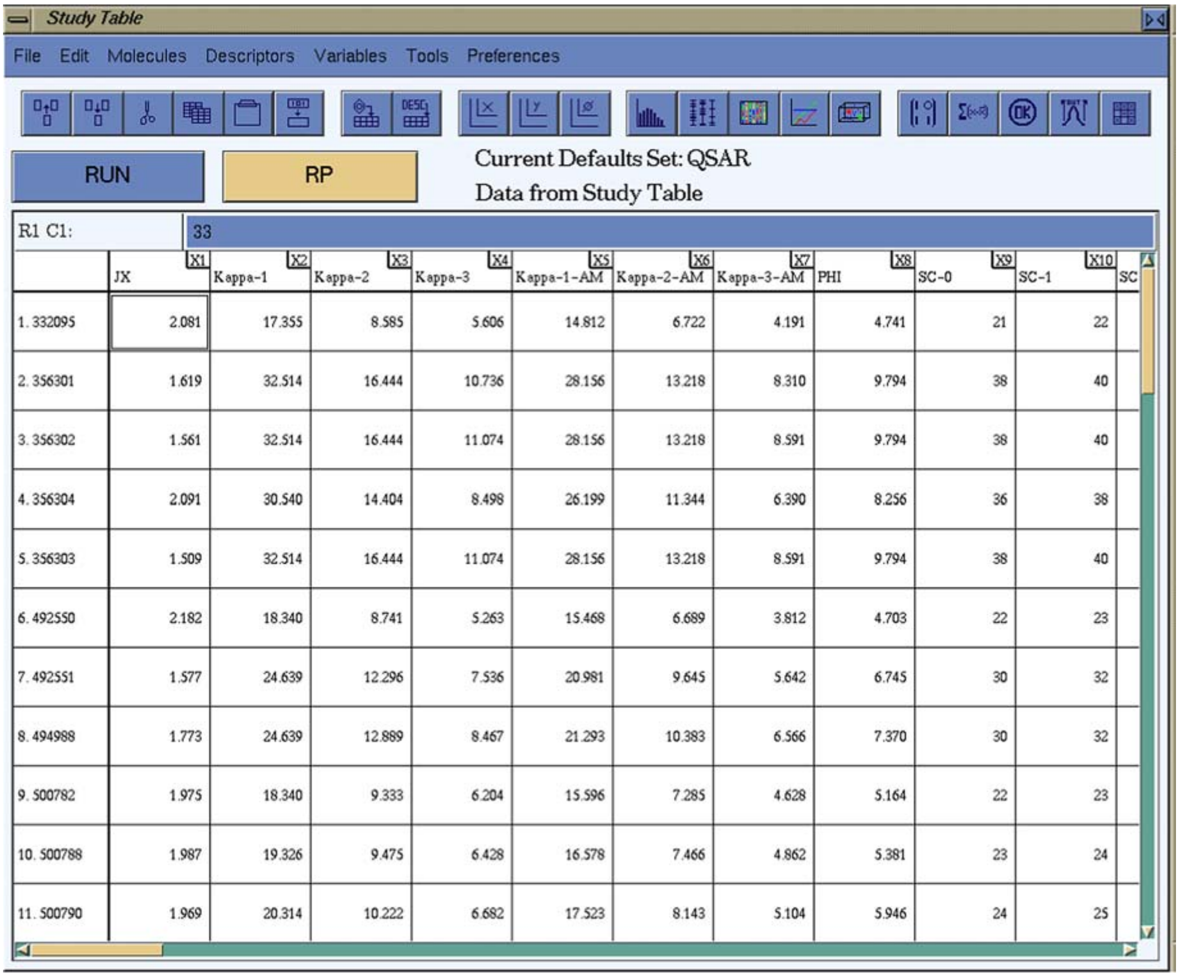


Figure 2. Plot of data in Table 1 used “safely” to build a QSAR model without compromising the integrity of the data. Model equation: $X_2 = -0.101 + 0.994{*}X_5 + 0.339{*}X_{56} - 0.008{*}X_{42} - 1.545{*}X_8$. $r^2 = 0.891$; $q^2 = 0.856$. LOF = 0.131.

one obtains molecules that satisfy the criteria. Those compounds that have the same Molecular Weight with the target molecule moves over to the second phase when all the descriptors are calculated for each of the compounds. Then these descriptors are compared to the target values. If none of the compounds match all of the descriptors then another algorithm kicks in and generates focused libraries for each of the compounds. The molecular weight and other descriptors are then computed once again. This process continues until a compound is invented that provides the exact computed descriptor values as the target one. Such a brute force approach will require smart structure-suggesting and structure-modification algorithms and very long compute times, and even then a match will require a certain amount of luck. We conclude that this approach is difficult, if not outright impossible, to lead to deciphering the molecular structures from the supplied numerical data.

To make the above task even more difficult, one can provide the computed data in fewer significant figures. By using fewer significant figures, for example, for the Molecular Weight (i.e., without any numbers after the decimal point), one can make it very difficult for a person to identify a single unique target compound due to the ambiguity that result with tens to hundreds of compounds that may have the same molecular weight in low resolution. Predictive models will not require data in large significant figures; hence, the quality of the predictive model will be maintained while guessing the structure is made significantly more difficult.

Table 2. Numerical data in a Cerius^2 study table provided as a proof-of-concept evaluation of an ADME/Tox *in silico* model.



Current Defaults Set: QSAR
Data from Study Table

R1 C1: 33

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	
	JX	Kappa-1	Kappa-2	Kappa-3	Kappa-1-AM	Kappa-2-AM	Kappa-3-AM	PHI	SC-0	SC-1	SC
1. 332095	2.081	17.355	8.585	5.606	14.812	6.722	4.191	4.741	21	22	
2. 356301	1.619	32.514	16.444	10.736	28.156	13.218	8.310	9.794	38	40	
3. 356302	1.561	32.514	16.444	11.074	28.156	13.218	8.591	9.794	38	40	
4. 356304	2.091	30.540	14.404	8.498	26.199	11.344	6.390	8.256	36	38	
5. 356303	1.509	32.514	16.444	11.074	28.156	13.218	8.591	9.794	38	40	
6. 492550	2.182	18.340	8.741	5.263	15.468	6.689	3.812	4.703	22	23	
7. 492551	1.577	24.639	12.296	7.536	20.981	9.645	5.642	6.745	30	32	
8. 494988	1.773	24.639	12.889	8.467	21.293	10.383	6.566	7.370	30	32	
9. 500782	1.975	18.340	9.333	6.204	15.596	7.285	4.628	5.164	22	23	
10. 500788	1.987	19.326	9.475	6.428	16.578	7.466	4.862	5.381	23	24	
11. 500790	1.969	20.314	10.222	6.682	17.523	8.143	5.104	5.946	24	25	

Conclusions

We have suggested several approaches for transferring data that not only protect the structural integrity of proprietary compounds, but also provide an easy means of exchanging such data between parties involved in *in silico* model development, or in proof-of-concept model evaluation. We have also shown that it will be difficult, if not outright impossible, to reverse-engineer molecular structure from numerical data; however, there will always remain the possibility that someone will eventually be able to do this. There are several expert ideas being developed to totally encode and/or mask proprietary structural data, while still permitting molecular modeling work to progress on such data. This is a very important issue to resolve for the overall benefit of the lead discovery and evaluation process in the pharmaceutical industry, as well as software development in building good robust predictive QSAR and/or *in silico* ADME/Tox models.

However, since we live in an imperfect world, chances are that with enough time, and effort, there will be attempts to reverse-engineer protected data in the public domain in order to decipher the proprietary structures. What are the likelihoods that such efforts will succeed ? According to Chris Lipinski [3], “if it costs more to crack the code of descriptors than the [actual] value of the compound, then it will probably not be done”. This remains an area of interest and challenge to the pharmaceutical industry, software vendors, academic software developers, and molecular modelers.

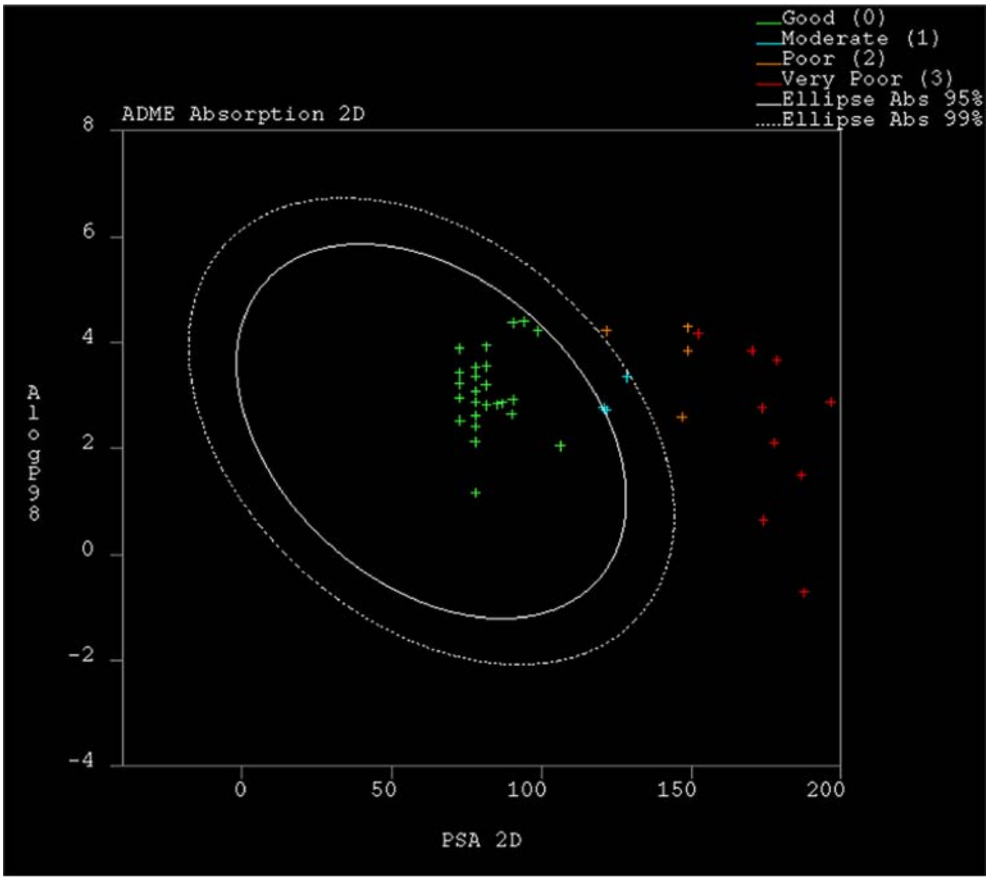


Figure 3. Human Intestinal Absorption (HIA) prediction for a small, focused library from Company-X as determined by the *in silico* HIA model in Cerius^2 [5].

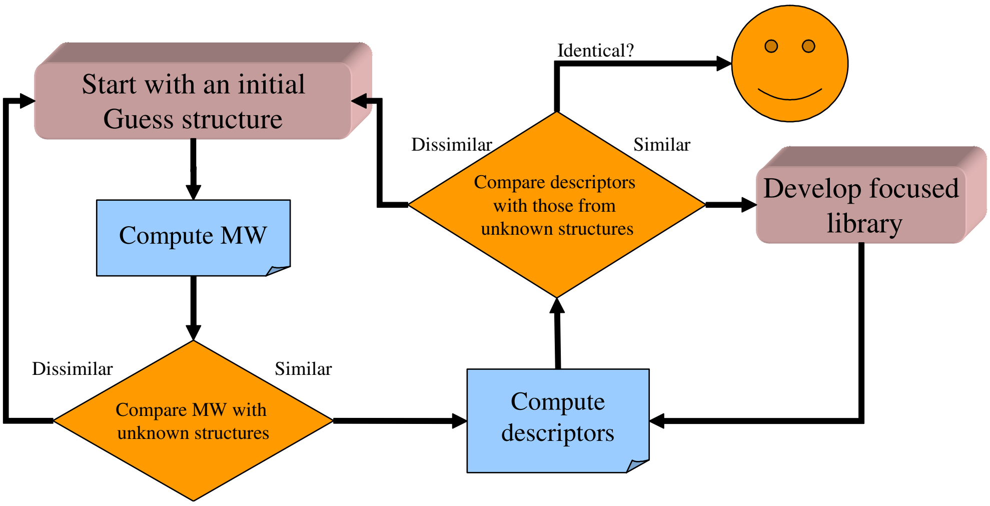

Start with an initial Guess structure
Identical?
Dissimilar
Similar
Compare descriptors with those from unknown structures
Develop focused library
Compute MW
Dissimilar
Similar
Compare MW with unknown structures
Compute descriptors

Figure 4. A process for “guessing” structures from descriptors (!?)

References

1. Mullin, R., Chem. Eng. News, 81 (2003) 21, ISSN 0009-2347 http://pubs.acs.org/cen/coverstory/8130/8130drugdiscovery.html.
2. Safe Exchange of Chemical Information Panel Discussion, CINF Division, 229th ACS Meeting, San Diego, CA. 2005, http://www.acsinf.org/cinf/meetings/229nm/229cinfprogram.htm.
3. Wilson, E.K., Chem. Eng. News, 83 (2005) 24, ISSN 0009-2347, http://pubs.acs.org/cen/science/83/8317sci1.html.
4. Safe Exchange of Chemical Information – The ChemMask Project by Tudor Oprea, BioComputing Division, University of New Mexico, New Mexico, USA. http://pimento.-health.unm.edu/.
5. Programs available in Cerius^2 release 4.10, Accelrys Inc., 10188 Telesis Court, San Diego, CA 92121. http://www.accelrys.com/products/cerius2/.
6. Cheng, A. and Merz, K.M. Jr., J. Med. Chem., 46 (2003) 3572.
7. (a) Egan, W.J., Merz, Jr., K.M. and Baldwin, J.J., J. Med. Chem., 43 (2000) 3867; (b) Egan, W.J. and Lauri, G., Adv. Drug Del. Rev., 54 (2002) 273.
8. Cheng, A., Dixon, S.L., J. Comput. Aided Mol. Des. (2003) 1.
9. Dixon, S.L. and Merz, K.M. Jr., J. Med. Chem., 44 (2001) 3795.
10. Susnow, R.G. and Dixon, S.L., J. Chem. Inf. Comput. Sci., 43 (2003) 1308.
11. Accelrys, Inc. has an ‘awk’ script written by Dr Thomas P. Stockfisch, which is being used as part of the process described in Figure 2.
12. Procedure adopted in the early part of 2000 at Pharmacopeia Labs, Princeton, NJ, for “proof-of-concept” ADME prediction experiments.