US 20040009536A1

(19) United States

(12) Patent Application Publication (10) Pub. No.: US 2004/0009536 A1

Grass et al. (43) Pub. Date: Jan. 15, 2004

(54) SYSTEM AND METHOD FOR PREDICTING ADME/TOX CHARACTERISTICS OF A COMPOUND

(76) Inventors: **George Grass**, Tahoe, CA (US); **Glen D Leesman**, Hamilton, CA (US); **Daniel Norris**, San Diego, CA (US); **Patrick Sinko**, Lebanon, NJ (US); **Jehangir Athwal**, San Diego, CA (US); **Carleton Sage**, Cardiff-by-the-Sea, CA (US); **Troy Bremer**, Dana Point, CA (US); **Kevin Holme**, San Diego, CA (US)

Correspondence Address:
ARENT FOX KINTNER PLOTKIN & KAHN
1050 CONNECTICUT AVENUE, N.W.
SUITE 400
WASHINGTON, DC 20036 (US)

(21) Appl. No.: **10/332,997**

(22) PCT Filed: **Jul. 30, 2001**

(86) PCT No.: **PCT/US01/23763**

Related U.S. Application Data

(60) Provisional application No. 60/221,548, filed on Jul. 28, 2000. Provisional application No. 60/267,435, filed on Feb. 9, 2001.

Publication Classification

(51) **Int. Cl.**[7] **G01N 33/53**; G01N 33/567; G06G 7/48; G06G 7/58; G06F 19/00; G01N 33/48; G01N 33/50; G01N 31/00

(52) **U.S. Cl.** **435/7.2**; 702/19; 702/22; 703/11; 703/12

(57) ABSTRACT

A method for developing a predictive model of a chemical compound property. The method includes obtaining at least one descriptor from structural data for each of a plurality of compounds. At least one chemical compound property is obtained for each of the plurality of compounds. The predictive model is developed by mapping the at least one descriptor to the chemical compound property. The chemical compound property may be an ADME property. The ADME property may be absorption. The chemical compound property may also be an toxicity property.

FIG.1

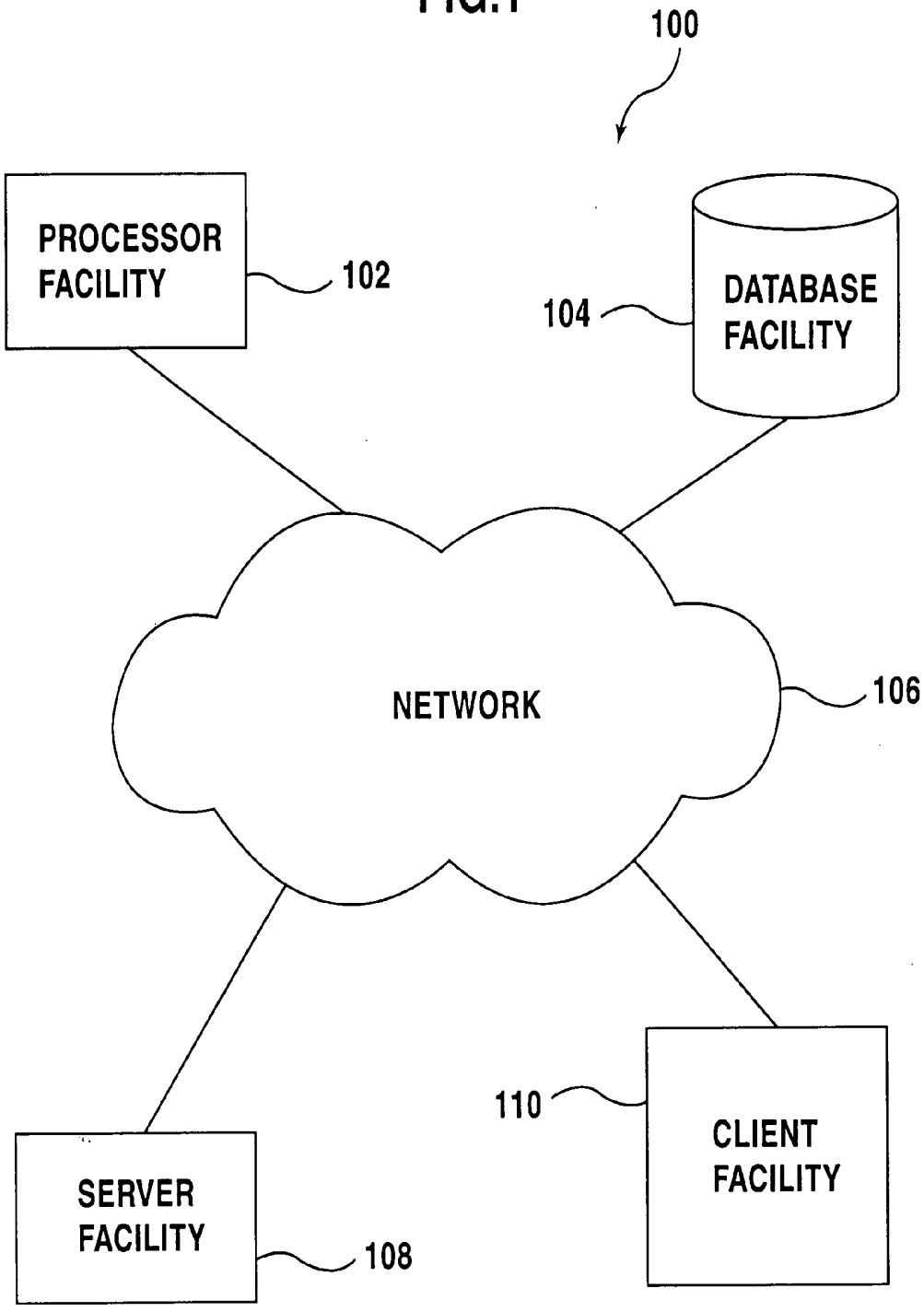

100
PROCESSOR FACILITY
102
104
DATABASE FACILITY
NETWORK
106
110
CLIENT FACILITY
SERVER FACILITY
108

FIG.2

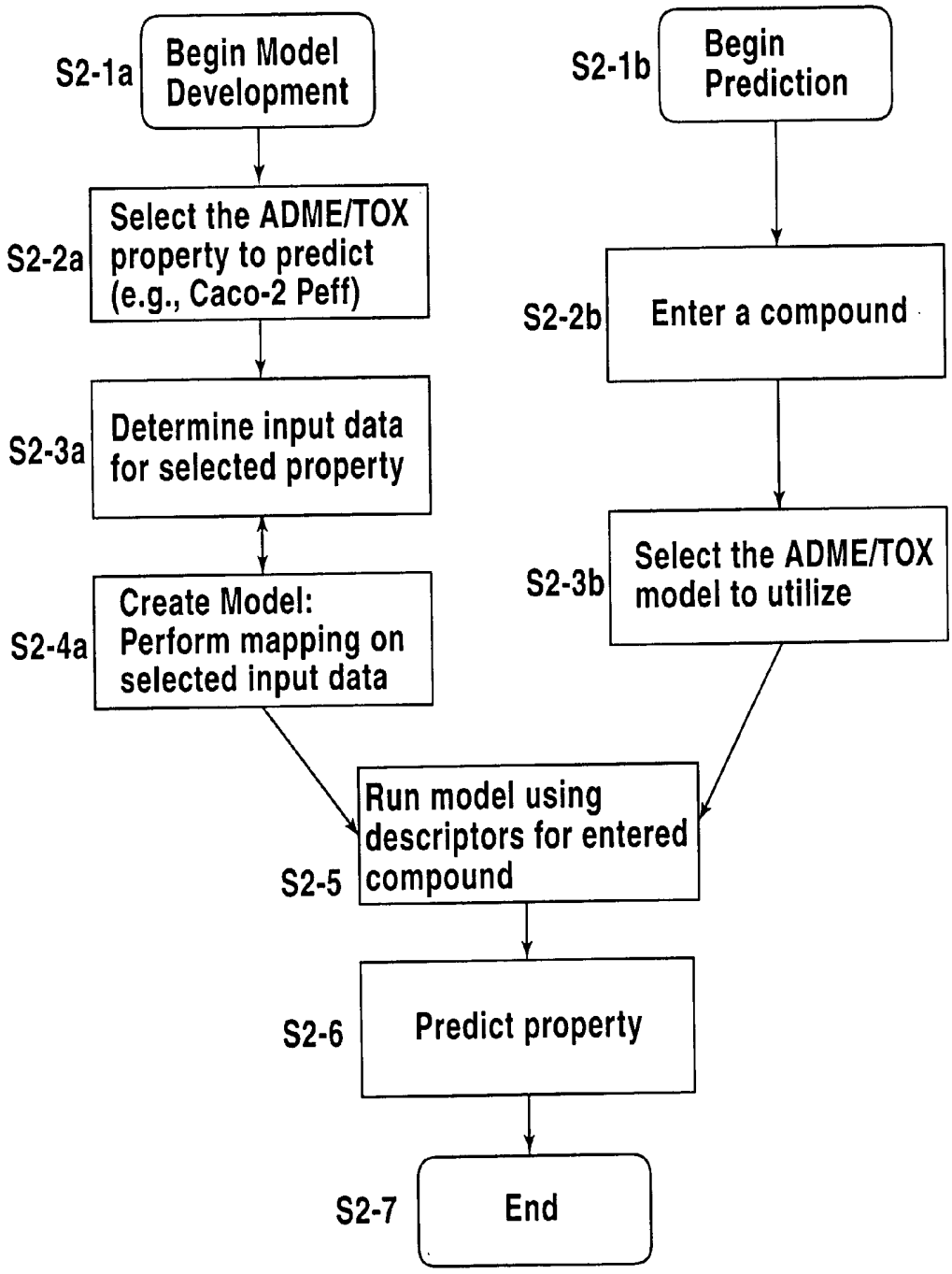

S2-1a
Begin Model Development
S2-1b
Begin Prediction
S2-2a
Select the ADME/TOX property to predict (e.g., Caco-2 Peff)
S2-2b
Enter a compound
S2-3a
Determine input data for selected property
S2-3b
Select the ADME/TOX model to utilize
S2-4a
Create Model: Perform mapping on selected input data
S2-5
Run model using descriptors for entered compound
S2-6
Predict property
S2-7
End

FIG.3

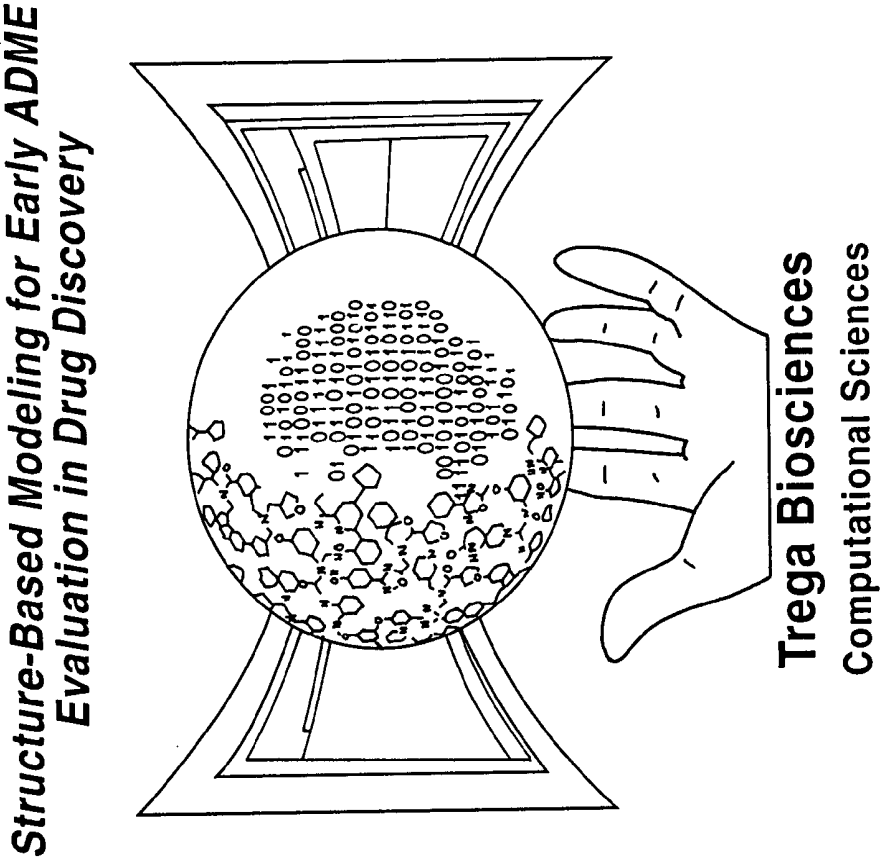

Structure-Based Modeling for Early ADME Evaluation in Drug Discovery
Trega Biosciences
Computational Sciences

FIG.4

Preface

- The methods and applications described in this presentation have been limited in scope to the ADME area. It should be understood that these methods are generally applicable to any research are where chemical structure is to be correlated with some experimental or otherwise determined property. Examples would be QSAR modeling for molecule potency and/or specificity, toxicological profiles of molecules, physicochemical properties of molecules (solubility, melting point), etc.

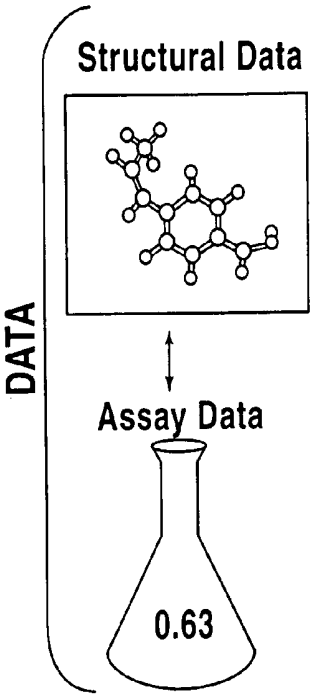
FIG.5

Structure-based Models for ADME/Tox

- Model Building Approach
  - Data
  - Chemical Description / Selection
  - Model Development / Testing
- iDEA Version 2.0 Models
  - Caco-2 Peff
  - FDP
- Future Models and Directions

FIG.6

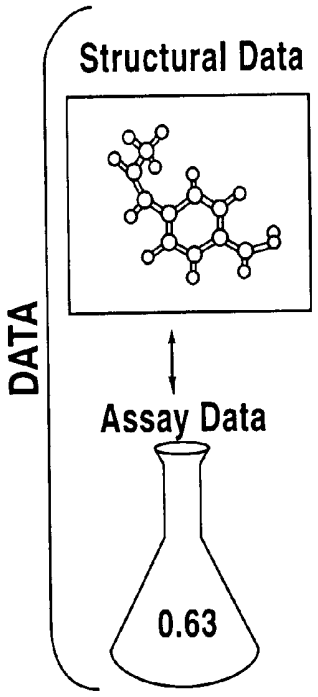
Experimental System



- Models are only as good as the input data
  - Sufficient data generated following a single, well-defined and controlled SOP are critical
- The simpler the experimental system, the more likely a good model can be built from it
- More (good) data is better

FIG.7

Developing the Dataset



Trega Drug Composite (TDC)

- iDEA Consortium Compounds
  - Human PK data known and in house
- Collection of Commercially available drugs / failures
- Currently > 4000 Compounds
- Growing!
  - Chem.Folio neighbors in Chemspace
  - Additional Collaborations
  - Literature mining

FIG.8

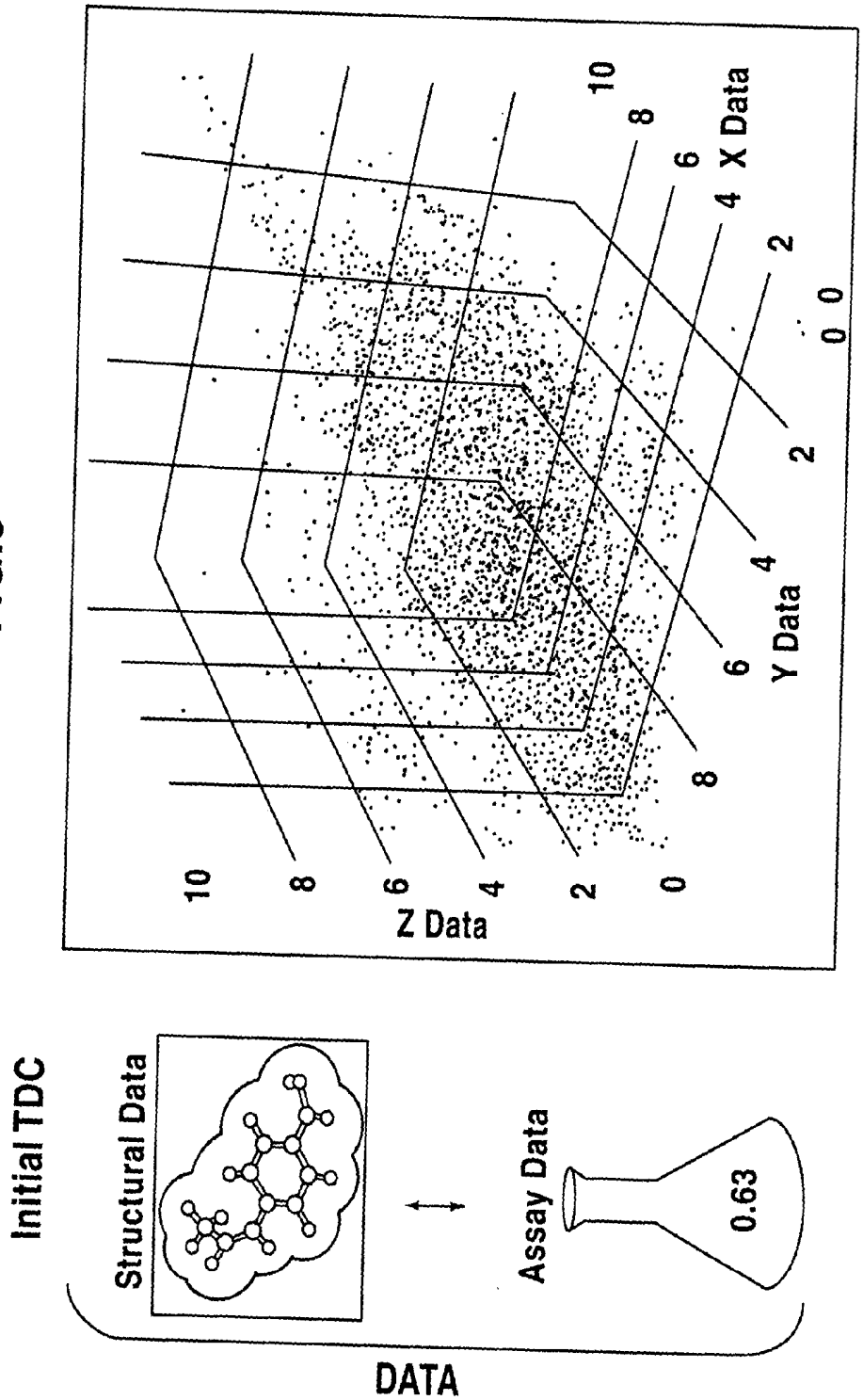

Z Data
0
2
4
6
8
10
Y Data
0
2
4
6
8
10
X Data
0
2
4
6
8
10
Initial TDC
Structural Data
Assay Data
0.63
DATA

FIG.9

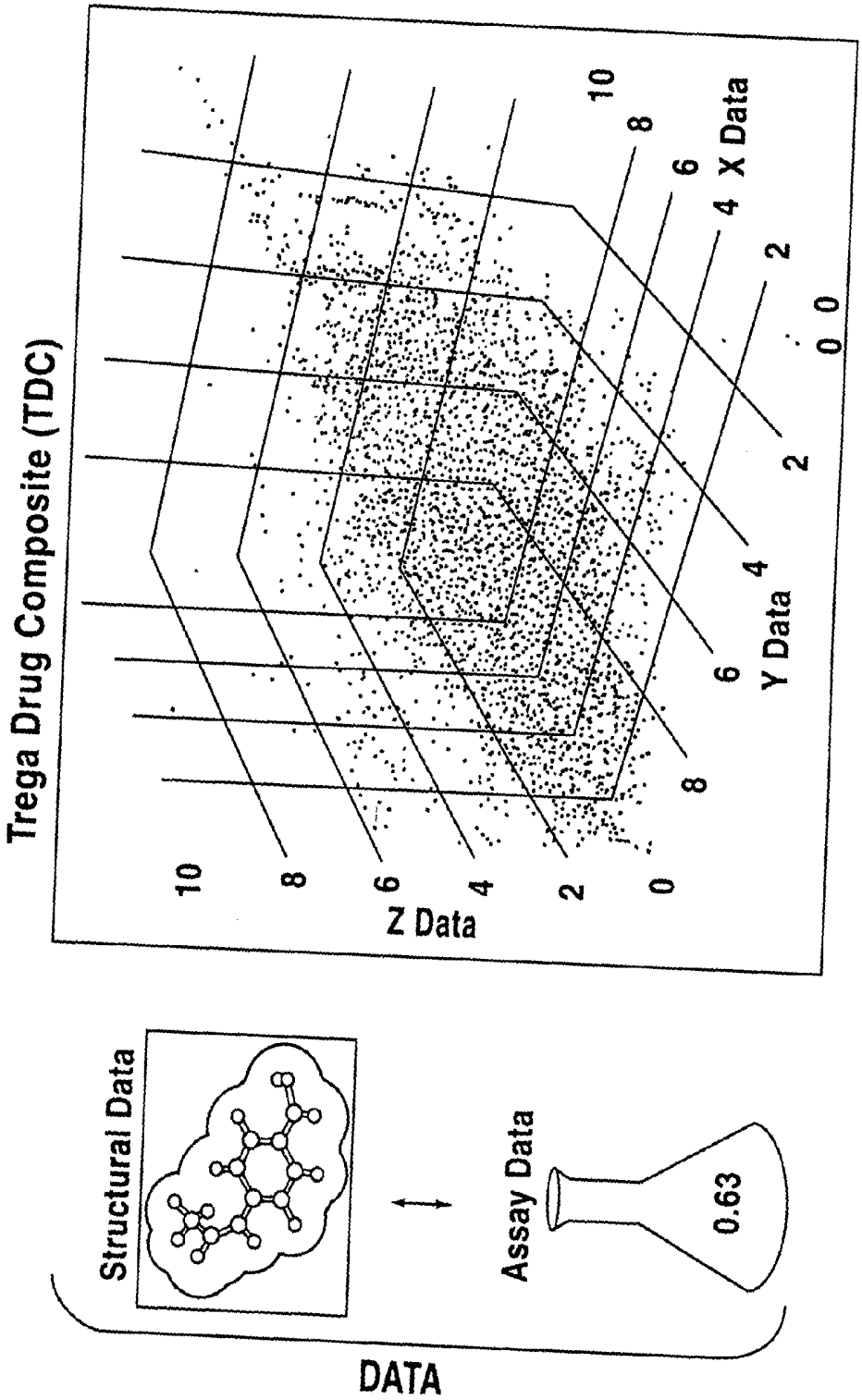

Trega Drug Composite (TDC)
Z Data
0
2
4
6
8
10
Y Data
0
2
4
6
8
10
X Data
0
2
4
6
8
10
Structural Data
Assay Data
0.63
DATA

FIG.10

Considerations for Experimental Data

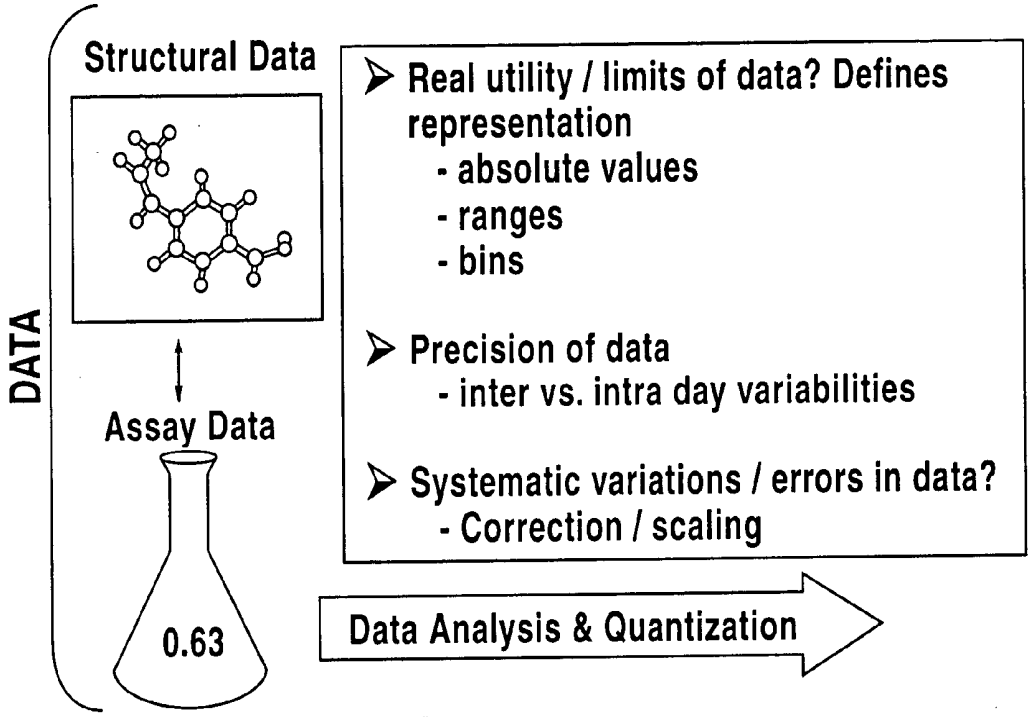


FIG.11

Describing Chemical Space

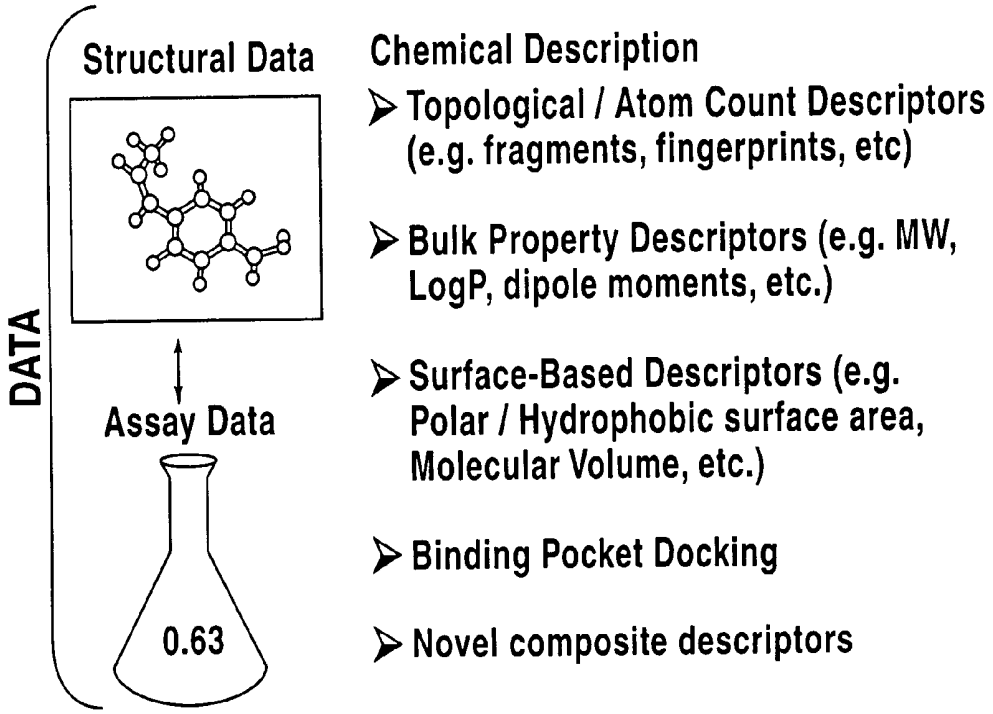


FIG.12

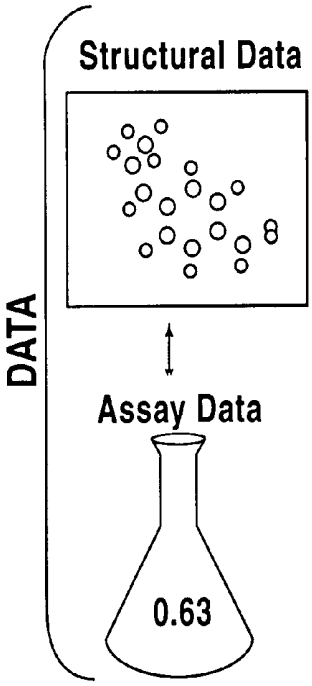
General Methods for Statistical Pattern Recognition in Computational Chemistry / ADME

- Affine Regression
- Fisher's Discriminate Analysis
- Principal Component Analysis
- Cross Validation
- Kernel Representations
- Support Vector Machines
- Neural Networks
- Genetic Algorithms
- Boosting

Note: The use of Kernel representation and Support Vector Machine methods in the Computational Chemistry / ADME area is novel

FIG.13

Descriptor Selection



Chemical Descriptor Selection

- Choose the smallest number of dimensions that effectively describe model space
  - Exhaustive search of 50 descriptors to pick best 10: ~3.7×10^{16} choices!
- Computational selection of variables
  - Iterative Selection
  - Simulated Annealing
  - Genetic Algorithms
  - Trega Implementations follow...

FIG.14

Feature Selection Methods

- Forward / Genetic Algorithm
  - Affine Regression
  - Kernel Methods
  - Neural Networks
  - Finite State Machine - MAP
  - Nearest Neighbors Methods
- Backward
  - Fisher's Linear Discriminate Analysis

FIG.15

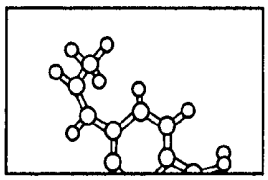
Data Compression Methods

- Non Targeted
  - Linear Principal Component Analysis: SVD
  - Kernel PCA: Kernel Eigenanalysis
- Targeted
  - Fisher's Linear Discriminate Analysis
  - Kernel Fisher's Discriminate Analysis

FIG.16

Model Building

Structural Data



Chemical Descriptor Selection

- Methods
  - Affine Regression
  - Nearest Neighbor Methods
  - Discriminant Analysis
  - Neural Networks
  - Novel Pattern Recognition Methods
    - Listed on Next Pages...
- Considerations
  - Model Bias
  - Overtraining

Statistical Pattern Recognition

FIG.17

Methods

- Kernel Methods
  - Joint Forward Feature Selection / Kernel Evaluation
  - Support Vector Classification and Regression
- Neural Networks
- Affine Regression
- Nearest Neighbor Methods
- Data Compression
  - Targeted
  - Non Targeted

FIG.18

Classification

- Fisher's Discriminate Analysis
- Kernel Fisher's Discriminate Analysis
- Neural Networks
- Nearest Neighbors
  - K-d Trees
  - Epsilon
- Support Vector Classification

FIG.19

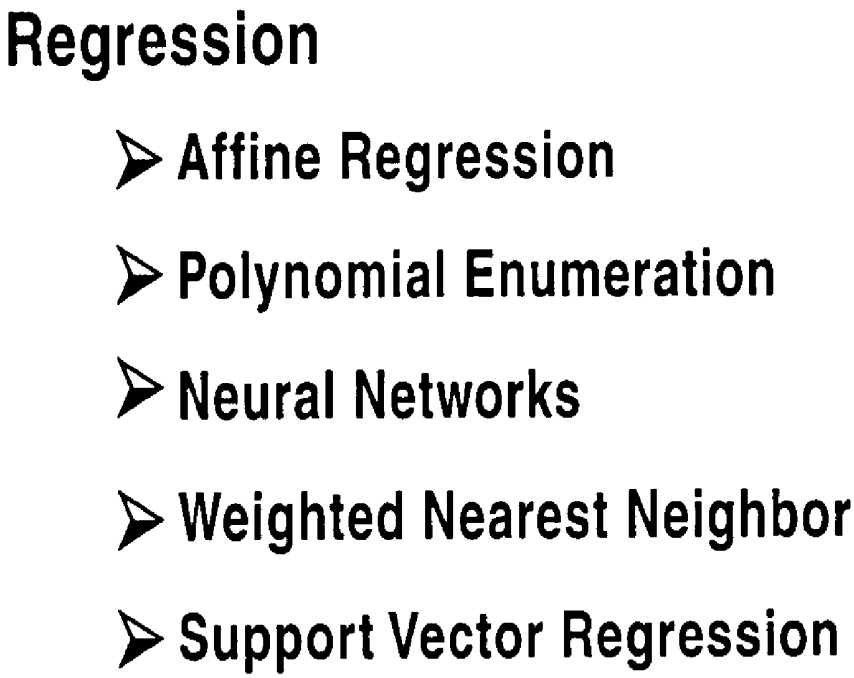

Regression
Affine Regression
Polynomial Enumeration
Neural Networks
Weighted Nearest Neighbor
Support Vector Regression

FIG.20

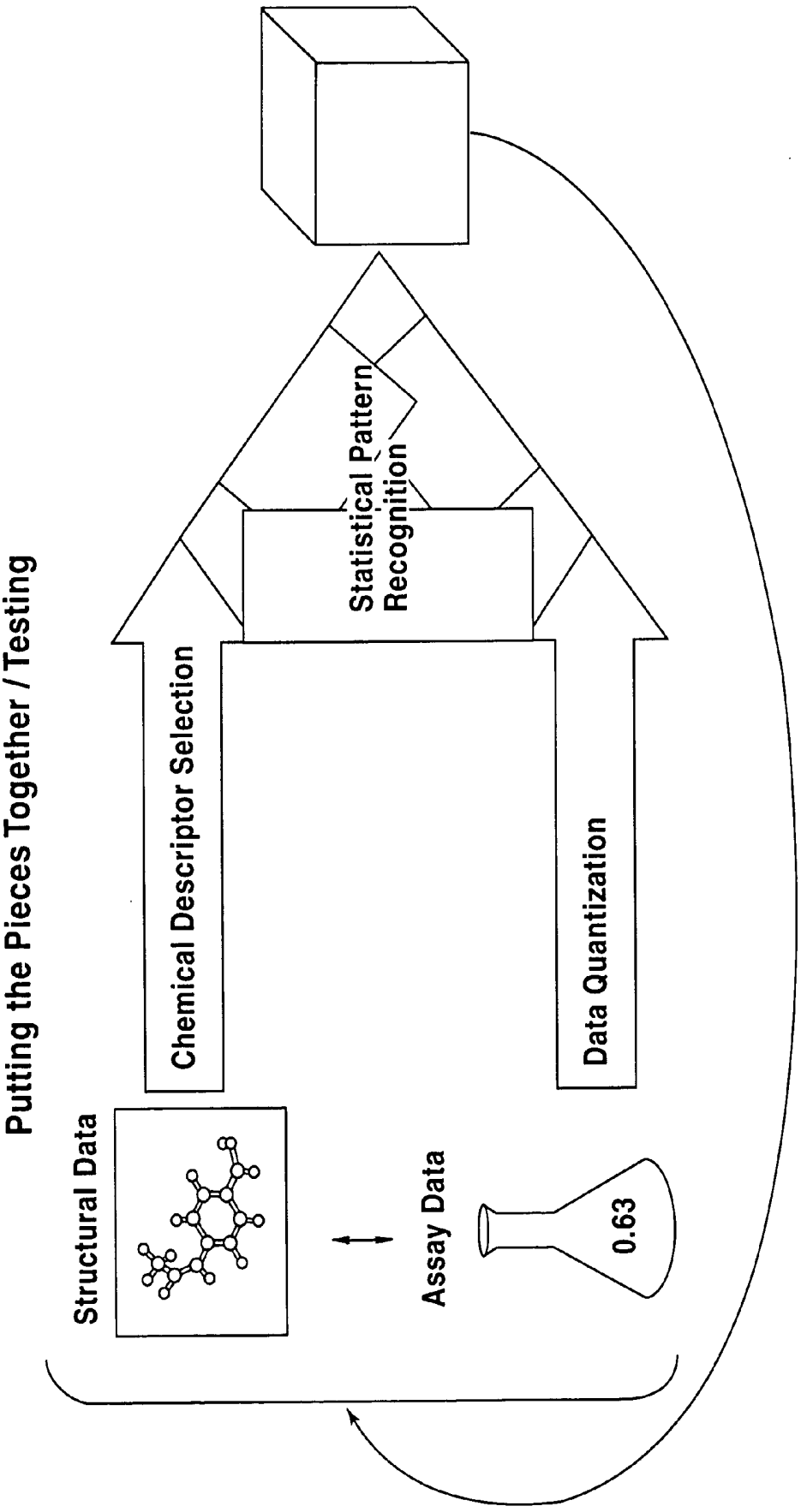

Putting the Pieces Together / Testing
Structural Data
Chemical Descriptor Selection
Statistical Pattern Recognition
Data Quantization
Assay Data
0.63

FIG.21

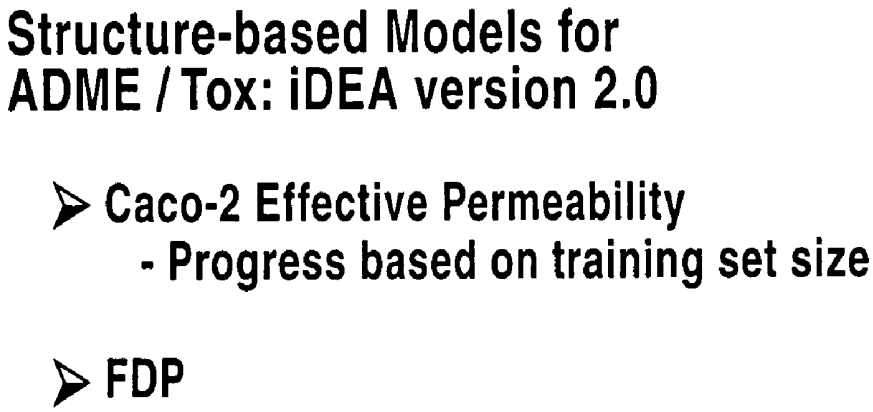

Structure-based Models for
ADME / Tox: iDEA version 2.0
Caco-2 Effective Permeability
- Progress based on training set size
FDP

FIG.22

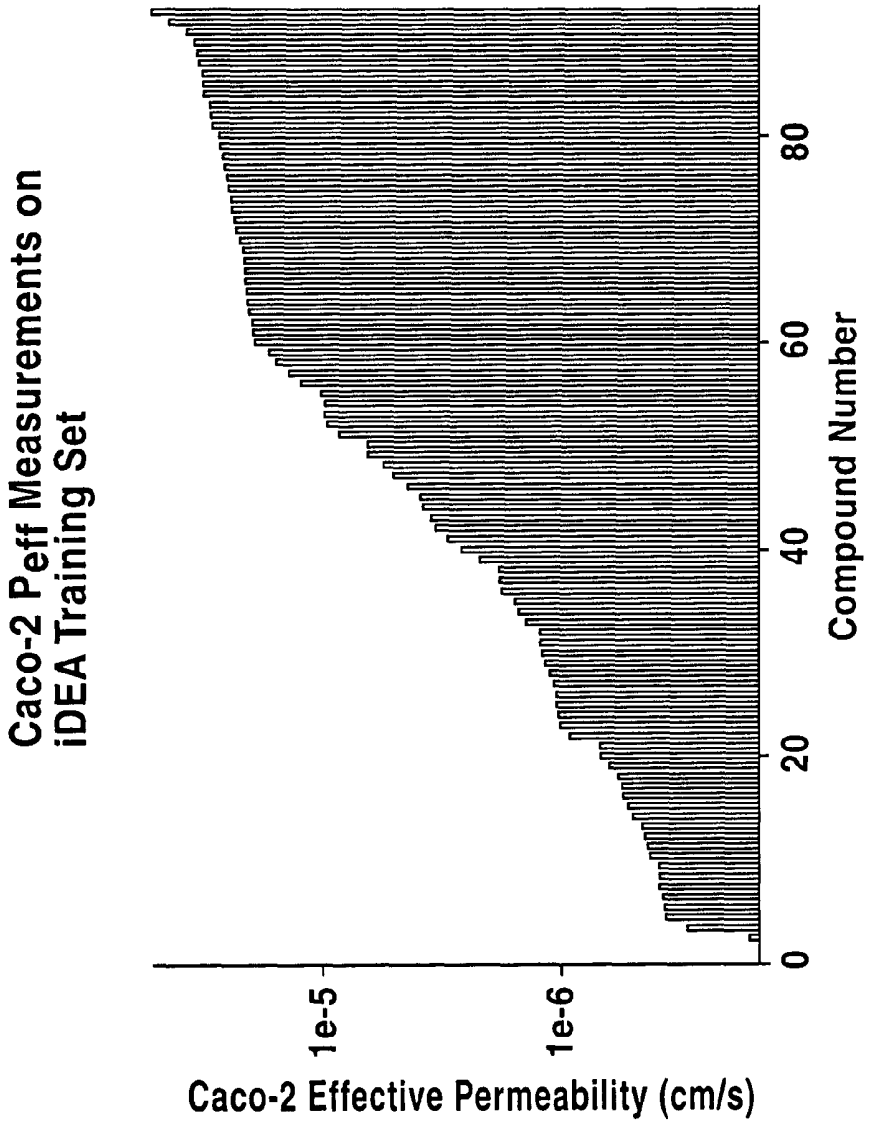

Caco-2 Peff Measurements on
iDEA Training Set
0
20
40
60
80
Compound Number
1e-5
1e-6
Caco-2 Effective Permeability (cm/s)

FIG.23

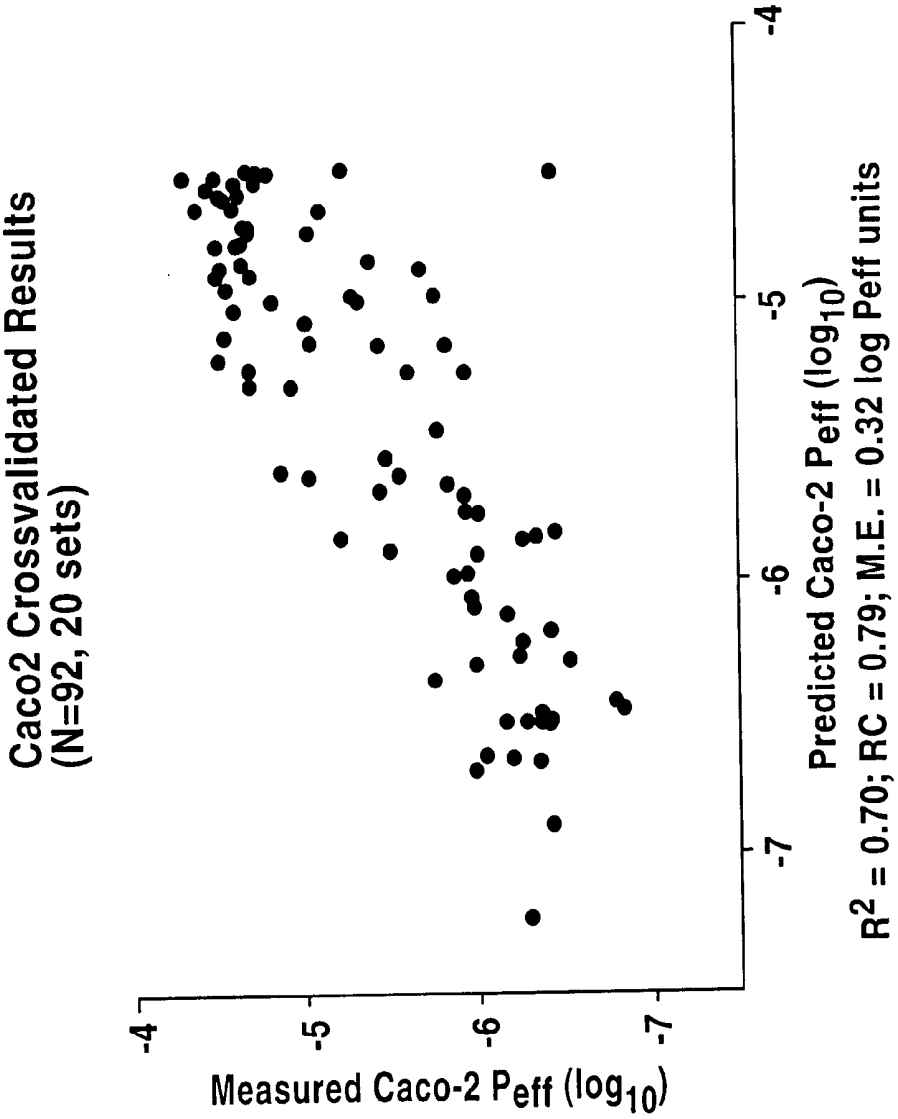

Caco2 Crossvalidated Results
(N=92, 20 sets)
Measured Caco-2 Peff (log10)
Predicted Caco-2 Peff (log10)
R2 = 0.70; RC = 0.79; M.E. = 0.32 log Peff units
-4
-5
-6
-7

FIG.24

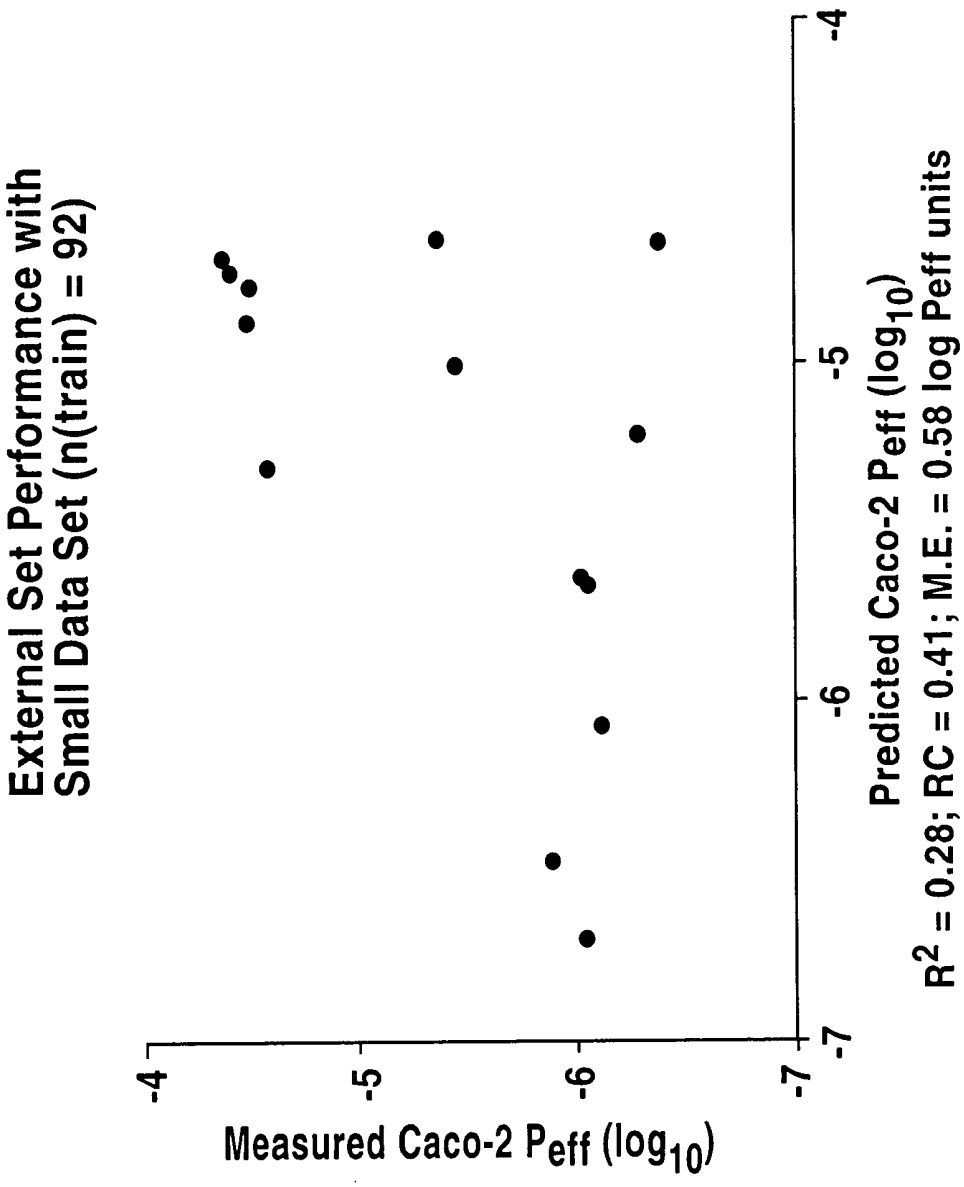

External Set Performance with
Small Data Set (n(train) = 92)
Predicted Caco-2 Peff (log10)
R2 = 0.28; RC = 0.41; M.E. = 0.58 log Peff units
Measured Caco-2 Peff (log10)
-4
-5
-6
-7

FIG.25

Structure -> *in vitro* Models: What more *good* data can do for you...

Exemplified by Caco-2 Effective Permeability version 2.0

FIG.26

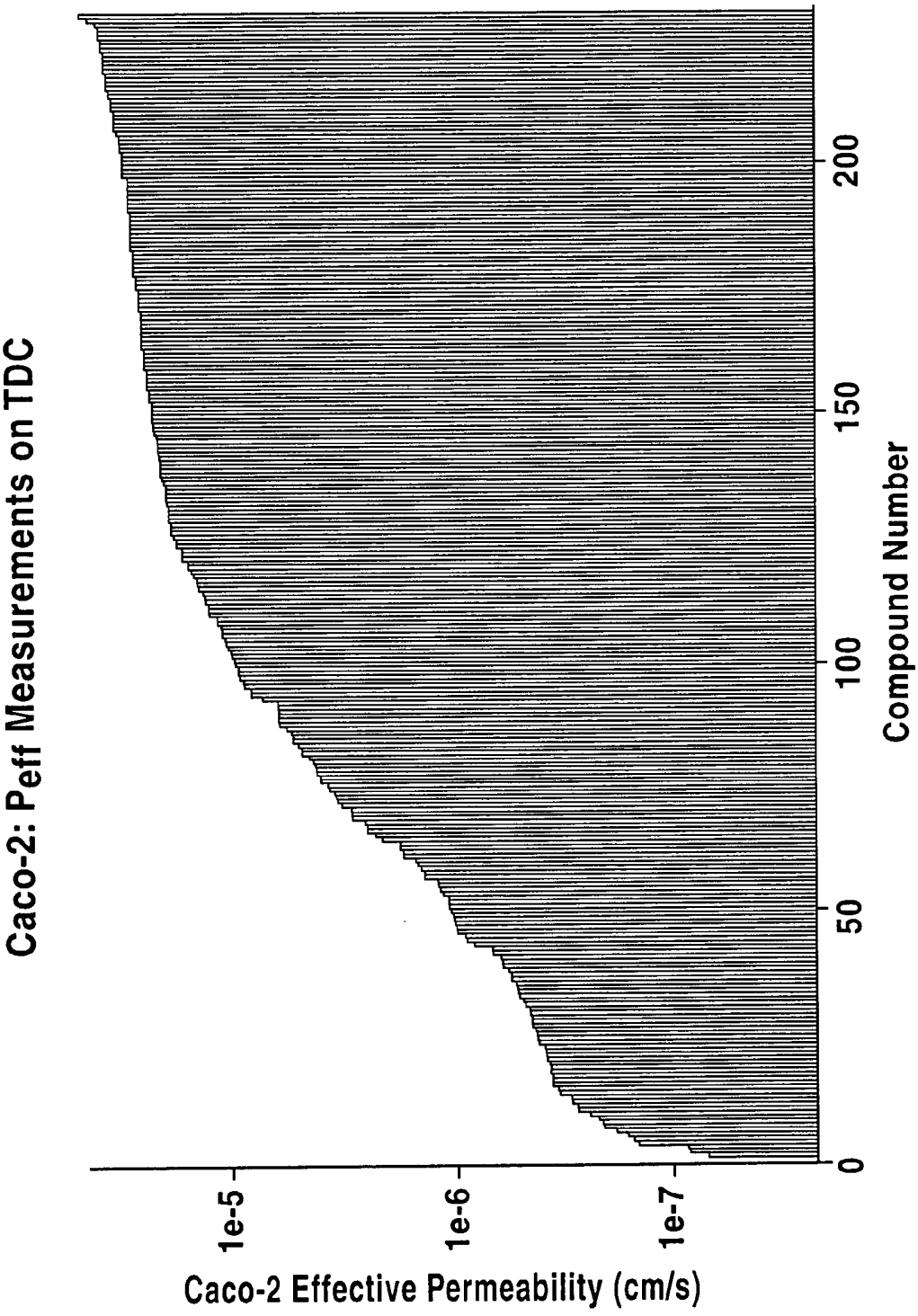

Caco-2: Peff Measurements on TDC
Caco-2 Effective Permeability (cm/s)
1e-5
1e-6
1e-7
0
50
100
150
200
Compound Number

FIG.27

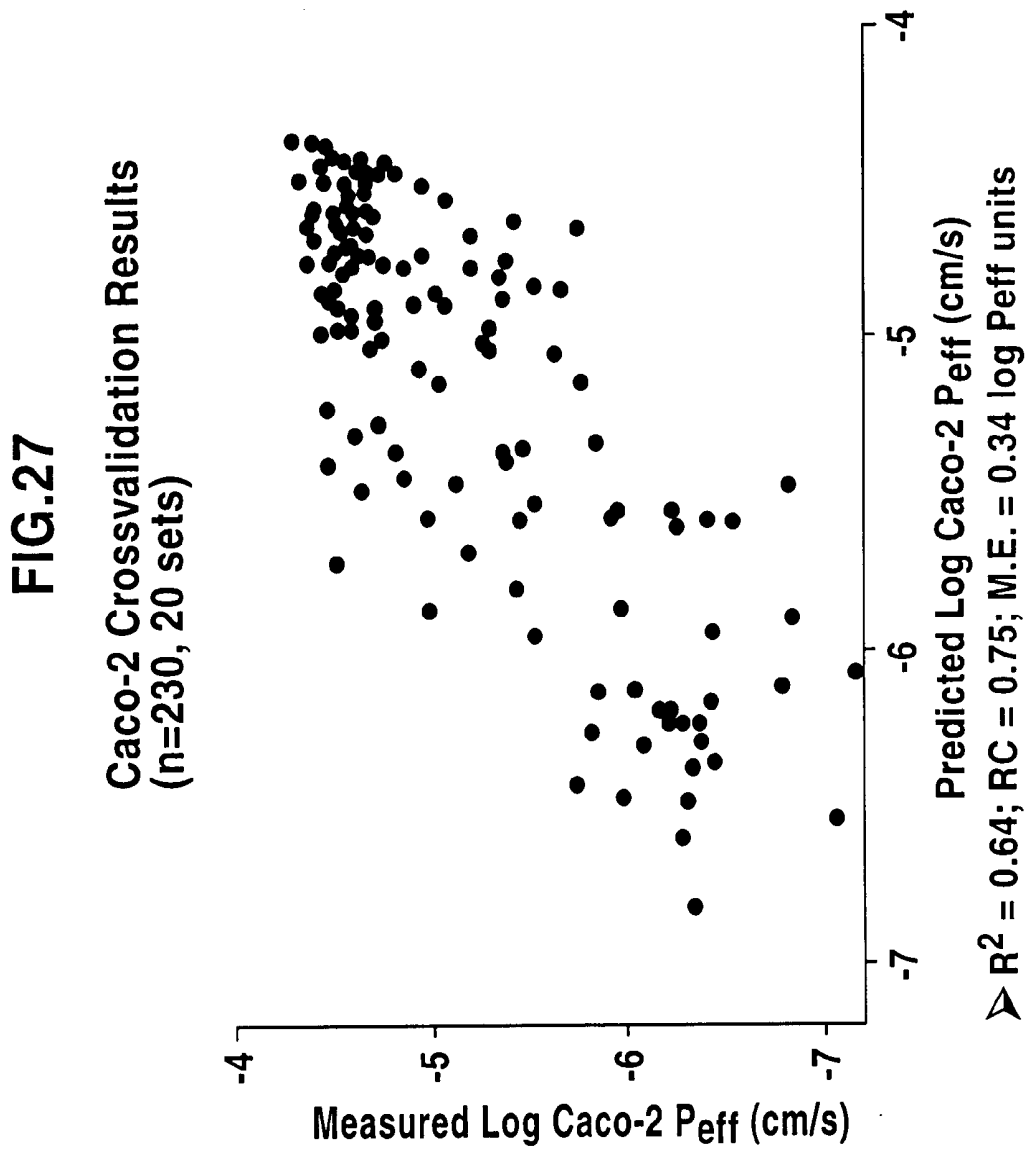

Caco-2 Crossvalidation Results
(n=230, 20 sets)
Measured Log Caco-2 Peff (cm/s)
Predicted Log Caco-2 Peff (cm/s)
R2 = 0.64; RC = 0.75; M.E. = 0.34 log Peff units
-4
-5
-6
-7

FIG.28

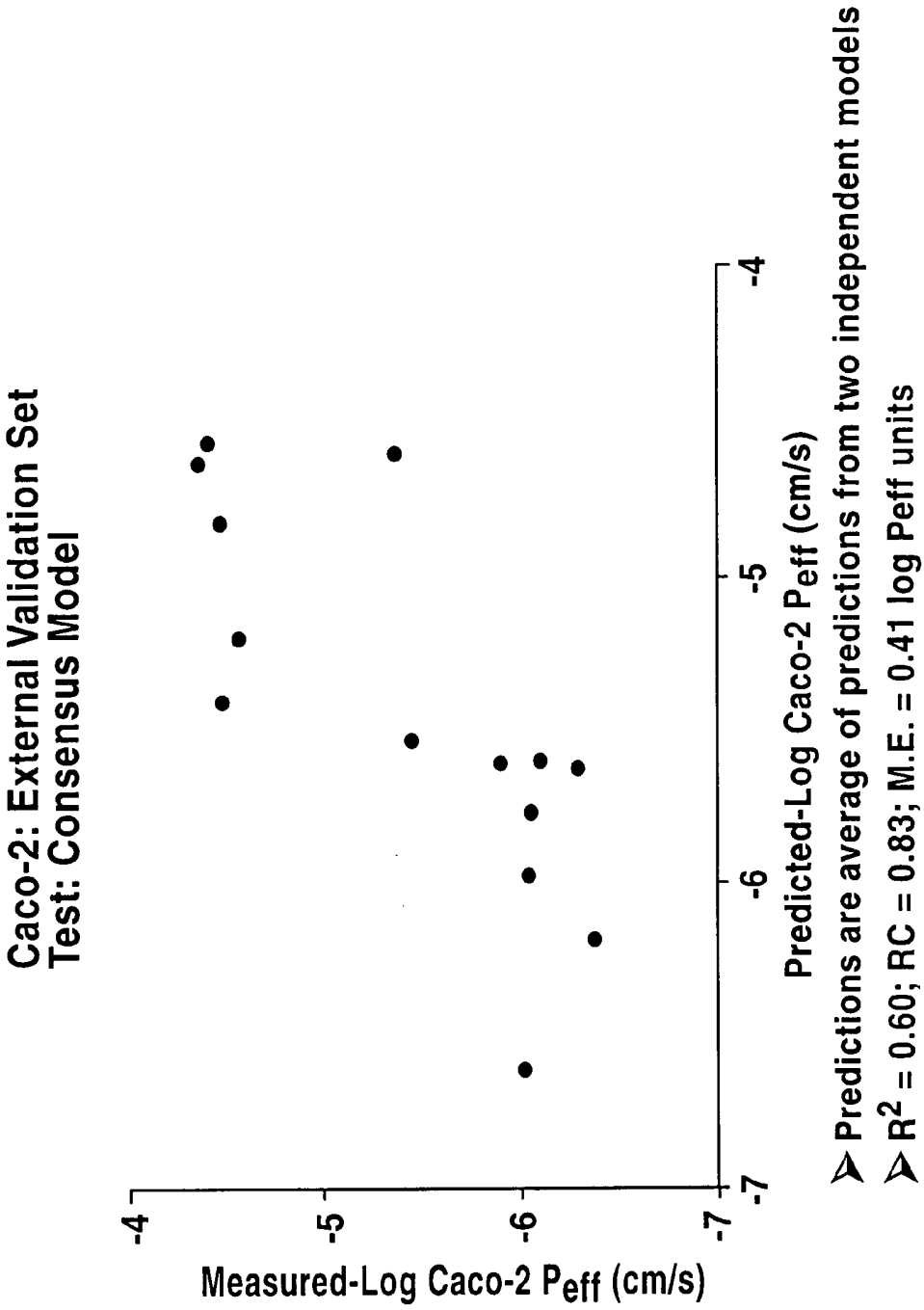

Caco-2: External Validation Set
Test: Consensus Model
Measured-Log Caco-2 Peff (cm/s)
-4
-5
-6
-7
Predicted-Log Caco-2 Peff (cm/s)
-4
-5
-6
-7
Predictions are average of predictions from two independent models
R2 = 0.60; RC = 0.83; M.E. = 0.41 log Peff units

FIG.29

External Validation: Test2- Chemist's Eye versus Computational Model

- Wanted as fair a test as possible with a fairly large sample size
- 40 compounds with new data-not seen before by model or (perhaps) chemists
- Make prediction (high, medium or low bins) using version 2.0 of iDEA Caco-2 Peff model.
- Let chemists as individuals, and as a group (by averaging the individuals) predict permeability bin of compounds
- Compare results

FIG.30

External Validation 2: Results

	Best Chemist	%	Worst Chemist	%	Avg Chemist	%	Computational Model	%
off-by-0	22	55	13	25	14	35	24	60
off-by-1	15	38	18	50	23	58	13	33
off-by-2	3	8	9	25	3	8	3	8
Score	21		36		29		19	

FIG.31

Structure -> FDP Model

version 2.0

FIG.32

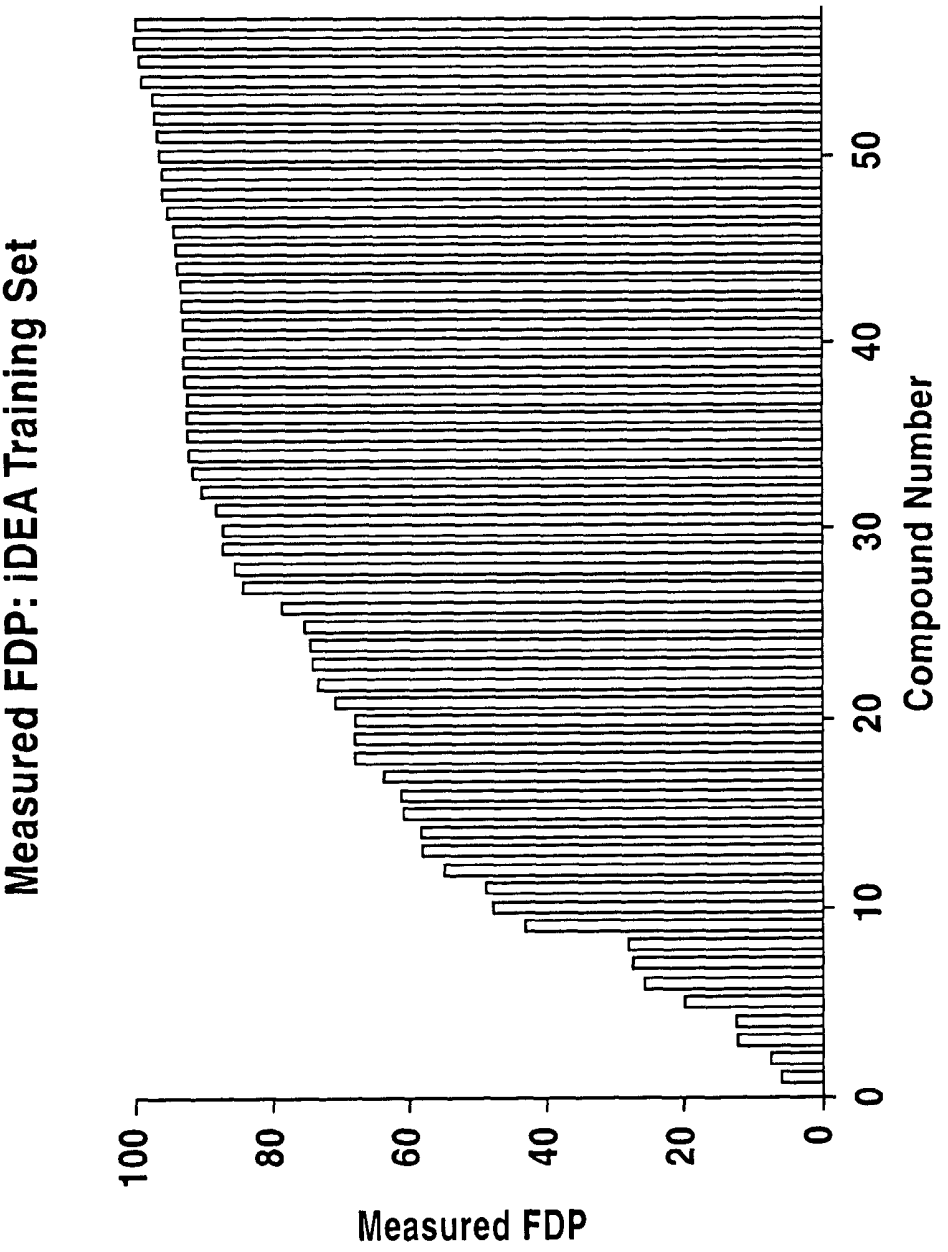

Measured FDP: iDEA Training Set
Measured FDP
100
80
60
40
20
0
0
10
20
30
40
50
Compound Number

FIG.33

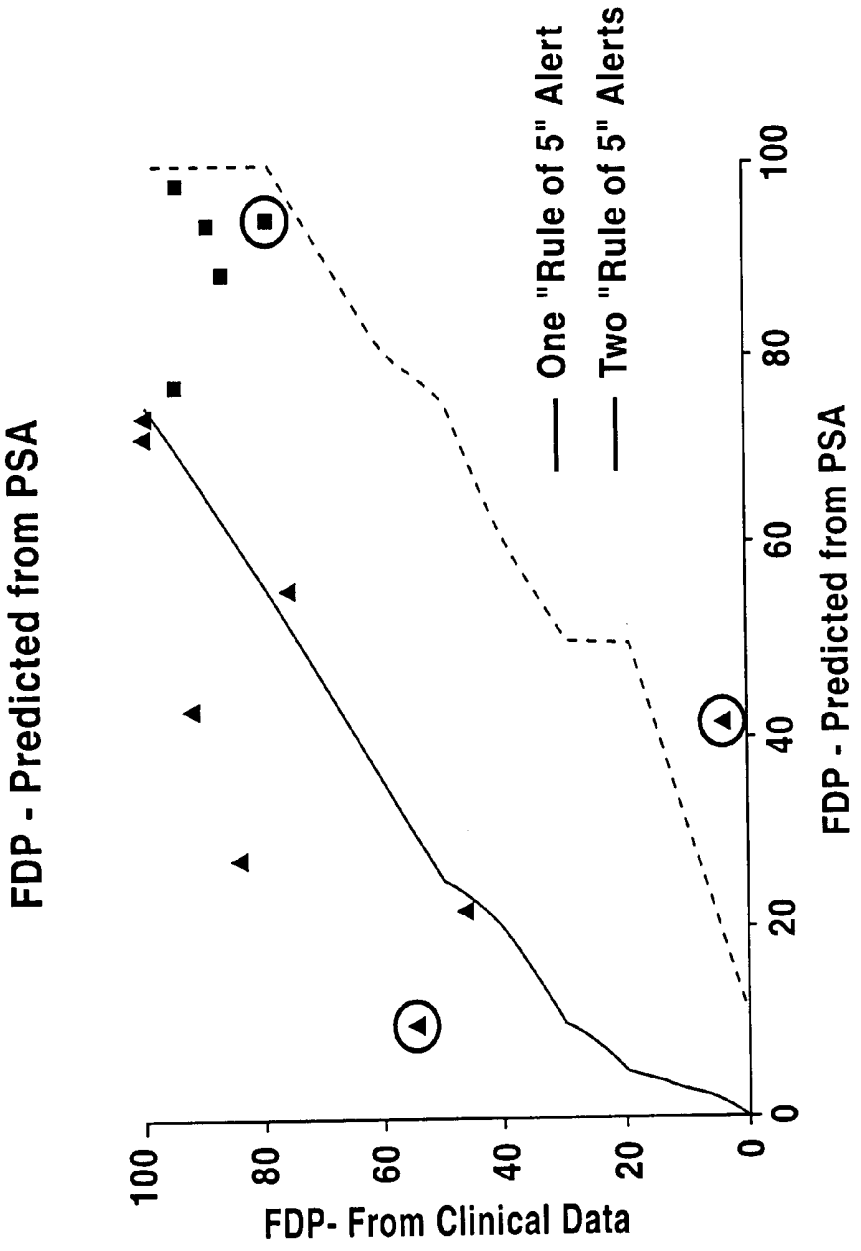

FDP - Predicted from PSA
One "Rule of 5" Alert
Two "Rule of 5" Alerts
FDP- From Clinical Data
FDP - Predicted from PSA
0
20
40
60
80
100
0
20
40
60
80
100

FIG.34

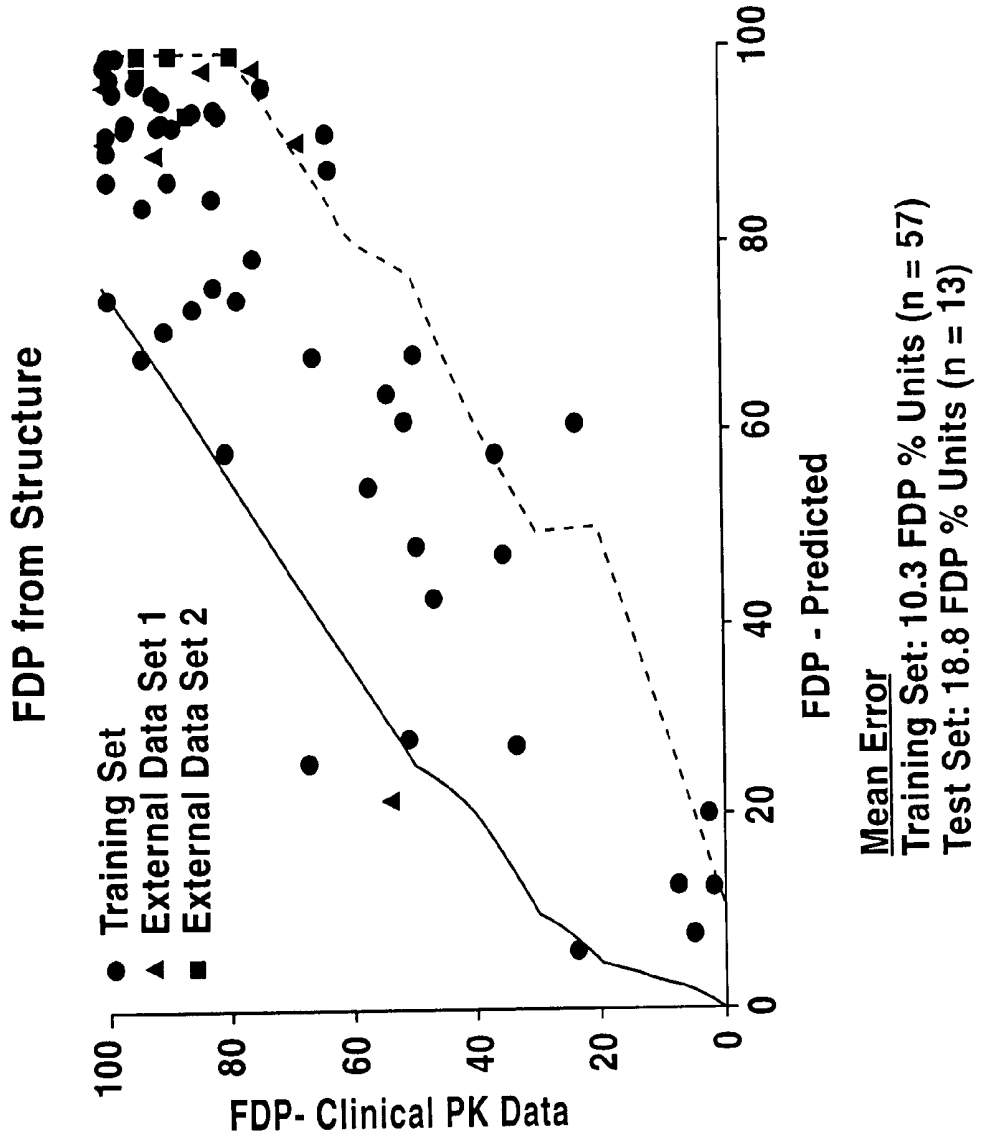

FDP from Structure
Training Set
External Data Set 1
External Data Set 2
FDP- Clinical PK Data
FDP - Predicted
0
20
40
60
80
100
Mean Error
Training Set: 10.3 FDP % Units (n = 57)
Test Set: 18.8 FDP % Units (n = 13)

FIG.35

FDP from Structure: A Screening Tool

Compound	FDP - Meas.	FDP - Pred.	Bin
1	4	100	off by 2
2	100	91	same
3	92	90	same
4	84	98	same
5	69	91	same
6	76	98	same
7	100	97	same
8	54	22	off by 1
9	87.5	94	same
10	95	98	same
11	95	100	same
12	80	100	same
13	90	100	same

- Bins-Low: 0-33; Medium: 34-66; High 67-100
- Results:
  - 11/13 (84.6%) same bin (Training set: 87%)
  - 1/13 (7.6%) off-by-one bin (Training set: 11%)
  - 1/13 (7.6%) off-by-two bins (Training set:2%)

FIG.36

Future Models in coming iDEA versions: Prototypes

- Cytochrome P450 Isozyme models
  - 2D6
  - 3A4

FIG.37

CYP450 Screening Models
Initial Models Developed Using Literature Data

- **Isozyme Data Sets:** 2D6, 3A4, 2C8, 2C9, 2C18, 2C19, 1A2

- **Purpose:** Develop a model that can discriminate between compounds that interact with CYP450s and those that do not.

- **Data Sets:** Collected list of subtrates/inhibitors/activator of CYPs
  - Positive Data: literature
  - Negative Data: if a compound was not identified as a substrate/inhibitor, then it was assumed not to interact with the enzyme

- More data in house (n=4000) now from an external collaboration (Confirmation of Negatives)

FIG.38

CYP2D6 Neural Network Initial Screening Model: Literature

Results from Crossvalidation (20 -sets):

Class	Num/Class	Correct	Error	%Correct	%Error
CYP2D6=No	193	166	42	79.8	20.2
CYP2D6=Yes	92	85	10	89.5	10.5
			Normalized	84.7	15.3

- **False positive rate**
  - Indicating a compound interacts when it really doesn't
  - 10.5% of compounds would be incorrectly discarded.
- **False negative rate**
  - Indicating a compound does not interact with CYP2D6 when it really does)-
  - 20.2% would be incorrectly promoted
- **Utility as a screening filter:**
  - Discriminates between molecules interacting with CYP2D6 at about 85% accuracy.

FIG.39

CYP P450 2D6 Inhibition Assay Validation

Results on known 2D6 substrates/inhibitors in dataset

- Expect better agreement with known substrates than 3A4, since 2D6 is a more specific enzyme
  - 94% of published 2D6 substrates/inhibitors were identified by screen

- Data used to develop a second CYP2D6 structure-based screening model

FIG.40

CYP2D6/AMMC:
Initial Structure-Based Screening Model

Results from Crossvalidation (n=359, 10-sets)

Class	Num/Class	Correct	Error	%Correct	%Error
CYP2D6=No	222	207	15	93.2	6.8
CYP2D6=Yes	137	125	12	91.2	8.8
			Normalized	92.2	7.8

- **False positive rate**
  - Indicating a compound interacts when it really doesn't
  - 8.8% of compounds would be incorrectly discarded.
- **False negative rate**
  - Indicating a compound does not interact with CYP2D6 when it really does
  - 6.8% would be incorrectly promoted
- **Utility as a screening filter:**
  - Discriminates between molecules interacting with CYP2D6 at about 92% accuracy.

FIG.41

CYP P450 3A4 Inhibition Assay Further Validation

Results on known 3A4 inhibitors in dataset

- Expect imperfect agreement with known substrates, since BzRez used as sole substrate, and 3A4 known to have multiple sites of interaction
  - 80% published 3A4 inhibitors were identified by screen

- Data used to develop a BzRez / 3A4 structure based screening model

FIG.42

CYP3A4/BzRez: Initial Structure-Based Screening Model

Results from Crossvalidation (n=323, 10-sets):

Class	Num/Class	Correct	Error	%Correct	%Error
CYP3A4=No	177	153	24	86.4	13.6
CYP3A4=Yes	146	128	18	87.7	12.3
			Normalized	87.1	12.9

- **False positive rate**
  - Indicating a compound interacts when it really doesn't
  - 12.3% of compounds would be incorrectly discarded.
- **False negative rate**
  - Indicating a compound does not interact with CYP3A4 when it really does)-
  - 13.6% would be incorrectly promoted
- **Utility as a screening filter:**
  - Discriminates between molecules interacting with CYP3A4 at about 87% accuracy.

FIG.43

Application of Computational Tools in Combichem Library Analysis and Design

- Chem.Folio Analysis: ~574, 000 Compounds
  - Caco-2 Peff
    - 13% High (>10^{-5} cm/s)
    - 56% Medium (10^{-5} - 10^{-6} cm/s)
    - 31% Low (<10^{-6} cm/s)
  - FDP
    - 77% High (67-100%)
    - 11% Medium (33-67%)
    - 12% Low (0-33%)
  - Predicted results can be experimentally confirmed, and allowed model extension

- Models being implemented early in library design and building block selection

FIG.44

Simultaneous ADME consideration in Compound Design

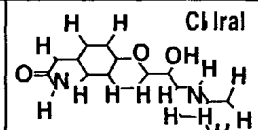
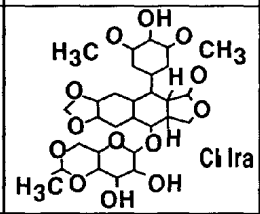
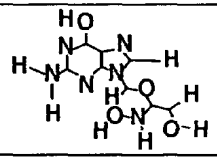
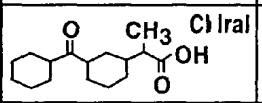
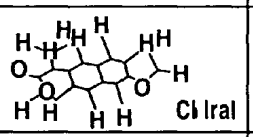
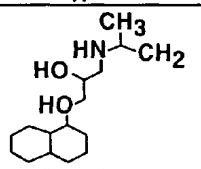
Molname	structure	MW	pFDP	pCaco2_Peff	CYP2D6 Potential	CYP3A4 Potential
Atenolol		266.34	Medium	Medium	0	1
Etoposide		588.57	High	Medium	0	0
Ganciclovir		255.24	Low	Low	0	1
Ketoprofen		254.29	High	High	0	0
Naproxen		230.27	High	High	0	0
Propranolol		259.35	High	High	1	1

FIG.45

Future Efforts

- Defining "Prediction Space" for all models
  - Can we know up front where we will predict badly?
- Structure to %F model
- Accumulate more data for existing systems to enhance and improve models
- Build models in new areas from experimental data:
  - BBB
  - PGP
  - More Extensive Collection of CYP Models
  - Additional PK models
- Relating model information back to structural features- Integration with Chem.Folio- Make models more interpretable

SYSTEM AND METHOD FOR PREDICTING ADME/TOX CHARACTERISTICS OF A COMPOUND

[0001] This application claims the benefit of U.S. Provisional Application Nos. 60/221,548 filed Jul. 28, 2000, entitled PHARMACOKINETIC-BASED DRUG DESIGN TOOL AND METHOD; and 60/267,435 filed Feb. 9, 2001 entitled SYSTEM AND METHOD FOR PREDICTING ADME CHARACTERISTICS OF A COMPOUND BASED ON ITS STRUCTURE.

BACKGROUND OF THE INVENTION

[0002] 1. Field of the Invention

[0003] The present invention relates to systems and methods for predicting the characteristics of a chemical compound. In particular, the present invention is related to pharmacokinetic systems and methods for predicting the Absorption, Distribution, Metabolism, Excretion and/or Toxicological (ADME/TOX) characteristics or properties of a chemical compound based on structural modeling of the chemical compound and mathematical analysis.

[0004] 2. Description of the Prior Art

[0005] Pharmacodynamics refers to the study of fundamental or molecular interactions between drug and body constituents, which through a subsequent series of events results in a pharmacological response. For most drugs, the magnitude of a pharmacological effect depends on the time-dependent concentration of drug at the site of action (e.g., target receptor-ligand/drug interaction). Factors that influence rates of delivery and disappearance of drug to or from the site of action over time include its ADME properties. The study of factors that influence how drug concentration varies with time is the subject of pharmacokinetics. Additionally, the toxicological properties of a drug should also be considered. These properties taken together represent the ADME/TOX properties of a compound.

[0006] In nearly all cases, the site of drug action is located on the other side of a membrane from the site of drug administration. For example, an orally administered drug must be absorbed through a series of physiological barriers at some point or points along the gastrointestinal (GI) tract. Once the drug is absorbed, and thus passes a membrane barrier of the GI tract, it is transported through the portal vein to the liver and then eventually into systemic circulation (i.e., blood and lymph) for delivery to other body parts and tissues by blood flow. Thus, how well a drug crosses membranes is of key importance in assessing the rate and extent of absorption and distribution of the drug throughout different body compartments and tissues. In essence, if an otherwise highly potent drug is administered extravascularly (e.g., oral) but is poorly absorbed (e.g., GI tract), a majority of the drug will be excreted or eliminated and thus cannot be distributed to the site of action.

[0007] The ADME/TOX properties of a candidate drug (chemical compound) are usually determined through conventional laboratory testing (in vitro or in vivo) combined with mathematical modeling. For instance, pharmacokinetic data analysis may be based on empirical observations after administering a known dose of drug to an animal and fitting of the data collected from the animal (e.g., from its liver cells) by either descriptive equations or mathematical (compartmental) models. Time-concentration data from a subject that has been given a particular dose of a drug may be collected followed by plotting the data points on a logarithmic graph of drug concentration versus time to generate one type of concentration-time curve. A mathematical equation is used to model what might happen to the drug as it is transported through a human body. Classical one, two and three compartment models used in pharmacokinetics require in vivo blood data to describe concentration-time effects related to the drug decay process, i.e., blood data is relied on to provide values for equation parameters. For instance, while a model may work to describe the decay process for one drug, it is likely to work poorly for others unless blood profile data and associated rate process limitations are generated for each drug in question. Thus, current models are very poor for predicting the in vivo fate of diverse drug sets in the absence of blood data and the like derived from animal and/or human testing (Lipinski et al. 1997. Advanced Drug Delivery Reviews. 23, 3-25; Palm et al. 1997. Pharm. Res. 14(5) 568-571). For this reason, animal testing is still very much used to predict the ADME/TOX properties of chemical compounds. However, several studies have shown that in general, such types of testing in animal models are poor surrogates for performance in humans (W. K. Sietsema, Int. J. Clin. Pharmacol, Therapy, and Toxicol., 27:179-211 (1989)). Furthermore, conventional laboratory testing and animal testing is very costly and time consuming.

[0008] Thus, there is a need for new and improved systems and methods for predicting the ADME/TOX characteristics of chemical compounds that can eliminate or reduce the need for animal testing as well as all other types of physical experimental testing. These new systems will also improve the correlation to the true needed endpoint, which, in most cases is man.

SUMMARY OF THE INVENTION

[0009] The present invention solves the aforementioned problems by providing new and improved systems and methods of predicting the ADME/TOX properties of candidate drugs (chemical compounds). Such systems and methods may use empirical statistical pattern recognition approaches to take known chemical structures and characteristics (e.g., ADME/TOX) of all compounds for which data has been generated (e.g., data is available from various labs, is published, etc.) and to relate the structures and their characteristics to experimental data in such a way to accurately predict the characteristics of a new proposed structure (compound).

[0010] According to an embodiment of the present invention, provided is a system for predicting the target data of a compound in a mammalian (actual descriptions are human related) body comprising a database facility and a processor facility. The database facility is configured to store input data. The processor facility is configured to allow the entry of input data relating to a new proposed chemical compound including structural data, to perform an analysis of the chemical compound by mapping the data entered to produce predicted target data for the chemical compound based on the analysis.

[0011] According to another embodiment of the present invention, provided is a method for creating or developing a model to be used for evaluating the ADME/TOX characteristics of a proposed compound. The method comprises the following steps:

[0012] (a) selecting training compounds based on the characteristics to be predicted of the proposed compounds (for which a complete set of input and target data exists)

[0013] (b) selecting descriptors applicable to the characteristic to be predicted based on an analysis of the training compounds selected in step (a), such as via a genetic algorithm or other appropriate mathematical analysis

[0014] (c) mapping the training set obtained in (b) to the target data resulting in a model which could predict the target data of a proposed compound.

[0015] Compounds should be selected for their applicability for the problem to be solved, for example, such as for Caco-2 effective permeability (Caco-2 cells possess many of the properties of the small intestine; as such, these cells represent a useful and well-accepted tool for studying the absorption and/or secretion of drugs/chemicals across the intestinal mucosa). Accordingly, drugs may be selected as compounds to be analyzed because of their proven permeability or absorption properties. Other compounds may similarly be selected and added to the data set. Once compounds have been analyzed for descriptors, they may be tested by conventional means (e.g., lab testing, etc.) to determine various characteristics to be predicted by the system above (e.g., CaCo-2 permeability). Once all data has been analyzed and collected, they are loaded into the database for use in predicting the ADME/TOX properties of proposed compounds.

[0016] In other embodiments, the method may include:

[0017] (a) receiving at least one proposed compound (e.g., the molecular structure, etc.) via a user input means (e.g., from a file, input via a form, etc.),

[0018] (b) selecting training compounds from the database facility based on the characteristics to be predicted of the proposed compounds (for which a complete set of input and target data exists)

[0019] (c) selecting the most meaningful descriptors applicable to the characteristic to be predicted based on an analysis of the training compounds selected in step (b), such as via a genetic algorithm or other appropriate mathematical analysis

[0020] (d) creating validation data subsets of the training data based upon the distribution of descriptors and target characteristics of compounds selected in (b/c)

[0021] (e) mapping the training set obtained in (d) to the target data resulting in a model which could predict the target data of a proposed compound.

[0022] (f) modifying (for example: boosting, bootstrap aggregation (bagging)), and other model enhancement methods, etc.) one or more models produced in (e) based upon performance on validation sets obtained in (d) to form a composite model

[0023] (g) combining (via boosting, committee machines etc,) a set of two or more models produced in (e or f) based upon performance on validation sets obtained in (d) to form a composite model

[0024] (h) running the model determined in either step (e), (f) or (g) using the required input data (the identity of the subset of input data itself. was determined in step (c)) to predict the required target data

[0025] According to another embodiment of the present invention, provided is a system for predicting the chemical properties of at least one proposed compound comprising: a database facility configured to store and to serve input data relating to the characteristics of training compounds (descriptor(s) (for example, structure and experimental data)) as well as target data (for example, chemical properties of selected compounds) for the training compounds; and a processor facility coupled to the database facility and configured to predict the characteristics of a proposed compound by:

[0026] (a) selecting training compounds from the database facility based on the characteristics to be predicted of the proposed compounds (for which a complete set of input and target data exists)

[0027] (b) selecting descriptors applicable to the characteristic to be predicted based on an analysis of the training compounds selected in step (a), such as via a genetic algorithm or other appropriate mathematical analysis

[0028] (c) mapping the training set obtained in (b) to the target data resulting in a model which could predict the target data of a proposed compound.

[0029] According to another embodiment of the present invention, provided is a system for predicting the chemical properties of at least one proposed compound comprising: a database facility configured to store and to serve input data relating to the characteristics of the proposed compound (descriptor(s) (for example, structure and experimental data)); and a processor facility coupled to the database facility and configured to predict the characteristics of a proposed compound by:

[0030] (a) receiving at least one proposed compound (e.g., the molecular structure, etc.) via a user input means (e.g., from a file, input via a form, etc.),

[0031] (b) running the model using the appropriate input data to predict the required target data

[0032] According to another embodiment of the present invention, provided is a system for predicting the chemical properties of at least one proposed compound comprising: a database facility configured to store and to serve input data relating to the characteristics of training compounds (descriptor(s) (for example, structure and experimental data)) as well as target data (for example, chemical properties of selected compounds) for the training compounds; and a processor facility coupled to the database facility and configured to predict the characteristics of a proposed compound by:

[0033] (a) receiving at least one proposed compound (e.g., the molecular structure, etc.) via a user input means (e.g., from a file, input via a form, etc.);

[0034] (b) selecting training compounds from the database facility based on the characteristics to be predicted of the proposed compounds (for which a complete set of input and target data exists);

[0035] (c) selecting the most meaningful descriptors applicable to the characteristic to be predicted based on an analysis of the training compounds selected in step (b), such as via a genetic algorithm or other appropriate mathematical analysis;

[0036] (d) creating validation data subsets of the training data based upon the distribution of descriptors and target characteristics of compounds selected in (b/c);

[0037] (e) mapping the training set obtained in (d) to the target data resulting in a model which could predict the target data of a proposed compound;

[0038] (f) modifying (for example: boosting, bootstrap aggregation (bagging)), and other model enhancement methods, etc.) one or more models produced in (e) based upon performance on validation sets obtained in (d) to form a composite model;

[0039] (g) combining (via boosting, committee machines etc,) a set of two or more models produced in (e or f) based upon performance on validation sets obtained in (d) to form a composite model; and

[0040] (h) running the model determined in either step (e), (f) or (g) using the required input data (the identity of the subset of input data itself was determined in step (c)) to predict the required target data.

[0041] Analysis used to select the most meaningful subset of input data (step (c)) for predicting target data may be performed via feature selection methods such as forwards or backwards selection and may include regression/classification methods. Such analyses should consider model bias and overtraining.

[0042] The preceding analyses may include various data compression techniques.

[0043] A particular model may be biased if the training data is poorly distributed (e.g. the distribution has sharp peaks, regions between nodes that are devoid of data, etc). Accordingly, compounds may be selected and tested to improve the distribution and enhance the model's ability to generalize. Furthermore, the input's and target's distributions along with the proposed compound's descriptors and characteristic values are used to calculate a confidence metric.

[0044] The methods and applications described herein have been limited in scope to the ADME/Tox area. It should be understood that these methods are generally applicable to any research area where chemical structure is to be correlated with some experimental or otherwise determined property. Examples would be QSAR modeling for molecule potency and/or specificity, toxicological profiles of molecules, physicochemical properties of molecules (solubility, melting point), etc.

BRIEF DESCRIPTION OF THE DRAWINGS

[0045] **FIG. 1**. is a block diagram of a system for predicting the ADME/Tox properties of a candidate drug;

[0046] **FIG. 2** is a flow chart of the method for developing a model that will predict the ADME/Tox properties of a candidate drug; and for predicting the ADME/Tox properties of a candidate drug.

[0047] FIGS. **3-45** are individual showings of particular points pertinent and important to the present invention and illustrate specific examples of an embodiment of the invention aimed at predicting human ADME data.

DESCRIPTION OF SPECIFIC EMBODIMENTS

[0048] 1. Definitions

[0049] The following bolded terms are used throughout this document with the following associated meanings:

[0050] Absorption: Transfer of a compound across a physiological barrier as a function of time and initial concentration. Amount or concentration of the compound on the external and/or internal side of the barrier is a function of transfer rate and extent, and may range from zero to unity.

[0051] Affine Regression: Linearly combining input data to approximate output data. This is essentially a linear regression that does not require the regression to go through zero.

[0052] Bioavailability: Fraction of an administered dose of a compound that reaches the sampling site and/or site of action. May range from zero to unity. Can be assessed as a function of time.

[0053] Boosting: A general method which attempts to increase the accuracy of a learning algorithm.

[0054] Compound: Chemical entity. Could be a drug, a gene, etc.

[0055] Computer Readable Medium: Medium for storing, retrieving and/or manipulating information using a computer. Includes optical, digital, magnetic mediums and the like; examples include portable computer diskette, CD-ROMs, hard drive on computer etc. Includes remote access mediums; examples include internet or intranet systems. Permits temporary or permanent data storage, access and manipulation.

[0056] Cross Validation: Used to estimate the generalization error. This method is based on resampling the data set, using randomly (or otherwise chosen) samples of the training set as test sets.

[0057] Data: Experimentally collected and/or predicted variables. May include dependent and independent variables.

[0058] Input Data: Data which is used as an input in the training or execution of a model. Could be either experimentally determined or calculated.

[0059] Target Data: Data for which a model is generated. Could be either experimentally determined or predicted.

[0060] Test Data: Experimentally determined data.

[0061] Descriptor: An element of the input data.

[0062] Committee Machine: A model that is comprised of a number of submodels such that the knowledge acquired by the submodels is fused to provide a superior answer to any of the independent submodels.

[0063] Regression/Classification: Methods for mapping the input data to the target data. Regression refers to the

methods applicable to forming a continuous prediction of the target data, while classification (or in general pattern recognition) refers the methods applicable to separating the target data into groups or classes. The specific methods for performing the regression or classification include where appropriate: Affine or Linear Regressions, Kernel based methods, Artificial Neural Networks, Finite State Machines using appropriate methods to interpret probability distributions such as Maximum A Posteriori, Nearest Neighbor Methods, Decision Trees, Fisher's Discriminate Analysis.

[0064] Mapping: The process of relating the input data space to the target data space, which is accomplished by regression/classification and produces a model that predicts or classifies the target data.

[0065] Feature Selection Methods: The method of selecting desirable descriptors from the input data to enable the prediction or classification of the target data. This is typically accomplished by forward selection, backward selection, branch and bound selection, genetic algorithmic selection, or evolutionary selection.

[0066] ADME: Properties of absorption, distribution, metabolism, and excretion and encompasses other measures related to absorption, distribution, metabolism, and excretion. For example, heptocyte turnover or Caco-2 effective permeability.

[0067] Dissolution: Process by which a compound becomes dissolved in a solvent.

[0068] Fisher's Discriminate Analysis: A linear method which reduces the input data dimension by appropriately weighting the descriptors in order to best aid the linear separation and thus classification of target data.

[0069] Genetic Algorithms: Based upon the natural selection mechanism. A population of models undergo mutations and only those which perform the best contribute to the subsequent population of models.

[0070] Input/Output System: Provides a user interface between the user and a computer system.

[0071] Kernel Representations: Variations of classical linear techniques employing a Mercer's Kernel or variations to incorporate specifically defined classes of nonlinearity. These include Fisher's Discriminate Analysis and principal component analysis. Kernel Representations as used by the present invention are described in the article, "Fisher Discriminate Analysis with Kernels," Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Muller, GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany, ©IEEE 1999 (0-7803-5673-X/99), and in the article, "GA-based Kernel Optimization for Pattern Recognition: Theory for EHW Application," Moritoshi Yasunaga, Taro Nakamura, lkuo Yoshihara, and Jung Kim, IEEE© 2000 (0-7803-6375-2/00), which are both hereby incorporated herein by reference.

[0072] Metabolism: Conversion of a compound (the parent compound) into one or more different chemical entities (metabolites).

[0073] Artificial neural networks: A parallel and distributed system made up of the interconnection of simple processing units. Artificial neural networks as used in the present invention are described in detail in the book entitled, "Neural networks, A Comprehensive Foundation," Second Edition, Simon Haykin, McMaster University, Hamilton, Ontario, Canada, published by Prentice Hall ©1999, which is hereby incorporated herein by reference.

[0074] Permeability: Ability of a barrier to permit passage of a substance or the ability of a substance to pass through a barrier. Refers to the concentration-dependent or concentration-independent rate of transport (flux), and collectively reflects the effects of characteristics such as molecular size, charge, partition coefficient and stability of a compound on transport. Permeability is substance and/or barrier specific.

[0075] Physiologic Pharmacokinetic Model: Mathematical model describing movement and disposition of a compound in the body or an anatomical part of the body based on pharmacokinetics and physiology.

[0076] Principal Component Analysis: A type of non-directed data compression which uses a linear combination of features to produce a lower dimension representation of the data. An example of principal component analysis as applicable to use in the present invention is described in the article, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," Bernhard Scholkopt, *Neural Computation,* Vol. 10, Issue 5, pp. 1299-1319, 1998, MIT Press., and is hereby incorporated herein by reference.

[0077] Simulation Engine: Computer-implemented instrument that simulates behavior of a system using an approximate mathematical model of the system. Combines mathematical model with user input variables to simulate or predict how the system behaves. May include system control components such as control statements (e.g., logic components and discrete objects).

[0078] Solubility: Property of being soluble; relative capability of being dissolved.

[0079] Support Vector Machines: Method which regresses/classifies by projecting input data into a higher dimensional space. Examples of Support Vector machines and methods as applicable to the present invention are described in the article, "Support Vector Methods in Learning and Feature Extraction," Berhard Scholkopf, Alex Smola, Klaus-Robert Muller, Chris Burges, Vladimir Vapnik, Special issue with selected papers of ACNN'98, *Australian Journal of Intelligent Information Processing Systems,* 5 (1), 3-9), and in the article, "Distinctive Feature Detection using Support Vector Machines," Partha Niyogi, chris Burges, and Padma Ramesh, Bell Labs, Lucent Technologies, USA, IEEE ©1999 (0-7803-5041-3/99), which are both hereby incorporated herein by reference.

[0080] 2. Preferred Embodiments

[0081] There are roughly four major properties involved in human pharmacokinetics: Absorption, Distribution, Metabolism, and Elimination (ADME). For example, when a drug is taken into the body orally, the first thing that has to happen is it has to get absorbed into the body in GI tract.

From there, the drug travels to the liver via the portal vein where it is either metabolized or not. After the drug passes through the liver it is distributed throughout the body. Once the drug is distributed throughout the body, it is transported to the kidney to get eliminated. The effectiveness of a drug (a chemical compound) is directly related to the way a body will absorb, distribute, metabolize and eliminate the compound. In addition to the ADME properties of a compound, the toxicological effects of the compound should also be considered. The present invention is directed to systems and methods for predicting various characteristics (ADME/Tox characteristics) related to the way a body will absorb, distribute, metabolize, eliminate, and respond to potential toxic effects of a compound based on the compound's chemical structure and/or associated experimental data.

[0082] The molecular structure of a proposed compound may be input as a 2-dimensional (2D) connection table, which is essentially a two-dimensional graph of how the atoms of a compound are arranged (the structures may actually be 3-dimensional (3D), but may be represented as 2D via well known methods). Alternatively, the structure may be input as a 3D structure. Either 2D or 3D structural representations are desirable inputs for models using structure to predict ADME/Tox characteristics.

[0083] There are really three fundamental properties of the molecule that decide whether or not it's a drug: the first is whether or not it actually interacts with a particular molecular target in the body (in most cases, some kind of protein); the second is whether or not the body can absorb, metabolize, distribute and eliminate the compound adequately, and third, whether or not the compound elicits a toxic response.

[0084] The present invention provides systems and methods for predicting the ADME/Tox properties (e.g., Caco-2 effective permeability or Caco-2 Peff), of a proposed compound through statistical analysis of compound data. By using the present invention, it is therefore possible to significantly reduce the need for expensive and time consuming testing, such as animal testing, because the ADME/Tox characteristics of an untested compound is predicted with a high level of accuracy.

[0085] The first section of the present invention employs mathematical analyses of a diverse compilation of training data (chemical compound data including conventional experimental results, chemical descriptor analysis, etc.) to determine what data relates to the ADME/Tox property to be predicted. Once the type or types of data that are applicable to the ADME/Tox property (descriptors) are determined, mathematical analyses of the selected training data to obtain the selected ADME/Tox characteristic for each training data compound are performed in order to create a model. The model can then be used to predict a proposed compound's ADME/Tox property by inputting the same type of data for the proposed compound into the model. Running the model with the proposed compound's descriptors produces the predicted ADME/Tox characteristic.

[0086] Models are only as good as the input assay and test data, and therefore, a key to producing highly accurate predictions is the use of well-defined standard operating procedures for generating data as well as insuring that the data has a good distribution. Therefore, the present invention provides a method for collecting and compiling a diverse training data set to be used to mathematically predict the ADME/Tox characteristics of a proposed chemical compound.

[0087] The input data is collected and/or calculated for a variety of chemical compounds preferably representing currently prescribed drugs as well as failed drugs and potential new drugs (this is a continual process, since as more data is collected, the resulting models will have improved performance). Assay data may be collected from well established sources or derived by conventional means. For instance, in vitro assays characterizing permeability and transport mechanisms may include in vitro cell-based diffusion experiments and immobilized membrane assays, as well as in situ perfusion assays, intestinal ring assays, incubation assays in rodents, rabbits, dogs, non-human primates and the like, assays of brush border membrane vesicles, and averted intestinal sacs or tissue section assays. In vivo assay data typically are conducted in animal models such as mouse, rat, rabbit, hamster, dog, and monkey to characterize bioavailability of a compound of interest, including distribution, metabolism, elimination and toxicity. For high-throughput screening, cell culture-based in vitro assays or biochemical assays from isolated cell components or recombinantly expressed components are preferred. For high-resolution screening and validation, tissue-based in vitro and/or mammal-based in vivo data are preferred.

[0088] Cell culture models are preferred for high-throughput screening, as they allow experiments to be conducted with relatively small amounts of a test sample while maximizing surface area and can be utilized to perform large numbers of experiments on multiple samples simultaneously. Cell models or biochemical assays also require fewer experiments since there is no animal to animal variability. An array of different cell lines also can be used to systematically collect complementary input data related to a series of transport barriers (passive paracellular, active paracellular, carrier-mediated influx, carrier-mediated efflux) and metabolic barriers (protease, esterase, cytochrome P450, conjugation enzymes).

[0089] Cells and tissue preparations employed in the assays can be obtained from repositories, or from any eukaryote, such as rabbit, mouse, rat, dog, cat, monkey, bovine, ovine, porcine, equine, humans and the like. A tissue sample can be derived from any region of the body, taking into consideration ethical issues. The tissue sample can then be adapted or attached to various support devices depending on the intended assay. Alternatively, cells can be cultivated from tissue. This generally involves obtaining a biopsy sample from a target tissue followed by culturing of cells from the biopsy. Cells and tissue also may be derived from sources that have been genetically manipulated, such as by recombinant DNA techniques, that express a desired protein or combination of proteins relevant to a given screening assay. Artificially engineered tissues also can be employed, such as those made using artificial scaffolds/matrices and tissue growth regulators to direct three-dimensional growth and development of cells used to inoculate the scaffolds/matrices. It will be understood that ideally any known test results could be added to a test data set in order to adjust the model or to provide a new property to solve towards.

[0090] The drugs (compounds) selected should be as diverse in character as possible. Therefore, the compounds

may be analyzed and defined in chemical space. Chemical space can be represented as an N-base coordinate system in which to plot compounds and may be used to show the diversity of a sample of compounds. The axes of N-base coordinate system may be selected from all or some of the input data. Drugs may be eliminated from a particular training data set (the training data may be grouped to solve for a particular ADME/Tox property) if it is determined that they bias the training data set.

[0091] In the present invention, a collection of drugs have been plotted in a six-base chemical space (see **FIG. 3**). The axes of the six-base are physicochemical descriptors that were selected so that the best separation of known drugs is maintained. Data is also selected from combinatorial libraries of chemicals which are near neighbors for each of the drugs creating an extended data set. The compounds are ideally each tested for various ADME/Tox characteristics or properties to be predicted, however it is not necessary to test every compound for actual results.

[0092] There are many considerations for the experimental data. Each data set of experimental data is analyzed to decide how it is going to be used in model building. For example, is it appropriate to use a certain data set to predict absolute values of compounds or is there too much error in the data set? If there is not enough data in a data set to cover a particular range (either coverage in the data space, representation in the data space, or certainty in the data space) it is possible to put the data into bins, such as 0 to 20, 21 to 40, 41 to 60, 61 to 80, 81 to 100. Alternatively, the data may require scaling correction to account for systematic variations in the data. One having ordinary skill in the art will readily understand the grouping of experimental data, scaling and systematic variations used to adjust a data set.

[0093] Next, a tool is used to calculate additional data by analyzing each compound and describing the compound with chemical descriptors. Chemical descriptors are well known in the art of modeling compounds, and may be determined by analyzing a 2D or 3D structure of a compound.

[0094] Finally, all the training data (input and target data) collected or created is compiled and preferably maintained in a relational database or other known means for making the data easily accessible and available to be manipulated and analyzed in accordance with the present invention.

[0095] The present invention is now described with reference to **FIG. 1**. In particular, system **100** includes a processor facility **102** and a data facility **104** coupled to a network **106**. The processor facility **102** may be a conventional computer, such as a PC, configured to access database facility **104** and to execute analytical software in accordance with the present invention. Database facility **104** may be a conventional database server running a database engine, such as SQLSERVER® or ORACLE 8i® and is configured to maintain and to serve data, such as the test data described above. The data may be stored and maintained by any means such as in a relational dataspace or an objected oriented dataspace.

[0096] The present invention includes analytical tools which may be executed on processor facility **102**. The analytical tools may be in the form of software that is loaded locally on processor facility **102** or may be served via a server **108** (e.g., an HTML form, JAVA program, etc. served on a web server), which optionally may be included. Accordingly, a client facility **110** may be connected to the network **106**, which may include parts of the Internet and World Wide Web (WWW), or local area networks (LANS). The client facility **110** could be a web browser or other terminal configured to access and run the analytical tools remotely or to download the analytical tools (e.g., via HTML, IIOP, etc.) via network **106** and run them locally.

[0097] The configuration of system **100** is merely exemplary and is not meant to limit the present invention. It will be appreciated that the present invention may take many forms and configurations. For example, the present invention may be implemented via a software solution including a database and forms configured to run on a stand-alone PC, or may alternatively be a combination of software and firmware, and may be implemented in a client-server, stand-alone or web configuration.

[0098] The operational aspects of the present invention are now described with reference to the flow chart in **FIG. 2**. The flow chart represents two independent starting pathways which meet at step **S2-5**, a model development pathway, and a model execution or prediction pathway, these two initial pathways will be described independently.

[0099] Model Development Pathway (**S2-1***a*->**S2-5**)

[0100] The model development pathway begins in step **S2-1***a* and immediately proceeds to step **S2-2***a*. At step **S2-2***a,* the ADME/Tox property to be predicted is selected. For example, it may be desired to predict the Caco-2 Peff of the compound, or the FDP (fraction of the dose administered that is absorbed at the portal vein). The system might allow for the selection to be from a table, radio group, pop-list, or by any known means. Also at step **S2-2***a*, a set of training compounds appropriate for developing the selected ADME/Tox property model is entered into the system. Many compound descriptors may be entered or calculated, such as molecular weight, structure, specific gravity, etc.

[0101] Next, at step **S2-3***a,* a group of meaningful input data is selected based on the property to be predicted or a related performance metric using feature selection methods. For example, a genetic algorithm coupled with a regression/classification method, such as a neural network, may be used to build many models predicting the Caco-2 Peff of a compound. Features are then selected from the resulting models with the objective of choosing the smallest number of dimensions that effectively describe the model space. One should keep in mind when performing the analyses to select a number of descriptors which avoids biased and non-predictive models (e.g., overtraining).

[0102] Once the descriptors have been selected, a model is created at step **S2-4***a* by using regression/classification methods to map the input data to the ADME/Tox property to be predicted. The modeling effort may involve Affine Regressions, Nearest Neighbor Methods, Discriminate Analysis, Support Vector Machines, Artificial neural networks, Data Compression techniques (targeted and non-targeted), Genetic Algorithms, and Boosting. In addition, a method for calculating a confidence metric is created by analyzing information related to the model such as the distributions and values of the input and target data and the methods involved in building the model.

[0103] It should be noted that instead of predicting continuous values for a specific ADME/Tox property, the present invention may be used to classify a particular compound (e.g., can it be absorbed, is it toxic, etc.). A compound is classified by the same method predicting a specific ADME/Tox property, except that the analyses performed may vary slightly, and the classifications are performed to solve for a "yes/no" or "high, medium, low" binning type solution (e.g., 1-bit).

[0104] The model resulting from step S2-4*a* is used in step S2-5 to predict new proposed compounds in the model execution pathway.

[0105] Model Execution Pathway (S2-1*b*->S2-7)

[0106] Once the model has been created/developed, then the model may be used to predict the ADME/Tox property of the proposed compound. The model execution pathway begins at step S2-1*b*, and proceeds directly to S2-2*b* where at least one proposed compound may be entered.

[0107] Next, at step S2-3*b*, the property to be predicted is selected. For example, it may be desired to predict the Caco-2 Peff of the compound, or the FDP. The system might allow for the selection to be from a table, radio group, pop-list, or by any known means.

[0108] Next, at step S2-5, the descriptors for the proposed compound (identified in step S2-3*a*)) are input into the model created in step S2-4*a*. The model is run and a result (e.g., a Caco-2 Peff or FDP prediction) is produced in step S2-6. As described above, a measure of confidence in the result may also be produced.

[0109] Processing terminates at step S2-7.

[0110] It should be readily apparent to one having ordinary skill in the art that the preceding method may be implemented via numerous configurations. For example, the preceding method and analysis therein may be implemented via a C++ program coupled to a data warehouse, or alternatively may be implemented via a combination of program components and databases.

[0111] Heretofore, only highly trained pharmacokinetic experts were capable of determining and therefore, estimating a compound's ADME/TOX. Moreover, such estimations usually included very time consuming and costly experimentation. The present invention now provides a less expensive and time consuming, and potentially more accurate means for predicting the ADME characteristics of proposed drugs, and therefore, by using the present invention, many individuals and entities will now be able to more affordably screen compounds for their applicability as drugs before any animal testing or other lab testing is necessary.

[0112] All publications and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication or patent application was specifically and individually indicated to be incorporated by reference.

[0113] The invention now being fully described, it will be apparent to one of ordinary skill in the art that many changes and modifications can be made thereto without departing from the spirit or scope of the invention.

We claim:

1. A method for developing a model to predict a chemical compound property, the method comprising:

obtaining at least one descriptor from structural data for each of a plurality of compounds;

obtaining at least one descriptor from experimental or predicted data for each of a plurality of compounds;

obtaining at least one chemical compound property for each of the plurality of compounds; and

developing the model by mapping the descriptors to the chemical compound property.

2. The method of claim 1, wherein the chemical property is an ADME property.

3. The method of claim 2, wherein the ADME property is absorption.

4. The method of claim 2, wherein the ADME property is Caco-2 Effective Permeability.

5. The method of claim 1, wherein the chemical property is a toxicity property.

6. The method of claims **1-5** wherein obtaining at least one descriptor comprises selecting the descriptors applicable to the characteristic to be predicted based on an analysis of the plurality of compounds.

7. The system of claim 6 wherein the analysis used to select the descriptors for predicting the characteristic is selected from at least one of the following: Affine Regressions, Kernel Methods, Artificial neural networks, Finite State Machines—Maximum A Posteriori, Nearest Neighbor Methods, Fisher's Linear Discriminate Analysis, or other regression/classification methods.

8. The system of claim 6 further comprising:

performing a chemical space analysis of the plurality of compounds;

if the chemical space analysis indicates that the plurality of compounds selected should be modified to improve diversity of the chemical space, then modifying the plurality of compounds by addition or deletion of a compound to improve the diversity of the chemical space covered by the plurality of compounds.

9. A system for predicting an ADME/Tox of a compound in a mammalian body, the system comprising:

a database facility, the database facility configured to store and to provide structural and experimental or predicted data; and

a processor facility, the processor facility configured to allow the entry of data relating to a new proposed chemical compound including structural data and experimental or predicted data, to perform an analysis of the chemical compound by mapping the data entered to produce a predicted ADME/Tox property of the chemical compound based on the analysis.

10. A method for compiling chemical compound data to be used for evaluating the characteristics of a proposed compound, the method comprising:

selecting a plurality of compounds;

obtaining a descriptor analysis for each of the plurality of compounds;

obtaining test results related to the characteristics being evaluated; and

loading the descriptor analysis and the test results into a database used to predict the characteristics of proposed compounds.

11. The method of claim 10 further comprising:

performing a chemical space analysis of the plurality of compounds;

if the chemical space analysis indicates that the plurality of compounds selected should be modified to improve diversity of the chemical space, then modifying the plurality of compounds by addition or deletion of a compound to improve the diversity of the chemical space covered by the plurality of compounds.

12. A system for predicting the chemical properties of a proposed compound comprising:

a database facility configured to store and to serve data relating to the characteristics of selected compound, including structure data, descriptor data, and test data; and

a processor facility coupled to the database facility and configured to predict the characteristics of a proposed compound by:

(a) receiving at least one proposed compound via a user input means;

(b) selecting training compounds from the database facility based on the characteristics to be predicted of the proposed compounds;

(c) selecting the most meaningful descriptors applicable to the characteristic to be predicted based on an analysis of the training compounds selected in step (b);

(d) creating validation data subsets of the training data based upon the distribution of descriptors and target characteristics of compounds selected in (b/c);

(e) mapping the training set obtained in (d) to the target data resulting in a model which could predict the target data of a proposed compound;

(f) modifying (for example: boosting, bootstrap aggregation (bagging)), and other model enhancement methods, etc.) one or more models produced in (e) based upon performance on validation sets obtained in (d) to form a composite model;

(g) combining (via boosting, committee machines etc,) a set of two or more models produced in (e or f) based upon performance on validation sets obtained in (d) to form a composite model; and

(h) running the model determined in either step (e), (f) or (g) using the required input data (the identity of the subset of input data itself was determined in step (c)) to predict the required target data.

13. The system of claim 12 wherein the analyses consider model biases and over training.

14. A method for predicting a characteristic of a chemical compound, the method comprising:

receiving as an input structure data for the compound; and

mapping the data to at least one chemical characteristic.

15. A predictive model of a chemical compound property produced according to the method of any of claims **1-3**.

16. A computer readable medium containing a chemical compound characteristic model, the medium comprising:

a computer readable medium; and

a data structure on the medium that generates at least one characteristic for a compound from structure data and experimental or predictive data for the compound.

17. The medium of claim 16, wherein the characteristic is an ADME property.

18. The method of claim 17, wherein the ADME property is absorption.

19. The method of claim 16, wherein the characteristic is a toxic property.

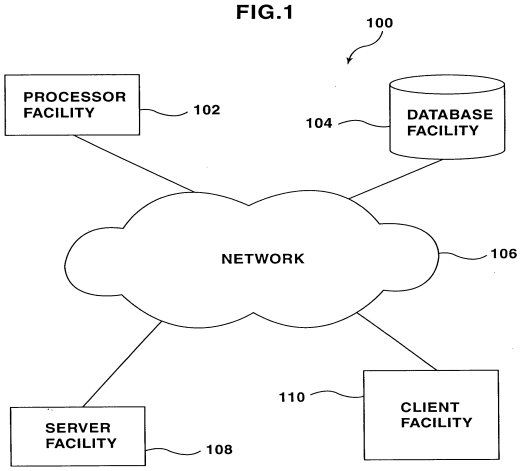
* * * * *

专利名称(译)	用于预测化合物的adme / tox特征的系统和方法		
公开(公告)号	US20040009536A1	公开(公告)日	2004-01-15
申请号	US10/332997	申请日	2001-07-30
[标]申请(专利权)人(译)	乔治·GRASS LEESMAN GLENÐ 诺里斯丹尼尔 SINKO PATRICK ATHWAL杰汉吉尔 SAGE卡尔顿 BREMER TROY HOLME KEVIN		
申请(专利权)人(译)	乔治·GRASS LEESMAN GLENÐ 诺里斯丹尼尔 SINKO PATRICK ATHWAL杰汉吉尔 SAGE卡尔顿 BREMER TROY HOLME KEVIN		
[标]发明人	GRASS GEORGE LEESMAN GLEN D NORRIS DANIEL SINKO PATRICK ATHWAL JEHANGIR SAGE CARLETON BREMER TROY HOLME KEVIN		
发明人	GRASS, GEORGE LEESMAN, GLEN D NORRIS, DANIEL SINKO, PATRICK ATHWAL, JEHANGIR SAGE, CARLETON BREMER, TROY HOLME, KEVIN		
IPC分类号	G01N31/00 G01N33/48 G01N33/50 G01N33/53 G01N33/567 G06F19/00 G06G7/48 G06G7/58		
CPC分类号	G06F19/707 G06F19/704 G16C20/30 G16C20/70		
外部链接	Espacenet USPTO		

摘要(译)

一种开发化合物性质预测模型的方法。该方法包括从多个化合物中的每一个的结构数据获得至少一个描述符。对于多种化合物中的每一种，获得至少一种化学化合物性质。通过将至少一个描述符映射到化合物属性来开发预测模型。化学化合物性质可以是ADME性质。 ADME属性可能是吸收。化学化合物性质也可以是毒性。

FIG.1


100
PROCESSOR FACILITY
102
104
DATABASE FACILITY
NETWORK
106
110
CLIENT FACILITY
SERVER FACILITY
108