IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 11, Issue 11, November 2021, ||Series -I|| PP 36-42

Data Mining and its Impact on Science

Hans-Jörg Schneider
FR Organische Chemie der Universität des Saarlandes
D 66123 Saarbrücken Germany

Received 05 November 2021; Accepted 20 November 2021

Abstract. Modern techniques make large masses of data available, and have gained an enormous importance in many fields of science. With several examplesfrom social sciences it is illustrated that an indiscriminate use of data as basis for data mining and machine learning can be misleading, including assumed correlations between intelligence and skin colour or climate. This is also the case with one the oldest applications of data mining in science by the analysis of many thousands crystal structures, as deposited in e.g. the Cambridge Structural Database. The observed large diversity in different crystals makes it for example difficult to define the geometric conditions for hydrogen bonds and to distinguish contributing binding mechanisms. Strong interactions can dominate the crystal structure, and interfere with the identification of weaker interactions. QSAR-type analyses, important for drug discovery, can on the basis of e.g. similarity approaches lead to binding free energy predictions of supramolecular complexes with a multitude of parameters withoften little physical meaning.

I. INTRODUCTION.

The systematic analysis of large amounts of data has gained increasing importance in science.[1] This applies not only to the social and economic disciplines[2] and part of humanities, where artificial intelligence can e.g. help machine translation, but also to physical sciences.[3] In this article perspectives, but in particular limitations of data mining in science will be discussed, as well as its consequences for the future development of experimental sciences.

Extensive collection and comparison of many observations has played a decisive role in biology already in the 18th century.Carl von Linné has laid 1735 the foundations of plant taxonomy. The underlying mechanisms of genetics were then clarified by experimental studies, beginning with the crossbreeding experiments of Gregor Mendel around 1866, and ending with the discovery of nucleic acids functions in the middle of the last century.

The pitfalls in the interpretation of large amounts of data and the underlying statistical methods[4] has been emphasized for fields such as economics or psychology, where data mining plays a likely even greater role than in sciences which are largely based on experiments.

The problems associated with the interpretation of psychological data are remarkably visible in heated controversies around a possible association of race and color with the mean intelligent quotient IQ across nations.Rushton and Lynn had contended 1991 and 1995 that greater intelligence is needed to adapt to a colder climate; over many generations the more intelligent members of a population would then be more likely to survive and reproduce.[5],[6] Templer and Arikawa[7] assumed that skin color is closely related to ambient temperature in a country, and found indeed an usual high correlation coefficient of 0.92 between skin color and IQ score;they noted, however,that of course it would be absurd to suggest that a lighter complexion makes people more intelligent. Indeed the correlation is a typical example of parameter intercorrelation: both skin colour and ambient temperature variations just reflect the sun impact in different countries. Subsequent discussion of the Templer and Arikawa paper maintained that these correlations are *"completely non-informative regarding any causal or functional connection between individual differences in skin pigmentation"*; [7b], and that it contains misleading conclusions,is based upon faulty collection and analysis of data, and that publication of the Templer and Arikawa article was "*unfortunate*"[7c].In a rebuttal Templer and Arikawa clarified that their study was primarily designed to test the contention of Rushton and Lynn that people in colder climates tend to have higher intelligence, which correlates with temperature.[7d]Later defendants of Rushton expressed that "*mob science works to discredit valid research and enforce collective ignorance*"[8]. A possible limitation of correlations with IQ could be the varying educational conditions in different countries; this effect was believed to be taken care of [5b] by application of the so-called Flynn-effect.[9] Another significant influence could be due to smaller chances for black or coloured people to work in intelligence-demanding jobs, due to discrimination or, particularly in Africa, to low technology levels. Controversies around the origin of IQ differences carry on until recently;[10]errors associated with the method of correlated vectors has been shown to lead eventually to nonsensical results.[11]

The correlation shown in Figure 1with a comparison of maternal mortality ratio and CO_2-emissions in different countries is another example of parameter intercorrelation. Obviously there is anexponential increase of maternal mortality with decreasing CO_2-emission in each country. A similar correlation is observed if one compares infant mortality ratios with CO_2-emissions. An ignorant machine could conclude that a high CO_2-content in the air helps to decrease mortality. Instead, both parameters have asmajor common origin the different degree of poverty in the countries. Such parameter inter-correlations can also play a role in natural sciences.

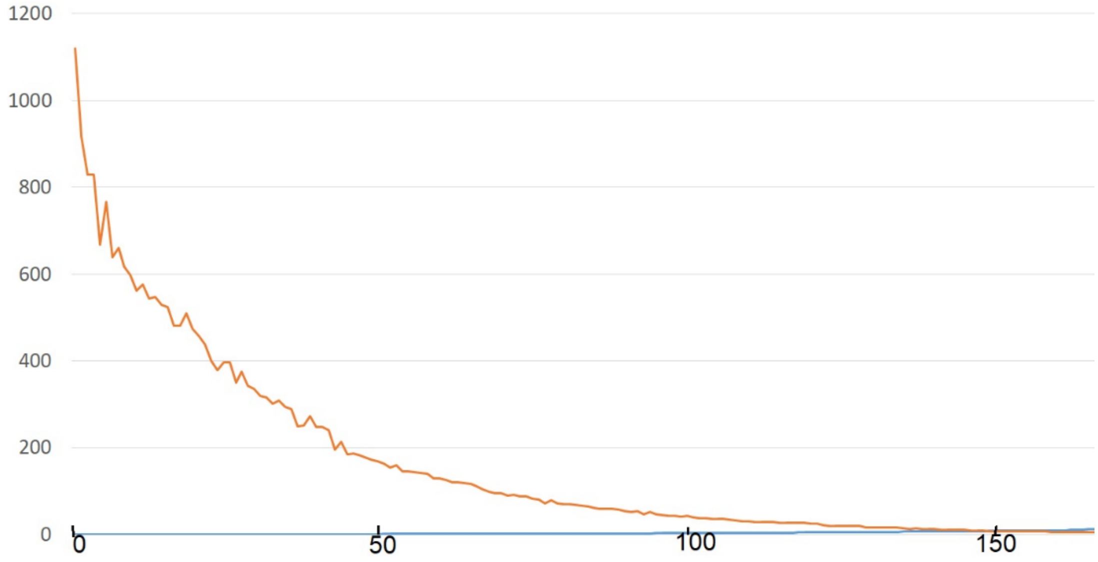

Figure 1. Maternal mortality ratio (number of death per 100.000 birth) and CO_2-emissions (metric tons per capita) in different countries.

Applications in Science. Probably the most important scientific applications of data mining are those in medical disciplines where experimental investigations are usually precluded.[1c,d,12]Epidemiological analyses are oftencontroversially discussed regarding their significance, like e.g. those on cell phone radiation risks.[13] Statistical analyses can help to identify e.g. the role of metal ions in diseases. Although the cause of Alzheimer's disease is not yet clear a multivariate statistical analysis has shown that four metals (Mn, Al, Li, Cu) in peripheral blood are strongly associated with the disease.[14]Figure 2 illustrates as typical example the role of selene in breast cancer development.[15]Better chances for establishing origin of metal ions for such diseases are given if the action of such ions can be secured by investigation of the interaction mechanism.

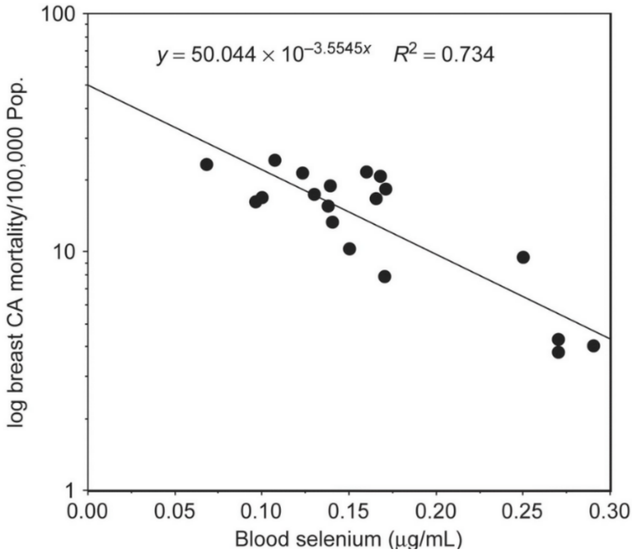

Figure 2.Correlation between selene concentration in blood and age-corrected female breast-cancer mortalities.[15a]

Drug discovery presents awidespread application of data mining.[16] In view of about 10^{60} molecules which are classified as drug-like molecules their prioritization remains a particular challenge, compared to finding ‘a needle in a haystack’. Quantitative Structure-Activity Relationship (QSAR) methods use machine learning tools, based essentially on Emil Fischer’s ‘lock-and-key’ model of the interaction between a ligand and a biological receptor.[17] The Protein Data Bank (https://www.rcsb.org/) provides >2 million structural data for proteins, nucleic acids and their complexes, which can be used for screening the interaction of molecules which are either experimentally available or can be designed by computational methods.High-Throughput Screening (HTS) experiments with automated machines deliver huge data mases on the activity of potential drugs with immobilized biomolecules or whole cells. Another promising strategy is based on similarity searches between known and new drugs.Although these screening strategies have led to large libraries of promising drug candidates, application of such methods has shown until now only limited therapeutic progress.[18]The problem is often the identification of the biological target. Alzheimer’s disease therapies based on the amyloid as target have for instance failed in clinical trials, as this disease is a rather multifactorial syndrome.[19] Genomic studies can support the development of new drug targets,[20] including those for psychiatry.[21]

In chemistry the endless number of compounds and of numerical data for their properties has made data mining to a particular promising approach. A large collection of experimental data has for example been the basis to characterize solvent polarity.[22] Abraham’s comprehensive scales of solute hydrogen-bond acidity and basicity[23] involve besides solute H-bond acidity and basicity numbers also terms for dispersion, polarizability , dipole-dipole or dipole-induced- dipole plus as well as an endoergic cavity term; some of them are within limits inter-correlated.Chemists have made use of structural databases derived from X-ray measurements long before the term data mining was coined. The Cambridge Structural Database (CSD)[24] (https://www.ccdc.cam.ac.uk/)comprises over a million structures of small compounds, and lends itself for analyses of typical geometries. In practice distances between atoms or groups below or near the sum of *van der Waals* distances are taken as evidence for noncovalent interactions, and the frequency of such occurrences is considered to be ameasureof their general significance.[25,26]Machine learning with structural data and techniques such as simulated annealing, genetic algorithms, and density functional theory are increasingly used for corresponding predictions.[27]

While such analyses have delivered essential numerical data for many interactions, they can be misleading particularly regarding weak interactions. Thus, the occurrence of only0.6 % crystal structures with short C-F···HX (X = O, N) distances in a total of 5947 C···F fragments has led to the conclusion that “*organic fluorine hardly ever makes hydrogen bonds*.”[28]However, ifone restricts the data basis to compounds which contain only X—Hand C—F bonds, no atoms heavier than chlorine, and carry no good hydrogen-bond acceptors such as oxygen or nitrogen one finds a sizeable number of structures which do indicate C-F bonds as hydrogen bond acceptor.[29]Many experimental and theoretical investigations have indeed shown that fluorine actually makes stronger hydrogen bonds than other halogens.[30]Obviously, an indiscriminate choice of data as basis for data mining can be quite misleading.

The identification particularly of weak non-covalent interactions in a multitude of crystals is often hampered by less well-defined geometric parameters and the occurrence of different binding mechanisms. The usual restriction for hydrogen bonds with A····H--D angles above at least 130 ^{o} cannot strictly be applied, with values even below 90 ^{o} considered to be acceptable.[31]For C-F····H-C bonds F····H distances between 2.3 and 3.0 Å, and F····H-C angles between 100 and 175^{o}have been observed.[32]With the aim to identify both dispersive interactions and hydrogen bonds a CSD analysis has been carried out with several thousand crystals which contain arenes and C-Hal fragmentswithin a distance that would allow both hydrogen bonds with the aromatic hydrogens and/ordispersiveinteractions with the π-cloud.[33]The latter would be characterized by shorter distances *d* and by small angles α between the arene plane and the aryl centroid –halogen vector line (Figure 3).About 20.000 of such structures showed considerable scatter; the majority exhibiting larger *d*and α values pointing to a preference of H-bonds,with a steep cutoff at $\alpha < 90^{0}$, just slightly off the value for linear H-bonds (Figure 4). As often no clearcut preferences can be observed. Dispersive contributions, characterized by smaller distances*d*and angles αseem to occur less frequently, and are in comparison to other halides barely favoured for iodine derivatives a more polarizable unit.

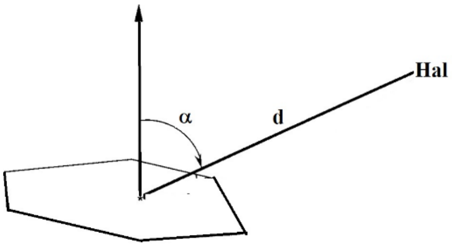

Figure 3. Interaction of R-Hal with benzene characterized by an angle α and distance *d*.

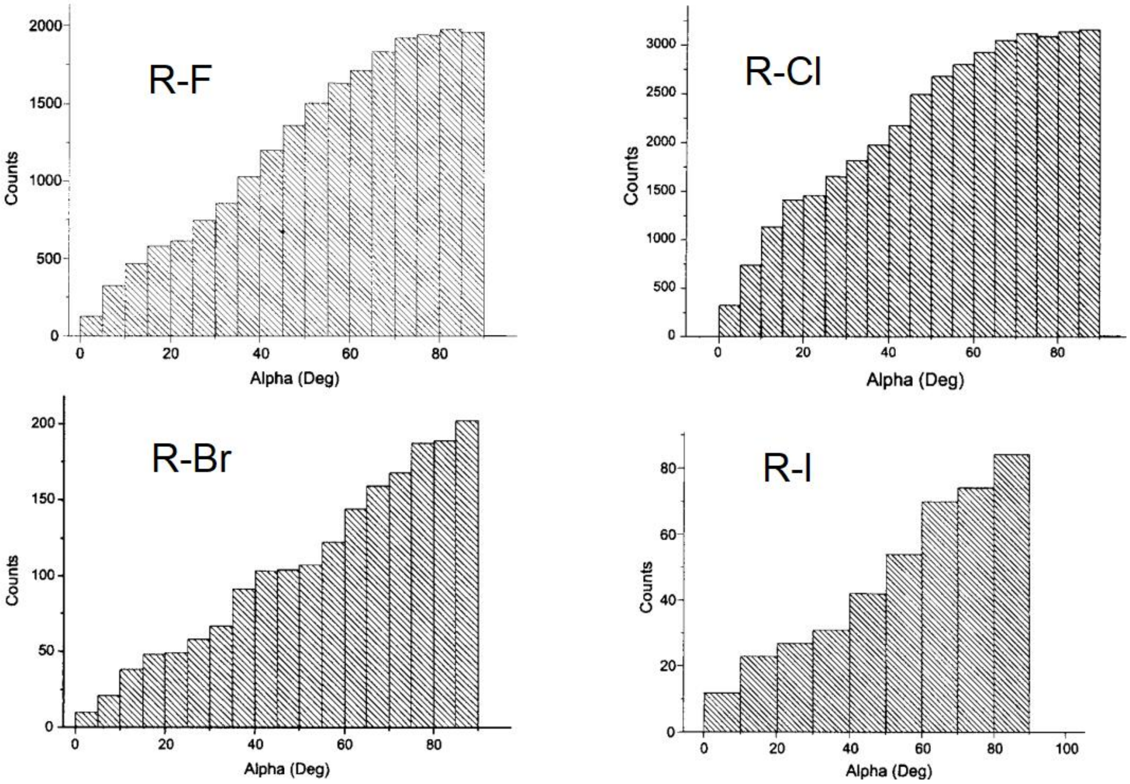

Figure 4. Histogram of crystals with interaction between R-Hal and arenes, characterized by an angle α.

The quantification of noncovalent chemical interactions is of importance for many biological and synthetic complexes, and is most often based on equilibrium measurements in solution. Parameters, or descriptors, which assess the strength of the underlying interactions by force field calculations are based on known affinity data as training sets and e.g. vibrational spectra; they are used also as scoring functions for the prediction of protein-ligand interactions or for drug discovery. Alternative QSAR methods use e.g. similarity approaches or learning sets for structure and binding energy prediction of supramolecular complexes.Figure 5 illustrates for complexes with ß-cyclodextrin in water that acceptable correlations with experimental numbers can need a large number of descriptors, which moreover have oftenbarely physicochemical meaning and can themselves exhibit inter-correlation.[34]Investigation of modified complexes and their structure in different solvents can help to disentangle the underlying noncovalent interactions.

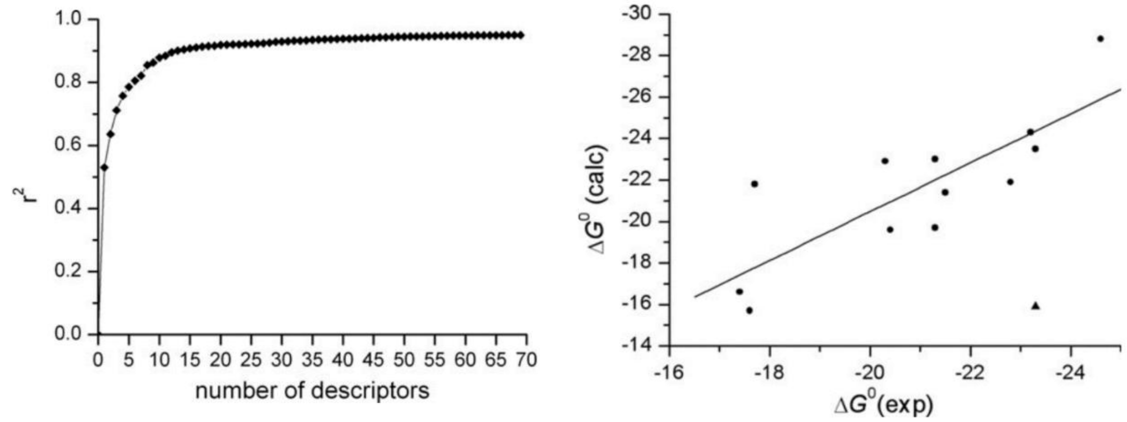

Figure 5. (a) Number of descriptors used for prediction of binding free energies ΔG° with ß-cyclodextrin complexes. (b) plot of the predicted and experimental ΔG° values (kJ mol^{-1}) of the screening hits. Reproduced with permission from Steffen, Lengauer, Wenzet al , *New J. Chem.*, **2007**, *31*, 1941.

II. CONCLUSIONS.

Modern computing allows to store and to compare unlimited amounts of data, automatized techniques can produce in shortest time huge data masses. This bears the danger of misleading conclusions regarding cause and effect, and, in short, of eventually leaving thought to machines.Good correlations do not imply causation, accurate predictions do not mean understanding cause and effect, while understanding can well allow concise predictions. The modern developments of data mining influencesincreasingly also natural sciences, where experiments are the traditional basis. In biology, phenomenological observations have for long been supplemented by analyses of metabolisms, signalling pathways and genetic mechanisms. Immunology relies essentially on experimental investigations of molecular and cellular components of the immune system. Environmental studies can be supported by laboratory simulations. Rational design of drugs, particularly those which are not only symptomatic, rely not on trial-and-error testing of large compound libraries but on the study of interactions between drug candidates and a biological target. Chemists can design and synthesize molecules in order to elucidate by physical measurements their properties and interactions. In particular noncovalent interactions can be quantified in different media with properly designed supramolecular complexes.[35]Experimental approaches should not be sacrificed in favour of certainly often more economical applications of machines.

[1] Charu, C.A. Data Mining: The Textbook Springer; **2015**; Tan, P.N., Steinbach, M, , Kumar, V. Introduction to Data Mining.Addison Wesley, **2006**, Pearson, **2018.**

[2] Russell,M.A. Mining the Social Web O'Reilly Media; 3rd edition **2019**.

[3] a) Handbook on Big Data and Machine Learning in the Physical Sciences Kalinin, V.S. et al, Eds, World Scientific Publishing Co, **2020**; b) Hsieh, W. Machine Learning Methods in the Environmental Sciences,Cambridge University Press; **2018**; c) Cleophas, T.J., Zwinderman, A.H. Machine Learning in MedicineSpringer; 2nd ed. **2020**; d) Chang, A.C. Intelligence-Based Medicine: Artificial Intelligence and Human Cognition in Clinical Medicine and HealthcareAcademic Press; **2020**; e) Schütt, K.T. et al Machine Learning Meets Quantum PhysicsSpringer; **2020**; f) Data Mining in Drug DiscoveryHoffmann, R.D. et al, Eds. Wiley-VCH; **2013**; g) Larson, R.S., Oprea, T.I. Bioinformatics and Drug Discovery Humana; 3rd ed. **2019**: h) Cartwright, H.M. Machine Learning in Chemistry: The Impact of Artificial IntelligenceRoyal Society of Chemistry,**2020**; i) Information Science for Materials Discovery and DesignLookman, T. et al, Eds., Springer, **2016**.

[4] Vigen, T. Spurious Correlations, Hachette Books, **2015**.

[5] a)Lynn, R. Race differences in intelligence: A global perspective. Mankind Quarterly, **1991**, 31, 255– 296; b) Lynn, R., & Vanhanen, T. IQ and the wealth of nations Westport7 Praeger Publishers, **2002**, New York7 Random House.

[6] Rushton, J. P. Race, evolution and behavior: A life history perspective. **1995**, New Brunswick, NJ7 Transaction; Rushton, J. P. Race, evolution, and behavior. A life history perspective (3rd edition). **2002,**Port Huron7 Charles Darwin Research Institute; Rushton, J. P., Rushton, W. W. Brain size, IQ, and racial group differences: Evidence from musculoskeletal traits. Intelligence, **2003,**31, 139–155.

[7] a)Templer, D.I. , Arikawa H. Temperature, skin color, per capita income, and IQ /An international perspectiveIntelligence**2006**, 34, 121–139; b) Arthur R. Jensen ibid , 128-131; c) Hunt, E., Robert J. Sternberg ibid 131- 137; d) Templer, D.I. , Arikawa , H. ibid. 138-139.

[8] Gottfredson,L.S, Resolute ignorance on race and RushtonPersonality and Individual Differences**2013,**55, 218–223.

[9] Flynn, J. R. Massive IQ gains in 14 nations. Psychological Bulletin,**1987**, 101, 1171– 1191.

[10] Carl, N., Woodley Menie, M. A.)Intelligence**2019,** 77, Article Number: 101397 DOI: 10.1016/j.intell.2019.101397.

[11] Jelte M. Wicherts,I. Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results)Intelligence**2017**, 60, 26–38.

[12] Yoo, I., Alafaireet, P., Marinov, M. et al. Data Mining in Healthcare and Biomedicine: A Survey of the Literature. J. Med. Syst.**2012** , 36**,** 2431–2448.

[13] Miller, A.B. et al Frontiers Pub Health**2019**, 7, Article Number: 223.

[14] Guan, C. et al Characterization of plasma metal profiles in Alzheimer’s disease using multivariate statistical analysis PLoS ONE. **2017**, 12, e0178271; see also Rochoy, M., Bordet, R., Gautier, S ,, Chazard, E. PLoS ONE**2019**, 14, e0220174.

[15] a) Schrauzer, G.N., Surai, P.F. Selenium in human and animal nutrition: Resolved and unresolved issues. Crit. Rev. Biotechnol.**2009**,29, 2–9; b) Stolwijk, J.M., Garje, R., Sieren, J.C., Buettner, G.R., Zakharia, Y. Understanding the Redox Biology of Selenium in the Search of Targeted Cancer Therapies Antioxidants**2020**,9, 420.

[16] Wassermann, A-M., Lounkine, L., Davies, J.W., Glick. M., Camargo, L.M. The opportunities of mining historical and collective data in drug discoveryDrugDiscovery Today**2015**20, Number 4

[17] Loging, W, Harland L, Williams-Jones B. High-throughput electronic biology: Mining information for drug discovery, Nature Rev. Drug Discovery.**2007**, 6, 220-230; McInnes, C. Virtual screening strategies in drug discovery. Curr. Opin. Chem. Biol., **2007**, 11, 494-502.

[18] Cruz-Monteagudo, M et al Drug Discovery Today **2017**, 22,994-1007; Achary, P.G.R. Mini Rev. Med. Chem.**2020**, 20 , 1375-1388.

[19] Mullane, K., Williams, M.Biochem. Pharmacol.**2020**, 177, Article Number 113945.

[20] Chu, X., Quan, Y, Zhang, H. Drug Discovery Today **2020**, 25,821-827; Penrod, N.M., Cowper-Sallari, R., Moore, J.H.Trends Pharmacol Sci. **2011**32, 623–630.

[21] Papassotiropoulos, A., de Quervain, D. Trends Cogn. Sci.**2015**, 19, 183-187.

[22] Katritzky, A.R. et al Quantitative measures of solventpolarity Chem.Rev.**2004,**104, 175-198; Reichardt, C, Solvatochromic Dyes as Solvent Polarity Indicators Chem. Rev.**1994**, 94, 2319–2358.

[23] Abraham, M.H. Scales of Solute Hydrogen-bonding: Their Construction and Application to PhysicochemicaI and BiochemicaI Processes Chem. Soc. Rev.**1993**, 22, 73-83; Abraham, M.H. et al , Application of hydrogen bonding calculations in property based drug design Drug. Discov. Today**2002**, 7, 1056-1063.

[24] Taylor, R., Wood, P.A. A Million Crystal Structures: The Whole Is Greater than the Sum of Its Parts, Chem. Rev.**2019**, 119, 9427-9477.

[25] Rzepa, H.S. Discovering More Chemical Concepts from 3D Chemical InformationSearches of Crystal Structure Databases J. Chem. Educ.**2016**, 93, 550-554.

[26] Wackerly, J.W. et al Using the Cambridge Structural Database To Teach Molecular Geometry Concepts in Organic Chemistry J. Chem. Educ.**2009**, 86, 460-464.

[27] a)Woodley, S.M., Day, G.M.,Catlow, R. Structure prediction of crystals, surfaces and nanoparticles. Phil. Trans. R.Soc. A**2020** , 378: 20190600. http://dx.doi.org/10.1098/rsta.2019.0600 b) Neumann, M.A. Tailor-Made Force Fields for Crystal-Structure Prediction J. Phys. Chem. B**2008**, 112, 32, 9810–9829 https://doi.org/10.1021/jp710575h; c) Graser, J. , Kauwe, S.K., Sparks**,** T.DMachine Learning and Energy Minimization approaches for Crystal Structure Predictions: A Review and New Horizons, Chem. Mater.**2018**, 30, 11, 3601–3612. https://doi.org/10.1021/acs.chemmater.7b05304 ; d) Day, G. M.; Motherwell et al A test of crystal structure prediction of small organic molecules Acta Crystallogr. B**2005**, 61, 511– 527, and references cited therein.

[28] Dunitz, J. D., Taylor, R.. Chem. Eur. J.**1997**, 3, 89–98; Dunitz, J.D.ChemBioChem**2004**, 5, 614 -621.

[29] Taylor, R. The hydrogen bond between N—H or O—H and organic fluorine: favourable yes, competitive no Acta Cryst.**2017,** B73, 474–488.

[30] Schneider, H.-J. Chem. Sci.**2012**, 3, 1381 – 1394.

[31] Desiraju, G.R., Steiner, T.The Weak Hydrogen Bond , Oxford University Press, Oxford etc, **1999**.

[32] see also Gilli P, Bertolasi, V. , Ferretti V. J. Am. Chem. Soc.**1994,** 116, 909-915.

[33] Swierczynski, D., Luboradzki, R., Dolgonos, G., Lipkowski, Schneider, H.-J. Non- Covalent Interactions of Organic Halogen Compounds with Aromatic Systems–Analyses of Crystal Structure DataEur. J. Org. Chem. **2005**, 1172-1177.

[34] Steffen A, Lengauer T, Wenz G, et al , New J. Chem., **2007**, 31, 1941.

[35] Biedermann, F., Schneider, H.-J. Experimental Binding Energies in Supramolecular Complexes, Chem. Rev., **2016**, 116 , 5216–5300, and references cite therein.