CHEMOINFORMATICS: PRECURSORS FOR "AI IN MATERIALS"

Logan Ward
Asst. Computational Scientist
Argonne National Laboratory

20 January 2022

Historical Perspective

"The application of informatics methods to solve chemical problems" – T. Engel (2003)

"Chemoinformatics" is "information technology" and chemistry

- Chemistry and computing have a 60+ year history
- Ex: Paper on typing notation for chemicals (1952 Wiswesser)
  - "to sort and list such information with standard tabulating machine or with the new IBM scanning machinery"
  - "The notation in fact has made possible the establishment of the Willson Toxicity Registry, a new activity to catalogue and correlate mammalian toxicity data"

list:

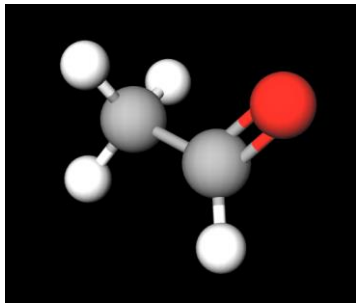
System	Symbols per Compound
Dyson	27.45
Gruber	29.00
Silk	19.36
Wiswesser	12.93

Modern uses of AI techniques in materials follow trends in "chemoinformatics"

Ref: T. Engel. J. Chem. Inf. Model. (2006), 2267.

Molecular Descriptors

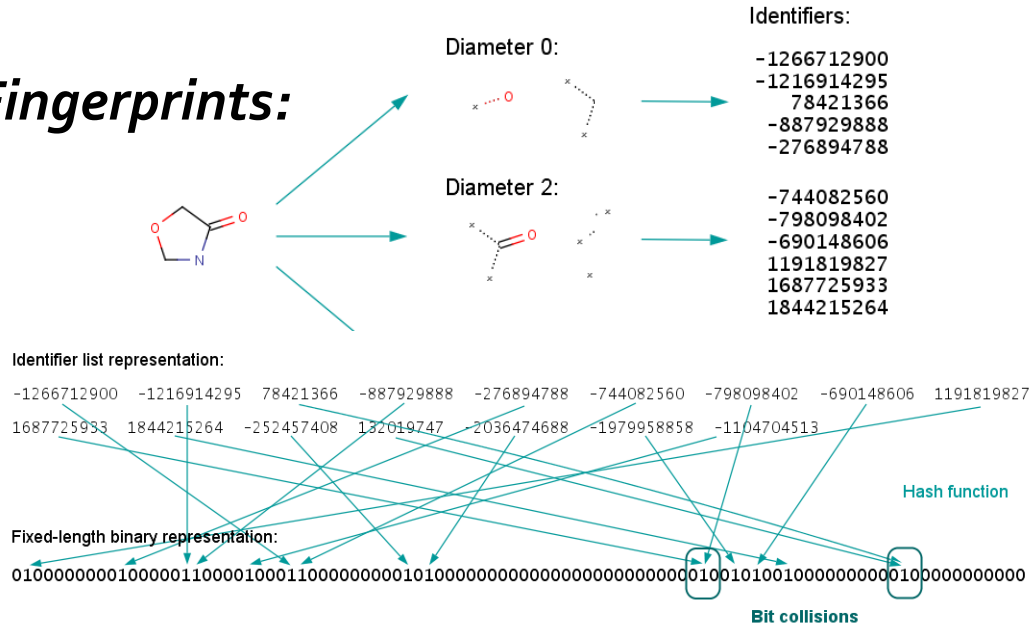
Discriminative



Line-notations for structure:

- SMILES (ex: "CC=O")
- InChI (ex: InChI=1S/C2H4O/c1-2-3/h2H,1H3)

Fingerprints:



Figures: docs.chemaxon.com

Descriptive

Extremely Well-Studied

Handbook of Molecular Descriptors

Author(s): Prof. Dr. Roberto Todeschini, Dr. Viviana Consonni
First published: 22 September 2000
Print ISBN: 9783527299133 | Online ISBN: 9783527613106 | DOI: 10.1002/9783527613106
Copyright © 2000 WILEY-VCH Verlag GmbH
Book Series: Methods and Principles in Medicinal Chemistry

Free Access

Bibliography (Pages: 524-667)

First Page | PDF | References | Request permissions

Many types of descriptors:

1. Constitutional (ex: "how many Ns?")
2. Structural (ex: Solvent-Accessible Surface Area)
3. Quantum-chemical (ex: partial charges)

Good Ref: T. Le et al. Chem. Rev. (2012), 2889

QSAR: "Quantitative Structure-Activity Relationships"

[Contribution from the Department of Chemistry, Pomona College, Claremont, California]

ρ–σ–π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure

By Corwin Hansch and Toshio Fujita[1]

Received August 19, 1963

Using the substituent constant, σ, and a substituent constant, π, defined as $\pi = \log P_X - \log P_H$ (P_H is the partition coefficient of a parent compound and P_X that of a derivative), regression analyses have been made of the effect of substituents on the biological activity of benzoic acids on mosquito larvae, phenols on gram-positive and gram-negative bacteria, phenyl ethyl phosphate insecticides on houseflies, thyroxine derivatives on rodents, diethylaminoethyl benzoates on guinea pigs, and carcinogenic compounds on mice.

Early examples of using "data" + "statistics" date to the 1960s!

General ingredients have not changed:

1. Trusted chemical data resource
2. Informative chemical descriptors
3. Appropriate regression algorithm

$$\log \frac{1}{C} = 0.519\pi + 1.540; \quad \overset{r^2}{0.955} \quad \overset{r}{0.977} \quad \overset{s^{12}}{0.130} \qquad (13)$$

Good Refs: T. Le et al. Chem. Rev. (2012), 2889, Dudek et al., Comb. Chem & HTS. (2006)

QSAR is still a tool you need to know

SCIENTIFIC REPORTS

OPEN

Quantitative design rules for protein-resistant surface coatings using machine learning

Received: 10 September 2018
Accepted: 22 November 2018
Published online: 22 January 2019

Tu C. Le[1], Matthew Penna[1,2], David A. Winkler[3,4,5,6] & Irene Yarovsky[1,2]

Preventing biological contamination (biofouling) is key to successful development of novel surface and nanoparticle-based technologies in the manufacturing industry and biomedicine. Protein adsorption is a crucial mediator of the interactions at the bio – nano -materials interface but is not well understood.

Science has created better descriptors, better supervised learning algorithms, but the basic idea is still the same.

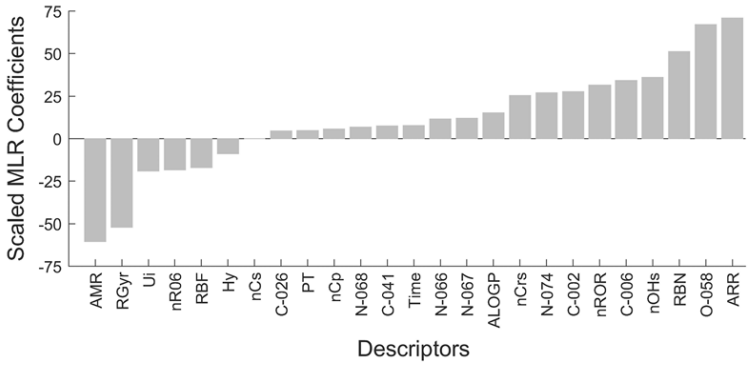


Figure 5. Scaled MLR coefficients of the most relevant descriptors selected from the pool 67 descriptors.

Ref: Le et al. Scientific Reports, (2019)

(Shallow) Machine Learning and QSPR

Table 2c. Melting-Point QSPR for Nitrocyanamide Salts

Eq	N	r^2	q^2	F	s^2	Term	Coefficient	t-test	Descriptor
4	7	0.960	0.894	120	3.71	0	-1.12×10^2 ($\pm 1.03 \times 10^1$)	−10.9	Intercept
						1	−7.11 (±0.649)	−10.9	$HOMO_{subst}$: Highest occupied molecular orbital located on the substituent.

Source: Trohalaki, Pachter. QSAR & Comb. Sci. (2005)

Sophisticated methods still useful,
but don't discount linear regression!

Why?

- Small dataset sizes
- Informative descriptors
- Interpretability
- Limited computational costs

Deep Learning is not Cure-All

Table 3: Summary of performances(test subset): conventional methods versus graph-based methods. Graph-based models outperform conventional methods on 11/17 datasets.

Category	Dataset	Metric	Best performances - conventional methods	Best performances - graph-based methods
Quantum Mechanics	QM7	MAE	KRR(CM): 10.22	**DTNN: 8.75**
	QM7b	MAE	KRR(CM): 1.05	**DTNN: 1.77***
		E	Multitask: 0.0150	**MPNN: 0.0143**
		E	Multitask(CM): 4.35	**DTNN: 2.35**
		SE	XGBoost: 0.99	**MPNN: 0.58**
		SE	XGBoost: 1.74	**MPNN: 1.15**
		SE	XGBoost: 0.799	**GC: 0.655**
		-PRC	Logreg: 0.129	**GC: 0.136**
Biophysics	MUV	AUC-PRC	**Multitask: 0.184**	Weave: 0.109
	HIV	AUC-ROC	**KernelSVM: 0.792**	GC: 0.763
	BACE	AUC-ROC	**RF: 0.867**	Weave: 0.806
	PDBbind(full)	RMSE	**RF(grid): 1.25**	GC: 1.44
Physiology	BBBP	AUC-ROC	**KernelSVM: 0.729**	GC: 0.690
	Tox21	AUC-ROC	KernelSVM: 0.822	**GC: 0.829**
	ToxCast	AUC-ROC	Multitask: 0.702	**Weave: 0.742**
	SIDER	AUC-ROC	**RF: 0.684**	GC: 0.638
	ClinTox	AUC-ROC	Bypass: 0.827	**Weave: 0.832**

* As discussed in section 4.4, DTNN outperforms KRR(CM) on 14/16 tasks in QM7b while the mean-MAE is skewed due to different magnitudes of labels.

For benchmark problems of learning from molecular data, conventional ML better for **6/17** cases

Ref: Wu *et al.* Chem Sci. (2017) 10.1039/C7SC02664A

An Example: Computational Toxicology

https://ntp.niehs.nih.gov/whatwestudy/niceatm/comptox/ct-opera/opera.html

Home » What We Study » NICEATM: Alternative Methods » Computational Toxicology » **OPERA**

Computational Toxicology

- Adverse Outcome Pathways
- Computational Models of Chemical Activity
- ICE: Integrated Chemical Environment
- In Vitro to In Vivo Extrapolation
- Integrated Approaches to Testing and Assessment
- OPERA

OPERA

SHARE THIS: 1

https://ntp.niehs.nih.gov/go/opera

Open Structure-activity/property Relationship App

Quantitative structure–activity/property relationship (QSAR/QSPR) models provide predictions of chemical activity that can augment non-animal approaches for predicting toxicity. However, the performance of QSAR models highly depends on the quality of the data and modeling methodologies used.

To provide robust QSAR/QSPR models for chemical properties of environmental interest that can be used for regulatory purposes, EPA NCCT created the Open Structure-activity/property Relationship App (OPERA) (Mansouri et al. 2018). OPERA is a free and open-source/open-data suite of QSAR models providing predictions for physicochemical properties, environmental fate parameters, and toxicity endpoints. All OPERA models were built on curated data and QSAR-ready chemical structures standardized using an open-source workflow (Mansouri et al. 2016).

OPERA is an ongoing collaboration between NICEATM and EPA. Recent additions to OPERA include predictions for:

QSAR is not going away any time soon.

PRACTICAL CONSIDERATIONS: WHAT TOOLS ARE OUT THERE?

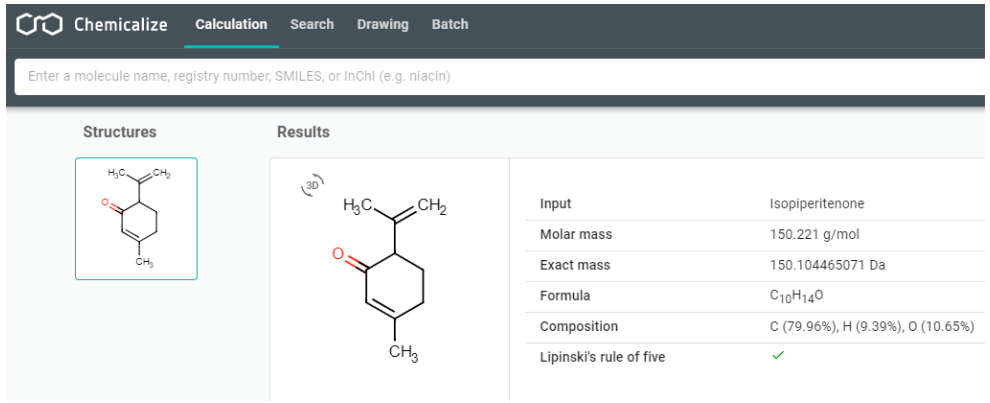
Chemical Databases Abound

ChEMBL

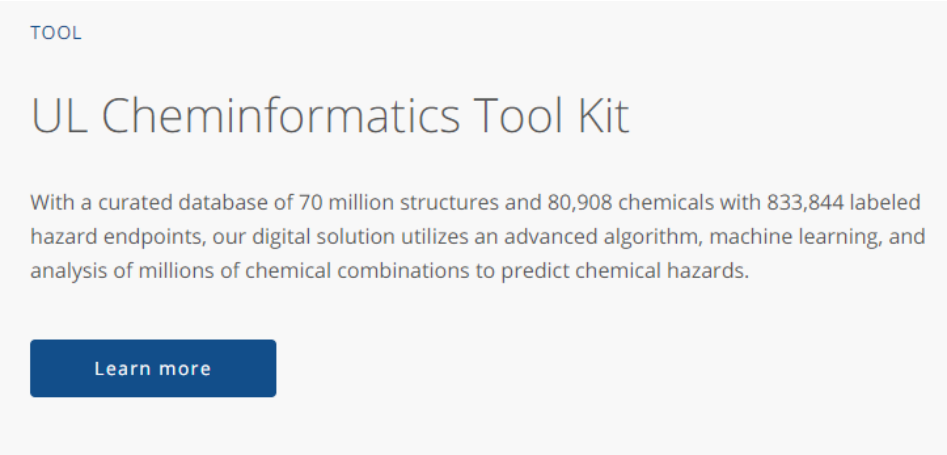
https://www.ebi.ac.uk/chembl/

ChemSpider
Search and share chemistry

http://www.chemspider.com/

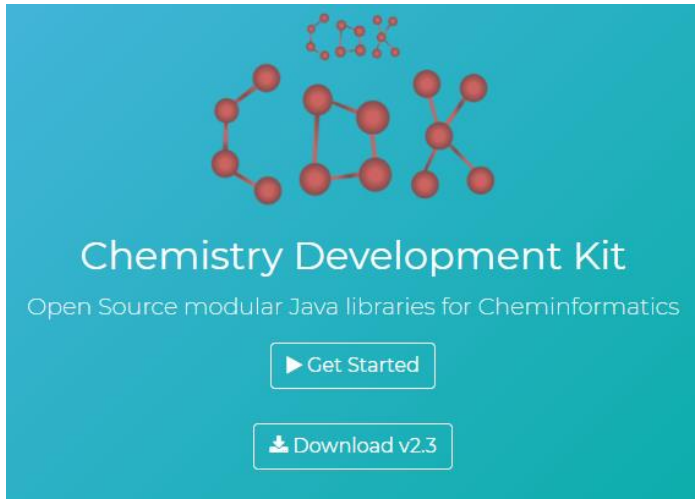


https://chemicalize.com/



https://www.ul.com/resources/apps/ul-cheminformatics-tool-kit

Toolkits for Building QSAR Models



https://cdk.github.io/

README.md

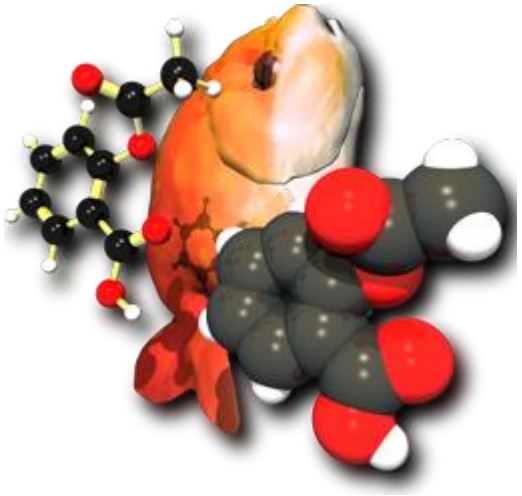
RDKit

Azure Pipelines succeeded | docs passing | DOI 10.5281/zenodo.4318717

RDKit is a collection of cheminformatics and machine-learning software written in C++ and Python.

- BSD license - a business friendly license for open source
- Core data structures and algorithms in C++
- Python 3.x wrapper generated using Boost.Python
- Java and C# wrappers generated with SWIG
- 2D and 3D molecular operations

https://rdkit.org



http://openbabel.org/wiki/Main_Page

kode chemoinformatics

https://chm.kode-solutions.net/

Final Note: Chemoinformatics is a Mature Field

There is still a lot materials informatics can learn from

- 60+ years of history
- Textbooks, more textbooks, codes and review papers
- Decades of "lessons learned"

Handbook of Molecular Descriptors

Author(s): Prof. Dr. Roberto Todeschini, Dr. Viviana Consonni
First published: 22 September 2000
Print ISBN: 9783527299133 | Online ISBN: 9783527613106 | DOI: 10.1002/9783527
Copyright © 2000 WILEY-VCH Verlag GmbH
Book Series: Methods and Principles in Medicinal Chemistry

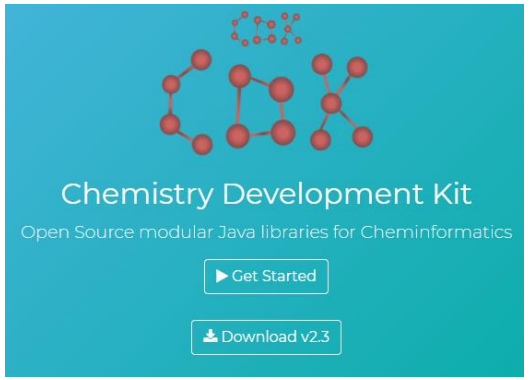


Table 3. Common QSPR modelling pitfalls and methods of avoiding them

pitfall	recommendation to minimize or avoid
use of uninformative descriptors	use descriptors that are related to the molecular structure where possible, use virtual screening methods when complex descriptors are necessary, and develop new materials descriptors
overfitting, and grossly underdetermined systems	reduce size of descriptor pool before building models,[24] monitor number of fitted parameters (descriptor weights or neural network weights) to ensure they are substantially less than the number of experiments, and check that training and test set statistics are similar
descriptor selection and chance correlations	use Topliss criteria[22,23] to estimate probability of chance correlations and descriptor scrambling; avoid methods where repeated sampling of a larger pool of descriptors is done to obtain the optimum subset of descriptors; and use sparse, context-dependent feature-selection methods[17,18]
modeling complex, nonlinear structure–property relationships	avoid overly complex nonlinear models, compare nonlinear model statistics with linear models, and use regularizing methods that attempt to optimize model complexity[16,17]
validating QSPR models	synthesize new materials that models predict to be superior and test if feasible, use independent test sets to assess model predictivity otherwise, and employ cross-validation methods with caution[27,28]
domain of applicability of models	calculate the range of all descriptors used to develop the model,[29−32] avoid extrapolations using descriptor space distant from that used in model, and use probabilistic modeling methods (e.g., Bayesian regularization[16]) that allow estimation of likely prediction error
incorrect handling of outliers	avoid removing outliers wherever possible, check whether outlier lies well within domain before removing it, remove outliers sparingly and describe why they were omitted, and retest properties for outliers to eliminate measurement of transcription errors

Good Refs: T. Le et al. Chem. Rev. (2012), 2889

Practical Exercises

First for the course!

WardLT/applied-ai-for-materials:molecular-property-prediction/chemoinformatics

Learn how to:

- Manipulate chemical data with RDKit and Pandas
- Train conventional machine learning models with *descriptors* and *fingerprints*