創源生技
GGA

AI於藥物開發的應用

創源生技 分子視算中心

經理 陳冠文 (Gene)

EXTEDO eRegulatory Affairs
DASSAULT SYSTEMES
BIOVIA
VALGENESIS
ELSEVIER
excelra
QIAGEN

About Gene

- 國立清華大學分子與細胞生物研究所畢
- 資格
  - BIOVIA Discovery Studio & Pipeline Pilot Trainer & Support
  - Qiagen Ingenuity Pathway Certified Analyst
- 工作經歷
  - 台灣藥學會翻譯
  - 創源生物科技股份有限公司
    - 專員 (2013-2014)
    - 主任 (2014-2015)
    - 副理 (2015~2018)
    - 經理 (2018~迄今)
  - Product Manager:
    - BIOVIA Discovery Studio
    - BIOVIA Pipeline Pilot
    - BIOVIA Registration
    - BIOVIA Insight
    - QIAGEN Ingenuity Pathway Analysis
    - QIAGEN Ingenuity Variant Analysis
    - Excelra GOBIOM / GOSTAR
    - Elsevier Pathway Studio
- 專長
  - 資料庫設計、科學資訊流程管理、酵素結構優化 & 抗體設計、小分子新藥開發

Copyright©2018 GGA Corp., All rights reserved.

Projects and Publications

- Projects
  - Optimize the Industrial Design Parameters by Machine Learning Techniques - 2018
  - A New Macrocyclin Reacts as Ribosome Inhibitor for Helicobacter *pylori* – 2018 paper revised
  - NOX2 Homology Modeling and Structural Analysis - 2018
  - Drug Discovery for CXCR4 antagonists - 2018
  - Epitope Prediction for the Complex of N-Glycan and mAB - 2018
  - Automatics Workflow for Building Pharmacophore Database – 2018
  - Molecular Docking & Mutation-Based Structural Modification Study - 2018
  - In-silico Computation for Antibody-Drug Conjugates (ADCs) Modeling and Simulation – 2017
  - Identification of Imidazole-derivative Like Inhibitors against X using Molecular Docking Approach - 2017
  - Structural Determination for Inhibitors of Dual-Specificity Tyrosine Phosphorylation-Regulated Kinases – 2017
  - Identification of GPCR Binding Site by 2D and 3D Simulation Tools - 2017
  - Toward to Precision Medicine: Three Different Computational Approaches for Promoting Therapeutic Quality - 2016
  - Modeling and Virtual High Throughput Screening for Discovery of Novel Human X Leads with BIOVIA Discovery Studio® - 2016

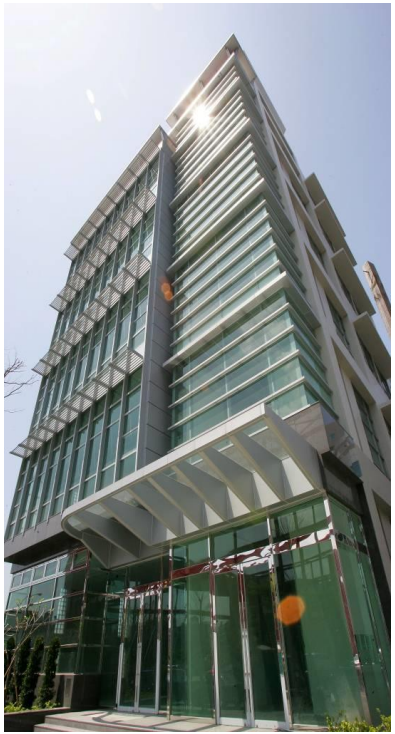
Copyright©2018 GGA Corp., All rights reserved.

Projects and Publications

- Projects
  - Predict the *Cyprinus carpio* Protein X Function by using Molecular Dynamics tools – 2016
  - A Novel Concept to Construct Antibody with “Arms” - 2016
  - Modeling & Structural Optimization of X Synthetic Enzyme X with BIOVIA® Discovery Studio® - 2015
  - Predict the Antibody Binding Epitopes on the Plasminogen structure with Biovia® Discovery Studio® - 2015
  - Modeling & Simulating Protein Pentamer with Accelrys® Discovery Studio® - 2015
  - Generating 3D-QSAR Pharmacophore Model for Human Na+/K+ - ATPase with BIOVIA® Discovery Studio® - 2015
  - In silico screening of lead compound against c-Raf using BIOVIA® Discovery Studio® - 2015
  - Structural Elucidation of the Human Ig gamma1 chain C region and Insight into the Identification of Ligand Binding Site with BIOVIA Discovery Studio® - 2015
  - Discover Protein Specific Binding Site and Potential Inhibitor - 2014
  - Development of eIRB and eCTO System – 2014 to Present
  - Project Member for building pediatrics-using drugs database and genetic-test platform. Responsible for surveying over 20 marketing NSAIDs, including drug mechanism, pharmacokinetics properties, doctor used prescription and labeled instructions -2014
  - Project Manager for developing ultrasound software - 2013
- Publications
  - Molecular Simulation of HER2/neu Degradation by HSP90 Inhibitors, 2008, JCCS.
  - A Novel Strategy for the Structure-based Drug Design of Heat Shock Protein 90 Inhibitors, 2008, IEEE CIM.

Copyright©2018 GGA Corp., All rights reserved.

GGA is part of the BIONET Group (訊聯生物科技)



CEO: Christopher Tsai, Ph.D. 蔡政憲 博士

Established: Nov. 2008

Main Product & Service Areas:

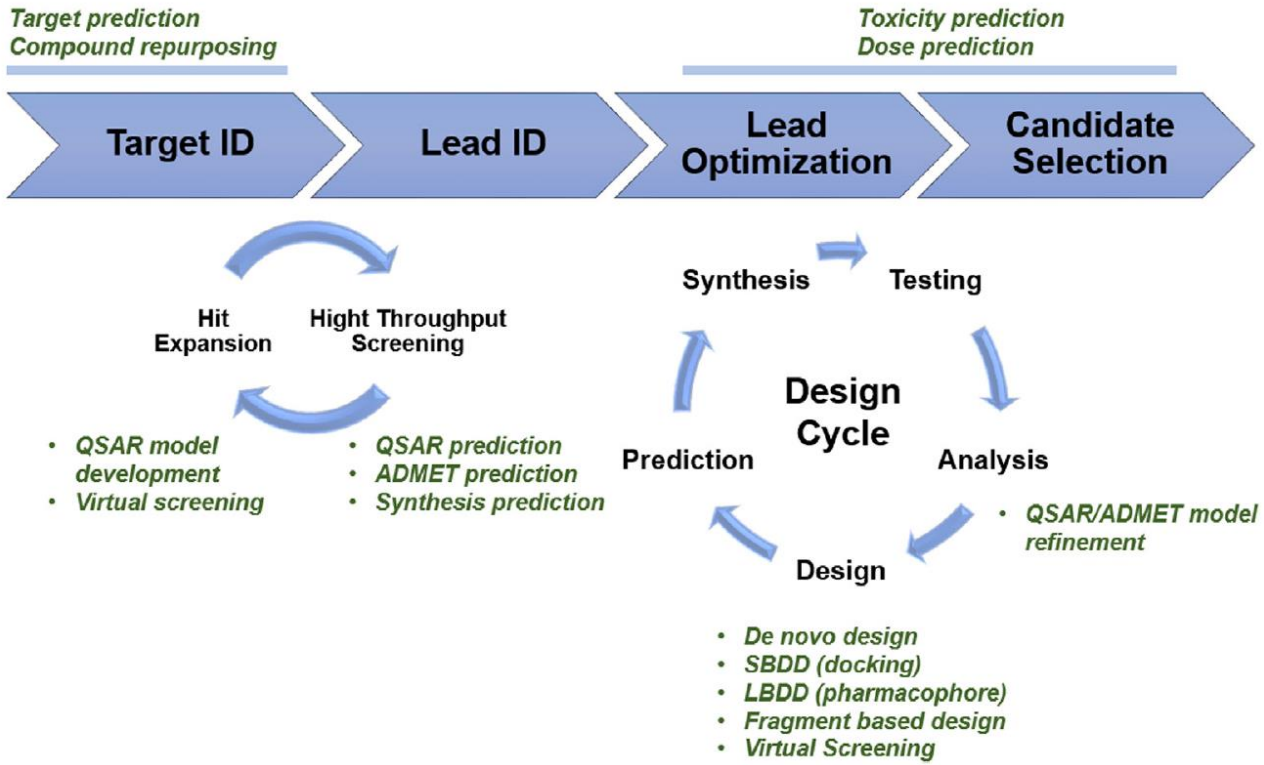
1. Genetic Testing
2. Molecular Diagnosis
3. **Scientific Informatics & Bio IT**

IPO Date: September 17, 2012

Stock Ticker: 4160 (Taiwan OTC)

Copyright©2018 GGA Corp., All rights reserved.

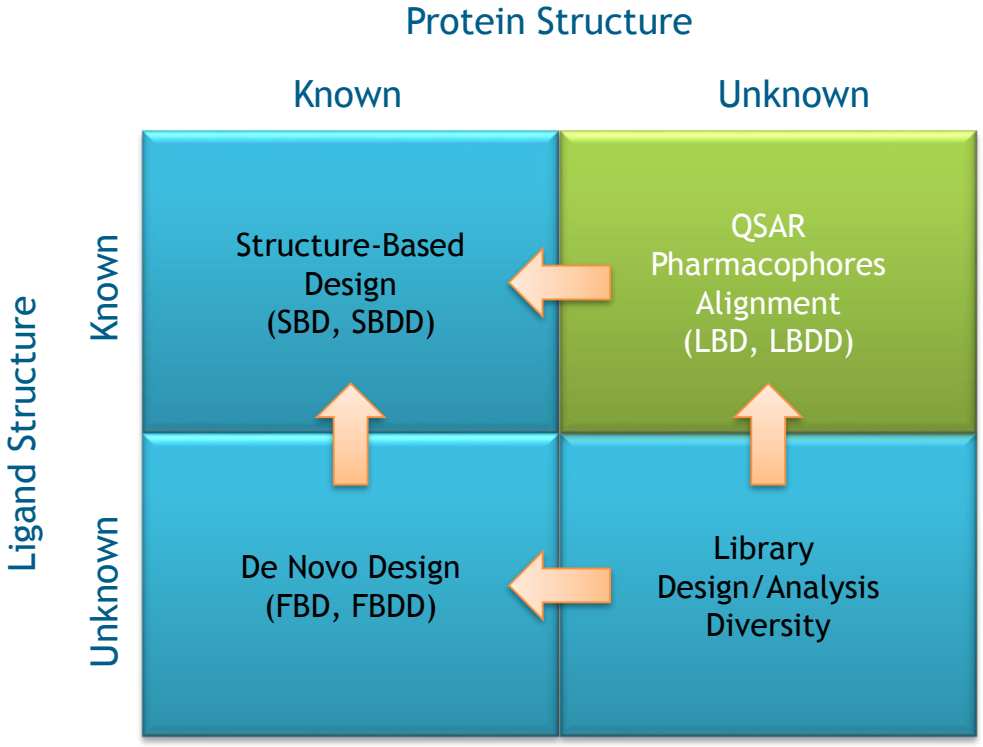
Application of artificial intelligence and machine learning tools



Bioorganic & Medicinal Chemistry Letters 28 (2018) 2807–2815

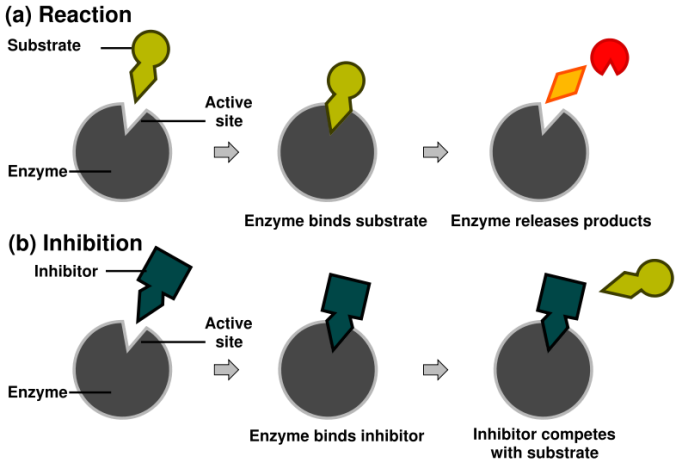
Copyright©2018 GGA Corp., All rights reserved.

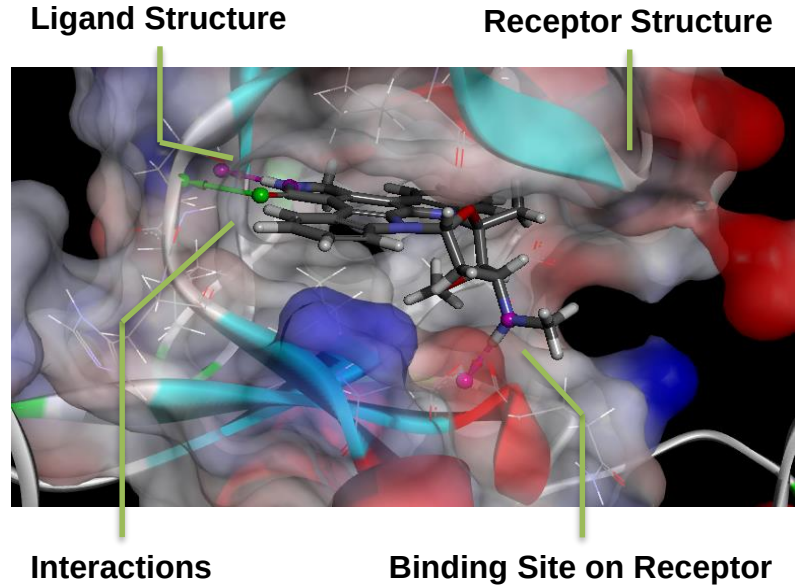
Drug Design Scenarios



Copyright©2018 GGA Corp., All rights reserved.

Docking & Scoring – Structure-Based Design





Copyright©2018 GGA Corp., All rights reserved.

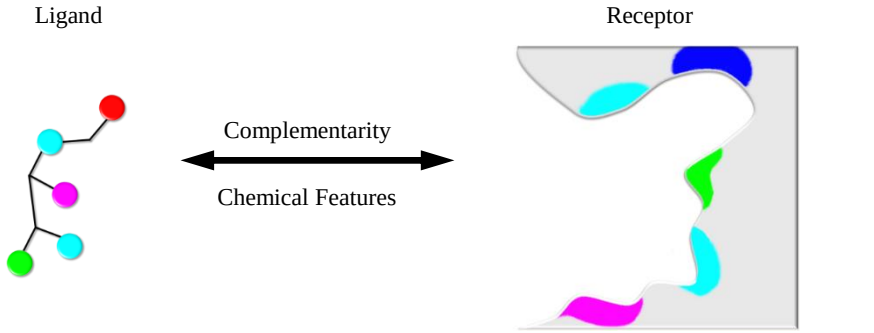
Ligand-Based Design

"The molecular framework that carries (*phoros*) the essential features responsible for a drug's (*pharmacon*)biological activity."

P. Ehrlich, *Dtsch. Chem. Ges.* **1909**, *42*, 17

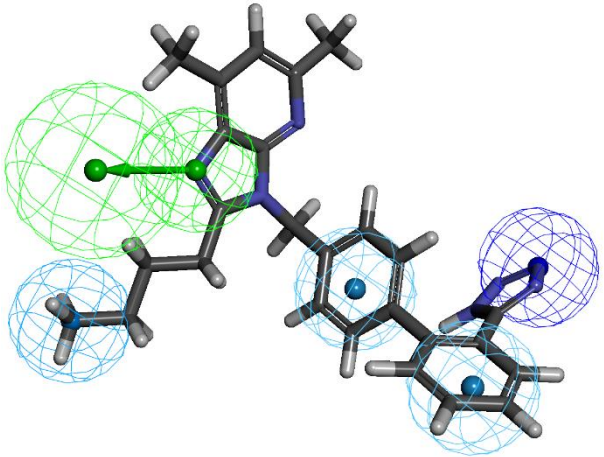
"A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response."

Wermuth et al. Pure Appl. Chem. 1998 70, 1129-43

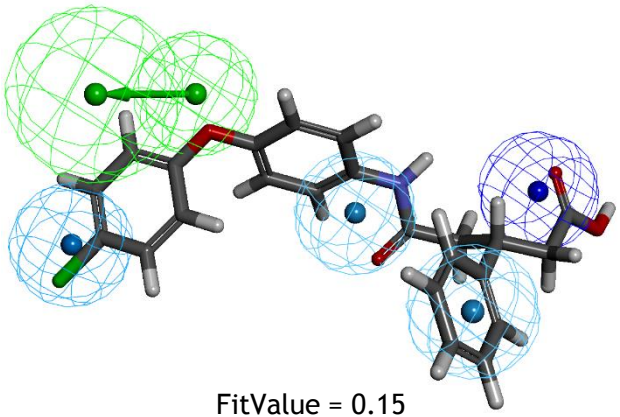


Copyright©2018 GGA Corp., All rights reserved.

Ligand-Pharmacophore Scoring



FitValue = 4.44

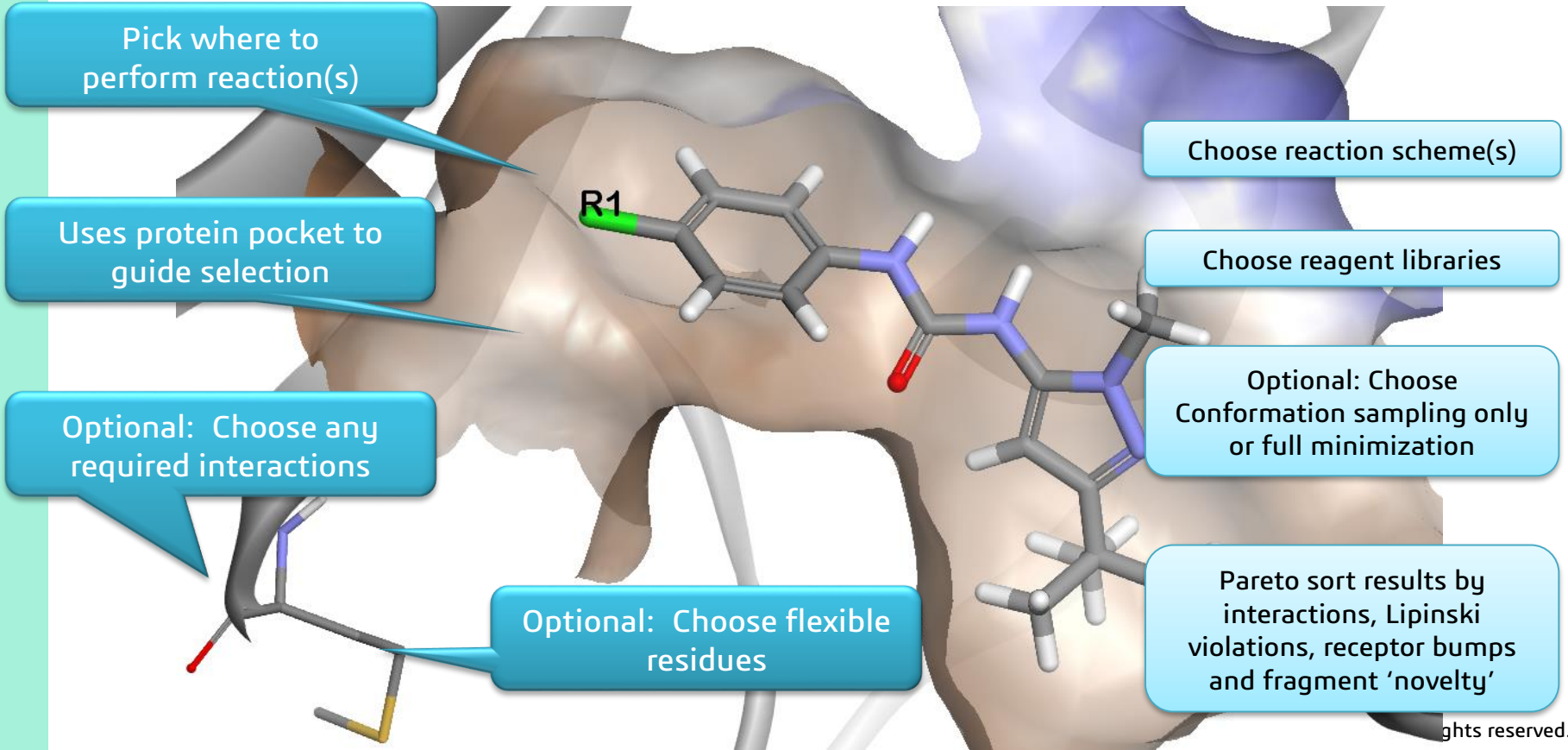


FitValue = 0.15

When comparing a ligand and a pharmacophore, how is the quality of the matching (mapping) determined and quantified?

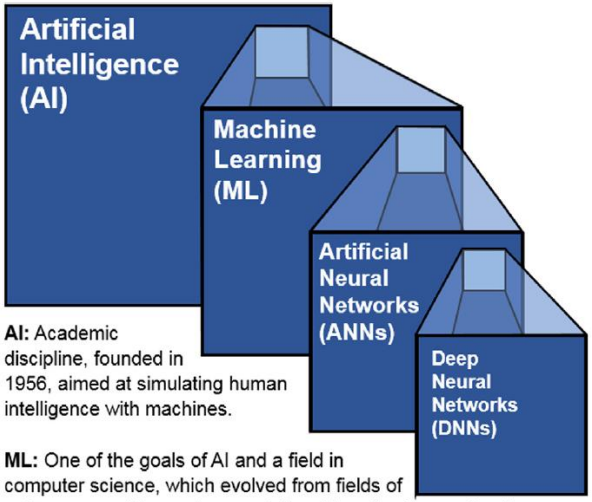
Copyright©2018 GGA Corp., All rights reserved.

Fragment-Based Design



ghts reserved.

Key definitions and relationships between artificial intelligence and deep learning



AI: Academic discipline, founded in 1956, aimed at simulating human intelligence with machines.

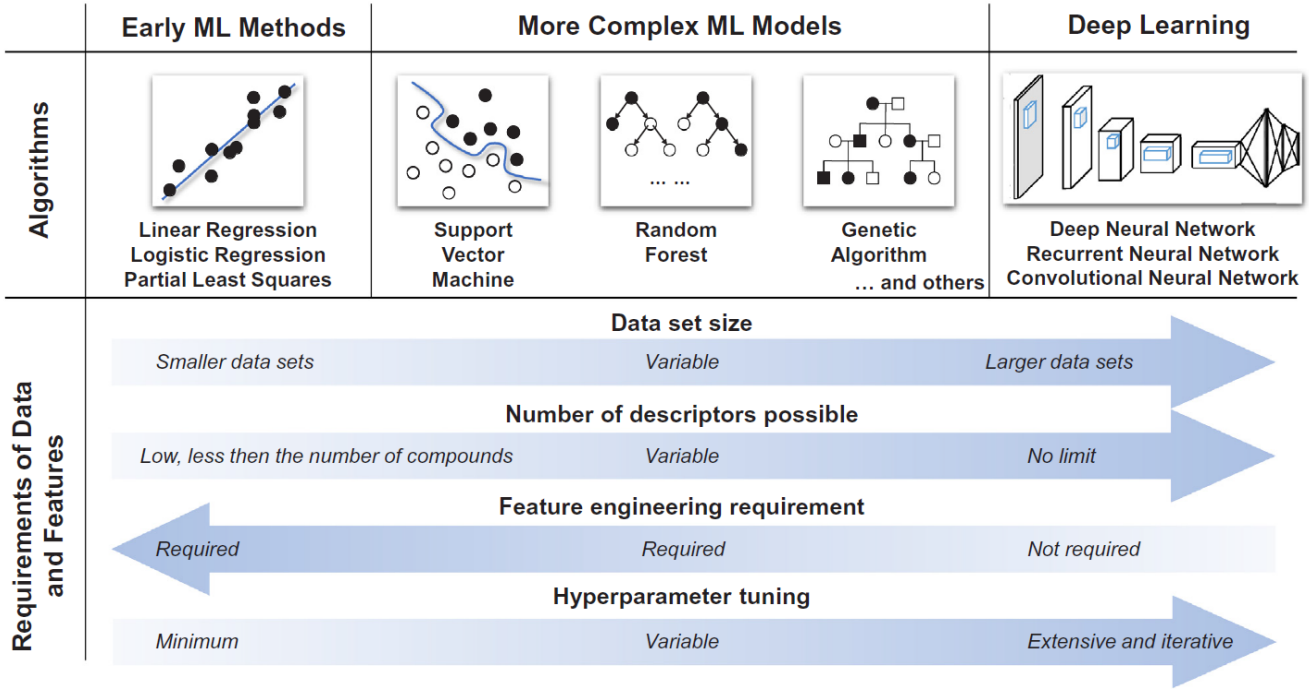
ML: One of the goals of AI and a field in computer science, which evolved from fields of pattern recognition and computational learning theory. Applies statistical techniques to allow models to progressively improve performance.

ANN: A learning algorithm loosely inspired by biological neural networks; a non-linear statistical data modeling tool applied in ML tasks.

DNN: Platforms composed of multiple neural networks interacting in different ways. **Deep learning (DL)** is a class of ML algorithms which use DNN architectures to accomplish learning tasks.

Copyright©2018 GGA Corp., All rights reserved.

Evolution of machine learning methods employed in predictive model development and associated data requirements

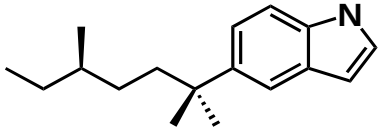


Bioorganic & Medicinal Chemistry Letters 28 (2018) 2807–2815

Copyright©2018 GGA Corp., All rights reserved.

Predict compound properties by Statistic method: QSAR

- Quantitative Structure-Activity Relationships



Chemical Data



IC_{50} 0.4 μM

Numerical Properties

Molecular Descriptor

MW	AlogP	Surface Area	Dipole	…
243.39	5.601	260.587	0.662	…

Define Correlation

$$IC_{50} = f([MW],[A\log P],\cdots)$$

Copyright©2018 GGA Corp., All rights reserved.

Machine Learning Creates Models from Data

Machine learning (ML) always starts with a set of data.

From the data set, an ML algorithm derives rules to predict one property (Y) from other properties (X's).

Molecule	tox	ALogP	Molecular_Weight	Num_H_Acceptors	Num_H_Donors	Num_Rings
	nontoxic	1.6910	171.28	2	1	1
	toxic	1.3990	222.33	4	2	2
	nontoxic	2.6680	382.18	4	4	0
	nontoxic	2.9060	307.31	4	4	0

Continuous Y → Regression model

Categorical Y → Classification model

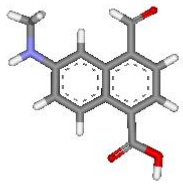
We call this set of rules (or equation) a model.

We build models to:

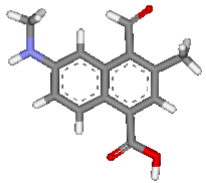
- Better understand our data
- Make predictions for new data

Copyright©2018 GGA Corp., All rights reserved.

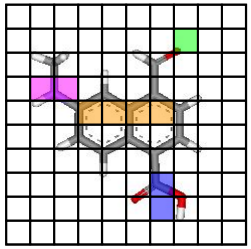
Example: Pharmacophore

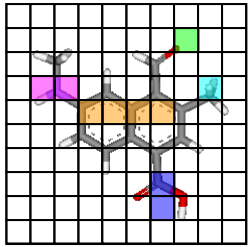


Active 1



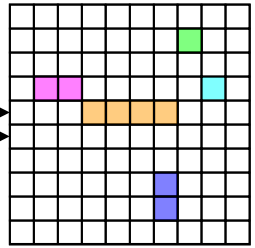
Active 2





01011010…

01011011…

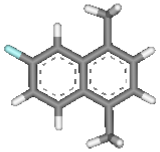


01011011…

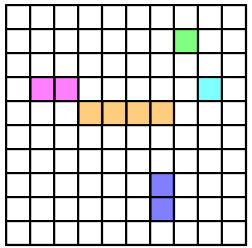
- HB Acceptor
- HB Donor
- Hydrophobe
- Ring Aromatic
- Negative Ionizable

Copyright©2018 GGA Corp., All rights reserved.

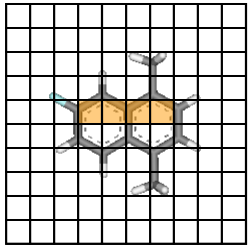
Example: Pharmacophore



Inactive

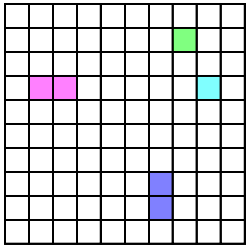


01011011…



00010001…

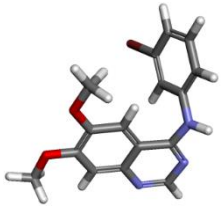




01001010…

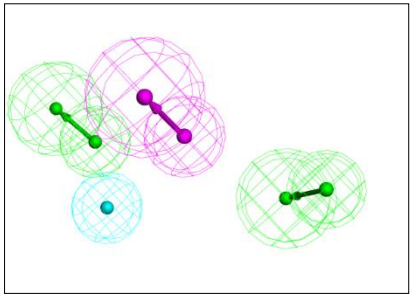
Copyright©2018 GGA Corp., All rights reserved.

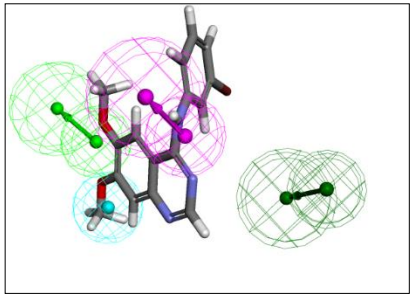
Model fitting



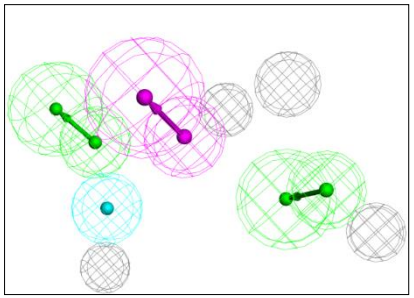
Activ = 450

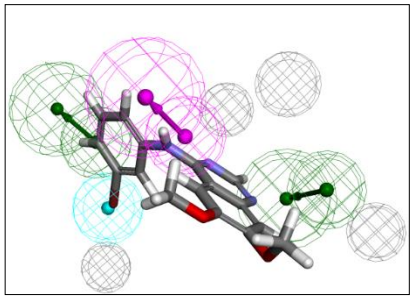
Ligand Pharmacophore Mapping (Maximum Omitted Features = -1)





Estimate = 19 / FitValue = 6.6



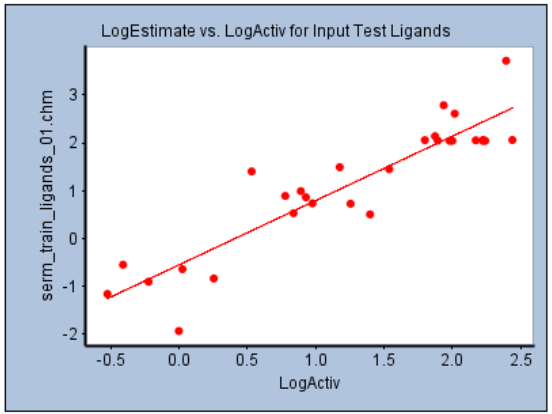


Estimate = 660 / FitValue = 5.1

Copyright©2018 GGA Corp., All rights reserved.

How to check if a model fit is good?

Validation Result Using Input Test Ligands			
Model Name	**q^2**	**RMS Error**	**Mean Absolute Error**
serm_train_ligands_01.chm	0.832	0.632	0.457



- The R^2 statistic has become the almost universally standard measure for model fit in linear models.
- What is R^2?

$$R^2 = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y})^2}$$

← Model error

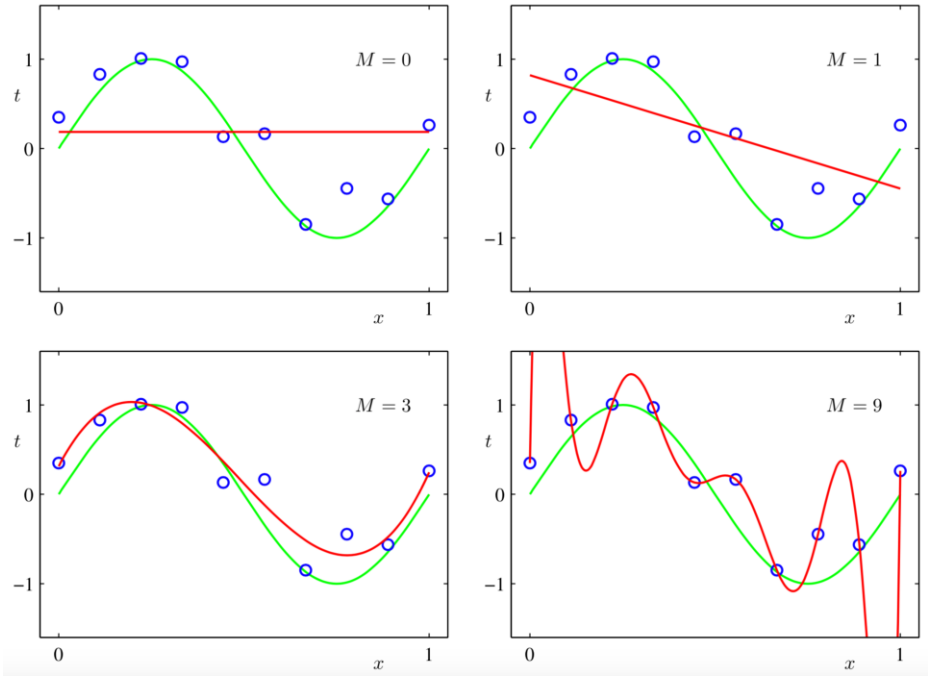
← Variance in the dependent variable

- It is the ratio of error in a model over the total variance in the dependent variable.
- Hence the lower the error, the higher the R^2 value.

Copyright©2018 GGA Corp., All rights reserved.

OVERFITTING

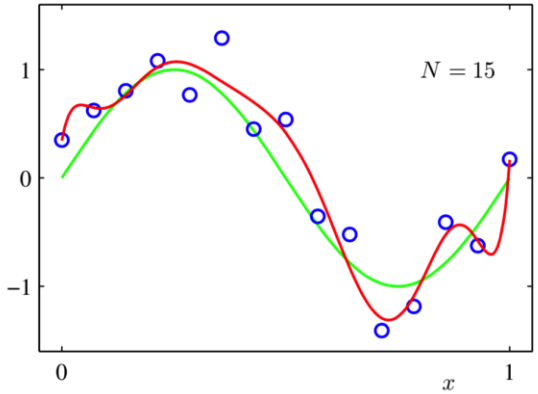
- What are the consequences of overfitting?
- The noise is caught, but the signal isn't.
- "*Overfitted models will have high R^2 values, but will perform poorly in predicting out-of-sample cases*"

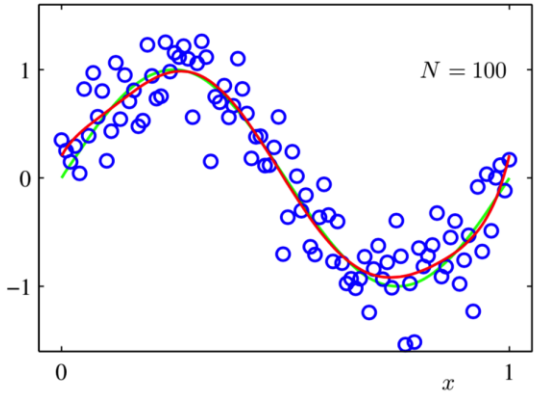


Copyright©2018 GGA Corp., All rights reserved.

OVERFITTING

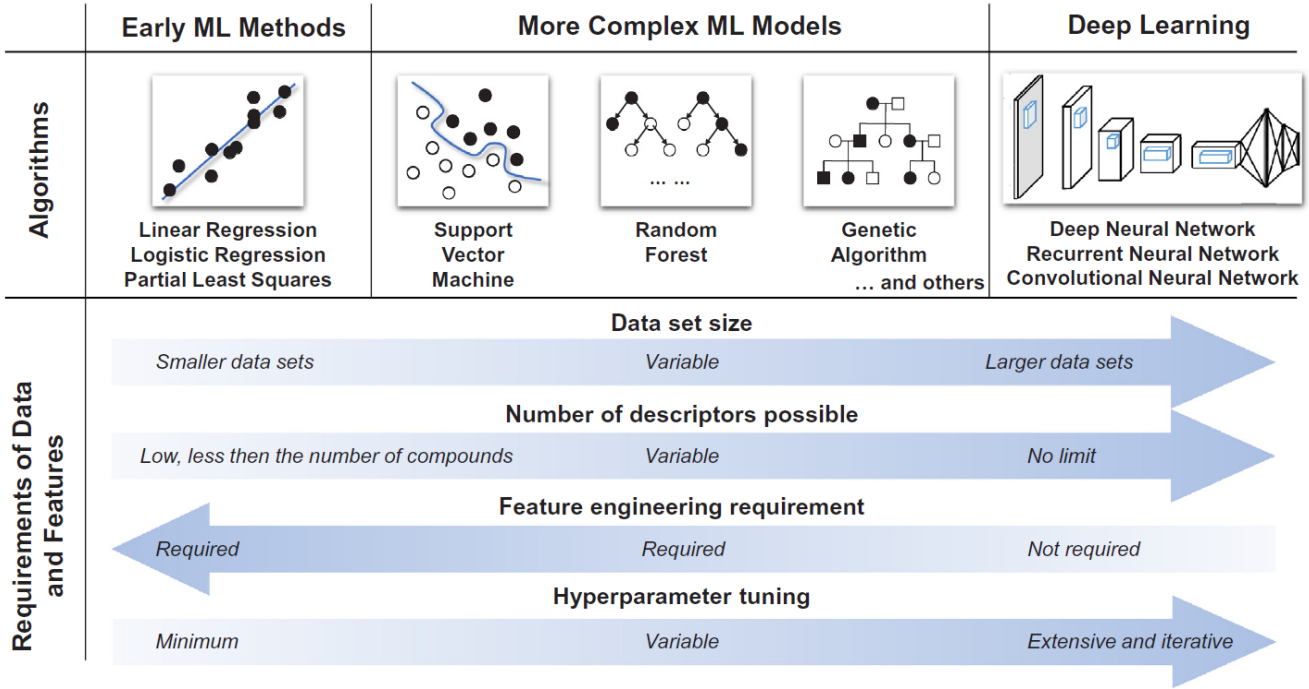
- What are the consequences of overfitting?
- *"Overfitted models will have high R^2 values, but will perform poorly in predicting out-of-sample cases"*
- Implement regularization
- Need more data





Copyright©2018 GGA Corp., All rights reserved.

Evolution of machine learning methods employed in predictive model development and associated data requirements

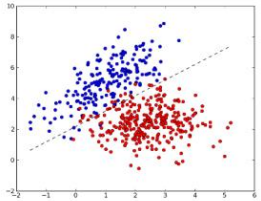


Bioorganic & Medicinal Chemistry Letters 28 (2018) 2807–2815

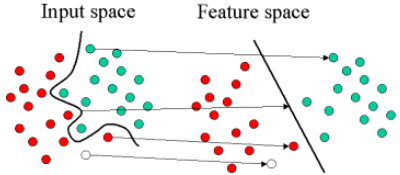
Copyright©2018 GGA Corp., All rights reserved.

Classification Algorithms

- Linear Discriminant Analysis (LDA): find line/hyperplane to separate classes



- Support Vector Machine (SVM): find separation plane in transformed space

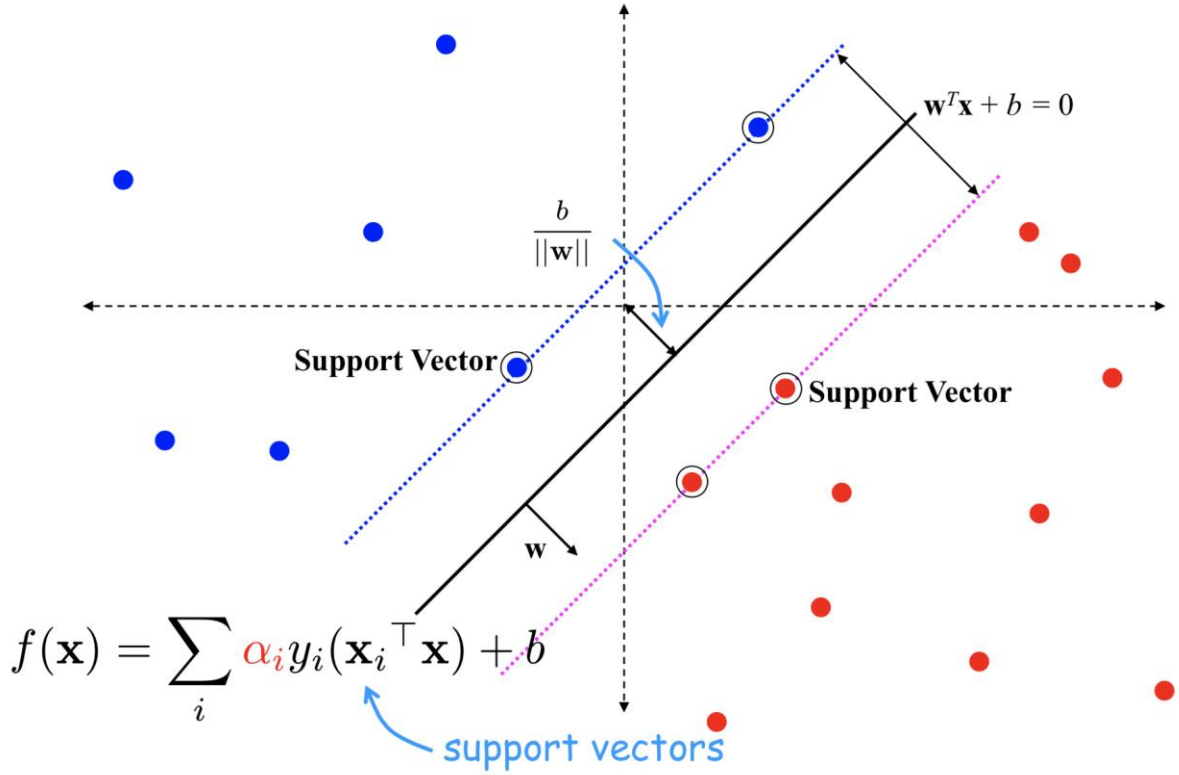


- Logistic Regression (LR): linear fit to give class probability (via logit transform)

$$\log\left(\frac{p_{disease}}{1-p_{disease}}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_J X_J$$

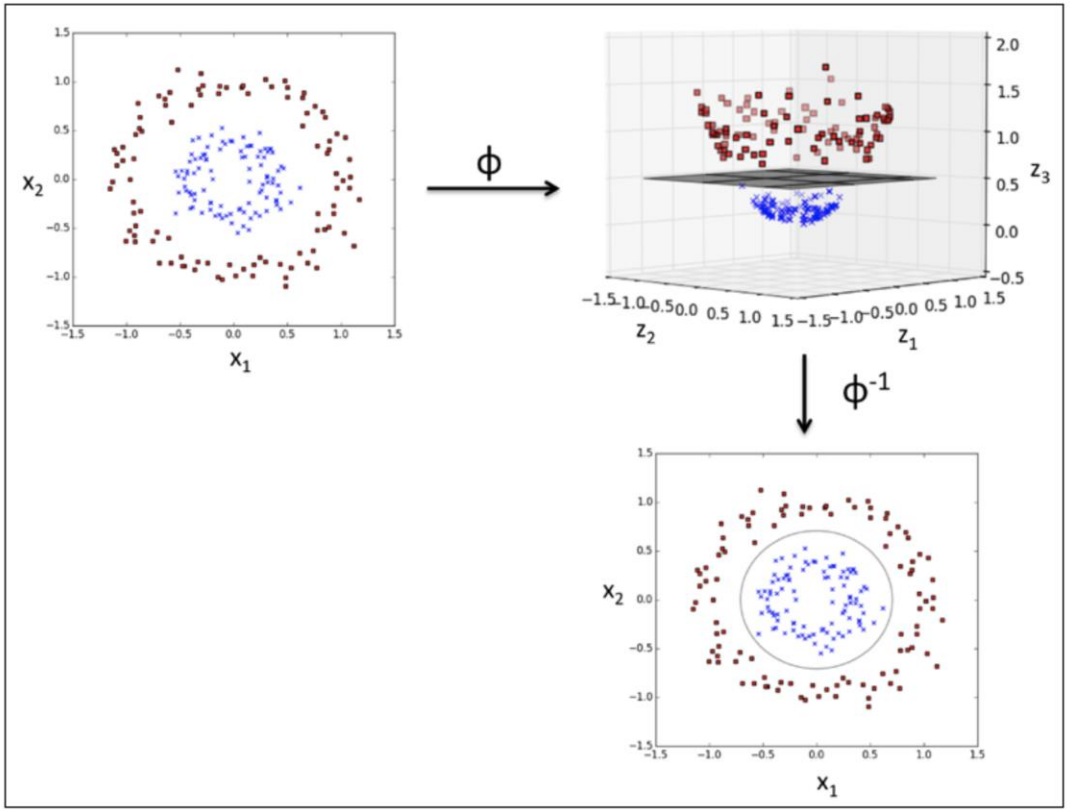
Copyright©2018 GGA Corp., All rights reserved.

Support Vector Machine (SVM)



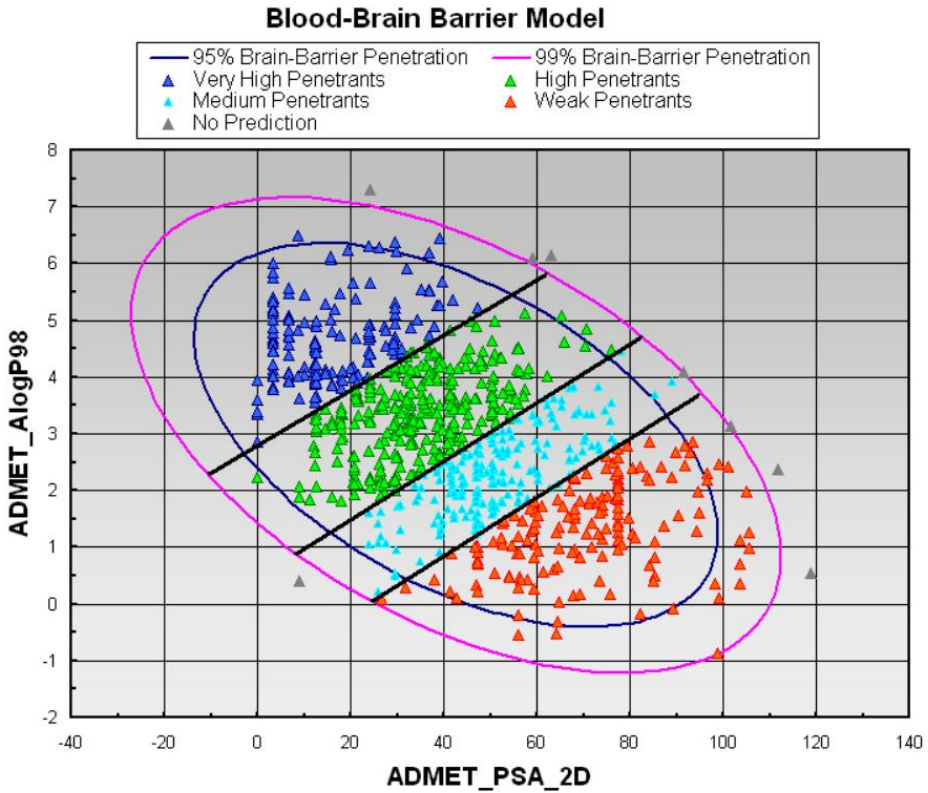
Copyright©2018 GGA Corp., All rights reserved.

Support Vector Machine (SVM)



Copyright©2018 GGA Corp., All rights reserved.

Classification Algorithms in ADMET prediction



Copyright©2018 GGA Corp., All rights reserved.

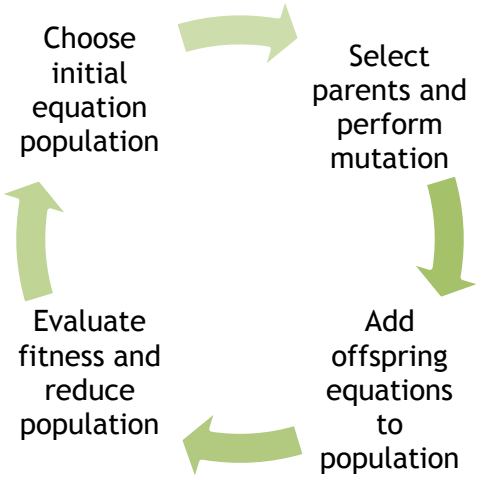
Genetic Functional Approximation

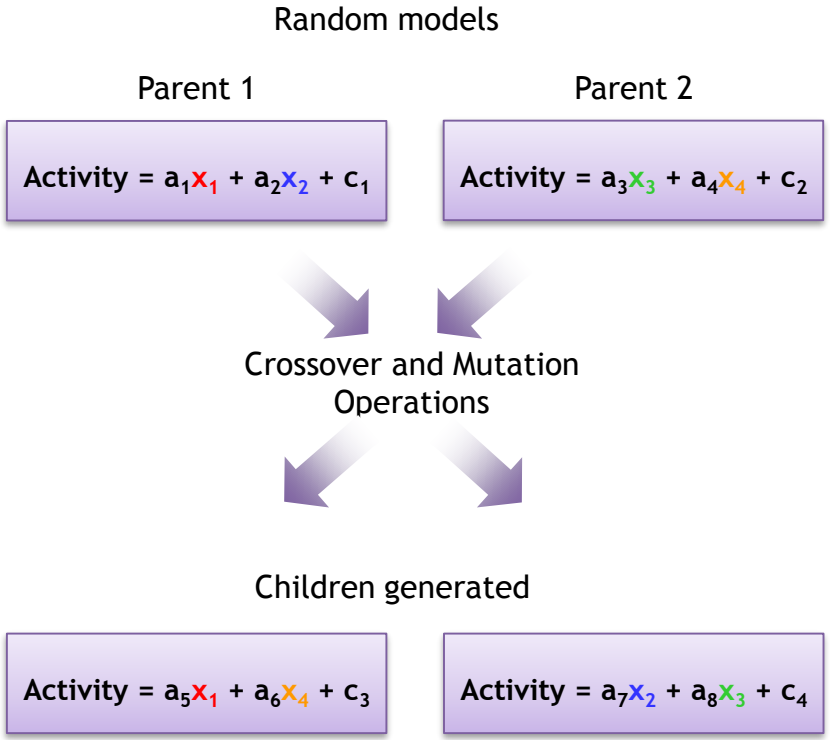
- Computer Simulation of Darwin's Theory
  - using genetic mutation of possible functions of the explanatory variable to get the best fit to the response data.
- Advantages
  - evolutionary algorithm
  - multiple models provide different insights into system
  - prevents overfitting



Copyright©2018 GGA Corp., All rights reserved.

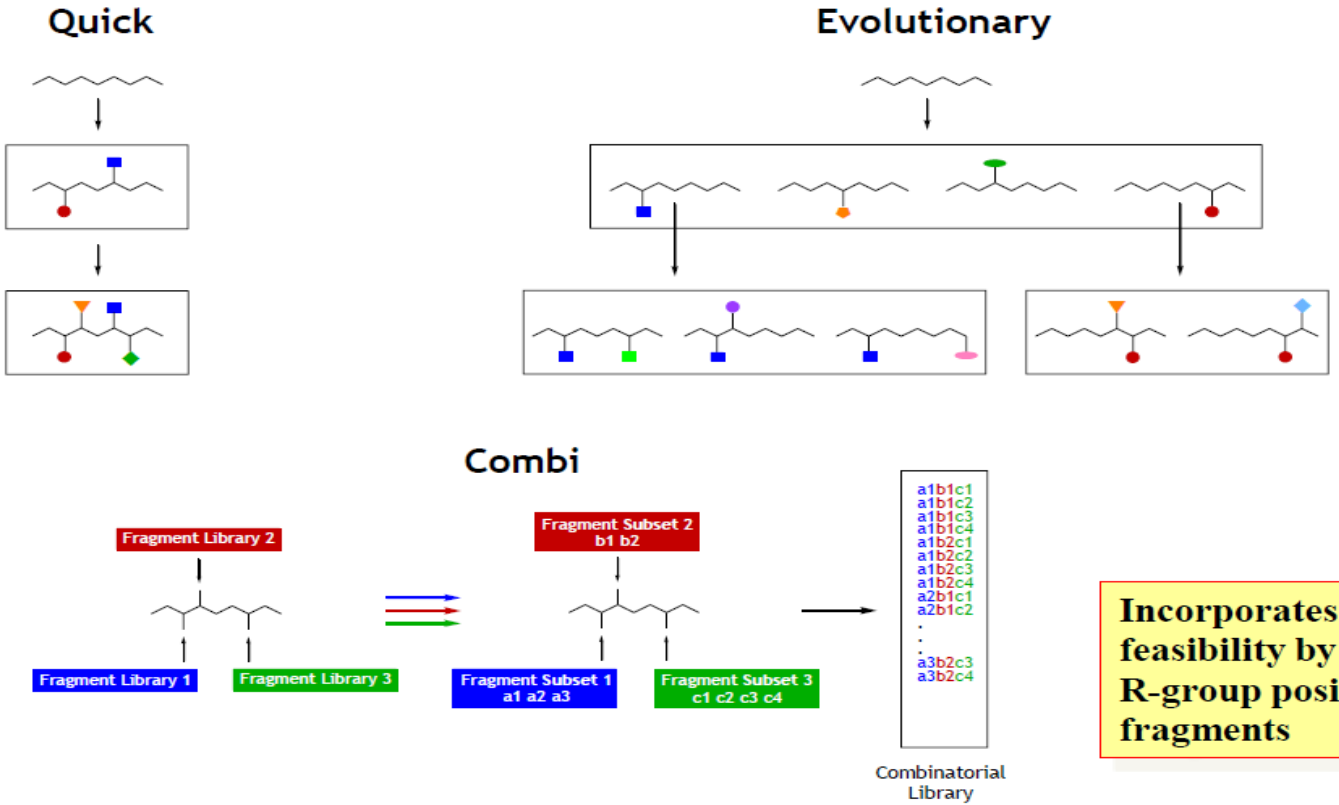
Genetic Function Approximation





Copyright©2018 GGA Corp., All rights reserved.

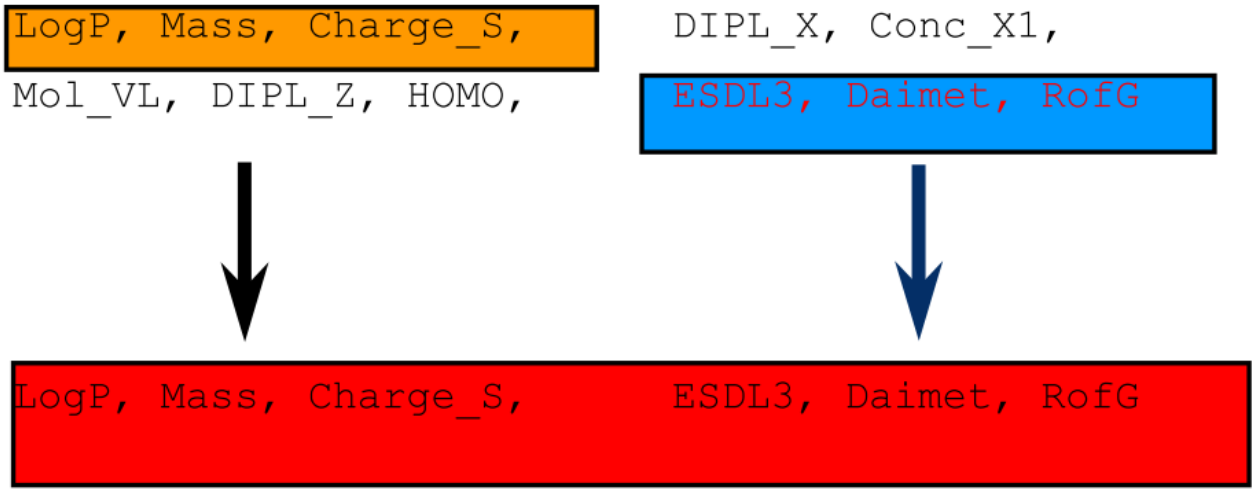
De Novo Evolution



Incorporates synthetic feasibility by selection of R-group positions and fragments

ved.

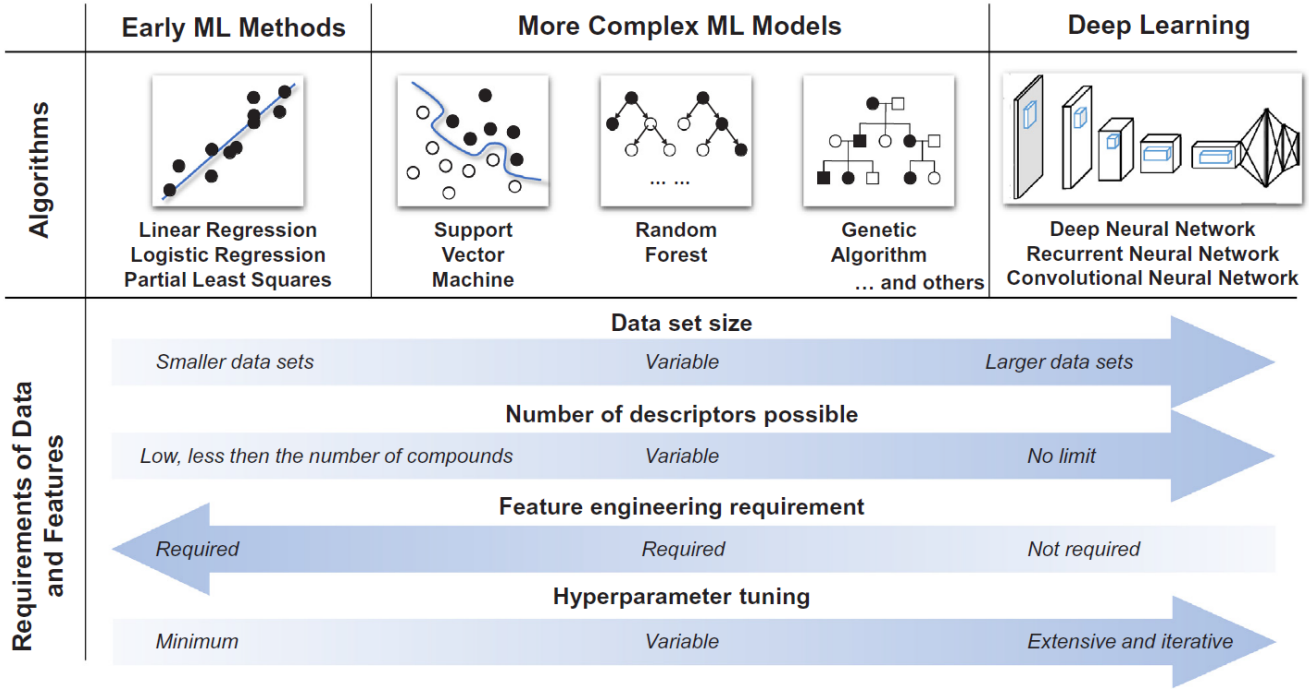
GFA: Crossover Operations



- The first two lines represent the parents
- The last line represents the first child of the next generation, formed by a crossover operation

Copyright©2018 GGA Corp., All rights reserved.

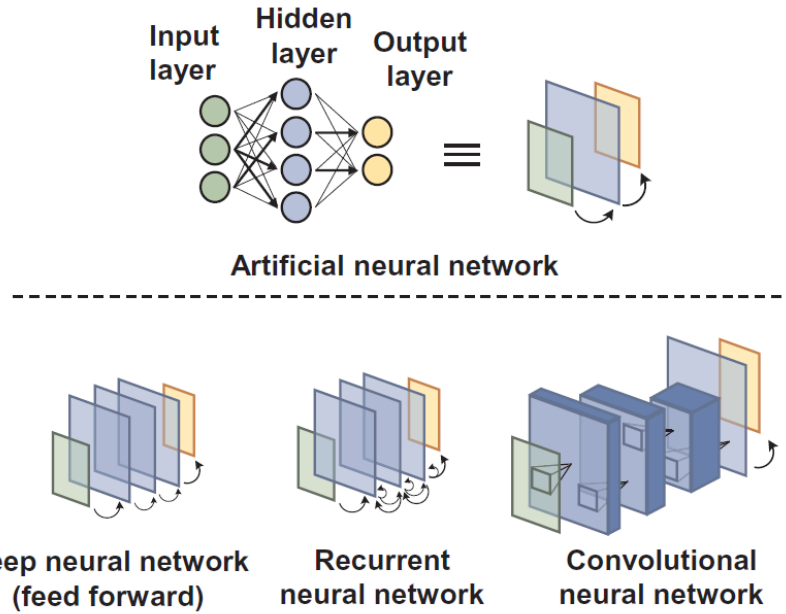
Evolution of machine learning methods employed in predictive model development and associated data requirements



Bioorganic & Medicinal Chemistry Letters 28 (2018) 2807–2815

Copyright©2018 GGA Corp., All rights reserved.

Representation of selected neural network architectures



Neural Network (NN): connected logistic units to predict 2+ class probabilities

Deep Neural Network* (DNN): NN with 2+ hidden layers, new training tricks

Copyright©2018 GGA Corp., All rights reserved.

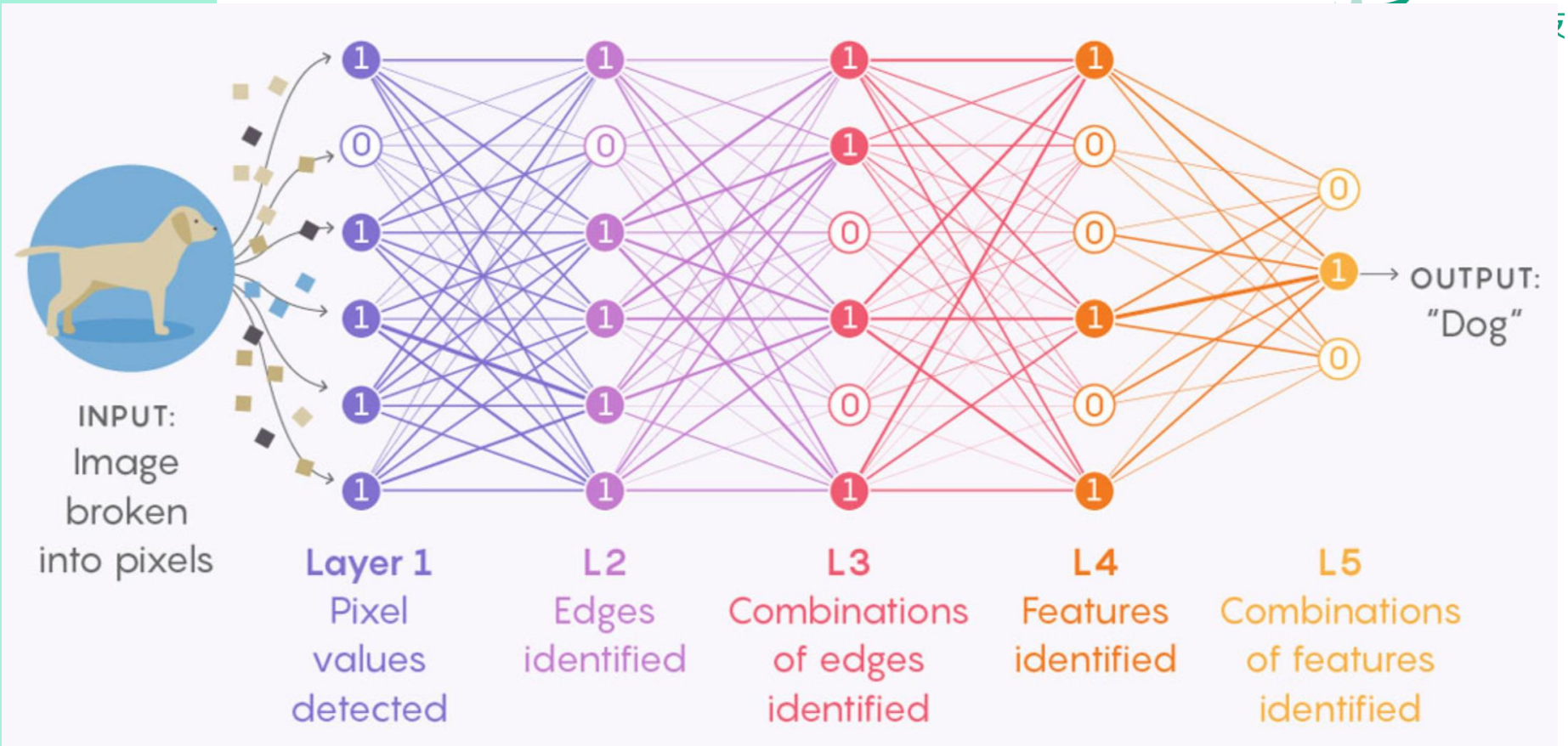
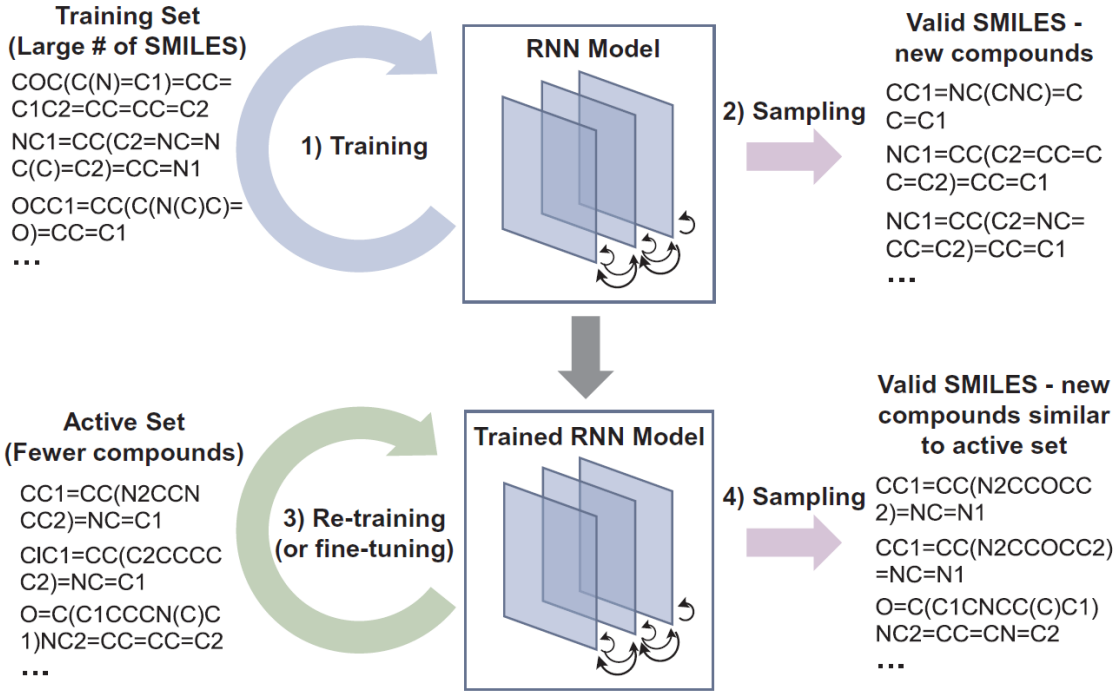


Figure from Lucy Reading-Ikkanda/Quanta agazine

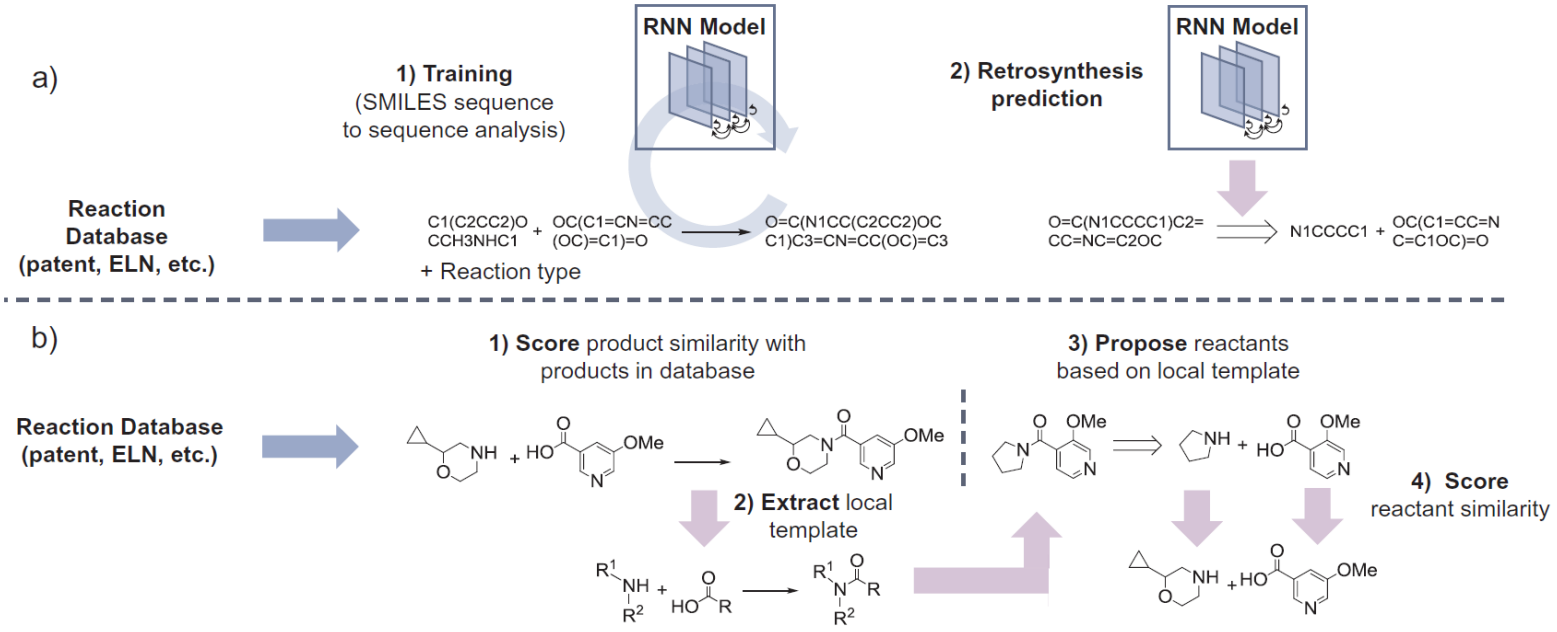
Copyright©2018 GGA Corp., All rights reserved.

General strategy for generation of new SMILES using RNN models



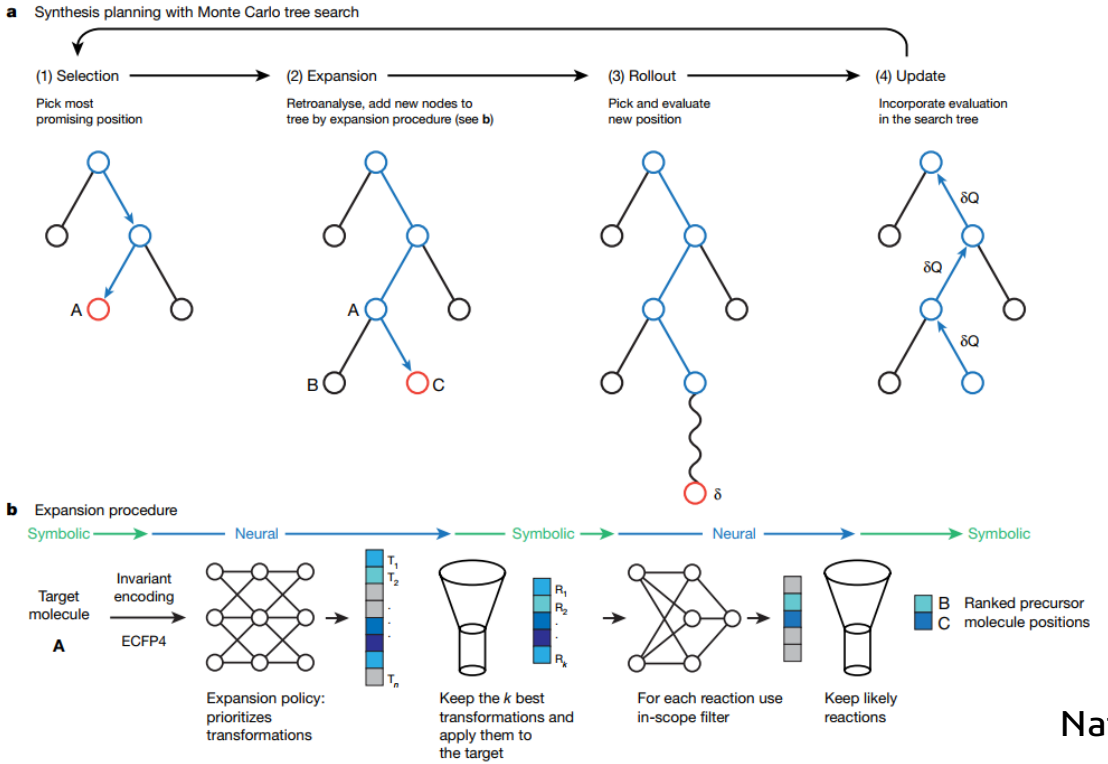
Copyright©2018 GGA Corp., All rights reserved.

Simplified schematic representation of retrosynthesis predictive algorithms not utilizing a template database



Copyright©2018 GGA Corp., All rights reserved.

Planning Chemical Syntheses with Deep Neural Networks



Nature 555, 604 (2018)

Copyright©2018 GGA Corp., All rights reserved.

Toxicity Prediction by Machine Learning

- Research has show that computerized chemical toxicity prediction beats animal testing.



OXFORD SOT Society of Toxicology www.toxsci.oxfordjournals.org ToxSci 20 Years

TOXICOLOGICAL SCIENCES, 2018, 1–15

doi: 10.1093/toxsci/kfy152
Advance Access Publication Date: July 11, 2018
Research Article

Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility

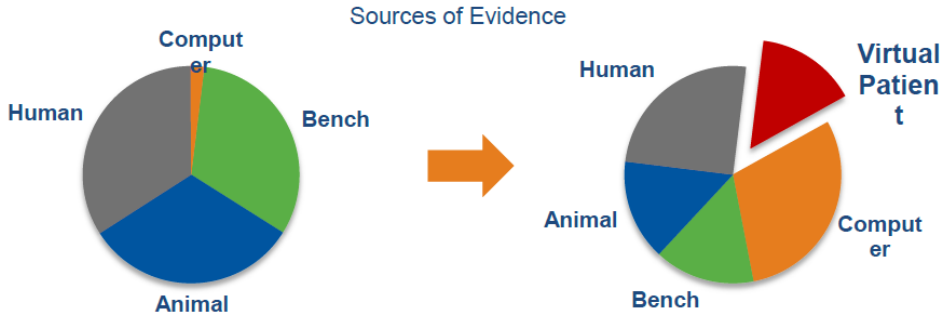
Thomas Luechtefeld,[*,†] Dan Marsh,[†] Craig Rowlands,[‡] and Thomas Hartung[*,§,1]

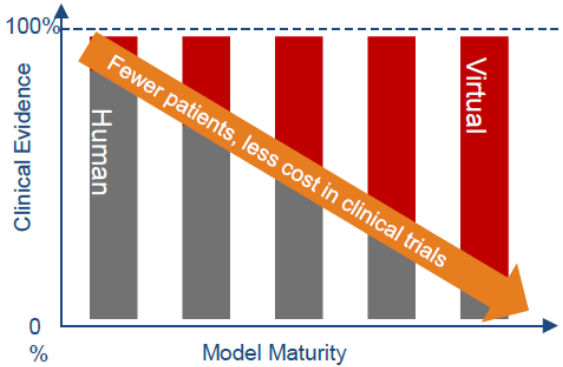
[*]Johns Hopkins University, Bloomberg School of Public Health, Center for Alternatives to Animal Testing (CAAT), Baltimore, Maryland 21205; [†]ToxTrack, Baltimore, Maryland 21209; [‡]UL Product Supply Chain Intelligence, Underwriters Laboratories (UL), Northbrook, Illinois 60062; and [§]University of Konstanz, CAAT-Europe, Konstanz 78464, Germany

The toxicity of new chemicals based on their structural similarity to known ones can be predicted by Read-Across Structure Activity Relationships. In the future, in silico virtual screening can save tremendous of time and costs for industry.

Copyright©2018 GGA Corp., All rights reserved.

Innovation in Clinical Trial Design: Pressures Impacting Clinical Research





Build on:

FDA Document: Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials

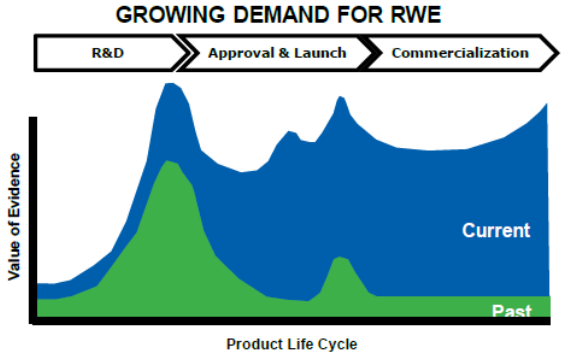
Maintain clinical endpoints with reduced human evidence sample size

Virtual Patient = computer models, engineering data, predicate and OUS data

Copyright©2018 GGA Corp., All rights reserved.

Effective integration of diverse information streams (Big Data) is driving fundamental shifts towards the collection and use of Real World Evidence

RWE goes beyond traditional controlled clinical trials. Medical device companies can use RWE to inform stakeholders about a **product's value, real world treatment patterns, benefits/risks and cost effectiveness**



Changing RWE models within the device

RWE Mandate	IT & Informatics	Source of Data &Tools	Value Creation
From observing and understanding…	*From tightly coupled data and analytics solutions for technical "experts"…*	*From data providers seen as vendors and suppliers…*	*From evidence generation primarily supports commercial objectives…*
…to shaping and Influencing	*…to standards based, modular architectures that empower domain experts*	*…to data providers seen as partners and collaborators*	*…to evidence informing product design and patient care*

Copyright©2018 GGA Corp., All rights reserved.

Early drug discovery which have been influenced by the most recent advancements in ML

- Structure generation
- de novo design
- Synthetic assessment
- Ligand binding and activity prediction
- ADMET predictions

Copyright©2018 GGA Corp., All rights reserved.

Q&A

Copyright©2018 GGA Corp., All rights reserved.



創源生技
GGA

謝謝指導

Improving the Quality of Life Through the Power of Biotechnology!

創源生物科技股份有限公司分子視算中心
地址: 台北市114內湖區新湖一路36巷28號
電話: 02-2795-1777#3014
Email: msc-support@gga.asia