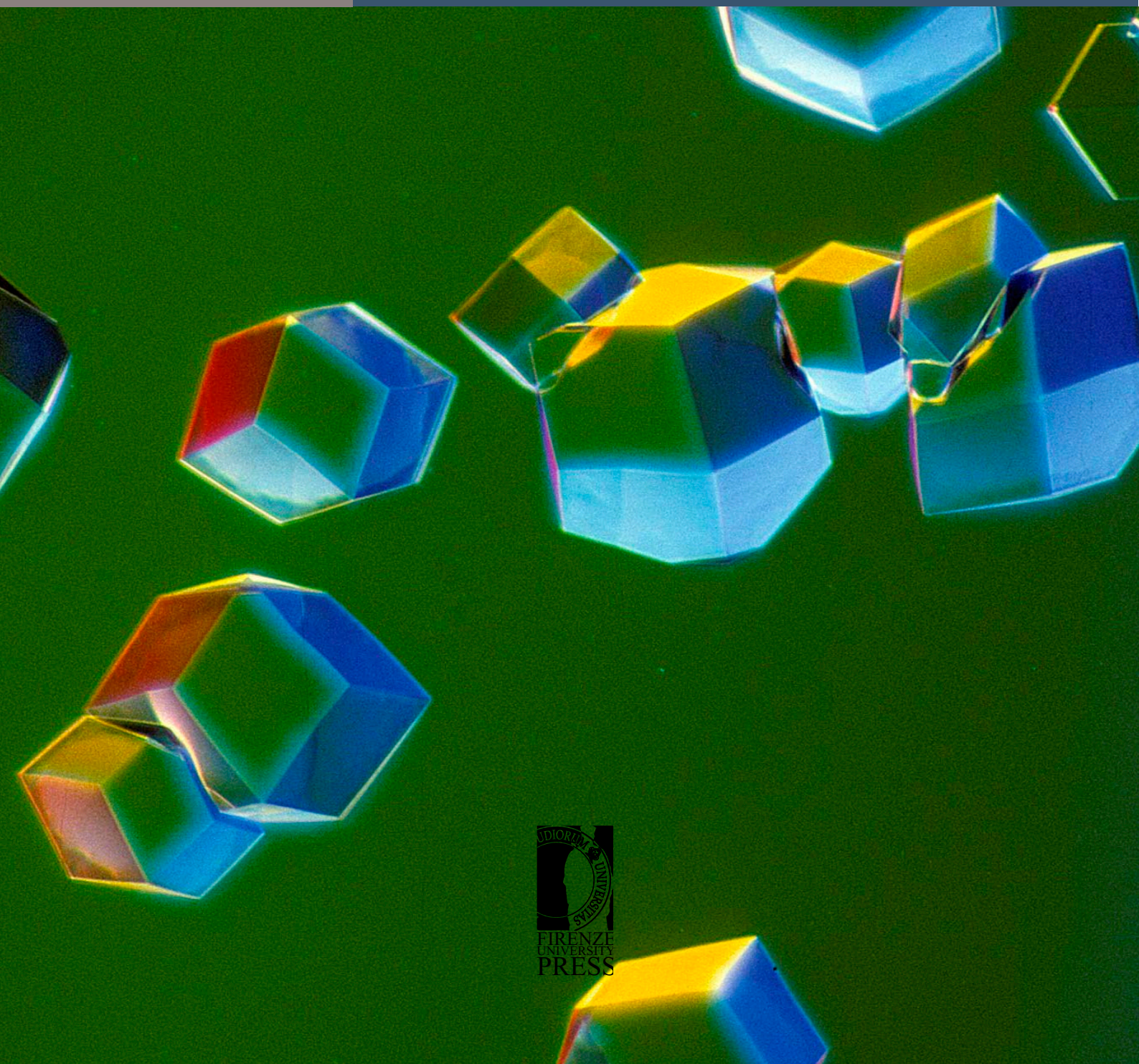
March 2019
Vol. 3 - n. 1

2532-3997

# Substantia

An International Journal of the History of Chemistry



FIRENZE UNIVERSITY PRESS

# *Substantia*

## An International Journal of the History of Chemistry

**Vol. 3, n. 1 - March 2019**

Firenze University Press

***Substantia*. An International Journal of the History of Chemistry**

*Published by*
**Firenze University Press** – University of Florence, Italy
Via Cittadella, 7 - 50144 Florence - Italy
http://www.fupress.com/substantia

Direttore Responsabile: **Romeo Perrotta**, University of Florence, Italy

Cover image: Brightfield micrograph with colored filters (magnification 14x) from crystals of influenza virus neuraminidase isolated from terns, by Julie Macklin and Graeme Laver, Australian Natl. Univ., Canberra. Courtesy of Nikon Small World (1st Place 1987 Photomicrography Competition) – www.nikonsmallworld.com



*Substantia* is honoured to declare the patronage of:

Museo Galileo - Institute and Museum of the History of Science
Piazza dei Giudici, 1 - 50122 Florence, Italy
http://www.museogalileo.it

Fondazione Prof. Enzo Ferroni - Onlus
Via della Lastruccia, 3
50019 Sesto Fiorentino, Italy
http://www.fondazioneferroni.it

UNIVERSITÀ DEGLI STUDI FIRENZE
DIPARTIMENTO DI CHIMICA "UGO SCHIFF"

With the financial support of:

Fondazione Cassa di Risparmio di Firenze,
Via M. Bufalini 6 - 50121 Firenze, Italy

Substantia
An International Journal of the History of Chemistry

# No walls. Just bridges

*Substantia* is a peer-reviewed, academic international journal dedicated to traditional perspectives as well as innovative and synergistic implications of history and philosophy of Chemistry.
It is meant to be a crucible for discussions on science, on making science and its outcomes.
*Substantia* hosts discussions on the connections between chemistry and other horizons of human activities, and on the historical aspects of chemistry.
*Substantia* is published *open access* twice a year and offers top quality original full papers, essays, experimental works, reviews, biographies and dissemination manuscripts.
All contributions are in English.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Editorial

# I won a project!

Juan Manuel García-Ruiz

*Laboratorio de Estudios Cristalográficos, Instituto Andaluz de Ciencias de la Tierra, CSIC-Universidad de Granada, Spain*

Yes, I know that playing the lottery is one way to pay taxes for those who do not know statistics. But in the future, we may "win in a raffle" to do many things, such as being a member of a board of directors, a councilman, or even a member of parliament because in the future it is very likely that councilmen, deputies and many other public positions will be chosen randomly. There is a controversial but solid theory supporting that randomness is one of the best mechanisms for optimizing selection processes[1,2,3,4]. Scientists have already begun to test this idea, and in fact, they may already get a project if they present it to an interesting program of the Volkswagen Foundation called EXPERIMENT![5]

The program "EXPERIMENT! In search of bold research ideas"[6] aims to fund radically new scientific ideas, ideas that go against the dominant thinking in a scientific discipline, crazy ideas or ideas of dubious feasibility that would have no or very little chance of being selected in the classic science funding program. Projects cannot formally last more than eighteen months and have maximum funding of one hundred and twenty-thousand euros. The program started in 2013 and is an absolute success. Every year, the Foundation receives around six hundred applications, prescriptively German. Six hundred and forty applications have been received this year.

The internal evaluation team of the Volkswagen Foundation selects one hundred and fifty of the most scientifically daring proposals, those best suited to the objectives of the program. Subsequently, these one hundred and fifty proposals are evaluated by a panel of ten scientists from different countries in the world, except Germany. This panel of experts rejects a few of those one hundred and fifty applications that for some important reason should not be funded by this program, mainly because they are not radically new or because they are obviously viable. Finally, out of all the others, the panel selects the fifteen that it considers the best, and which will be financed by EXPERIMENT! It is easy to see that selecting fifteen proposals, out of a hundred and fifty that have been selected from more than six hundred applications, is very complicated for an expert, not to mention agreeing on them with the other nine colleagues on the panel. To avoid endless discussions, each member of the panel has a joker, a wild card – which can only be used once – to approve a specific project, thus putting an end to the discussion about that project.

The Volkswagen Foundation tries to ensure that the selection is as impartial as possible. For example, the system is double-blind: neither the candidates know the panel members nor the panel members know who the candidates are. There are no names of people or institutions on the forms, and the foundation itself takes care of deleting any possible data from the proposal that could be used to identify the candidates' names, age, genre, or university of origin. But even so, the existence of a problem of equanimity derived from the enormous competitiveness of the program has been detected.

When experts evaluate and compare those ca. one

[1] B. Henning, *The end of politicians: Time for a real democracy,* 2017.
[2] L. Carson, P. F. L. Carson, B. Martin, *Random selection in politics*. Greenwood Publishing Group, 1999.
[3] O. Dowlen, *The political potential of sortition: A study of the random selection of citizens for public office*. Andrews UK Limited, 2017.
[4] G. Delannoi, O. Dowlen, *Sortition: Thoery and Practice*. Andrews UK Limited, 2016.
[5] The non-profit Volkswagen Foundation is the largest private foundation for research and academic teaching in Germany, spending more than 200 million euros in 2018. Despite its name, it is independent and not affiliated with the automaker company.
[6] https://www.volkswagenstiftung.de/en/funding/our-funding-portfolio-at-a-glance/experiment

hundred and forty research proposals that they have considered, in principle, eligible for funding by the program, they always find some of them outstanding, which should be clearly funded. Let us say there are five of them. However, when it comes to selecting the other ten that can still be funded, they find that there are many more than ten proposals that are so good that it is technically impossible to decide which of them is better than the others. And that's when problems arise. When the differences between projects are small, when it is difficult for an expert to assess the superiority of one project over another objectively, aspects come into play that are subjective to the evaluator and that cause the rational evaluation system to fail. Among these factors is the tribal instinct of scientists, that is, the irresistible tendency to support those projects that are closer to their discipline and their way of thinking, what we could call intellectual nepotism. In addition to introducing injustice in the evaluation, this bias favors the most common disciplines over the rare ones, reducing the thematic diversity of the selected proposals.

In order to tackle this problem EXPERIMENT! has, for the last two years, launched an experiment that may seem too daring to some. But that's what this program is all about! The experiment consists in selecting not only the fifteen projects by the panel of experts but also an identical number of projects by lottery. Not among all the projects submitted, but among all the projects considered eligible for funding by the panel, including the fifteen approved for their technical quality in the opinion of the evaluators. That is to say, fifteen projects are selected by technical evaluation of the experts and fifteen projects by pure chance, by lottery. A total of twenty-five projects have been selected this year because, during the lottery, projects already approved by the panel can be awarded. Only a list of the twenty-five projects is made public without revealing which were selected by the panel and which by lottery, and the follow-up and treatment that the Foundation will make of all of them will be identical. The comparative study of the benefits of the two selection systems will be carried out by an external evaluation company. We will see what comes out of this trial, the first to be conducted with a significant number of projects.

The idea of raffling project funding repels the academic world. Accustomed to peer review, i.e., decisions about the quality of a paper (to be published) or a project (to be funded) or a researcher or professor (to fill a position) are made by experts of the same rank as the candidates, the proposal that an entire academic effort be the subject of a lottery draw, abandoned at random, seems unfair, irrational, even obscene. However, precisely one of the stronger points of the lottery system is the cost/benefit ratio for the researcher as well and for the advancement of science.

A study has recently been published which concludes that when calls for funding research projects are very competitive, the effort researchers waste in writing their proposals may be comparable to the total scientific value of the research they intend to support[7]. The authors of the study themselves suggest that it would be more effective to replace peer review with a partial system of lotteries – such as EXPERIMENT! or to fund on the basis of researchers' past scientific successes rather than on their research proposals for the future.

Of course, many considerations can be made about the goodness of a lottery funding system. It depends on the external framework in which the researcher operates, the type of research program, the length and difficulty of the application forms, the number of calls to which a researcher can apply in a given country, the reasons for which it is presented, whether merely scientific or rather promotional, etc. But, in my opinion, the draw system is not unworthy and it should be investigated on which context its effectiveness depends and which modifications would optimize it. It should be explored as what it is, as a complex system, and its behavior analyzed with numerical simulations and the analysis of real cases such as the EXPERIMENT! program. And, of course, the equations "selection by peer review = fair and rational" and "selection by lottery = unfair and capricious" should be forgotten: the lottery comes into play when the technical evaluation system by peer review ceases to be fair and effective, and not to replace it but to improve it.

Nowadays, the use of chance in the management of public affairs is reduced to popular juries in some countries. However, the lottery selection mechanism has been used in many moments of history by political systems that have worked well, from classical Greece to the prosperous and stable republics of Venice or Florence[8]. In the outstanding Greece of the 6th century B.C., practically all public positions were chosen by lottery. Even army positions, excluding, for reasons of efficiency, those of the highest rank. The lottery system was widely used in the selection of public offices in Florence in the fourteenth and fifteenth centuries, and even the *doge* of Venice, as well as many of the public and elective offices of the city of the Signoria, were chosen by a complicated

[7] K. Gross, C. T. Bergstrom, Contest models highlight inherent inefficiencies of scientific funding competitions. *PLoS biology*, 2019, vol. 17, no 1, p. e3000065

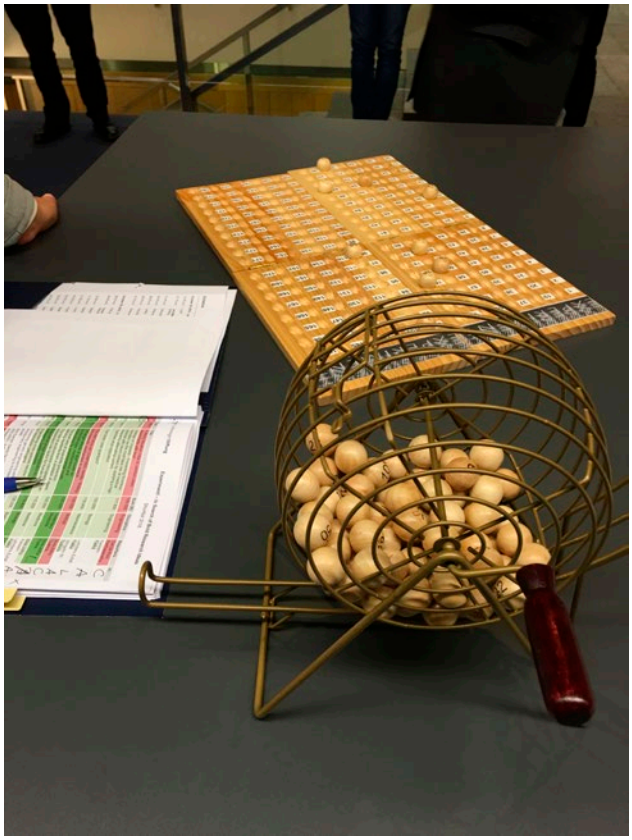
[8] B. Manin, *The Principles of Representative Government*. Cambridge University Press, 1997.

system that included largely random selection[9].

The advantages of the random selection system are many, since, for example, it complicates corruption and bribery, makes factions useless, makes unnatural agreements impossible, disqualifies long-term promises, and reduces electoral expenditure to almost zero. Imagine a congress in which deputies were elected at random. Imagine a lady from Spain, a farmer, a lesbian chosen by pure chance to be member of the European Parliament. She could not say "we lesbians think", nor "we women farmers believe", nor "we Spanish want", because she would realize, or they would make her realize, that she is not there representing anyone except herself and that the strength of the system is that each of the raffled seats in the Parliament votes and decides in their own conscience, for their own interests. That sum of non-prostituted interests is what gives strength to the lottery election mechanism. But let's leave the management of public affairs for another time, and let's return, to finish, to the academy, that is what interests me now.

In my opinion, the most worrying thing about the evaluation of EXPERIMENT! is how to make an objective and relevant comparison between the two groups of projects funded, those selected by the panel of experts and those selected by lottery. As we made clear at the beginning, this program is looking for bold, daring, doubtlessly viable projects based on ideas that move in the diffuse and changing frontier of knowledge. How to evaluate the results of projects that by their very nature should fail in most cases? What criteria should be used to qualify the productivity of a project that is going to explore a niche not yet trodden by science? This problem is totally new in evaluation and its solution is nothing trivial.

On the other hand, the result of the comparison will be very dependent on the composition of the panel, on the selection criteria of its components. When we had to design the evaluation system for the EXPLORA Program – dare to discover, dare to be wrong – a pioneering Spanish program in the financing of bold ideas, it became clear that the database of the National Evaluation Agency should not be used. The reason is that this task requires colleagues who are open-minded, non-egocentric, intellectually generous, with excellent scientific culture, and if possible with a certain sense of smell to detect in a proposal the semi-hidden potential that straddles the genius and the naive. We have to look for evaluators who would have bet on Columbus, on Marconi, on Wegener. That is not easy. Only nine years ago, during the evaluation of a program for bold ideas, an advanced facial recognition project and another one about crypto currency were rejected, because they were useless (who's going to be interested in that?). The role of the panel of experts is crucial because the final list of projects selected by these programs where the intellectual risk is assessed is the only, or more precisely, the best message that can be sent to future candidates to convince them that, fortunately, there are programs that don't care about financing failure if the frontier of knowledge is explored with audacity.



ACKNOWLEDGMENTS

The author acknowledges the team of the Experiment! Program of the Volkswagen Foundation, and Dr. Enrique Perez (Institute of Astrophysics of Andalucía) for useful discussions of this subject. Dr. Alfonso García-Caballero is also acknowledged for help with the English version of the manuscript.

[9] J. S. Coggins, C. F. Perali. 64% Majority rule in Ducal Venice: Voting for the Doge. *Public Choice*, 1998, 97(4), 709-723. https://doi.org/10.1023/A:1004947715017

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Research Article

# Similarities and contrasts in the structure and function of the calcium transporter ATP2A1 and the copper transporter ATP7B

Giuseppe Inesi

*California Pacific Medical Center Research Institute, 32 Southridge Rd West, Tiburon CA 94920, USA*
E-mail: giuseppeinesi@gmail.com

OPEN ACCESS

**Citation:** G. Inesi (2019) Similarities and contrasts in the structure and function of the calcium transporter ATP2A1 and the copper transporter ATP7B. *Substantia* 3(1): 9-17. doi: 10.13128/Substantia-68



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Competing Interests:** The Author(s) declare(s) no conflict of interest.

**Abstract**. $Ca^{2+}$ and $Cu^{2+}$ ATPases are enzyme proteins that utilize ATP for active transport of $Ca^{2+}$ or $Cu^{2+}$ across intracellular or cellular membranes.[1-4] These enzymes are referred to as P-type ATPases since they utilize ATP through formation of a phosphorylated intermediate (E-P) whose phosphorylation potential affects orientation and affinity of bound cations by means of extended conformational changes. Thereby specific cations are transported across membranes, forming transmembrane gradients in the case of $Ca^{2+}$, or accepting $Cu^{2+}$ from delivering proteins on one side of the membrane and releasing it to carrier proteins on the other side. Binding of $Ca^{2+}$ or $Cu^{2+}$ is required for enzyme activation and utilization of ATP by transfer of ATP terminal phosphate to a conserved aspartate residue. The ATPase protein is composed of a transmembrane region composed of helical segments and including the cation binding site (TMBS), and a cytosolic headpiece with three domains (A, N and P) containing the catalytic and phosphorylation site. The number of helical segments and the cytosolic headpieces present significant differences in the two enzymes. In addition, details of transmembrane cation extrusion are different. The $Ca^{2+}$ and $Cu^{2+}$ ATPase sustain vital physiological functions, such as muscle contraction and relaxation, activation of several cellular enzymes, and elimination of excess cation concentrations. A historic review of studies on chemical and physiological mechanisms of the $Ca^{2+}$ and $Cu^{2+}$ ATPase is presented.

**Keywords**. Calcium ATPase, Copper ATPase, Cation Active Transport.

## THE CALCIUM TRANSPORT ATPASE

The $Ca^{2+}$ATPase (SERCA) is a mammalian membrane bound protein sustaining $Ca^{2+}$ transport and involved in cell $Ca^{2+}$ signaling and homeostasis. It is made of a single polypeptide chain of 994 amino acid residues distributed in ten trans-membrane segments (M1 – M10) and a cytosolic headpiece including three distinct domains (A, N and P) that are directly involved in catalytic activity (Fig 1).[5]

The N domain contains residues such (Phe-487) interacting with the adenosine moiety of ATP whereby the ATP substrate is cross-linked to the P domain.

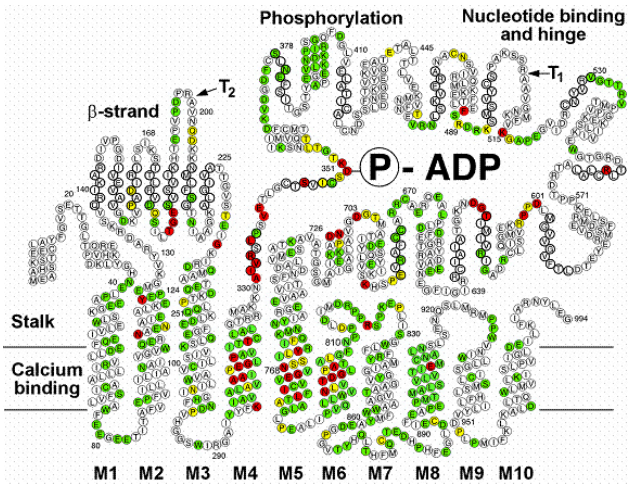



**Figure 1.** Aminoacid sequence and two dimensional folding model of the SERCA1 $Ca^{2+}$ ATPase.[5] See text for explanations.

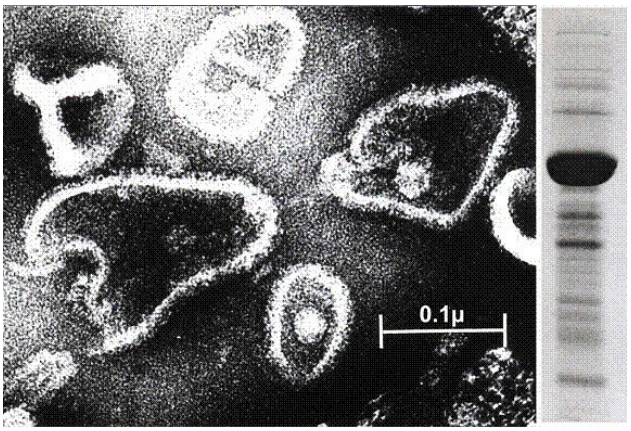



**Figure 2.** Purified vesicular fragments of sarcoplasmic reticulum membrane shown by negative staining on electron microscopy. On the right side, electrophoretic analysis demonstrates that the protein composition consists almost entirely of $Ca^{2+}$ ATPase.[8]

The P domain contains a residues (Asp-351) undergoing phosphorylation to yield a phosphorylated intermediate (E-P), a residue (Asp-703) coordinating $Mg^{2+}$, and other features characteristic of P-type ATPases. The A domain contains the signature sequence $^{181}$TGE that provides catalytic assistance for final hydrolytic cleavage of (E-P). Cooperative and sequential binding of $Ca^{2+}$ involved in catalytic activation and transport (Figs. 1 and 3) occurs on sites I and II located within the trans-membrane region.[6,7]

$Ca^{2+}$ATPases (SERCA1 and SERCA2) are associated with intracellular membranes of skeletal and cardiac muscle (sarcoplasmic reticulum: SR), and especially high concentrations with the skeletal muscle SR. Therefore, isolation of vesicular fragments of skeletal SR yields concentrated and fairly pure protein, shown by very frequent particles corresponding to ATPase protein visualized by electron microscopy (Fig 2 left panel), and prominent ATPase component visualized by electrophoresis (Fig 2, right panel).

This preparation is very convenient for functional and structural characterization of the ATPase.[8] In fact, it was demonstrated (8) with this preparation that, at equilibrium and in the absence of ATP, SERCA binds two $Ca^{2+}$ per mole, with high cooperativity and high affinity ($2.3 \times 10^6$ $M^{-1}$) (Fig. 3a) at neutral pH, although the affinity is lower at low pH and higher at higher pH.[9] When ATP is added to SR vesicles pre-incubated with $Ca^{2+}$ in rapid kinetic experiments (Fig 3b), the bound $Ca^{2+}$ facing the outer medium disappears soon (becomes non available to isotopic exchange, i.e., occluded), indicating that the outer opening of the binding cavity closes to the outside medium as soon as a first reaction product with ATP is formed.

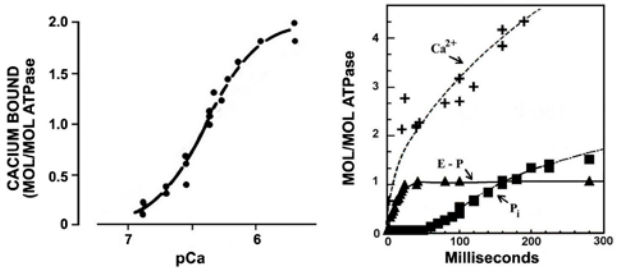



**Figure 3a.** $Ca^{2+}$ binding to SR ATPase under equilibrium conditions, in the absence of ATP. The stoichiometry of binding is 2 $Ca^{2+}$ per ATPase, with a binding constant of $2.3 \times 10^6$ $M^{-1}$, and high cooperativity.[9] Figure 3b. Pre-steady state activity of SR vesicles started by addition of ATP in the presence of $Ca^{2+}$. Note that upon addition of ATP a rapid burst of EP formation occurs and, at the same time, 2 $Ca^{2+}$ per ATPase become occluded. Steady state Pi production and further $Ca^{2+}$ uptake then follow, with a ratio of 2 $Ca^{2+}$ per Pi produced.[10]

Pi release and further $Ca^{2+}$ uptake then occur following a delay, indicating that trans-membrane $Ca^{2+}$ release and hydrolytic cleavage of EP occur after a slow step and, soon after that, further cycles contribute to steady state activity.[10]

Based on these kinetic observations, a diagram is shown in Fig. 4, where the basal enzyme is indicated as $2H^+{\cdot}E_2$. Following $2Ca^{2+}$ binding in exchange for $2H^+$, the active enzyme is referred to as $E_1.2Ca^+$. Following binding and utilization of ATP, the resulting phosphoenzyme is indicated as $ADP{\cdot}E_1$-$P.2Ca^{2+}$. Upon release of ADP, the free energy associated with this intermediate is utilized for a slow conformational change yielding transmembrane release of bound $Ca^{2+}$ in exchange for 2 $H^+$,

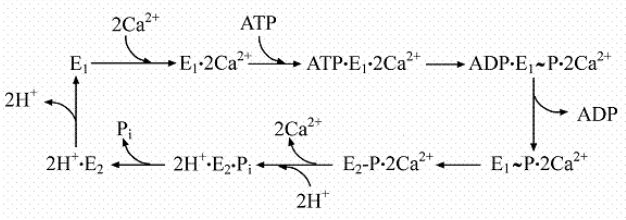



**Figure 4.** Diagram outlining the sequential reactions of a $Ca^{2+}$ ATPase cycle at neutral pH. The cycle starts with the enzyme in basal conformation, with $H^+$ bound at the specific $Ca^{2+}$ exchange site ($2H{\cdot}E_2$). Upon $H^+$ dissociation, $2Ca^{2+}$ bind and the enzyme is activated ($E_1{\cdot}2Ca$). ATP then leads to formation of the high potential phosphorylated intermediate ($ADP{\cdot}E_1{*}P{\cdot}2Ca$). Following dissociation of ADP, the phosphorylated intermediate uses its potential for a conformational change reducing affinity and orientation of bound calcium. 2 $Ca^{2+}$ are then dissociation in exchange for 2 $H^+$. The residual phosphoenzyme then undergoes hydrolytic cleavage with release of Pi, and returns to the basal conformation with $H^+$ bound ($2H{\cdot}E_{2)}$. The stoichiometry of $H^+$ binding is 2 per E at neutral pH. At high pH, less or no H+ exchanges for $Ca^{2+}$. Thereby $Ca^{2+}$ is not released before Pi cleavage, and the enzyme undergoes an uncoupled cycle.

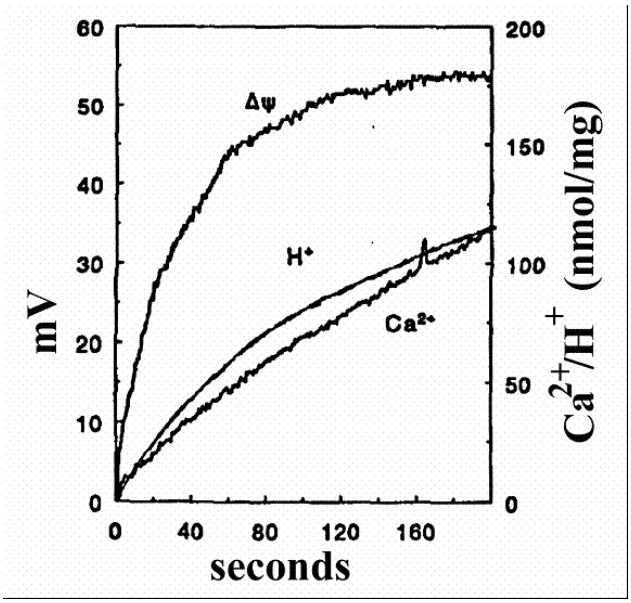



**Figure 5.** ATP dependent $Ca^{2+}$ uptake, $H^+$ countertransport and development of transmembrane electrical potential in reconstituted SERCA proteoliposomes. The proteoliposomes were placed in a neutral pH medium, containing 100 mM K2SO4, 50 microM $CaCl_2$, and color reagents for detection of $Ca^{2+}$, pH and electrochemical gradients. The reaction was started by the addition of 0.2 mM ATP, and followed by differential absorption spectrometry.[11]

followed by hydrolytic cleavage of Pi and return to the basic $2H^+{\cdot}E_2$ state.

The reaction scheme in Fig. 4 outlines a specific exchange of 2 $Ca^{2+}$ for $2H^+$ in the $2H^+{\cdot}E_2$ state and $2H^+$ for 2 $Ca^{2+}$ in the $E_2$-P·2$Ca^+$ state.

Clear evidence for this exchange was obtained with SERCA reconstituted in phospholipids vesicles that do not allow trans-membrane passive leak of charge which occurs in native SR membranes (except for transported $Ca^{2+}$). $Ca^{2+}$ and $H^+$ concentrations and electrical potential were then measured with appropriate sensors (Fig. 5). It was found that addition of ATP was accompanied by $Ca^{2+}$ uptake and stoichiometric $H^+$ extrusion, as well as formation of electrical potential.[11]

The important role of $H^+$ at the $Ca^{2+}$ sites was also demonstrated in experiments with native membrane vesicles, as it was found that phosphorylation of ATPase with Pi can be obtained only at acid pH. This indicates that upon 2 $H^+$ binding to $E_2$ (in exchange for $Ca^{2+}$ if present) the resulting $2H^+{\cdot}E_2$ acquires a specific conformation and free energy to allow phosphorylation with Pi, i.e. reversal of the $E_2$-P·2$Ca^{2+}$ to $2H^+.E_2$ step in the ATPase reaction cycle.[12]

Pioneering and highly informative crystallography by Toyoshima et al. revealed detailed structural information on the molecular structure of the entire molecule.[13] Nucleotide and phosphorylation domains of the $Ca^{2+}$ ATPase, relative to different stages of the enzyme cycle, are represented in Fig 6.[14] In the figure, the structure and conformational states of the $Ca^{2+}$ ATPase in the presence and absence of $Ca^{2+}$, substrate and product analogs are represented, with reference to $E_1$. 2$Ca^{2+}$, $E_1{\cdot}$AMPPCP, E-2·AlF4·(TG), and $E_2{\cdot}$(TG)·ATP, where TG (thapsigargin, a highly specific and potent SERCA inhibitor) is used to stabilize $E_2$.[13, 15, 16, 17] Color changes gradually from the N-terminus (blue) to the C-terminus (red). The two $Ca^{2+}$ (I and II) bound to the high affinity transmembrane site are circled when present. The two bound $Ca^{2+}$ undergo vectorial release in E2.AlF4·(TG), as the binding sites undergo a change in affinity and orientation. Three key residues (E183 in the A domain, D351 and D703 in the P domain) are shown in ball-and-stick. Note the positional change of headpiece domains in the various conformations. Note the nucleotide binding to the N domain, and variable relationship of the nucleotide phosphate chain (and $Mg^{2+}$) with the P and A domains.

As described above, kinetic and structural information yields a detailed understanding of the $Ca^{2+}$ ATPase catalytic and transport cycle as outlined in Fig. 4.

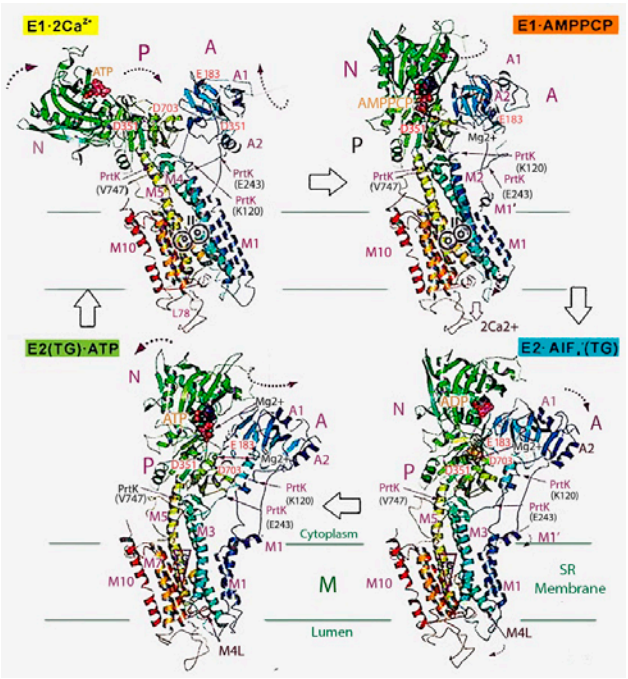



**Figure 6.** Sequence of conformational states of the calcium ATPase in the presence ($E1.2Ca^{2+}$), following nucleotide (analog) substrate binding (E1·AMPPCP), following enzyme phosphorylation (AlF4 analog) and $Ca^{2+}$ release ($E2.AlF_4$·TG) and in absence of $Ca^{2+}$ with bound ATP (E2(TG).ATP).[14]

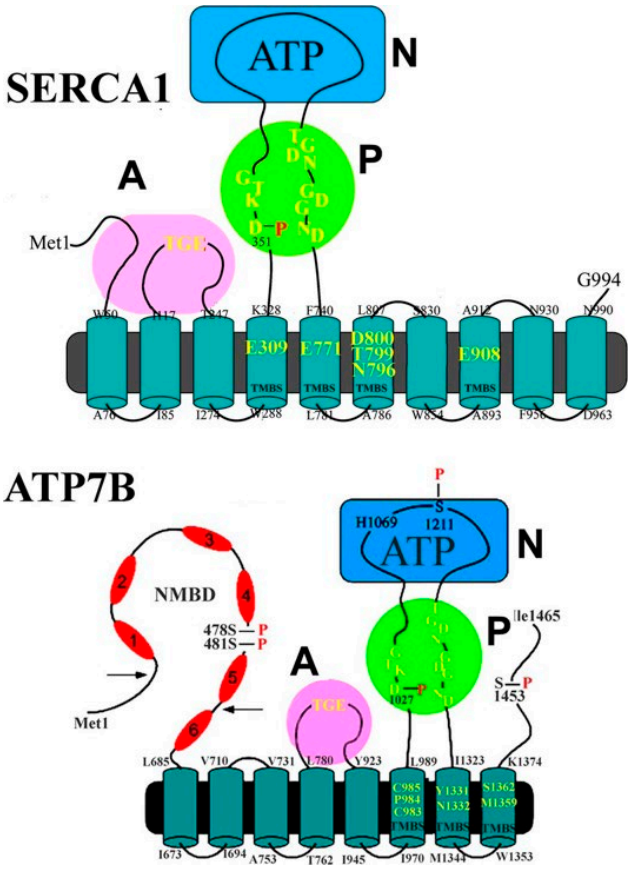



**Figure 7.** Two-dimensional folding models of the $Ca^{2+}$ ATPase (SERCA1) and $Cu^{+}$ ATPase (ATP7B) sequence. The diagram shows ten SERCA or eight ATP7B transmembrane domains including the calcium or copper binding sites (TMBS) involved in enzyme activation and cation transport. The extra-membranous regions of both enzymes comprises a nucleotide binding domain (N), the P domain with several conserved residues (in yellow) including the aspartate (Asp351 and Asp1027) undergoing phopsphorylation to form the catalytic intermediate (EP), and the A domain with the TGE conserved sequence involved in catalytic assistance of EP hydrolytic cleavage. The His1069 residue whose mutation is frequently found in the Wilson disease is shown in the ATP7B N domain. Specific features of ATP7B are the N-metal Binding Domain (NMBD) extension with six copper binding sites, and serine residues undergoing Protein Kinase assisted phosphorylation (Ser478, Ser 481, Ser1211, Ser1453).[23]

## THE COPPER TRANSPORT ATPASE

Bacterial and mammalian copper ATPases sustain active transport of copper by utilization of ATP . The mammalian $Cu^{+}$ ATPases include isoforms (ATP7A and ATP7B) that are involved in copper transfer from enterocytes to blood, copper export from the liver to the secretory pathways for incorporation into metalloproteins, and general copper homeostasis.[18,19] Genetic defects of ATP7A and ATP7B are related to human Menkes and Wilson diseases.[20, 21, 22]

$Cu^{2+}$ ATPases present functional analogies to the $Ca^{2+}$ ATPases, but specific differences as well. A comparison of SERCA and ATP7B bidimentional folding models (Fig. 7) shows that ATP7B comprises eight (rather than ten) transmembrane segments that include the copper binding site (TMBS) for catalytic activation and transport, and a headpiece comprising the N, P and A domains with conserved catalytic motifs analogous to SERCA.

A specific feature of ATP7B (less prominent in ATP7A, and absent in the bacterial copper ATPase) is an amino-terminal extension (NMBD) with six copper binding sites in addition to those in the TMBD. An additional feature is the presence of serine residues (Ser-478, Ser-481, Ser1211, Ser-1453 in ATB7B) undergoing kinase assisted phosphorylation.[23]

The native abundance of copper ATPase is quite low and, in order to accomplish biochemical experimentation, larger quantities were obtained by heterologous expression in insect or mammalian cells.[24, 25] It was found that addition of ATP to microsomes expressing heterologous ATP7B yields two fractions of phosphoryl-

ated ATPase protein, one acid labile corresponding to phosphoenzyme intermediate, and the other acid stable and dependent on kinase assisted phosphorylation. Acid labile phosphoenzyme is faster, and is not observed following mutation of the conserved aspartate (S1024) at the catalytic site, or following mutation of the transmembrane copper binding site (TMBD). Kinase assisted formation of alkali resistant phosphorylation is slower, involves Ser478, S481, Ser1121 and Ser1453, and is not observed in the presence of protein kinase inhibitors. Interestingly, it is not observed following mutation of the trans-membrane copper binding site (TMBD), indicating a dependence on enzyme activation ($E_2$ to $E_1$) transition.

Specific features of copper ATPase following addition of ATP are shown in Fig 8, to demonstrate the difference in phosphorylation of aspartate and serines in the copper ATPase. The time course of ATP7B following addition of ATP is shown in Fig 8A, with total phosphoenzyme (black squares) including acid and alkaline resistant (dark squares, including aspartate phosphoenzyme intermediate and phospho-serines), acid resistant (dark circles, i.e. aspartate phosphoenzyme intermediate) and alkaline resistant (light squares, i.e. phosphorserines). It is shown in Fig 8B that no alkaline resistant phospoenzyme (i.e. phospho-serines, light squares) is observed if protein kinase inhibitor is present, and in Fig 8C no acid resistant aspartate phosphoenzyme (dark circles) is observed when D1027N ATP7B is used. By comparison, it is shown in Fig 8D that WT SERCA undergoes only acid stable aspartate phosphorylation, and no alkali resistant serine (light squares) phosphorylation, i.e. the acid stable accounts for total phosphorylation .[25]

An estimate of $Cu^{2+}$ transport following phoshorylation of ATP7B with ATP was obtained by comparing microsomes of COS-1 expressing $Ca^{2+}$ ATPase (SERCA) or $Cu^{+}$ ATPae (ATP7B) absorbed on a solid supported membrane (SSM). The SSM consists of an alkanethiol monolayer covalently bound to a gold electrode via the sulfur atom and a phospholipids monolayer on top of it.[26, 27, 28] The adsorbed protein is activated by addition of ATP in the presence of a medium supporting ATPase activity. Related electrogenic events are recorded as current transients due to flow of electrons along the external circuit toward the electrode surface, as required to compensate for the potential difference across the vesicular membrane produced by displacement of positive charge upon vectorial translocation in the direction of the SSM electrode. When ATP is added to the membrane bound ATPase absorbed on the SSM in the presence of $Ca^{2+}$ or $Cu^{2+}$, a current transient is obtained due to vectorial translocation of bound $Ca^{2+}$ or $Cu^{2+}$ in the direction of the SSM electrode after phoshoenzyme

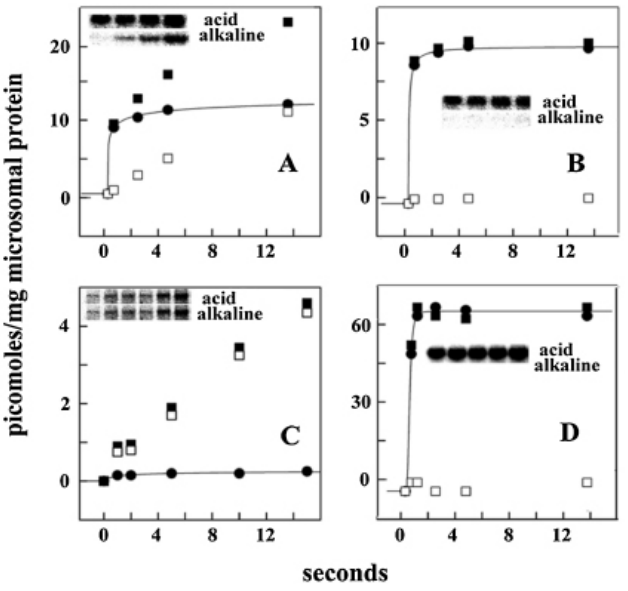



**Figure 8.** Phosphorylation of WT ATP7B (A, B), ATP7B D1027N mutant (C), and WT SERCA (D), in the absence (A, C and B) and in the presence of Proteinase K inhibitor. Microsomes obtained from COS-1 cells sustaining expression of the various ATPases were incubated with 50 microM (gamma -$^{32}$P)ATP in a reaction mixture sustaining enzyme activiy at 30 $^{o}$C, in the absence (A, C and D) or in the presence (B) of PKD inhibitor. Electrophoresis was then performed at acid pH to measure total phosphoproteins (black squares), or alkali pH to eliminate alkali labile phosphoenzyme and assess alkali resistant serine phosphorylation (empty squares). The difference (given in in the absence (A, C and D) or in the presence (B) of PKD inhibitor. Electrophoresis was then performed at acid pH to measure total phosphoproteins (black squares), or alkali pH to eliminate alkali labile phosphoenzyme and assess alkali resistant serine phosphorylation (solid black circles) corresponds to the phosphorylated aspartate enzyme intermediate.[25]

formation by utilization of ATP. In these experimental conditions, the electrogenic signal generated within the first enzyme cycle is observed.[29] It is shown in Fig 9A that in experiments with SERCA that the charge transfer observed at neutral pH is much reduced at acid pH. On the other hand, the charge transfer observed with ATP7B is significantly slower, and is not changed by alkaline or acid pH (Fig 9B). This difference is due to the lack of $Cu^{2+}/H^{+}$ exchange in the cation binding and release sites of the copper ATPase, as opposed to the requirement of $Ca^{2+}/H^{+}$ exchange in the calcium ATPase.

A crystallographic view of the copper ATPase protein and of the copper transport pathway across the membrane was obtained through LpCopA crystallization, trapped in the $E_2{\cdot}P_i$, as compared with E2P state.[31] The two states show the same conduit, appearing equivalent and open to the extracellular side, in contrast to the

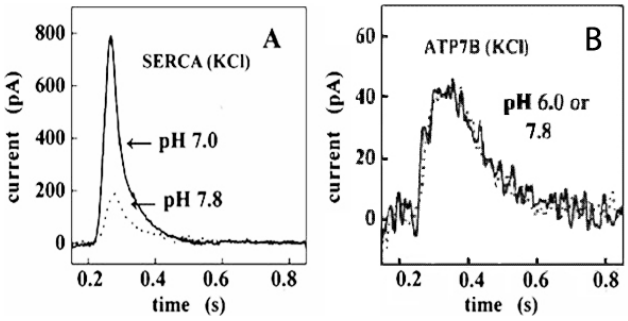



**Figure 9.** Charge measurements measured with enzymes absorbed on solid supports member (SSM). A: Current transients following addition of ATP to SERCA in a reaction mixture including 10 microM free Ca2+ and 100 mM KCl at pH 7.0 (solid line) or pH 7.8 (dotted lines). C: Current transients after addition to ATP on ATP7B in a reaction mixture containing 5 microM $Cu^+$ and 100 mM KCl at pH 6.0 (solid line) or 7.8 (dotted line).[28]

calcium ATPase where the $E_2.P_i$ state is occluded. In Fig 10a the A, P and N domains are colored in yellow, blue and red, respectively. The black arrows mark the copper transport pathway. In Fig 10b the E2P (pink) and E2.Pi (green) states are compared, showing movements of the extracellular domains (arrows), while the transmembrane domain remains rigid in two states, in contrast to the calcium ATPase where the E2.Pi becomes occluded. Fig 10c shows a close up of the extrusion pathway with the opening from the copper high affinity coordinating residues Cys382, Cys384, and Met717 shown as a red surface, with crystallographic water molecule shown as red spheres.

A diagrammatic comparison of the calcium and copper ATPases is shown in Fig 11, where the sequential conformational transitions of the catalytic and transport cycle are compared for calcium and copper ATPases.[30] We then see that the two calcium ions exit the ATPase from the E2P state, and the ion exit pathway closes concomitantly to hydrolytic cleavage of Pi and transition to the E2.Pi state. On the other hand, the copper ions exit the ATPase from the E2P state, but the exit pathway remains open in the E2.Pi state, and closes only in the E2 state after release of Pi.

Considering experimental results and modeling shown above, there seems to be a clear parallel between the difference in cation/proton exchange, and the conformational outcome in the exit pathways following cation release in the two ATPases. It is apparent that the closure of the release pathway in the calcium ATPase is due to $H^+$ binding in exchange for $Ca^{2+}$, and a consequent conformational effect on the E2.Pi state. The pathway closure in the copper ATPase occurs only following release of Pi and acquisition of the E2 conformation.

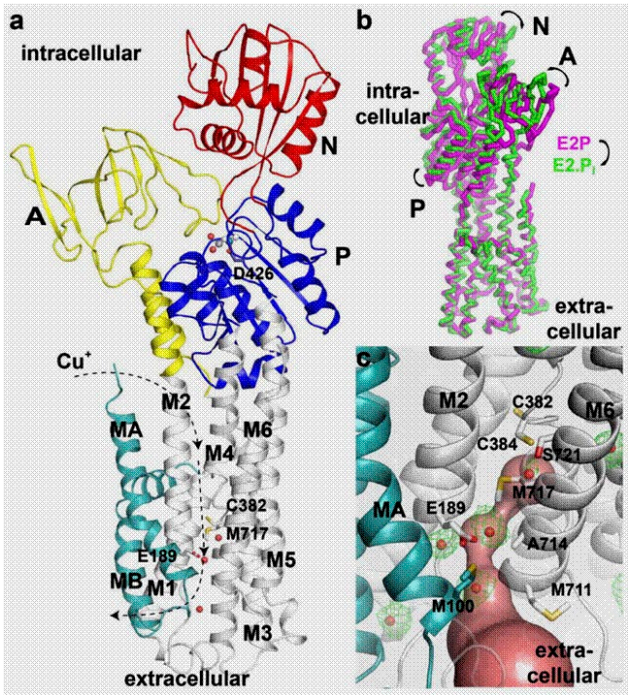



**Figure 10.** Diagrammatic representation showing that crystal waters of the E2-$BeF_3^-$ structure support the copper release pathway.[30]

A further distinctive feature of the copper ATPase is the effect of phosphorylation of serine residues catalyzed by Proteine Kinase D.[25] In experiments with microsomes of COS1 cells or hepatocytes expressing ATP7B it was found that utilization of ATP by ATP7B includes autophosphorylation of an aspartyl residue serving as the specific catalytic intermediate, as well as phosphorylation of serine residues catalyzed by Protein Kinase D. It is shown in Fig 12 A that ATP7B (stained in green) interacts first with TransGolgi network (blue) in perinuclear (nuclei red) location and, in the presence of $Cu^{2+}$, is transferred to intracellular trafficking vesicles. It is shown in Fig12 B that the trafficking is not interfered with by mutation of the TMBD Asp1027 (whose phosphorylation serves as phosphoenzyme intermediate). On the other hand, trafficking is interfered by Ser478, 481, 1121 and 1453 mutations in the NMD (Fig12C), by TMBS copper site mutation (Fig 12D), and by mutation of the 6th NMBD copper site mutation (Fig 12 E). This demonstrates that the NMD, absent in the calcium ATPase, plays a determinant role in conformational adaptations required for functions of the copper ATPase.

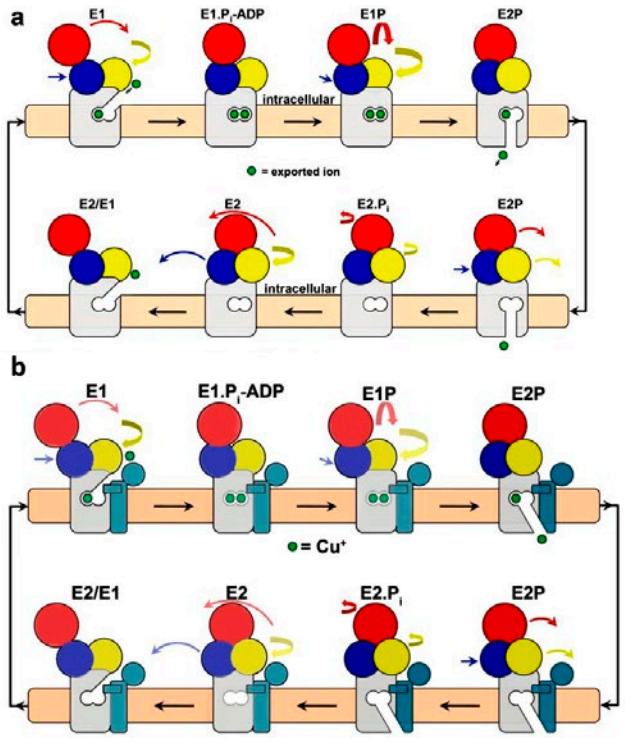



**Figure 11.** Diagrammatic comparison of the calcium and copper ATPases, showing the sequential conformational transitions of the catalytic and transport cycle. The two transported calcium ions exit the ATPase from the E2P state, and the ion exit pathway closes concomitantly to hydrolytic cleavage of Pi and transition to the E2.Pi state. On the other hand, the copper ions exit the ATPase from the E2P state, but the exit pathway remains open in the E2.Pi state, and closes only in the E2 state after release of Pi.[30]

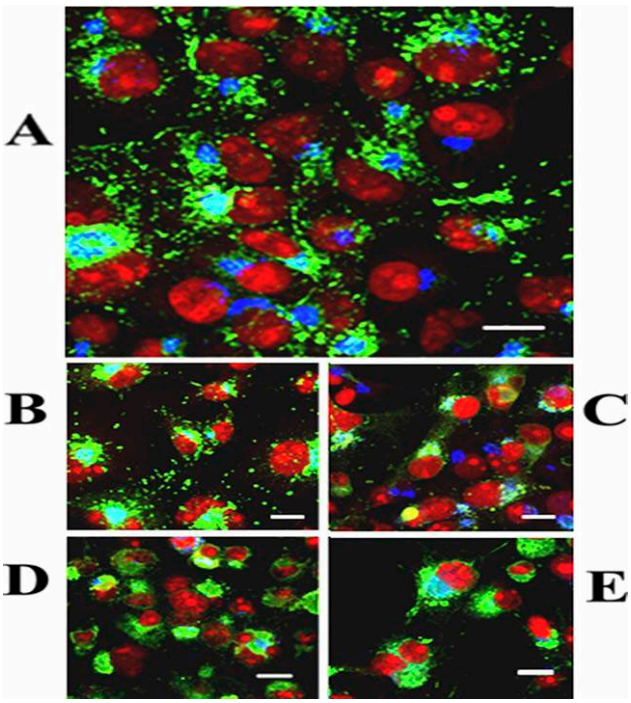



**Figure 12.** Intracellular distribution of ATP7b in COS1 cells expressing WT enzyme (A), subjected to mutation at Asp-1027 (B), Ser-478, Ser-481, Ser-1121 and SER-1453 (C), at the transmembrane (TMBD) copper site (D), or at the sixth NMBD copper site (E). Note the presence of cytosolic trafficking vesicles with WT enzyme (A), and even and even following Asp-1027 mutation (B), but no trafficking following serine, NMBD or TMBD copper sites (C, D and E).[25, 29]

## PHYSIOLOGICAL ROLES OF $CA^{2+}$ AND $CU^+$ ATPASES

$Ca^{2+}$ is a specific activator of muscle fibrils. Activation of contraction depends on $Ca^{2+}$ delivery and, in turn, and relaxation depends on reduction of $Ca^{2+}$ concentration in the cytoplasm of skeletal and cardiac muscle cells. At rest, the cytosolic concentration of $Ca^{2+}$ is much lower than in extracellular fluids and in the intracellular vesicles of sarcoplasmic reticulum. Muscle activation occurs when plasma membrane electrical action potentials open passive $Ca^{2+}$ channels, allowing flux of $Ca^{2+}$ in the cytosol for activation of myofibrils. Following the end of action potential, passive channels close, and cytosolic $Ca^{2+}$ is returned to extracellular fluids and to the sarcoplasmic reticulum interior through active transport by the $Ca^{2+}$ ATPase. Due to time limits and quantities of $Ca^{2+}$ available, passive fluxes and active transport across the sarcoplamic reticulum membrane are much prevalent over those across the outer plasma membrane. In the diagram on Fig 13, a cardiac myocyte is shown with the $Ca^{2+}$ ATPase (ATP) inserted in the plasma membrane (sarcolemma) and the sarcoplasmic reticulum membrane, collecting $Ca^{2+}$ to induce relaxation, and to be then released upon membrane excitation to induce contraction upon binding to myofibrils.[31] The inset shows the time course of an electrical action potential, $Ca^{2+}$ release, and occurrence of contraction. Channels for passive diffusion of $Ca^{2+}$, and mitochondria are also shown.

Copper is a required metal for homeostasis of plants, bacteria and eukaryotic organisms, determining conformation and activity of many metalloproteins and enzyme such as cytochrome oxidase and superoxide dismutase. Furthermore, due to possible reactivity with non-specific proteins and toxic effects, elaborate systems of absorption, concentration buffering, delivery of specific protein sites and elimination, require a complex system including small carriers, chaperones and active transporters. The P-type copper ATPases provide and important system for acquisition, active transport, distribution and elimination of copper. A diagram of copper distribution in eukaryot-

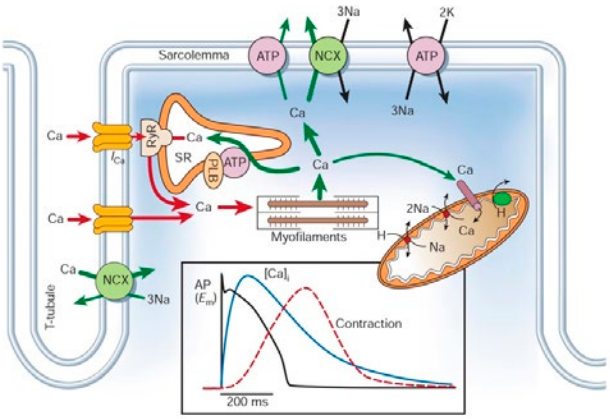



**Figure 13.** $Ca^{2+}$ transport in ventricular cardiac myocytes. ATP: ATPase. NCX: $Na^+/Ca^{2+}$ exchange, SR: sarcoplasmic reticulum. Bers DM. Cardiac exitation-contraction coupling.[31] Nature 2002; 415 :198-205

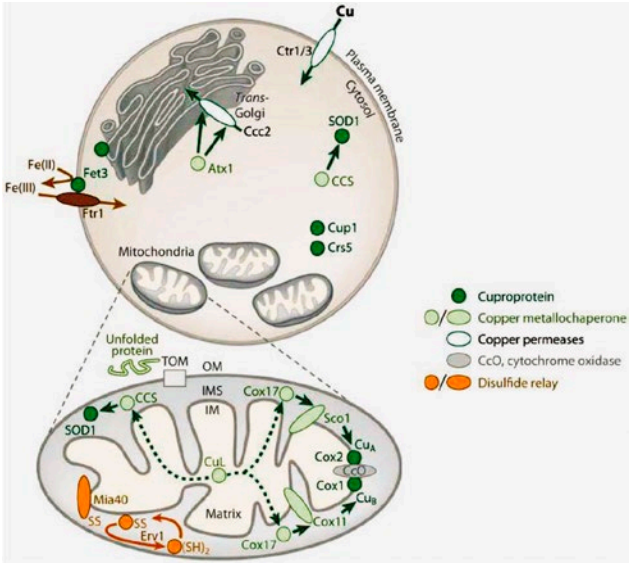



**Figure 14.** Diagram of copper eukaryoyic cellular distribution. See text for explanations.[32]

ic cells is given in Fig. 14, where it is shown that copper is imported into the cells copper permeases (Ctr1: oval cell membrane bound circles).[32]

Incoming $Cu^{2+}$ does not remain free in the cytosol, but is rather bound to various chaperones delivering it to specific proteins and secretory pathway. The $Cu^{2+}$ ATPase (Ccc2 in the figure with the trans-Golgi-Network) binds $Cu^{2+}$ through the intervention of the Atx1 chaperone, for delivery and transport across the cell membrane, or other destination depending on cell specificity. $Cu^{2+}$ delivery to the cytochrome c oxidase complex (CcO) involves Cox11, Sco1 and Cox 17 chaperones. Nuclear encoded chaperone proteins are imported unfolded across the mitochondrial membrane by a translocase, and then acquired in the inner mitochondrial space following introduction of disulphide bonds with the intervention of specific coupled enzyme.

In summary, it is evident that $Ca^{2+}$ and $Cu^{2+}$ ATPases are indispensable components of physiological systems, and the chemistry of their catalytic and transport mechanism is linked to biological function. Transport ATPases are required to regulate the concentrations of $Ca^{2+}$ and $Cu^{2+}$ within cells and cellular compartments, utilizing the energy of ATP to sustain appropriate concentrations across membranes. Appropriate cation concentrations are required to activate specific enzymes in one direction, and to produce relaxation and avoid toxic consequences in the other direction.

## REFERENCES


1. R.W. Albers, *Ann. Rev. Biochem*. **1967**, 36, 727-756.
2. R.L. Post, C. Hegyvary, S. Kume, *J. Bio. Chem*. **1972**, 247, 6530.
3. L. de Meis, A. Vianna, *Ann. Rev. Biochem*. **1979**, 48, 275.
4. G. Inesi, *J. Cell Commun Signal* **2011**, 5, 227.
5. D.H. MacLennan, C.J. Brandl, B. Korczak, N.M. Green, *Nature* **1985**, 316, 696.
6. D.M. Clarke, T.W. Loo, G. Inesi, D.H. MacLennan *Nature* **1989**, 339, 476.
7. J. P. Anderssen , B. Vilsen, *FEBS Letters* **1995**, 359: 101.
8. D. Scales, G. Inesi, *Biophys J*. **1976**, 16, 735.
9. G. Inesi, M. Kurzmack, C. Coan, D. Lewis, D., *J. Biol. Chem*. **1980**, 255, 3025.
10. S. Verjovski-Almeida, M. Kurzmack, G. Inesi, *Biochemistry* **1978**, 17, 5006.
11. L. Hao, J.L. Rigaud, G. Inesi, *J. Biol. Chem*. **1994**, 269, 14268.
12. L. deMeis, G. Inesi, *Biochemistry* **1985**, 24, 922.
13. C. Toyoshima, M. Nakasako, H. Nomura, H. Ogawa. *Nature* **2000**, 405, 647.
14. G. Inesi, D. Lewis, H. Ma, A. Prasad, C. Toyoshima, *Biochemistry* **2006**, 45, 13769.
15. C. Toyoshima, T. Mizutani, *Nature* **2004**, 430, 529.
16. C. Olesen, T.L. Sorensen, R.C. Nielsen, J.V. Moller, P. Nissen, *Science* **2004**, 306, 2251.
17. A.L. Jensen, T.L. Sorensen, V. Olesen, J.V. Moller, P. Nissen, *The EMBO Journal* **2006**, 25, 2305.
18. J.M. Arguello, S.J. Patel, J. Quintana, *Matallomics* **2016**, 8, 906.

19. S. Lutsenko, N.L. Barnes, M.Y. Bartee, O.Y. Dmitriev, *Physiol Rev*. **2007**, 87, 1011.
20. C. Vulpe, B. Levinson, S. Whitney, S. Packman, J. Gitschier, *Nat. Genet* **1993**, 7, 13.
21. J.D. Gitlin, *Gastroenterology*. **2003**, 125, 1868.
22. M. Harada, *Med. Electron. Microsc*. **2002**, 35, 61.
23. R. Pilankatta, D. Lewis, C. M. Adams, G. Inesi, *J. Biol. Chem*. **2009**, 284, 21207.
24. Y.H. Hung, M.J. Layton, I. Voskoboinik, J. F. Mercer, J. Camakaris, *Biochem J*. **2007**, 401: 569.
25. R Pilankatta, D. Lewis, G. Inesi. J.Biol Chem **2011**, 286, 7389.
26. J. Pintschovius, K. Fendler, E. Bamberg, *Biophys. J.* **1999**, 76, 827.
27. F. Tadini-Buoninsegni, G. Bartolommei, M.R. Moncelli, R. Guidelli, G. Inesi, *J. Biol. Chem*. **2006**, 281, 37720.
28. G. Tadini-Buoninsegni, G. Bartolommei, M.R. Moncelli, R. Pilankatta, D. Lewis, G. Inesi, *FEBS Lett*. **2010**, 584, 4519.
29. D. Lewis, R. Pilankatta, G. Inesi, G. Bartolommei, M. R. Moncelli, F. Tadini-Buoninsegni, *J. Biol. Chem*. **2012**, 287, 32717.
30. M. Andersson, D. Mattle, O. Sitsel, A.M. Nielsen, S. H.White, P. Nissen, P. Gourdon, *Nat. Struct. Mol. Biol*. **2013**, 21: 43-48.
31. M.D. Bers, *Nature* **2002**, 415, 198.
32. N.J. Robinson, D.R. Winge, *Annual Rev. Biochem*. **2010**, 79, 537.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Research Article

# Finding Na,K-ATPase II - From fluxes to ion movements

Hans-Jürgen Apell

*Dept. of Biology, University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Germany*
E-mail: h-j.apell@uni-konstanz.de

OPEN ACCESS

**Citation:** H.-J. Apell (2019) Finding Na,K-ATPase II - From fluxes to ion movements. *Substantia* 3(1): 19-41. doi: 10.13128/Substantia-207



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Competing Interests:** The Author(s) declare(s) no conflict of interest.

**Abstract.** After identification of the Na,K-ATPase as active ion transporter that maintains the $Na^+$ and $K^+$ concentration gradient across the membrane of virtually all animal cells, a long history of mechanistic studies began in which enzyme activity and ion-transport were intensively investigated. A basis for detailed understanding was laid in the so-called Post-Albers pump cycle. Developing new experimental techniques allowed the determination of different flux modes, the analysis of the kinetics of enzyme phosphorylation and dephosphorylation as well as of the transport of $Na^+$ and $K^+$ ions across the membrane. The accumulation of results from transport studies allowed the proposal of the gated channel concept that turned out to be a successful approach to explain the transport-related experimental findings. Eventually, it found its counterpart in the high-resolution structure of the ion pump. Recently it turned out that simple mutations of the Na,K-ATPase are the cause of several diseases.

**Keywords**. Ion transport, enzyme activity, flux modes, structure-function relation, electrogenicity, gated-channel concept, pump-related diseases.

**Dedicated to the late Prof. David C. Gadsby (1947-2019), a brilliant physiologist and biophysicist**

## I. DEVELOPMENT OF A FUNCTIONAL CONCEPT

In the 1950s the need for active ion transport through membranes was recognized. A number of concepts of the molecular mechanism of active transport had been proposed and discussed before the identification of the Na,K-ATPase. During that time James F. Danielli reviewed five possible mechanism that summarized the ideas.[1] They were adaptations of the carrier mechanism, which at that time had already been introduced as concept for passive ion transport. To perform active transport contractile proteins were coupled to the carrier to enable appropriately directed substrate transport. A different approach was proposed in 1957 by Peter Mitchell. His idea was substrate binding in a transporter to specific sites that experience translocation across the membrane by a rocking mechanism.[2] This proposal was published the same year as when Jens P. Skou identified the Na,K-ATPase as protein in crab nerve cell membranes.[3]

## II. THE POST-ALBERS CYCLE

The consequence of Skou's identification of the Na,K-ATPase and the fact that the ion pump could be selectively inhibited by ouabain (or other cardiac steroids) led directly to numerous target-oriented studies that provided a wealth of characteristical details.[4] A first proposal of the pump mechanism was published in 1963 by R. Wayne Albers and colleagues.[5] They discussed, as a possible pump mechanism of the "adenosine phosphatase" in the electrophorus electric organ $Na^+$-ATPase, a transphosphorylation in which phosphates were transferred along a chain of sites for phosphorylation from the cytoplasmic to extracellular side. $Na^+$ transport was suggested to be a by-product of the oriented transphosphorylation by acting as counter ion to the phosphate. $K^+$ transport might have been coupled with phosphate uptake.

Compilation of the continuously increasing experimental findings led Robert L. Post and Amar K. Sen in 1965 to a first guess of a reaction cycle with seven states.[6] In 1967 it was followed by the presentation of a reaction cycle by Albers and collaborators that described the enzymatic reactions, in which phosphorylation by ATP required the presence of $Na^+$ and dephosphorylation required $K^+$.[7] The four steps of the cycle are shown in Fig. 1A. In its $E_1$ forms the enzyme had inwardly oriented cation sites of high $Na^+$ affinity. The $E_2$ forms were characterized by outwardly oriented cation sites of high $K^+$ affinity.

While Albers and collaborators focused their view on the enzyme activity of the ion pump, Post et al. included five years later detailed information on the ion transport and presented the first pump cycle that assigned enzyme and transport activity together in a unified reaction cycle.[8] His proposal accounted not only for the physiological Na,K-ATPase function but also for Na-ATPase activity observed under (two) unphysiological conditions (Fig. 1B). This scheme, the so-called Post-Albers cycle, has become the prototypical reaction cycle of all P-type ATPases studied so far. Two basic features of the pump mechanism of the Na,K-ATPase are captured in this scheme: (1) The transport is performed in a consecutive (or "Ping-Pong") mode, which means that at first one ion species is translocated in one direction, then, after an exchange of ions, the second species is conveyed in the opposite direction. Since it was impossible to establish that $Na^+$ and $K^+$ were bound to the ion pump at the same time, it was suggested that they bind alternatingly to the same spatial sites which exhibit different binding affinities when accessible from one or the other side of the membrane. (2) $Na^+$ transport is connected to enzyme phosphorylation by ATP, $K^+$ transport takes place when the enzyme runs through the dephosphorylation half cycle. Two non-physiological pump modes, included in Post's proposal, were identified when the substrate conditions were modified appropriately: The first was observed when $K^+$ was removed from the extracellular medium. Nevertheless, a ouabain-sensitive but significantly reduced ATPase activity was detected. This finding led to the suggestion of a ATPase that is able to transport $Na^+$ out of the cell without a counter-transport of $K^+$, named "Na-ATPase". The second modi-

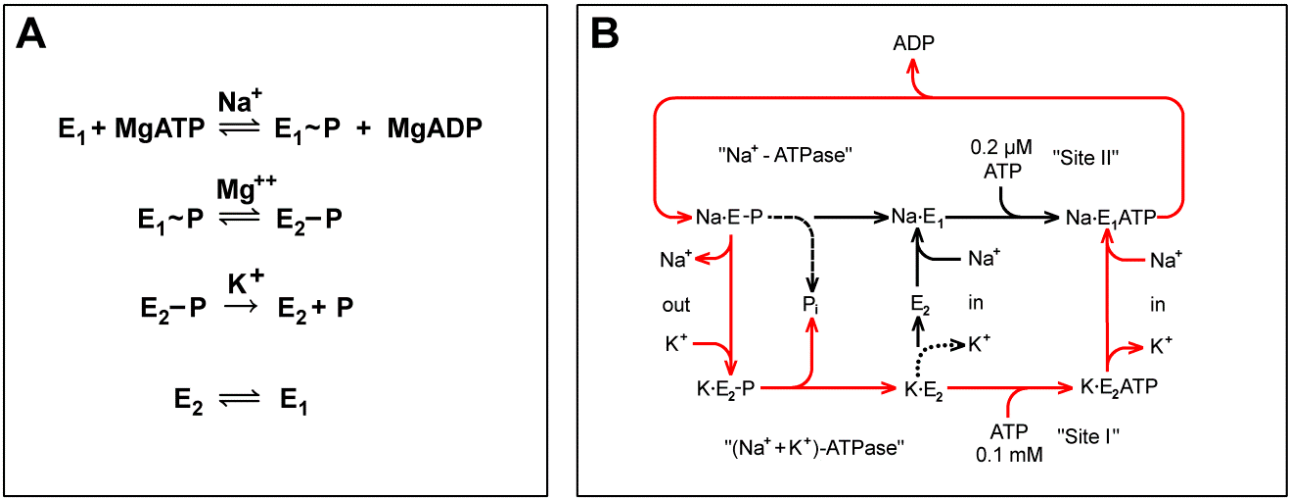



**Figure 1.** First representation of the pump cycle of the Na,K-ATPase. **A:** Reaction cycle of the enzyme activity with the ion substrates needed to enable the respective reaction step, published in 1967 by Albers and collaborators.[7] **B:** Reaction cycle with merged enzyme and transport functions, the so-called Post-Albers cycle, introduced 1972 by Post and collaborators.[8] The cycle marked in red represents the physiological mode.

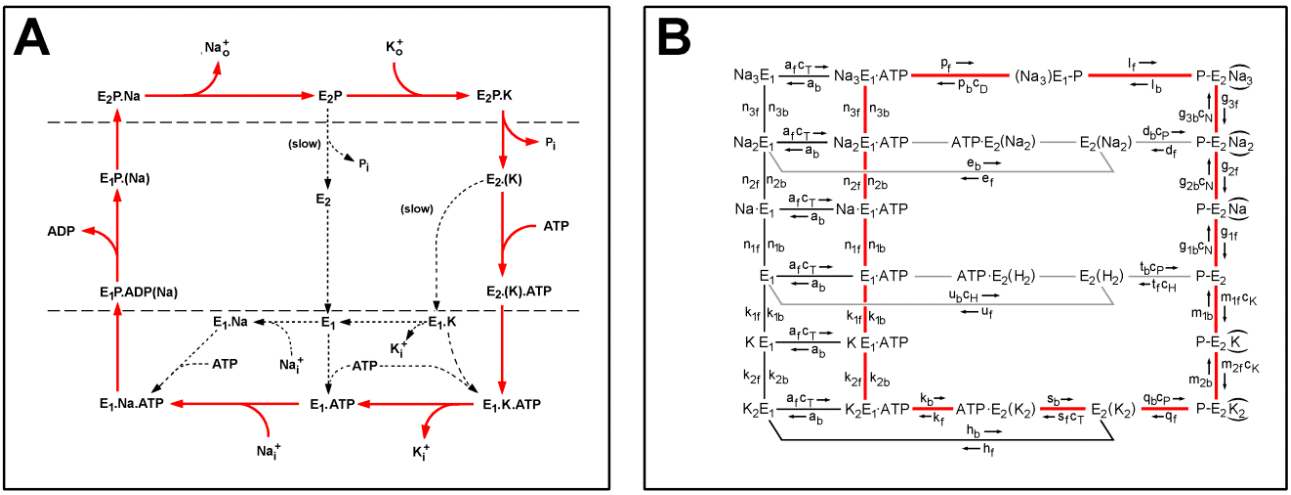



**Figure 2:** Development of the Post-Albers cycle with enhanced complexity provoked by increasing experimental insights. **A:** Inclusion of $Na^+$ and $K^+$ occluded states, indicated by framing the ions in parentheses, (Na) and (K). This scheme was adapted from Karlish and collaborators.[11] **B:** Pump scheme composed of all reaction steps that were determined experimentally from rabbit kidney ATPase until 1994, and which was used for successful numerical simulations of the experimental results.[12] The cycles drawn in red represent the physiological pump mode.

fication of the pump mechanism was found when the ATP concentration in the cell was reduced to values far below the physiological millimolar range. The dependence of the ATPase activity on the ATP concentration revealed the existence of two distinct binding affinities, "low affinity binding" in the range above 10 µM, and a high affinity binding in the 100 nM range. The low-affinity binding was associated with the $Rb^+$ (or $K^+$) bound $E_2$ conformation. Binding of ATP accelerated the transition from the $E_2$ to the $E_1$ conformation and deocclusion of the ion sites. But this process was not accompanied by phosphorylation of the enzyme.[8] High affinity binding of ATP occurred in the $E_1$ conformation and the presence of $Na^+$.[9] It was also shown that the ATP-concentration dependence revealed both low and high affinity binding of the nucleotide in the presence of $Na^+$ and $K^+$. In the absence of $K^+$ only high-affinity binding was detected[10] that took place in the $Na^+$ exporting half cycle, which was passed through in both pump modes.

In the years after 1972 extensive series of kinetic studies were published by several authors, in which a whole range of research activities were covered, such as studies on conformation transitions, enzymatic and transport activities. Based on these findings Karlish and collaborators presented in 1978 an extended Post-Albers scheme (Fig. 2A) which summarized all observed functional properties known at that time.[11] In their pump cycle an additional fundamental characteristic of the transport process, ion occlusion, was included: In the $E_1$ conformation the ion-binding sites were accessible from the cytoplasm, but after 3 $Na^+$ ions were bound, the enzyme became phosphorylated by ATP, and simultaneously the access to the ion-binding sites was locked and the ions were trapped inside the protein. Only thereafter the access of the ion sites to the extracellular side was unlocked, a process related to the conformation transition from $E_1$ to $E_2$, and the $Na^+$ ions were released. With respect to the $K^+$-transporting half cycle a corresponding reaction sequence occurred: Enzyme dephosphorylation caused occlusion of the $K^+$-loaded ion sites, and release of the $K^+$ ions to the cytoplasm only happened after deocclusion which was coupled to the conformation transition from $E_2$ to $E_1$.

Systematic and particularly time-resolved kinetic measurements led to a further extended Post-Albers scheme that allowed a successful simulation of those experiments.[12] The pump scheme shown in Fig. 2B is adapted from Heyse et al. and includes all six known flux modes of the Na,K-ATPase (see below). In this reaction scheme those reaction steps which have been identified in experiments with rabbit-kidney Na,K-ATPase are labeled with rate constants. The rate constants were either directly measured, determined from experiments or calculated from theoretical constraints. The reaction cycle shown in red represents in clockwise direction the Post-Albers cycle under physiological conditions. The counterclockwise reaction sequence describes the performance of the Na,K-ATPase as ATP synthase.

## III. FLUX MODES

Under diverse specific substrate conditions at least six additional transport modes ("non-canonical flux modes") were detected besides the physiological transport mode in which 3 $Na^+$ were removed from the cytoplasm in exchange against 2 $K^+$ taken up from the extracellular medium.[13,14] These flux modes are:

(1) Pump reversal, which can be observed at high intracellular concentrations of $K^+$, ADP and inorganic phosphate, $P_i$ as well as low concentration of ATP, high extracellular concentration of $Na^+$ and in the absence of $K^+$.[15,16] In this substrate condition the pump cycle is run through backwards and ATP in synthesized.

(2) Isostoichiometric exchange of $Na^+$ across the cell membrane was found to have taken place in the absence of $K^+$ and the presence of cytoplasmic ADP. In this mode the Na,K-ATPase acted as $Na^+$ shuttle by which the $Na^+$-translocating half cycle was executed forward and backward. First, 3 $Na^+$ were transferred out of the cell under consumption of ATP, then the 3 $Na^+$ were exchanged on the outside and transported back into the cell while ATP was produced from ADP and $P_i$, i.e. no net consumption of ATP took place in this mode.[17]

(3) Isostoichiometric exchange of $K^+$ operated also as shuttle service in which the $K^+$-translocating half cycle was executed forward and backward. 2 $K^+$ were bound extracellularly and transported into the cell via enzyme dephosphorylation and binding of ATP. In the absence of intracellular $Na^+$ and the presence of $P_i$ the physiological process was reversed by $K^+$ binding from the cytoplasmic side, release of the bound ATP and enzyme phosphorylation by $P_i$. ATP was bound but not hydrolyzed.[18] ATP was required only to promote the $E_2/E_1$ conformation transition.

(4) Uncoupled $Na^+$ efflux consuming ATP could be measured when neither $Na^+$ nor $K^+$ were present extracellularly.[19,20] In this mode it was assumed for a long time that after external release of $Na^+$ the pump cycle was completed by a return from the $E_2$-P to the $E_1$ conformation with empty binding sites. Not so long ago it was revealed, however, that this rather small flux (compared to the $Na^+$,$K^+$ mode) was only apparently uncoupled, but a $Na^+$,$H^+$ exchange in which protons were transported into the cell as $K^+$ congeners much less effectively but with the standard stoichiometry of 3 $Na^+$/2 $H^+$/ATP.[21]

(5) $Na^+$ exchange consuming ATP was detected in the absence of external $K^+$ but in the presence of $Na^+$ on both sides of the membrane.[22,23] This mode evolved from the uncoupled $Na^+$ efflux with increasing external $Na^+$ concentration.[14] An obvious mechanistic explanation for this flux mode was that the extracellular $Na^+$ acted as (less well fitting) congener of $K^+$ with a stoichiometry of 3 $Na^+$/2 $Na^+$/ATP.[10]

(6) Finally, an uncoupled $K^+$ efflux from red blood cells was found in the absence of extracellular $Na^+$ and $K^+$ that did not require the presence of ATP.[24] In the light of $H^+$ acting as congener of $K^+$ this flux mode may be explained also as shuttle mechanism exchanging $K^+$ and $H^+$ in homology to mode (3). This concept would avoid the necessity of an energetically less favorable return of the pump from state $E_2$-P to $E_1$ with empty ion-binding sites, as proposed in the originally published mechanism.[24]

## IV. ELECTROGENICITY

An important feature of the Na,K-ATPase (and of biological ion transporters in general) is the transfer of ions from one side of the membrane to the other, because ions are charged particles and well soluble in water but not in the membrane. A main task of the cell membrane is to exactly prevent unfacilitated permeation of ions between different compartments of the cell.

Therefore, structure and properties of biological membranes are optimized to reduce diffusion of ions through the membrane to a minimum. Repulsive electrostatic interactions are the predominant reason for this effect.[25] The charge of an ion is the origin of an electric field that influences the surrounding matter by attracting charges of opposite sign, repelling charges of the same sign, and reorienting electric dipoles. The energy needed to promote these responses in matter is provided by the kinetic energy of the moving ion. Because of the long range of electrostatic interactions a considerable portion of the surrounding matter is involved, and the amount of energy needed is dependent on a property of the matter called polarizability. The higher the energy is to displace a charge or reorient a dipole and make way for ion movement, the less probable it is that an ion will be able to permeate through the matter. In Fig. 3A a schematic representation of the energy profile of a lipid membrane is shown. The rise of the potential energy close to the water-membrane interface indicates the amount of energy needed to transfer the ion from the water into the hydrophobic and "apolar" core of the membrane formed by the fatty acids of the lipid molecules. This amount is large compared to the thermal energy of the ions. Therefore, a common property

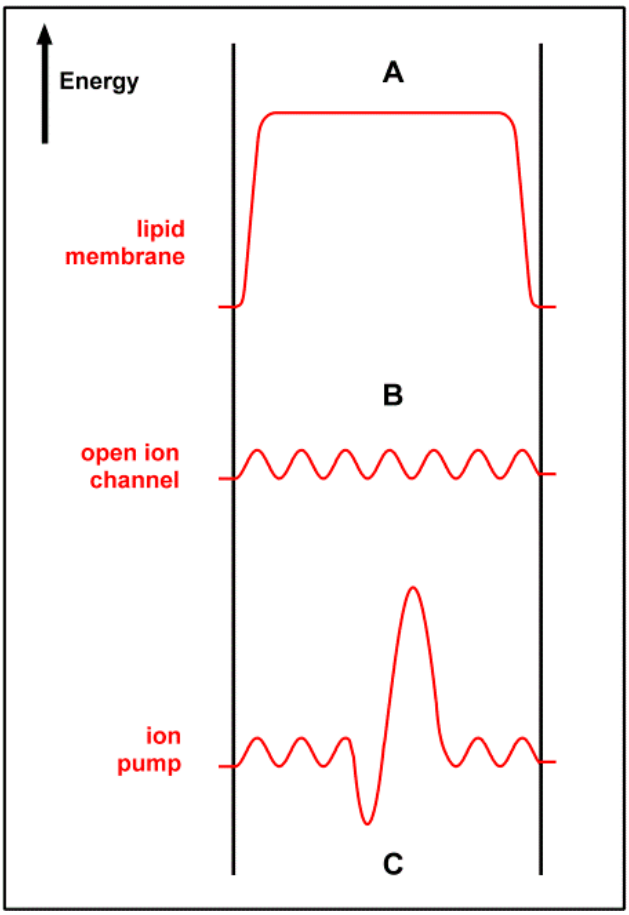



**Figure 3.** Schematic representation of potential energy profiles as detected by an ion along a pathway across a membrane in three different cases: (A) a simple lipid membrane, in which specific effects of the membrane-water interface are neglected, (B) an ideal ion channel without binding sites, and (C) an ion pump with an internal ion-binding site accessible from the left side and an energy barrier preventing propagation to the right-hand aqueous phase.

of all ion transporters in membranes has to be that they provide a pathway through the membrane that requires only a low amount of energy to be passed by an ion. This is accomplished by a pathway with a diameter that exceeds an ion-species dependent minimum and a lining of the pathway by molecules or parts of molecules such as amino-acid side chains that are easily polarizable (or "polar"). In Fig. 3B the energy profile of an ideal ion channel without ion-binding sites is depicted. Thus, the energy expenditure is low enough to allow an easy diffusion of ions along the pathway.

That can be implemented, as one extreme, by a wide, water-filled corridor or, as the other extreme, by a narrow channel lined with polar groups. An example for the latter is the gramicidin channel with a diameter of 4 Å in which carbonyl groups mimic the hydration shell which the passing monovalent cations have to leave behind for the most part at the entrance into the channel.[26] Numerous variations of channel shapes were found in between both extremes throughout the "channel kingdom". In the case of ion pumps, a pathway must exist that allows ion movement at low energy cost, however, it may not be continuous between both sides of the membrane because that would create a counterproductive bypass for ions. A promising proposal, the gated channel concept, which will be discussed in the subsequent chapter, is outlined in one of its states in Fig. 3C. Here an ion-binding site, indicated by a dimple in the energy profile, is accessible from the left side. A discharge of the ion from its binding site to the right side of the membrane is prevented by a high energy barrier.

The fact that a biological membrane consists primarily of a core of hydrophobic, apolar and insulating matter enclosed between conducting aqueous phases, allows the representation of a membrane as physical (plate) capacitor, which has in principle exactly this composition. Since the layer between both conductive plates of a capacitor is named "dielectric", the layer of the membrane formed by fatty acids is called membrane dielectric. It is characterized by a "dielectric constant", ε, that is low in the case of apolar matter (e.g. lipids, ε = 3-4) and high in polar phases (water, ε = 80). It controls the membrane capacity, $C = \varepsilon \cdot A/d$, where $A$ is the membrane area and $d$ its thickness. A fundamental consequence of this membrane property is that the transfer of an ion across the membrane is an "electrogenic" process.

Electrogenic transport is defined as the movement of electric charge through a medium with a low dielectric constant such as a biological membrane.[27,28] Electrogenic transport is characterized by two basic properties that were and are exploited constantly to study details of ion transport in the Na,K-ATPase and other ion transporters. The first impact of electrogenicity is that ion transport through the membrane dielectric produces an electric current and affects the electric membrane potential, $V_m$ (Fig. 4). Therefore, electrogenic ion pumps act as current generators, and charge movements can be detected as current signals with an external measuring device.[27]

The second impact is that the activity of an electrogenic transporter is affected by the membrane potential. When charges are moved inside the membrane in the course of voltage-dependent reaction steps, they move 'uphill' (as in Fig. 4) or 'downhill' on the electric membrane potential, Δφ. This generates an additional energy term for the process, $\Delta E = \Delta q \cdot \Delta\varphi$, which in turn modifies the rate constant of this reaction step and can be detected as altered kinetic behavior. In consequence,

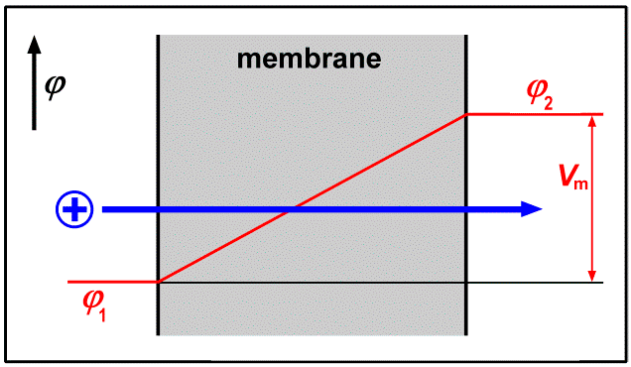



**Figure 4.** Schematic representation of electrogenic transport. The red line indicates the course of the electric membrane potential, here in case of a homogeneous membrane dielectric. The difference of the electric potentials on both sides, $\varphi_1 - \varphi_2$, is the membrane potential $V_m$. In cells it is always inside negative. According to basic principles of electrostatics, the movement of a charge, $\Delta q$, from one side of a capacitor to the other alters the electric potential difference, $\Delta V_m = C_m \cdot \Delta q$, proportionally to the membrane capacitance, $C_m$.

the externally measured pump current becomes voltage dependent.[29]

When the electrogenicity of an ion pump is investigated one may find an overall electrogenic behavior as in the case of the Na,K-ATPase or the sarcoplasmic reticulum (SR) Ca-ATPase, when after a complete pump cycle net charge is transferred across the membrane. In the case of the H,K-ATPase, two $K^+$ are exchanged against two $H^+$, therefore, no net charge is transferred after a pump cycle, the overall transport is electroneutral. But when the pump cycle is subdivided into single reaction steps, in some of these partial reactions charge is moved within the membrane dielectric, and these steps show electrogenic behavior.[30] The experimental concept to confine the activity of the Na,K-ATPase to specific partial reactions by appropriate experimental conditions, has turned out to be a powerful approach to identify electrogenic reaction steps in the pump cycle and to analyze their kinetic behavior.[12]

## V. THE GATED CHANNEL CONCEPT

As indicated above, the initial ideas of the molecular mechanism of ion translocation across the membrane by ion pumps were influenced by the carrier concept[1] or a rocking mechanism.[2] In both cases ions bind in a first step to sites provided by the protein and then these sites, imbedding the ion in a cage, are moved through the membrane. Already in 1957 Clifford S. Patlak introduced another mechanistic proposal on a more general level, the "gate type non-carrier mechanism".[31] He assumed that the transporter had a substrate-chelating moiety that was not physically displaced during pumping. Initially, it was accessible only from one side of the membrane at a time (Fig. 5A). Then the transporter "closed" on the approachable side and "opened" subsequently on the opposite side, where the substrate was released. After the site is empty, the conformational arrangement is reversed and the transporter returned to its initial state.

In 1979 Peter Läuger published an enhancement of this concept as "channel mechanism for electrogenic ion pumps"[28] in which he assumed an ion channel traversing the membrane inside the transport protein with varying energy barriers that were able to separate the ion-binding site from the external aqueous phases. He applied this approach first to the light-driven proton pump bacteriorhodopsin. A few years later he introduced this concept to ATPases (Fig. 5B),[32] and used it to provide a detailed microscopic model to analyze the current-voltage behavior of the Na,K-ATPase.[27,33] In this concept the ion pump was represented by a channel which consisted of a sequence of shallow energy dimples and two barriers that were able to change their height when the pumps ran through its multiple conformational states (Fig. 5B). Those variable barriers correspond to the gates of the channel.

There is a variety of designs possible to construct the ion pathway in the gated channel concept. The ion-binding sites can be arranged asymmetrically, i.e. close to one interface of the ion pump with the aqueous outside, or symmetrically buried deep inside the hydrophobic core of the protein. The implications would be the existence of one or two access channel, respectively, through which the ions have to move. For the sake of the gating mechanism, which is needed on either side of the binding sites, these must not be located on the protein's surface. Access channels may, however, differ in their shape. Two principal cases have to be distinguished, a narrow, even ion-selective ion channel (or "ion well") in contrast to a wide funnel (or "vestibule") that is filled with water molecules and various ions. In case of a narrow channel, the ions moving through it may be partly stripped of their hydration shell and interact with the wall-forming amino-acid side chains. The diffusion of ions through this structure resembles the process taking place in a typical ion channel. An important feature in this case is that part of the transmembrane voltage drops along the length of the channel and this action would be electrogenic (see above).[27] In the case of a wide open vestibule ion movement, it occurs more or less as free diffusion in a solution, and correspondingly, the electric conductance in this environment is high. This fact entails that

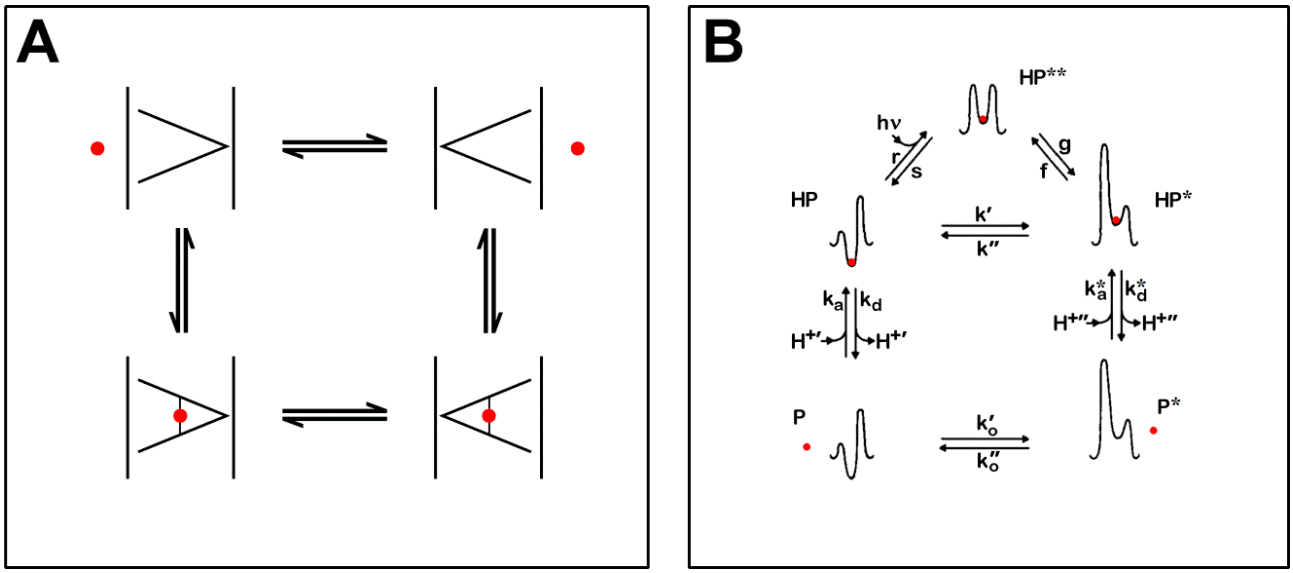



**Figure 5.** Original mechanistic concepts in which the ion-binding sites are not displaced during the transport process rather than the protein structure that controls access to these sites. **A:** The first proposal was the so-called gate type non-carrier mechanism by C.S. Patlak[31] in which two different conformational states of the protein generate alternatingly access to immobile ion sites from either side of the membrane. (Scheme adapted from Ref. 31). **B:** The second proposal is a channel mechanism in which ions diffuse through a low-resistance access channel (or 'ion well') to a binding site inside the membrane domain of the transporter, which is framed by mobile barriers on either side that control access from the outside. Here it is applied to the light-driven proton pump bacteriorhodopsin. In this representation the energy profiles of the access channels were omitted on both sides.(Scheme adapted from Ref. 32).

electric-field strength is low in such a vestibule, and no (significant) drop of the transmembrane voltage occurs. Correspondingly, both appearances were described in the literature as high-field and low-field access channels, respectively.

Based on the assumption that in the Na,K-ATPase the transported ions have to pass through access channels on either side of the ion-binding sites, the transfer of the ions from one aqueous phase to the other can be subdivided into at least four different reaction steps which are ion binding from one side, ion occlusion, deocclusion on the opposite side of the membrane, and ion release, as depicted in Fig. 6 for the $Na^+$-translocating half cycle. For $K^+$ transport a corresponding series of transport steps is valid.

In principle, all partial reactions indicated in color in the simplified Post-Albers cycle (Fig. 6A) may be accompanied by ion movements within (or through) the Na,K-ATPase. When, as in this case, a $Na^+$ ion is at the beginning in the cytoplasm, it resides at the electric potential, $V_m$, of the cell. At the end of the transport process, it is located outside the cell, where the level of the electric potential is 0 (per definition). Therefore, at each of the indicated reaction steps the ion may move through a fraction of the membrane potential, $V_m$. In the cartoon of Fig. 6B, these fractions of $V_m$ were indicated by the parameters α', β', β", and α". For each ion that traverses $V_m$ completely the condition, α' + β' + β" + α" = 1 must hold. These parameters were termed as "dielectric coefficients".[33] By determining their magnitude experimentally, important information can be achieved on the molecular mechanism of the ion transport. If the dielectric coefficient is zero, no charge is moved through the electric field within the membrane domain of the Na,K-ATPase. This has to be expected if the ion moves in a wide water-filled vestibule or if it is sterically fixed within the protein, e.g. in an immobile binding site. A high dielectric coefficient of a specific reaction step indicates a movement through a narrow channel. As will be shown later in detail, in case of the Na,K-ATPase the coefficients β' and β" were zero (or not significantly different from zero) which means that during enzyme phosphorylation and ion occlusion as well as during the conformation transition and ion deocclusion the ions were not shifted within the ion pump, their binding sites were sterically immobile.[34,35]

## VI. EXPERIMENTAL APPROACHES

After the identification of the Na,K-ATPase the first functional study was restricted to monitoring of the

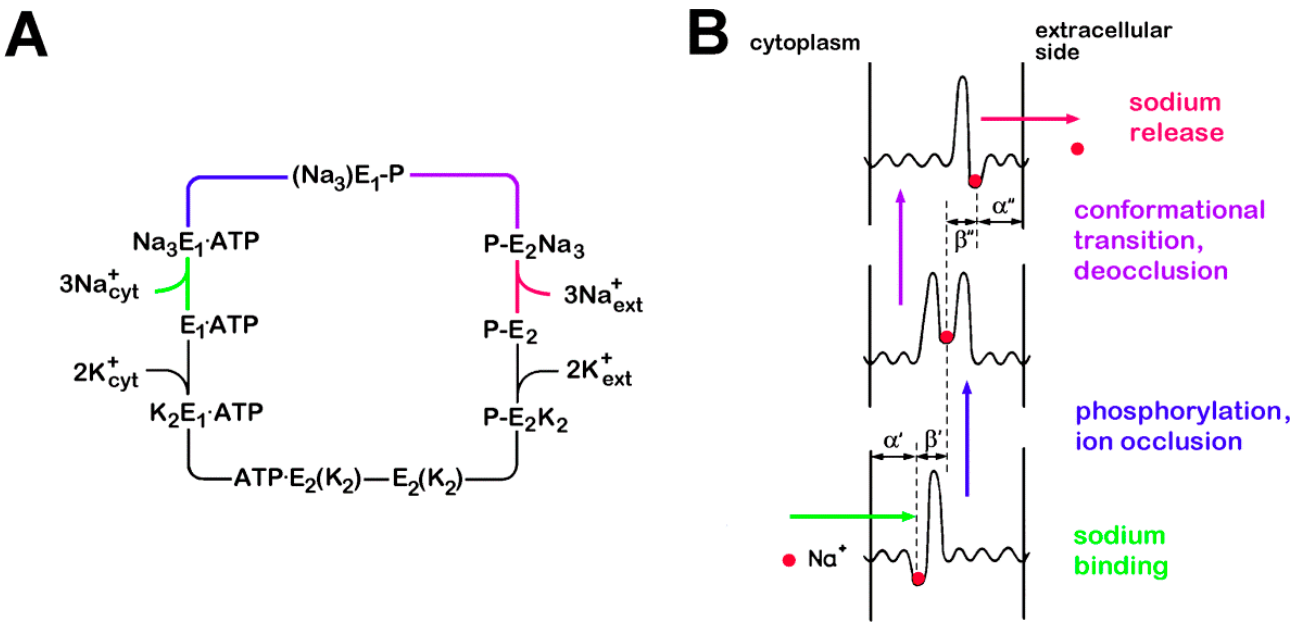



**Figure 6. A:** Post-Albers cycle of the Na,K-ATPase under physiological conditions. The $Na^+$-transporting half cycle is subdivided in four partial reactions, $Na^+$ binding (green), enzyme phosphorylation by ATP, occlusion of 3 $Na^+$, and release of ADP (blue), conformation transition, $E_1$-P to P-$E_2$, and deocclusion of the binding sites to the extracellular side (magenta), and release of the $Na^+$ ions (red). **B:** Sequence of schematic energy profiles as detected by the ions during $Na^+$-transport in three consecutive conformations as indicated in the pump cycle (**A**). The height of the energy barriers is schematic. This representation shall indicate that the high barriers are virtually impregnable for the ions with their available (thermal) energy. The quantities α', α", β', and β" are so-called dielectric coefficients that describe the fraction of the electric potential traversed by an ion in the respective reaction step (for details see text).

enzymatic activity because open membrane preparations from crab nerve were used, which did not separate both aqueous compartments that are needed to detect ion transport.[3] ATPase activity was measured as amount of inorganic phosphate released per volume of "enzyme solution." $P_i$ was determined by the colorimetric method introduced by Fiske and Subbarow in 1925,[36] and in later years by variations derived therefrom.[37,38] When Post studied broken erythrocytes he calculated a specific enzyme activity "per mg dry weight".[39] After introduction of a method to isolate and purify active Na,K-ATPase from kidney medulla preparations by Peter Jørgensen in 1969,[40] two methods were used to determine the amount of the enzyme in an assay, a micro-Kjeldahl method that quantifies the nitrogen content in a solution,[41] and the Lowry method (1951)[42] which became the standard method of protein determination in the years following. In 1978 a modified assay was introduced by Markwell et al. that allowed protein determination also in membranes without prior solubilization of the membrane-bound proteins.[43] An elegant method to determine the ATPase activity was introduced by Schwartz and collaborators in 1971.[44] A coupled pyruvate kinase/lactate dehydrogenase assay allowed 'real time' monitoring of ATP consumption by the Na,K-ATPase (or other ATPases) in buffers within a reasonable range around physiological conditions, which was and is widely used. Post et al. published in 1965 a study with enzyme isolated from guinea-pig kidneys in which they used radioactive [γ-$^{32}$P]ATP as further technique to study enzyme phosphorylation and dephosphorylation.[45] Measuring bound and released radioactive $P_i$ for a long time became the 'gold standard' to investigate enzyme phosphorylation and dephosphorylation. A thorough and comprehensive review on enzymatic properties of the Na,K-ATPase was published by Ian M. Glynn in 1985.[14]

The first ion-transport studies of the Na,K-ATPase were performed with intact erythrocytes by Post and Jolly in 1957 who measured changes of $Na^+$ and $K^+$ in the cells by flame photometry and determined a transport ratio of 2 $K^+$/3 $Na^+$ for the strophanthin sensitive flux.[46] This method, at that time was state of the art, but was improved about ten years later by Garrahan and Glynn who introduced the use of a radioactive sodium isotope, $^{24}$Na, to measure $Na^+$ transport in red cells.[20] In 1970 Paul De Weer applied tracer ions in experiments with squid giant axons. He measured $^{22}Na^+$ and $^{24}Na^+$ fluxes, determined rate constants and studied substrate dependencies,[47] followed by the first direct measurement of the electrogenicity of the Na,K-ATPase in axons with

electrodes in 1973.[48] Rhoda Blostein introduced in 1976 the use of inverted erythrocytes[49] to combine [$\gamma^{32}$P]ATP phosphorylation studies with $^{22}$Na and $^{42}$K fluxes in order to identify the sidedness as well as interaction of fluxes and enzymatic activities.[50] Important contributions could be provided also by the use of human resealed red cell ghosts.[51-53] Another chapter on transport studies with cells was opened by David Gadsby in 1979 who used Purkinje fibers from dog hearts and performed voltage clamp experiments with a microelectrode set-up.[54] He determined strophanthidin sensitive outward currents through the membranes of these cells. A further development was the whole-cell patch clamp with isolated cells from guinea pig ventricles that allowed the determination of the current-voltage dependence to the Na,K-ATPase.[55] In 1994 Don Hilgemann supplemented the set of electrophysiological techniques by the giant membrane patch method.[56] Applied to guinea pig myocytes he obtained in kinetic experiments, a time resolution of 4 µs, and analyzed external $Na^+$ binding. Six years later electrophysiological equipment was even further developed so that high-speed voltage jump experiments could be performed with squid giant axons, and time-resolved $Na^+$ release in the P-$E_2$ conformation was measured at a 3 µs time resolution.[57]

A different electrophysiological approach was introduced in 1988 by Bob Rakowski and Cheryl Paxson.[58] They were able to measure the current-voltage dependence of the Na,K-ATPase in *Xenopus laevis* oocytes in a membrane potential range between -120 mV and +60 mV by a conventional two-microelectrode voltage-clamp circuit. These cells were of interest in several ways: In the maturation state, in which they typically were used, they had low passive membrane conductance and a significantly reduced set of ion transporters compared to other cells. They were easy to investigate with electrophysiological techniques, and most of all, they were very suitable for heterologous expressions of Na,K-ATPase mutants.[59] Since then this technique has been used in numerous projects such as to study the interaction mechanism with ouabain[59], the role of glycolysation[60] or the properties of disease-inducing mutations.[61]

When transport activity is tracked by ion fluxes and overall electrogenicity in cellular systems, other ion transporters present in the membrane have to be accounted for. This is typically achieved by performing two identical experiments successively, once in the absence and once in the presence of a saturating concentration of a Na,K-ATPase-specific inhibitor, mostly ouabain. The difference of both recorded currents (or fluxes) is the contribution of the Na,K-ATPase. In 1974 a new approach was introduced when Stanley Goldin and Siu Tong reconstituted purified Na,K-ATPase from canine kidney in lipid vesicles.[62] They demonstrated that it was possible to incorporate active ion pumps into the lipid membrane by a dialysis method, and at least a fraction of pumps was oriented in a way that they could be activated by externally added ATP. Active and passive fluxes could be monitored by the use of tracer ions, $^{22}Na^+$, $^{42}K^+$, and $^{36}Cl^-$. At almost the same time, Lowell E. Hokin and collaborators published a study on purified shark enzyme reconstituted in vesicles. They showed that ouabain inhibited the pump activity only from inside the vesicles when ATP was added on the outside and confirmed pump-mediated Na-Na exchange as it was found in erythrocytes.[63] Later on, Beatrice Anner and collaborators used reconstituted vesicles to measure $^{22}$Na uptake and $^{86}$Rb export (as congener of $K^+$, more suitable because of its appropriately longer radioactive half-life time), and determined a transport ratio of approximately 3 $Na^+$ against 2 $Rb^+$.[64] In 1980 Elisabeth Skriver and collaborators published an electron-microscopical study in which they reported vesicle diameters of 90 ± 20 nm with randomly oriented intramembranous particles which were assigned as Na,K-ATPase molecules.[65] A few years later, Bliss Forbush introduced a rapid sampling technique of tracer fluxes across vesicle membranes that allowed a determination of rate constants in the order of below 10 ms.[66] This approach was very successfully applied to analyze the kinetics of $^{86}Rb^+$ or $^{42}K^+$ release from the occluded $E_2P$ conformation of the Na,K-ATPase in the presence of other substrates and inhibitors.[67,68] In 1985, an alternative method to the use of radioactive substrates was introduced by Apell and collaborators when the electrogenicity of the transport was exploited by a membrane-potential sensitive fluorescence dye, DiIC1(5), that was used together with valinomycin, to determine $K^+$ fluxes out of the vesicles upon pump activation by addition of ATP.[69] Two years later a further voltage sensitive fluorescence dye, oxonol VI, was introduced by Apell and Bersch which became a frequently used fluorescent probe to directly record the electrogenic pump activity of reconstituted Na,K-ATPase.[70] The detection mechanism was analyzed and it was shown that this technique may be used to measure a significant part of the current-voltage curve of the reconstituted ion pumps in a single experiment.[71]

A potent tool to gain access to details of the transport kinetics of the Na,K-ATPase was provided by Jack Kaplan and collaborators in 1978 when they introduced "caged ATP" that allowed triggering of the ATPase activity by production of an ATP-concentration jump.[72] Caged ATP is a photolabile 2-nitrobenzyl derivative of ATP that cannot be metabolized. By a short intensive UV flash in

the nano- to microsecond time range photolysis is activated and ATP released in millisecond time range.[73] Based on this concept of the synchronized activation of the Na,K-ATPase, Peter Läuger proposed an assay of adsorbing Na,K-ATPase-containing open "membrane fragments" onto an artificial lipid bilayer (BLM), creating a capacitive coupling between both membranes and then trigger the pumping process by a UV-flash induced release of ATP from its caged precursor in the buffer. Thus current transients may be detected in an external current pickup system by electrodes on both sides of the BLM. A first implication of this technique was published by Klaus Fendler and collaborators in 1985.[74] They verified that current transients could be recorded by this method and information on the kinetics of the $Na^+$ translocation through the Na,K-ATPase may be determined from the current transients. Two years later Borlinghaus et al. provided a detailed mechanistic analysis of the compound membrane system and the current transients activated by the ATP-concentration jumps.[34,75,76] They showed that enzyme phosphorylation, ion occlusion and the conformation transition, $E_1$-P to P-$E_2$, were not electrogenic. The time resolution of this technique was, however, limited by the photochemistry of ATP release with a pH-dependent limit of about 4 ms (at pH 7.2).[73] To overcome the pH-dependent limitations a modified caged ATP was introduced with a (pH-independent) ATP release rate of $>10^5$ $s^{-1}$, and it could be successfully applied to experiments with the Na,K-ATPase.[77,78] Since it turned out that in the pump cycle much faster reaction steps follow the rate-limiting conformation transition, a modified technique was developed that allowed the use of the compound membrane system to obtain kinetic parameters of those fast $Na^+$-moving reaction steps. In 1995, this charge-pulse technique was applied to measure the kinetics of $Na^+$ release in the $E_2P$ conformation and the rate constants in the submillisecond time range could be determined.[35] Further modifications of the compound membrane techniques were used to determine $Na^+$ binding and release on the cytoplasmic side by monitoring membrane-capacitance changes,[79,80] and by correlation of capacity changes and RH421 fluorescence signals (see below).[81] The problem of the fragility of the BLM was circumvented by the introduction of so-called solid supported membranes onto which Na,K-ATPase containing membrane fragments were adsorbed. These very robust compound membranes allowed fast buffer exchange. The possibility to freely choose buffer compositions had the advantage that ion-concentration changes could be performed in both directions.[82-84]

Since 1976, fluorescence techniques have been introduced to gain detailed information on the kinetics of conformation transitions in the Na,K-ATPase. Steven Karlish established several approaches with different collaborators. In stopped-flow experiments he used intrinsic tryptophan fluorescence to monitor and analyze the rate of the conformational transition $E_2(K) \rightarrow E_1Na$ and its dependence on ATP.[85] Formycin triphosphate, a fluorescent analog of ATP, was used to detect binding and dissociation of the nucleotide at different states of the pump cycle and the substrate dependence of these reactions.[11,86,87] While the enzymatic activities of the Na,K-ATPase were not or not significantly affected by these two techniques, a third assay, labeling the enzyme with fluorescein[88] or with fluorescein isothiocyanate (FITC),[89] confined the possibilities to investigate functional properties significantly. Labeling of the enzyme with these fluorescent compounds occurred close or in the nucleotide binding site, therefore, ATPase activity, phosphorylation by ATP, and nucleotide binding were abolished, but phosphorylation from inorganic phosphate and K-phosphatase activity were only partially inactivated. Advantage of these fluorescent labels was that they reported transitions between the $E_1$ and $E_2$ conformation.[89-91] FITC-labeled enzymes showed high fluorescence levels in $E_1$ and lower in $E_2$. FITC was found to bind covalently to a specific lysine of the cytoplasmic domain related to ATP binding, and was correspondingly affected by ATP (if present).[92] The molecular mechanism of the conformational sensitivity was that fluorescein is a pH sensitive dye and conformation transitions of the Na,K-ATPase include spatial rearrangements of the N domain with its nucleotide binding site, and thus minor local pH changes in the binding-site environment modulated the detected fluorescence (Stürmer & Apell, unpublished data). In 1982 a different fluorescein derivative, 5'-isothiocyanate fluorescein (5-IAF), was introduced by Kapakos and Steinberg.[93] This dye bound covalently to Cys457 on the cytoplasmic surface of the protein without inhibiting the enzyme activity.[94] It was used to study conformational changes, especially, $E_2 \rightleftarrows E_1$,[95] as well as $Na^+$-binding and ATP-induced partial reactions.[12,96]

Another conformation-sensitive fluorescent dye was eosin, whose application was introduced by Skou and Esmann in 1981.[97] They demonstrated that eosin bound reversibly to the Na,K-ATPase, with low affinity in the presence of $K^+$ and showed the same fluorescence emission as the free form in solution. In the presence of $Na^+$ it bound with high affinity and exhibited enhanced fluorescence. Its competition with ATP indicated that it bound to the ATP site. Eosin was used to monitor changes of enzyme conformations in the presence of a wide variety of substrate conditions.[97-99]

In the 1980s electrochromic styryl dyes[100] were introduced to gain further access to details of the ion-transport mechanism of the Na,K-ATPase. Originally these dyes were used to detect changes of the membrane potential of nerve cells in brain tissue.[101] The extremely hydrophobic compounds insert into the lipid phase of membranes, parallel to the fatty acids with their polar head facing the interface to the water phase. Because of an electrochromic effect their fluorescence properties change with the electric field in the hydrophobic part of the membrane. Due to this mechanism the response time upon changes of electric field was µs or faster. In 1988 Klodos and Forbush applied the dye RH160 to Na,K-ATPase containing membrane fragments and detected fluorescence changes upon addition of various substrates, although no transmembrane voltage was able to build up across the open membrane fragments.[102] They could not discriminate whether the detected response was caused by changes of local electric fields or due to interaction between dye and protein. The Läuger lab started in 1989 comparable studies with the dye RH421. They provided evidence that local electric fields induced by ions within the membrane domain of the Na,K-ATPase were the predominant cause of the observed fluorescence changes and presented a detection mechanism that correlated ion movement into and out of the membrane domain of the ion pump with the fluorescence changes.[103,104] The advantage of an easy application of styryl dyes with ion pumps in membrane fragments led to a frequent use of this technique by several groups.[105-110] It allowed the investigation of electrogenic and rate-limiting reaction steps around the Post-Albers cycle. The initial limitation that this technique required membrane fragments with a high density of ion pumps was conquered by its adaption to single Na,K-ATPase complexes solubilized in lipid/detergent micelles. This allowed an extension of its use even to recombinant Na,K-ATPase expressed in yeast which could be solubilized only as single enzyme molecules and not be isolated in form of purified membrane fragments.[111]

## VII. ENZYME FUNCTION

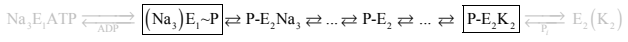
As mentioned above, ATP hydrolysis was the very feature to identify the enzyme.[3] It was found that $Na^+$, $K^+$, and $Mg^{2+}$ were essential cofactors to control the enzyme activity and assumed that it was involved in the active extrusion of $Na^+$ from the cell. [3] Three years later, in 1960, an ADP-ATP exchange was detected with $^{32}$P-labeled ADP, which was phosphorylated by an enzyme that was phosphorylated beforehand. The ($\beta$-$^{32}$P) labeled ATP was formed although no adenylate kinase was present.[112] Since 1962 it was known that the γ-phosphate of ATP forms an acid-stable phosphoenzyme.[113] In 1963 it was shown by the use of $^{32}$P-labeled ATP that the enzyme is phosphorylated in the presence of $Na^+$, and dephosphorylated subsequently by addition of $K^+$, the latter reaction was inhibited by the presence of ouabain.[114] In 1965, it was demonstrated that an acyl phosphate was formed by $^{32}$P-labeled ATP.[115,116] At the same time 'E1P' and 'E2P' were discussed as different conformations with respect to dephosphorylation,[7,117,118] as well as their decomposition by $K^+$,[45,115,119] or ADP.[120,121] In 1967 Garrahan and Glyn reported for the first time a backward running enzyme by formation of ATP from ADP and inorganic phosphate.[15] Three years later it was shown that the MgATP complex was the effective compound needed for enzyme activity,[122] although free ATP could bind and subsequent addition of $Mg^{2+}$ was able to start the reaction.[9] In 1972 it was found that ATP accelerated at high concentrations (> 400 µM) the conformation transition $E_2K \rightarrow E_1K$ in a non-phosphorylating fashion.[8] In 1973 it could be shown that an aspartyl side chain was phosphorylated by ATP.[123,124] It lasted until 1985, when the first amino-acid sequence became available, that the phosphorylated aspartate was identified in the large cytoplasmic loop between the fourth and fifth transmembrane helix.[125] This aspartate formed together with the next three amino acids a characteristic motif, Asp-Lys-Thr-Gly, that turned out to be invariant in the phosphorylation site of all P-type ATPases.[126]

The striking observation that the phosphorylated enzyme could be dephosphorylated by ADP or $K^+$ attracted a lot of interest and was considered to be a useful property to gain more insight into the molecular mechanism of the enzymatic machinery. The first concept to describe the enzymatic behavior was that of a two-pool model of the phosphorylated enzyme, $E_1$~P and $E_2$P. The first pool was filled upon binding of cytoplasmic $Na^+$ and phosphorylation from ATP. Enzyme in this pool had the $Na^+$ ions occluded[127,128] and could be dephosphorylated by ADP ("ADP-sensitive EP"). This pool was discharged by spontaneous conformation transition into the second pool, in which the ion-binding sites were extracellularly accessible. Therefore, $Na^+$ was released, and subsequently the enzyme could be dephosphorylated upon addition of $K^+$ ("K-sensitive EP"). The experimentally observed bi-phasic time course of the dephosphorylation of both pools and its dependence of the $Na^+$ concentration could, however, not be explained by this simple model.[129] In particular, experiments showing that the amount of enzyme dephosphorylat-

ed by either ADP or $K^+$ together was larger than 100% made the two-pool model obsolete.

The consequence was the introduction of an extended concept, a three-pool model that included an additional pool intercalated between $E_1$~P and $E_2$P.[130] The additional pool was thought to be drained via both dephosphorylation modes, by $K^+$ and ADP. Extensive experiments and discussion of $Na^+$, $K^+$ and ADP dependences as well as the sizes of the proposed three pools under various substrate conditions led to different concepts according to whether the pool 'in the middle' might "effectively be both ADP- and K-sensitive"[131] or might be "not sensitive to both ADP and $K^+$ but has to be converted to $E_a$~P, the first pool, which is $E_1$-P."[132]

When in retrospect the concept of enzyme dephosphorylation is revisited in terms of the detailed Post-Albers pump cycle available nowadays, the (linear) sequence of phosphorylated states of the Na,K-ATPase consists of (derived from Fig. 2B):




Only the (boxed) first and last state in the respective reaction sequence may be dephosphorylated directly. All states in between are inert to dephosphorylation. They are $E_2$-P states in which the ion-binding sites are accessible from the extracellular side of the membrane and are able to bind or release ions and are occupied by 1 to 3 $Na^+$, by 1 $K^+$ or no ion. The occupancy of these states rapidly achieves a steady state distribution in a diffusion controlled manner, depending on the current ion concentrations in the extracellular medium. The low $Na^+$ affinity of the ion-binding sites in the P-$E_2$ conformation and high affinity for $K^+$ leads under physiological conditions preferentially to dephosphorylation upon binding of a second $K^+$, P-$E_2K_2$ → $E_2(K_2)$ + $P_i$. Only at high $Na^+$ concentrations and very low (or no) $K^+$ a considerable amount of enzyme will undergo the (backward) conformation transition, P-$E_2$ $Na_3$ → $(Na_3)E_1$~P, and the resulting state can be dephosphorylated in the presence of ADP, $(Na_3)E_1$~P + ADP → $Na_3E_1$ATP.[53,133] This approach to a mechanistic description makes the introduction of pools of phosphorylated states unnecessary to explain the various dephosphorylation experiments.

In 1965 Post and Sen showed that it was possible to produce a $K^+$ influx into cells in the absence of ATP but in the presence of $Mg^{2+}$ and inorganic phosphate[6] which indicated binding of $P_i$ to the unphosphorylated enzyme, a reaction step later called 'backdoor phosphorylation'. Only two years later Glynn and Garrahan demonstrated that the thermodynamic requirement that the enzyme runs backwards could be fulfilled experimentally under appropriate substrate condition.[15] In the presence of $K^+$, $Mg^{2+}$ and $P_i$, addition of ouabain induced rapid backdoor phosphorylation.[134] In the absence of ouabain less steady-state phosphorylation was obtained because of $K^+$ promoted dephosphorylation.[135] The identical proteolytic digestion pattern obtained from the $P_i$-induced and ATP-induced phosphoenzyme was understood as strong indication that both phosphoenzymes were the same.[136,137]

Another interesting question was the nucleotide specificity of the enzyme. In 1968 Matsui and Schwartz studied the effect of nucleotides other than ATP, namely CTP, ITP, GTP, UTP.[138] Subsequently, their dissociation constants were determined.[9] After the method was introduced to reconstitute enzyme functionally in lipid vesicles, active transport of $Na^+$ and $K^+$ energized by CTP (almost as effective as ATP) und UTP (relatively ineffective) was reported.[63] From the $Na^+$ transport with various nucleotides and two synthetic ATP analogs a correlation was found between the proton-accepting properties of the nucleotides and their ability to provide active transport.[139]

Considerable attention was paid to the role of $Mg^{2+}$ for the enzyme activity after it was noticed from the very beginning that this ion was indispensable for function.[13,112] Detailed studies revealed that the MgATP complex was the activating substrate of the Na,K-ATPase.[122,140] This is not really surprising since physicochemical investigations yielded equilibrium dissociation constant of the MgATP complex in a pH-dependent manner between 1.5 μM (pH 8) and 10 μM (pH 6)[141], while the free $Mg^{2+}$ concentration in cells typically is in the order of 200 μM[142]. This implies that under physiological conditions more than 95% of total ATP is present as Mg complex. Free $Mg^{2+}$ was reported to bind to a low-affinity site where it caused inhibition of the enzyme activity.[143] It was shown that $Mg^{2+}$ is released from the enzyme only after its dephosphorylation in the $E_2$ conformation.[144] Therefore, it could be considered to be a product inhibitor when high $Mg^{2+}$ concentrations in the buffer impeded dissociation from its site and thus affected the $E_2$ → $E_1$ transition.[145] This concept was confirmed by Forbush[67,68] and complemented by the proposal that only one site for $Mg^{2+}$ per enzyme was required for both phosphorylation by ATP and enzyme inhibition by stabilizing the $E_2$ conformation. Binding of ATP at the low-affinity site in $E_2$ promoted $Mg^{2+}$ release and the site was reoccupied only after enzyme phosphorylation in $E_1$ by MgATP. In 2000, a conserved segment in the P domain of the α subunit was identified in which Asp710 contributed to the coordination of $Mg^{2+}$.[146]

Another tool to enlarge the insight into enzyme functions were various inhibitors that allowed the arrest

of the Na,K-ATPase in defined states or a restriction of possible reaction sequences to specific parts of the pump cycle. A review presenting a comprehensive survey was published by Glynn in 1985.[14] The most important group of inhibitors is that of cardiac steroids. Compounds in which a sugar is attached to the steroid are so-called cardiac glycosides (CGs), of which the most prominent is ouabain. Although it was clear that CGs interact with the extracellular side of the Na,K-ATPase, the molecular mechanism of inhibition was unknown until the 1990s. With the progress of molecular-biological methods mutagenesis of numerous (and 'suspect') amino acids was performed and the effect of the mutations and their resistance against different CGs was investigated to identify the binding site of the inhibitor.[147,148] Crucial amino acids were found in transmembrane and extracellular domains. At the same time, in 1996, functional studies revealed that $K^+$ accelerated enzyme dephosphorylation and thus antagonized ouabain binding. In the presence of high concentrations of ouabain (in the mM range), however, ouabain was able to bind even when 2 $Rb^+$ (as congeners of $K^+$) were bound, $E_2Rb_2$, and the inhibitor stabilized this state.[149] Major progress in mechanistic understanding was made when detailed structural information became available. In 2009, a first crystal structure of Na,K-ATPase at 2.8 Å resolution was published with a low-affinity bound ouabain in a state analogous to P-$E_2K_2$.[150] Ouabain was deeply inserted into the transmembrane domain with the lactone ring close to both $K^+$ ions bound to their sites. Most of the mutagenesis data, obtained with high-affinity bound ouabain, could be explained by this arrangement, which suggested that the CG binding site should be essentially the same in both conditions. Two and then four years later structures with high-affinity bound ouabain became available with a resolution of 4.6 Å and 3.4 Å.[151,152] These structures made visible that ouabain was bound to a site in the α subunit formed by transmembrane segments M1 to M6 and thus blocked the ion pathway from the extracellular side to the ion-binding sites. In the structure with the higher resolution it was found that a $Mg^{2+}$ ion was present in the cation transport site II when ouabain was bound with high-affinity. Comparison of the position of ouabain in the low- and high-affinity bound state showed that both were indeed mostly not significantly different. Prominent was only a difference in the location of the lactone ring of the inhibitor in the $Mg^{2+}$-bound and $K^+$-occluded condition. Altogether, functional and structural findings allowed a consistent explanation of the inhibition mechanism by preventing the conformation transition from $E_2$ to $E_1$ while clogging the extracellular access channel of the Na,K-ATPase. The well-known antagonistic effect of $K^+$ on ouabain (or any other cardiac glycoside) binding could be attributed to $K^+$-induced low-affinity ouabain binding.[152]

When the effect of CGs was studied with enzyme from many animals, typical binding affinities of the Na,K-ATPase were found in the range of µM.[153] Such a high affinity raised the question of why the enzyme should have evolved such a specific binding site for an exogenous compound and whether there were endogenous inhibitors aimed at this site. John M. Hamlyn and collaborators reported in 1982 the existence of a circulating inhibitor of the Na,K-ATPase,[154] in 1991 they identified a ouabain-like factor,[155] in 1999 it was confirmed that it was ouabain,[156] and in 2000 Wilhelm Schoner introduced ouabain as new steroid hormone.[157] In a recent review the story of discovery, advances and controversies of endogenous ouabain was published.[158]

Ouabain also plays a role as a signal messenger. Regulatory effects of the Na,K-ATPase inhibition by ouabain were initially assigned to changes in intracellular $Na^+$ and $K^+$ concentrations.[159] From research in cardiac hypertrophy crucial information was collected over a couple of years and it was established that ouabain stimulated myocyte growth and protein synthesis, comprised the induction of a number of early response proto-oncogenes and activated transcription factors already at low, nontoxic concentrations. Finally, experimental observations were published that ouabain binding activated signaling cascades.[160] The ouabain-stimulated signal transduction was mediated by the Na,K-ATPase but was apparently independent of ion transport function.[161] The signaling function of Na/K-ATPase controlled by CGs has been gradually appreciated over the last 20 years as can be followed in several reviews.[161-164]

Another potent inhibitor of the Na,K-ATPase (and all other P-type ATPases) is orthovanadate, $VO_4^{3-}$. It was identified 1977 by Lewis C. Cantley and collaborators as a potent inhibitor of the sodium pump with a $K_I$ of 40 nM. Inhibition was reversed to 100% by millimolar additions of norepinephrine. Vanadate was initially found as contamination in commercial "Sigma grade" ATP, isolated from horse muscle. From its tetrahedral structure it was concluded that it may bind to "a phosphate site".[165] From their study of interaction with the Na,K-ATPase Cantley concluded that "the unusually high affinity for vanadate is due to its ability to form trigonal bipyramidal structure analogous to the transition state for phosphate hydrolysis."[166] $Mg^{2+}$ was required as cofactor for inhibition. Its inhibitory action can be attributed to its high-affinity binding to the phosphate binding site, a condition in which it stabilizes the $E_2$ conformation of the enzyme.

One further inhibitor should be mentioned in the framework of this presentation, oligomycin, which was found to inhibit the mitochondrial ATP synthase. It is an antibiotic originally isolated from *Streptomyces*. In 1962, first reports were published that this macrolide also inhibited the Na,K-ATPase.[167,168] It was demonstrated that oligomycin is a potent inhibitor, however, it did not inhibit the enzyme completely, e.g. ADP-ATP exchange was unaffected. Eventually, all experimental findings were explained by the mechanistic concept that oligomycin blocked the conformation transition $E_1$-P → P-$E_2$.[14,169] This was supported by findings that the inhibitor shifted the equilibrium from a $Na^+$-deoccluded form to a $Na^+$-occluded form,[128] and stabilized $Na^+$ occlusion but not $K^+$ occlusion.[170] In 2013 a structure of the Na,K-ATPase complex with 3 $Na^+$ and an oligomycin molecule was published at a resolution of 2.8 Å.[171] Therein the inhibitor was bound close to helix 1' on the cytoplasmic side adjacent to the membrane surface.

## VIII. TRANSPORT FUNCTION

As can be seen from Figure 2, ion transport is a complex process even when restricted to the pump cycle under physiological conditions. By means of experimental studies at least ten reaction steps were identified that form the complete pump cycle and embrace the interplay of enzyme and transport functions (Fig. 2A). In the following, the focus will be set on molecular processes investigated in experimental studies to enlighten the transport mechanism of both ion species across the membrane. (An earlier review was published in 2004.[172]) To discuss the transport function of the Na,K-ATPase in detail, the pump cycle will be divided into four partial reactions: (1) Cytoplasmic ion exchange, (2) access transfer from the cytoplasmic to the extracellular side, (3) extracellular ion exchange, and (4) access transfer from the extracellular to the cytoplasmic side.

(1) $K_2E_1{\cdot}ATP \rightleftarrows E_1{\cdot}ATP \rightleftarrows Na_3E_1{\cdot}ATP$

The cytoplasmic ion exchange occurs in the $E_1$ conformation of the Na,K-ATPase. Under physiological conditions, in which the cytoplasmic $K^+$ concentration is high, $Na^+$ low, and the binding affinity for $K^+$ is higher than for $Na^+$, a significant fraction of the enzyme is found in the $K_2E_1{\cdot}ATP$ state.[173] Consequently, under steady-state conditions only a small fraction of the Na,K-ATPase populates $Na_3E_1{\cdot}ATP$, the state which is the one capable of being phosphorylated. Despite this unfavorable displacement of the occupation of the states in this reaction sequence, the pump is obviously able to perform its task of extruding $Na^+$ from the cytoplasm. This fact has to be assigned to the finding that the exchange of ions between binding sites and aqueous bulk phase is fast (and for the most part diffusion controlled) compared to the subsequent phosphorylation step so that a quasi-equilibrium distribution between the differently occupied states may be assumed. Therefore, the drain of the $Na_3E_1{\cdot}ATP$ state by the phosphorylation step is instantaneously compensated. It is known, however, that the Na,K-ATPase runs way below their kinetically possible maximum turnover due to this limiting condition.

When initial studies of ion binding and release in the $E_1$ conformation were performed, it was observed that $K^+$ release and binding of the first two $Na^+$ ions were not electrogenic.[174,175] This observation was interpreted as indication that both binding sites were negatively charged and located close to the cytoplasmic surface in a wide, water filled vestibule. This electroneutrality turned out to be, however, only an apparent effect[176]. In the $E_1$ conformation, the affinity of both binding sites for protons is so high that at physiological pH and in the absence of $Na^+$ and $K^+$ the sites are mostly protonated. Therefore, binding of both $K^+$ ions and of the first two $Na^+$ ions in titration experiments was effectively an electroneutral exchange against bound protons. At unphysiologically high pH, the electrogenicity of $K^+$ and $Na^+$ binding became very well visible in titration experiments when beginning with a cation-free electrolyte[176].

While $K^+$ and the first two $Na^+$ ions compete for the same sites in the pump, binding of the third $Na^+$ occurs to an exclusively $Na^+$-specific site in the $E_1$ state. This process was found to be electrogenic and the dielectric coefficient was shown to be in the order of 0.25.[12,35,83,177] It was demonstrated that electrogenic binding of the third $Na^+$ could be detected by styryl dye RH421 and simultaneously by a directly measured charge movement with identical results.[81] That means that $Na^+$ traverses 25% of the electric-potential drop across membrane to reach its binding site from the cytoplasm. (Which does not necessarily imply that the spatial distance is also 25 % of the membrane thickness.[27]) A study of cytoplasmic $Na^+$ binding and detailed analysis of the binding affinities revealed that the third $Na^+$ binds to a site with a higher affinity for $Na^+$ than the second site.[173] Such an observation could be explained by the assumption that the third binding site became available only after the first two sites were already occupied by $Na^+$. A possible mechanism was a conformational rearrangement in the transmembrane helices of the membrane domain upon binding of the second $Na^+$ which then assembled the third $Na^+$ site or opened access to it. Furthermore, bind-

ing of the third $Na^+$ was monitored also with a fluorescence change of the conformation-sensitive label FITC, which was linked to a highly conserved lysine in the nucleotide-binding site at the cytoplasmic N domain.[175] The accordance of results from the FITC and RH421 experiments was a strong indication that binding of the third $Na^+$ and conformational rearrangement at the cytoplasmic N domain of the protein were concurrent events.

In the composition of a mechanistic concept it may be concluded that binding of only the third sodium triggers a rearrangement of the cytoplasmic N domain with an appropriately bound MgATP, and thus arms the protein to make way for enzyme phosphorylation at the specific aspartate in the cytoplasmic P domain. Such a mechanism would also be in agreement with the finding that binding of the third $Na^+$ needed significantly higher activation energy (63.4 kJ/mol) than binding of the first two $Na^+$.[178] An activation energy of such a high magnitude points as well to a conformational rearrangement related to binding/release of the third $Na^+$.

An alternative proposal was suggested by Kanai et al., based on their structure of the Na,K-ATPase in an $E_1$ conformation with 3 $Na^+$ occluded.[171] Their concept was that the first $Na^+$ is bound to the $Na^+$-specific site III followed by occupation of sites I (in the middle) and II (outermost). They assumed from their crystal structure, which represented a $Na^+$-occluded phosphorylated intermediate, that there is just one access pathway to all three sites and the "innermost" site III can be reached only through (empty) sites I and II. If so, a possible reconciliation with the results from functional studies could be a single-file push-on mechanism in which sites I and II are occupied first, but with the arrival of a third $Na^+$ both ions in sites I and II are moved forward into sites III and I to make way for binding of the third ion into site II. To clarify the actual mechanism a crystal structure of the pump in the antecedent, non-occluded (and preferably only partly occupied) state would be very helpful.

(2) $Na_3E_1{\cdot}ATP \rightleftarrows (Na_3)E_1\text{-}P \rightleftarrows P\text{-}E_2Na_3$

When all three ion sites are occupied by $Na^+$ and MgATP is bound to the nucleotide binding site, the pump is able to perform auto-phosphorylation associated with a simultaneous occlusion of the ions. During this reaction step no charge movements were detected.[34] This observation indicated that the three ions in their binding sites are not displaced (at least not perpendicular to the membrane plane). This behavior is in agreement with the observation that the ion-binding sites of the closely related SR Ca-ATPase also are not significantly relocated throughout the complete pump cycle as can be established by comparison of the crystal structures obtained in numerous different states of the pump.[179] In the absence of oligomycin the gained occluded state, $(Na_3)E_1\text{-}P$, is transient and followed by a spontaneous transition to the $P\text{-}E_2$ conformation with deoccluded ion binding sites. This step is the rate-limiting process in the $Na^+$-translocating half cycle,[12,35,57,78] with a rate constant of about 22 $s^{-1}$ at 20 °C, and has the highest activation energy of all reactions of the pump cycle. With purified membrane preparations from rabbit kidney a value in the order of 115 kJ/mol was determined.[180]

The conformation transition showed only a minor dielectric coefficient (0 – 0.1),[35,57] and it could not be determined whether this was caused by ion movements or (more probably) movements of charged side chains in the helices of the membrane domain which underwent considerable reorientations during the transition. Besides unclasping the access channel between binding sites and extracellular aqueous phase, another major functional consequence of the transition was the reduction of the binding affinity for $Na^+$ by a factor of about 500.[12] This dramatic change was caused by minor movements of transmembrane helices which in turn modified the coordinating interactions between amino-acid sidechains and the $Na^+$ ions.

(3) $P\text{-}E_2Na_3 \rightleftarrows P\text{-}E_2 \rightleftarrows P\text{-}E_2K_2$

The extracellular sodium release is the best investigated partial reaction of the Na,K-ATPase pump cycle.[12,35,56,57,181-183] Release of the three $Na^+$ occurred sequentially and with different kinetic and electrogenic properties, which allowed an assignment to the respective reaction steps. The first $Na^+$ released had the highest dielectric coefficient of the whole pump cycle. The ion traversed 65 – 70 % of the electric potential in the membrane, and its release process had the lowest rate constant of the three ions in the order of 1000 $s^{-1}$ at 20 °C.[35,57,183] The activation energy of this partial reaction was found to be about 80 kJ/mol.[184] Dissociation of ions from a binding site and diffusion through a narrow pore-like structure typically would have activation energies below 20 kJ/mol. The observed high activation energy was, therefore, an indication of a conformational rearrangement of the participating membrane domain. The commonly accepted release mechanism is an initial rate-limiting deocclusion process for the first ion as associated consequence of the conformation transition from $E_1\text{-}P$ to $P\text{-}E_2$. The high dielectric coefficient of the first $Na^+$ released has to be explained by an ion

migration through a narrow access channel. Thereafter, another conformational relaxation was required before the second and third $Na^+$ ion exited the membrane domain, since the electrogenicity of these reaction steps was found to be just 10 – 20 % even though the distance between the ion binding sites was small. Because it is expected that the binding sites are not (significantly) dislocated during this partial reaction, a rearrangement of the α helices in the protein's membrane domain has to take place in a way that they form a wide access structure being filled with water molecules, as was proposed.[35,185,186] Such a vestibule would remodel the dielectric shape of the protein so that the bound ions would be able to reach the polar aqueous phase within a short "dielectric" distance of < 20 %.[35,57] The existence of such a structural rearrangement was supported by the high activation energy of 70 kJ/mol that was determined for this partial reaction.[57] The subsequent release of the second $Na^+$ was found to be fast with a rate constant in the order of 10,000 $s^{-1}$.[57] Thereafter, the release of the last $Na^+$ occurred with a similarly low dielectric coefficient, and with a rate so fast ($\geq 10^6$ $s^{-1}$)[57] that it could not be resolved with the experimental techniques available.

In the resulting P-$E_2$ state the ion-binding sites had a significantly lower binding affinity for protons than in the $E_1$ conformation[176] and were virtually empty at low $Na^+$ concentrations and in the absence of $K^+$. The following $K^+$ binding and transport into the cell have been studied extensively.[12,187-191] Sequential binding of $K^+$ (or its congener $Rb^+$) was resolved and a mechanism described as "flickering gate model" was introduced which implied that the first $K^+$ is slowly bound (or released) while the second $K^+$ bound was able to exchange fast with the aqueous phase[187]. The equilibrium dissociation constants for the first and second $K^+$ differed by a factor of 5 to 6 at a level in the sub-millimolar range. Besides $K^+$, congeneric monovalent cations were also able to be transported, such as $Rb^+$, $Cs^+$, $Tl^+$, $NH_4^+$, $H^+$ or even $Na^+$. Quaternary organic amines, which are large monovalent cations of different size, were used to probe the extracellular access channel and, in addition, they could be used as inhibitors of the Na,K-ATPase.[186,192-195]

(4) P-$E_2K_2$ ⇄ *ATP*·$E_2(K_2)$ ⇄ $K_2E_1$·*ATP*

While $K^+$ binding (or release) on either side of the membrane was electrogenic, occlusion, conformation transition and deocclusion on the opposite side of the membrane were electrically silent.[104,196,197] Release of $K^+$ to the cytoplasmic aqueous phase was actually a $K^+/Na^+$ exchange or a $K^+/H^+$ exchange in the absence of $Na^+$ (see above), and therefore, also only apparently electroneutral.[176]

When the second $K^+$ ion was embedded in its binding site in the P-$E_2$ state, a spontaneous conformational rearrangement occurred that caused ion occlusion and dephosphorylation of the enzyme, resulting in state $E_2(K_2)$.[187] The available experimental evidence indicated that dephosphorylation and occlusion go hand in hand. In 1988, the rate constant could only be estimated, and Forbush reported that it has to be much larger than 100 $s^{-1}$.[198] A few years later fits to kinetic data led to values of $>10^3$ $s^{-1}$ (all at room temperature).[12] For the reverse reaction, the backdoor phosphorylation, rate constants of $> 10^5$ $M^{-1}s^{-1}$ were determined.[68,199] Under physiological conditions, the dephosphorylated $E_2(K_2)$ state was only transient. From there, the pump cycle could advance in two different ways, depending on the ATP concentration present (Fig. 2).

At physiological ATP concentrations ("high ATP"), low-affinity binding of the nucleotide occurred,[8] and the resulting state, *ATP*·$E_2(K_2)$, underwent an accelerated transition to the $E_1$ conformation, $K_2E_1$·*ATP*. Rate constants of about 60 $s^{-1}$ were obtained for this reaction step at saturating [ATP] and room temperature.[200,201] Nevertheless, this reaction step was rate-limiting in the $K^+$-transporting half cycle. In analogy to the $E_1/E_2$ transition, it was linked up with a deocclusion of the ion sites, then accessible from the cytoplasmic side. In the presence of low ATP (< 100 nM), the occluded $E_2(K_2)$ state was able also to perform a transition to the $E_1$ conformation, $E_2(K_2) \rightarrow K_2E_1$, with a dramatically lower rate constant of < 0.3 $s^{-1}$ at room temperature.[67,201] The rate constant of the reverse step was much larger (290 $s^{-1}$)[89] so that in the absence of ATP and presence of $K^+$ the equilibrium of both conformational states was strongly shifted to the $E_2$ conformation when two ions occluded.

Summarizing all these findings, the pump mechanism can be represented schematically by the cartoon shown in Figure 7. It is based on the gated channel concept, in which the ion sites are embedded deep inside the membrane domain of the Na,K-ATPase. The ion sites are accessible only from one side at the same time. The observation that electrogenic ion movements were found only in one of both access channels at the same time may be interpreted as indication that the respective other channel is completely blocked. Only reaction steps in which ions are taken up from the aqueous phase or released from their binding sites to the outside of the protein are electrogenic and produce a detectable electric signal. Ion movements in the access channels are diffusion controlled, which leads to the consequence that under physiological conditions the exchange of both

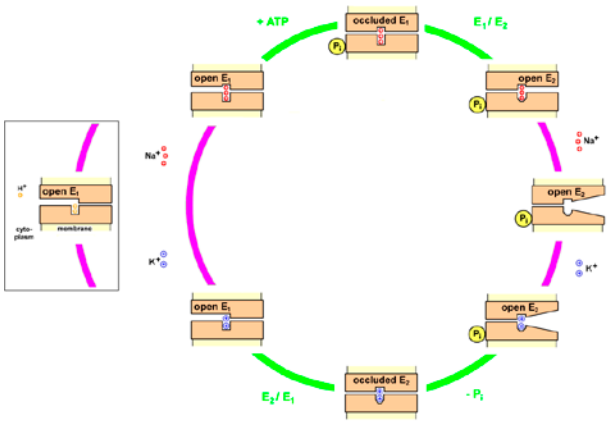



**Figure 7.** Schematic representation of the functional behavior of the Na,K-ATPase membrane domain (the cytoplasmic domain is omitted for clarity) in characteristic states of the pump cycle. The ion binding sites are located almost in the center of the membrane domain and are accessible from the cytoplasm in the $E_1$ conformation and from the extracellular aqueous phase in P-$E_2$. The open states in both conformations are separated by occluded states in which no ions may move within the access channels. In the P-$E_2$ conformation the initially narrow access channel widens up after release of the first $Na^+$. The pink fractions of the cycle indicate electrogenic (ion uptake and release), the green ones represent electroneutral processes (phosphorylation/dephosphorylation, conformation transition). While in the P-$E_2$ conformation the ion-binding sites may be empty, in the $E_1$ conformation the binding sites are occupied by protons in the absence of other monovalent cations (due to the high affinity for protons in $E_1$) as insinuated by the inset.

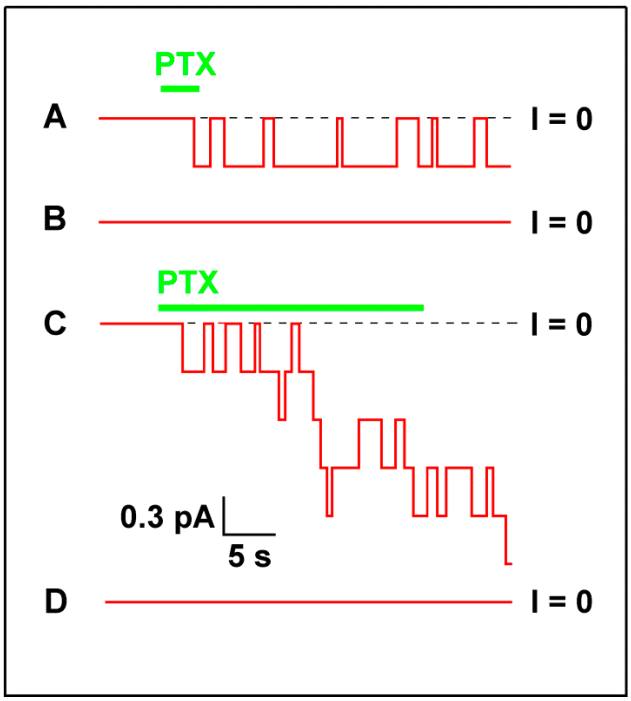



**Figure 8.** Schematically represented palytoxin (PTX) induced ion-channel behaviour of the Na,K-ATPase. **A.** In the presence of $Na^+$ and Mg·ATP and 25 pM PTX for a few seconds typical opening and closing of a single cation-selective channel were observed. **B.** After washout of PTX no longer channel events could be recorded and ion-pump activity was restored. **C.** Upon prolonged exposure to 25 pM PTX more and more Na,K-ATPase molecules were transformed into ion channels. **D.** In the presence of high concentrations of ouabain channel activity was completely suppressed. Figure adapted from Artigas & Gadsby[209].

$K^+$ against (the first) two $Na^+$ and correspondingly the reverse reaction occur so fast that the respective electric current contributions cancel each other. Therefore, it was assumed for some time that $K^+$ transport by the Na,K-ATPase was electroneutral until experiments were performed under $K^+$-limiting conditions.[176,188,191]

In the open P-$E_2$ conformation two different conformational arrangements of the access were found. Initially after the transition from $E_1$ a narrow channel with high electrogenicity was formed, and after the release of the first $Na^+$ a wide water-filled funnel developed and induced low electrogenicity. In the $E_1$ conformation, no such significant changes were found. The question whether the third $Na^+$ that binds to the $Na^+$-specific site, enters through an access different from that of the first two $Na^+$ or both $K^+$ ions could not be answered so far.

## IX. TRANSMEMBRANE CHANNEL FORMATION

The gated channel concept received convincing experimental support when the molecular mechanism of the interaction of palytoxin with the Na,K-ATPase was investigated. Palytoxin is a lethal marine toxin extracted from polyps of the genus Palythoa.[202] It was found that addition of palytoxin to mammalian cells caused the occurrence of rather nonselective cation channels with a single-channel conductance of about 10 pS.[203,204] Scrutinizing the membranes led to the discovery that those ion channels were formed by the Na,K-ATPase.[205-208] An important step forward was obtained when Artigas and Gadsby used outside-out or inside-out excised membrane patches to detect the effect of palytoxin on the level of single Na,K-ATPase molecules.[209,210] They recorded typical single-channel events upon addition of palytoxin with conductance of 7-10 pS[210] (Fig. 8), and proposed that palytoxin modified in the P-$E_2$ conformation the gate between the cytoplasm and the ion-binding sites, when the access channel to the extracellular side was already open. Thus a continuous pathway was established that formed a relatively non-selective cation channel. At a low palytoxin concentration, when only one or a few Na,K-

ATPase molecules were modified, a characteristic toggling between conducting and non-conducting states of the channel could be observed (Fig. 8). Therefore, it was concluded that the effect of palytoxin is reversible. In addition, when the toxin was washed out, the channel activity ceased. The fast reversibility of the open-channel formation suggested that no major conformational reorganization or even denaturation of the protein occurred but a simultaneously open condition of both occlusion gates was induced.[209] This concept was supported by the fact that common blockers of the access channels were able to clog the continuous ion pathway on both sides of the Na,K-ATPase.[211] By mutation of more than sixty amino acids in transmembrane helices TM1 to TM6, those could be identified by which the conductance of the ion channel could be affected, and a comparison with the crystal structure of the Na,K-ATPase allowed the proposal of the channel's shape and position.[212,213]

## X. COUPLING OF ENZYME AND TRANSPORT ACTIVITYIES

One of the major unresolved issues of the function of the Na,K-ATPase is the mechanism of energy conversion by the Na,K-ATPase (or any other P-type ATPase). From basic thermodynamic principles it is known that hydrolysis of ATP in the presence of known concentrations of ATP, ADP and $P_i$ provides under physiological conditions a Gibbs free energy in the order of -55 kJ/mol.[27] This energy is transferred to the ion pump by a chemical reaction, the phosphorylation of the specific aspartate in the P domain. Terrell L. Hill showed that energy transduction in molecular machines does not occur in a single reaction step of the reaction cycle (here: the pump cycle of the Na,K-ATPase) but is distributed over the whole cycle.[214,215] Therefore, to analyze the energetics of the Na,K-ATPase, it has to be determined to what extent single reaction steps contribute to the storage and consumption of the system's free energy in terms of changes of the so-called "basic free energy levels".[27,180] It was found that there was indeed no single "power stroke" reaction step in the pump cycle. Under physiological conditions many steps were even close to thermodynamic equilibrium such as the $Na^+$ binding and release steps. Extracellular $K^+$ binding and ATP binding in the $E_2(K_2)$ state were distinct "down-hill" steps in which energy was dissipated, where in contrast, release of both $K^+$ to the cytoplasm were the most prominent energy consuming steps. The overall energy consumption during a pump cycle could be calculated from the definition of the electrochemical potential gradients of $Na^+$ and $K^+$ across the membrane using the known concentrations of both ion species on either side of the membrane, and the electrical membrane potential.[27] At typical values of mammalian cells it was calculated that about 80 % of the energy provided by ATP hydrolysis was utilized in ion transport.[180] Compared to macroscopic machines such a yield is impressive.

While it is possible to calculate basic free energy levels, these numbers do not provide insight into the molecular processes of how energy is transferred from the initial "high energy phosphate" in state $(Na_3)E_1$-P to other moieties of the protein with the result that ions are eventually transferred from one side of the membrane to the other, a vectorial process. It can be assumed that the provided energy is distributed over the protein by rearrangements of amino-acid side chains in response to the coordination of the high-energy phosphate, thus creating changes in spatial alignments, mechanical tension and torque of helices, modified electrostatic interaction and dipole movements. Such transiently enhanced potential energy is buffered in various subdomains of the protein structure. Subsequently, it may drive a meticulously concerted sequence of relaxation processes that perform ion pumping by promoting specific reactions such as ion binding, occlusion and release to the opposite side of the membrane. So far the whole process did not, however, advance from the level of hand-waving arguments. Perhaps, additional structural details with atomic resolution of closely neighboring states of the pump cycle will produce fuel for thoughts, or an inspiring idea may be triggered by revisiting the available wealth of experimental data. Here, but not only here, an exciting terra incognita in the world of the Na,K-ATPase is waiting for exploration.

## XI. PUMP-RELATED DISEASES

A final chapter shall reveal and summarize how the knowledge on the location of single amino acids and their role in the functional context provided understanding of specific diseases. Small mutations on the molecular level of the Na,K-ATPase were found to significantly affect pump functions with far-reaching organismic impact. As in the case of ion-channel induced pathology, it was found also for the Na,K-ATPase that errors in the genetic code may provoke malfunctions of the ion pump that lead to phenotypes of explicit diseases.

From early studies of the Na,K-ATPase it has been known already that cardiac glycosides inhibit this enzyme.[216] Because these compounds were applied to treat congestive heart failure and cardiac arrhythmias,

it stood to reason that improper regulation of the Na,K-ATPase activity may correlate with various clinical conditions. Until the year 2000, the focus has been commonly set onto investigations with alteration of endogenous or xenobiotic factors. The cause of several diverging diseases such as cardiovascular, neurological, metabolic or renal disorders were traced back to a dysfunction in salt and water homeostasis of cells that is controlled by the Na,K-ATPase.[217]

In 2004, a specific mutation in the α2 isoform of the Na,K-ATPase was found to cause familial hemiplegic migraine,[218] and in the years that followed, further mutations were discovered to provoke various forms of migraine.[219] More recently it has been reported that mutations in the neuron-specific Na,K-ATPase α3 subunit are linked to rapid-onset dystonia Parkinsonism[220] and that a mutated α3 subunit may play a role in the neurodegeneration of Alzheimer patients.[221] A further disease, primary aldosteronism, was also attributed – among other causes – to malfunction of the Na,K-ATPase. Modifications of pump activity caused secondary hypertension by overproduction of aldosterone, which is initiated by single mutations of the α1 isoform of the Na,K-ATPase in adenomas within the zona glomerulosa of the adrenal cortex. Each of at least five single mutations in the α1 subunit has been found to induce overproduction of aldosterone.[222-224] Very recently, it was reported that the CAPOS (cerebellar ataxia, areflexia, pes cavus, optic atrophy, and sensorineural hearing loss) syndrome is caused by the single mutation, E818K, of the α3-isoform of Na,K-ATPase.[225] Mutations in the gene ATP1A1, which encodes the α1 subunit of the Na,K-ATPase, were identified as a cause of autosomal-dominant Charcot-Marie-Tooth Type 2 disease. A missense change was found that induced loss-of-function defects, resulting in peripheral motor and sensory neuropathies.[226] A missense mutation of the α2 subunit of the Na,K-ATPase was found in a patient with hypokalemic periodic paralysis and CNS symptoms.[227] An informative review on structure and function of the Na,K-ATPase isoforms in health and disease that contains an overview of the currently known disease-causing mutations was published by Clausen et al. in 2017.[228] In this rapidly developing field, the abundance of experimental methods and mechanistic studies collected in recent decades will surely promote further progress and provide invaluable benefits.

## ACKNOWLEDGEMENTS

This work was supported by the University of Konstanz (AFF 4/68).

## REFERENCES

1. J. F. Danielli; *Symp. Soc. Exp. Biol.,* **1954**, *8*, 502.
2. P. Mitchell; *Nature,* **1957**, *180*, 134.
3. J. C. Skou; *Biochim. Biophys. Acta,* **1957**, *23*, 394.
4. H.-J. Apell; *Substantia,* **2018**, *2*, 17.
5. R. W. Albers, S. Fahn, G. J. Koval; *Proc. Natl. Acad Sci U. S. A,* **1963**, *50*, 474.
6. R. L. Post, A. K. Sen; *J. Histochem. Cytochem.,* **1965**, *13*, 105.
7. R. W. Albers; *Ann. Rev. Biochem.,* **1967**, *36*, 727.
8. R. L. Post, C. Hegyvary, S. Kume; *J. Biol. Chem.,* **1972**, *247*, 6530.
9. C. Hegyvary, R. L. Post; *J. Biol. Chem.,* **1971**, *246*, 5234.
10. H.-J. Apell, V. Häring, M. Roudna; *Biochim. Biophys. Acta,* **1990**, *1023*, 81.
11. S. J. D. Karlish, D. W. Yates, I. M. Glynn; *Biochim. Biophys. Acta,* **1978**, *525*, 230.
12. S. Heyse, I. Wuddel, H.-J. Apell, W. Stürmer; *J. Gen. Physiol.,* **1994**, *104*, 197.
13. I. M. Glynn, S. J. D. Karlish; *Annu. Rev. Physiol.,* **1975**, *37*, 13.
14. Glynn, I. M., in *Membrane Transport* (Martonosi, A. N., Ed.), Plenum Press, New York, **1985**, p. 35.
15. P. J. Garrahan, I. M. Glynn; *J. Physiol.,* **1967**, *192*, 237.
16. P. De Weer, R. F. Rakowski; *Nature,* **1984**, *309*, 450.
17. J. D. Cavieres, I. M. Glynn; *J. Physiol.,* **1979**, *297*, 637.
18. T. J. B. Simons; *J. Physiol.,* **1974**, *237*, 123.
19. P. J. Garrahan, I. M. Glynn; *J. Physiol,* **1967**, *192*, 175.
20. P. J. Garrahan, I. M. Glynn; *J. Physiol,* **1967**, *192*, 159.
21. H.-J. Apell, G. Benz, D. Sauerbrunn; *Biochemistry,* **2011**, *50*, 409.
22. K. H. Lee, R. Blostein; *Nature,* **1980**, *285*, 338.
23. M. Forgac, G. Chin; *J. Biol. Chem.,* **1982**, *257*, 5652.
24. J. R. Sachs; *J. Physiol.,* **1986**, *374*, 221.
25. A. Warshel, S. T. Russell; *Q. Rev. Biophys.,* **1984**, *17*, 283.
26. B. A. Wallace; *Annu. Rev. Biophys. Biophys. Chem.,* **1990**, *19*, 127.
27. P. Läuger, *Electrogenic Ion Pumps*, Sinauer Associates, Inc., Sunderland, MA, **1991**, p. 313.
28. P. Läuger; *Biochim. Biophys. Acta,* **1979**, *552*, 143.
29. P. Läuger, H.-J. Apell; *Biochim. Biophys. Acta,* **1988**, *945*, 1.
30. P. Läuger, H.-J. Apell; *Biochim. Biophys. Acta,* **1988**, *944*, 451.
31. C. S. Patlak; *Bull. Math. Biophys.,* **1957**, *19*, 209.

32. P. Läuger; *Biochim. Biophys. Acta,* **1984**, *779*, 307.
33. P. Läuger, H.-J. Apell; *Eur. Biophys. J.,* **1986**, *13*, 309.
34. R. Borlinghaus, H.-J. Apell, P. Läuger; *J. Membr. Biol.,* **1987**, *97*, 161.
35. I. Wuddel, H.-J. Apell; *Biophys. J.,* **1995**, *69*, 909.
36. C. H. Fiske, Y. Subbarrow; *J. Biol. Chem.,* **1925**, *66*, 375.
37. L. F. Leloir, C. E. Cardini; *Meth. Enzymol.,* **1957**, *1*, 840.
38. O. Vagin, S. Denevich, K. Munson, G. Sachs; *Biochemistry,* **2002**, *41*, 12755.
39. R. L. Post, C. R. Merritt, C. R. Kinsolving, C. D. Albright; *J. Biol. Chem.,* **1960**, *235*, 1796.
40. P. L. Jørgensen, J. C. Skou; *Biochem. Biophys. Res. Commun.,* **1969**, *37*, 39.
41. J. Shaw, L. C. Beadle; *J. exp. Biol.,* **1949**, *26*, 15.
42. O. H. Lowry, A. L. Rosebrough, A. L. FARR, R. J. Randall; *J. Biol. Chem.,* **1951**, *193*, 265.
43. M. A. Markwell, S. M. Haas, L. L. Bieber, N. E. Tolbert; *Anal. Biochem.,* **1978**, *87*, 206.
44. A. K. Schwartz, M. Nagano, M. Nakao, G. E. Lindenmayer, J. C. Allen; *Meth. Pharmacol.,* **1971**, *1*, 361.
45. R. L. Post, A. K. Sen, A. S. Rosenthal; *J. Biol. Chem.,* **1965**, *240*, 1437.
46. R. L. Post, P. C. Jolly; *Biochim. Biophys. Acta,* **1957**, *25*, 118.
47. De Weer P.; *J. Gen. Physiol,* **1970**, *56*, 583.
48. De Weer P., D. Geduldig; *Science,* **1973**, *179*, 1326.
49. T. L. Steck, R. S. Weinstein, J. H. Straus, D. F. Wallach; *Science,* **1970**, *168*, 255.
50. R. Blostein, L. Chu; *J. Biol. Chem.,* **1977**, *252*, 3035.
51. I. M. Glynn, J. F. Hoffman; *J. Physiol,* **1971**, *218*, 239.
52. I. M. Glynn, S. J. D. Karlish; *J. Physiol.,* **1976**, *256*, 465.
53. J. H. Kaplan, R. J. Hollis; *Nature,* **1980**, *288*, 587.
54. D. C. Gadsby, P. F. Cranefield; *Proc. Natl. Acad. Sci. U. S. A,* **1979**, *76*, 1783.
55. D. C. Gadsby, J. Kimura, A. Noma; *Nature,* **1985**, *315*, 63.
56. D. W. Hilgemann; *Science,* **1994**, *263*, 1429.
57. M. Holmgren, J. Wagg, F. Bezanilla, R. F. Rakowski, P. De Weer, D. C. Gadsby; *Nature,* **2000**, *403*, 898.
58. R. F. Rakowski, C. L. Paxson; *J. Membr. Biol.,* **1988**, *106*, 173.
59. P. Artigas, D. C. Gadsby; *Proc. Natl. Acad. Sci. U. S. A,* **2006**, *103*, 12613.
60. K. L. Dürr, N. N. Tavraz, D. Zimmermann, E. Bamberg, T. Friedrich; *Biochemistry,* **2008**, *47*, 4288.
61. N. N. Tavraz, T. Friedrich, K. L. Dürr, J. B. Koenderink, E. Bamberg, T. Freilinger, M. Dichgans; *J. Biol. Chem.,* **2008**, *283*, 31097.
62. S. M. Goldin, S. W. Tong; *J. Biol. Chem.,* **1974**, *249*, 5907.
63. S. Hilden, H. M. Rhee, L. E. Hokin; *J. Biol. Chem.,* **1974**, *249*, 7432.
64. B. M. Anner, L. K. Lane, A. Schwartz, B. J. R. Pitts; *Biochim. Biophys. Acta,* **1977**, *467*, 340.
65. E. Skriver, A. B. Maunsbach, B. M. Anner, P. L. Jørgensen; *Cell Biol. Int. Rep.,* **1980**, *4*, 585.
66. B. Forbush, III; *Anal. Biochem.,* **1984**, *140*, 495.
67. B. Forbush, III; *J. Biol. Chem.,* **1987**, *262*, 11104.
68. B. Forbush, III; *J. Biol. Chem.,* **1987**, *262*, 11116.
69. H.-J. Apell, M. M. Marcus, B. M. Anner, H. Oetliker, P. Läuger; *J. Membr. Biol.,* **1985**, *85*, 49.
70. H.-J. Apell, B. Bersch; *Biochim. Biophys. Acta,* **1987**, *903*, 480.
71. H.-J. Apell, B. Bersch; *Prog. Clin. Biol. Res.,* **1988**, *268A*, 469.
72. J. H. Kaplan, B. Forbush, III, J. F. Hoffman; *Biochem.,* **1978**, *17*, 1929.
73. J. A. McCray, L. Herbette, T. Kihara, D. R. Trentham; *Proc. Natl. Acad. Sci. U. S. A,* **1980**, *77*, 7237.
74. K. Fendler, E. Grell, M. Haubs, E. Bamberg; *EMBO J.,* **1985**, *4*, 3079.
75. H.-J. Apell, R. Borlinghaus, P. Läuger; *J. Membr. Biol.,* **1987**, *97*, 179.
76. R. Borlinghaus, H.-J. Apell; *Biochim. Biophys. Acta,* **1988**, *939*, 197.
77. H. Thirlwell, J. E. T. Corrie, G. P. Reid, D. R. Trentham, M. A. Ferenczi; *Biophys. J.,* **1994**, *67*, 2436.
78. V. S. Sokolov, H.-J. Apell, J. E. Corrie, D. R. Trentham; *Biophys. J.,* **1998**, *74*, 2285.
79. V. S. Sokolov, S. M. Stukolov, N. M. Gevondyan, H.-J. Apell; *Ann. N. Y. Acad. Sci.,* **1997**, *834*, 364.
80. V. Yu. Tashkin, A. N. Gavrilchik, A. I. Ilovaisky, H.-J. Apell, V. S. Sokolov; *Biochemistry (Moscow) Supplement Series A,* **2015**, *9*, 92.
81. W. Domaszewicz, H.-J. Apell; *FEBS Lett.,* **1999**, *458*, 241.
82. J. Pintschovius, K. Fendler; *Biophys. J.,* **1999**, *76*, 814.
83. J. Pintschovius, K. Fendler, E. Bamberg; *Biophys. J.,* **1999**, *76*, 827.
84. F. Tadini-Buoninsegni, P. Nassi, C. Nediani, A. Dolfi, R. Guidelli; *Biochim. Biophys. Acta,* **2003**, *1611*, 70.
85. S. J. Karlish, D. W. Yates; *Biochim. Biophys. Acta,* **1978**, *527*, 115.
86. S. J. Karlish, D. W. Yates, I. M. Glynn; *Nature,* **1976**, *263*, 251.
87. S. J. Karlish, D. W. Yates, I. M. Glynn; *Biochim. Biophys. Acta,* **1978**, *525*, 252.

88. S. J. Karlish, L. A. Beauge, I. M. Glynn; *Nature,* **1979**, *282*, 333.
89. S. J. D. Karlish; *J. Bioenerg. Biomembr.,* **1980**, *12*, 111.
90. C. Hegyvary, P. L. Jørgensen; *J. Biol. Chem.,* **1981**, *256*, 6296.
91. C. T. Carilli, R. A. Farley, D. M. Perlman, L. C. Cantley; *J. Biol. Chem.,* **1982**, *257*, 5601.
92. R. A. Farley, C. M. Tran, C. T. Carilli, D. Hawke, J. E. Shively; *J. Biol. Chem.,* **1984**, *259*, 9532.
93. J. G. Kapakos, M. Steinberg; *Biochim. Biophys. Acta,* **1982**, *693*, 493.
94. P. A. Tyson, M. Steinberg, E. T. Wallick, T. L. Kirley; *J. Biol. Chem.,* **1989**, *264*, 726.
95. M. Steinberg, S. J. Karlish; *J. Biol. Chem.,* **1989**, *264*, 2726.
96. W. Stürmer, H.-J. Apell, I. Wuddel, P. Läuger; *J. Membr. Biol.,* **1989**, *110*, 67.
97. J. C. Skou, M. Esmann; *Biochim. Biophys. Acta,* **1981**, *647*, 232.
98. J. C. Skou, M. Esmann; *Biochim. Biophys. Acta,* **1983**, *727*, 101.
99. J. C. Skou; *Biochim. Biophys. Acta,* **1982**, *688*, 369.
100. L. M. Loew, S. Scully, L. Simpson, A. S. Waggoner; *Nature,* **1979**, *281*, 497.
101. A. Grinvald, L. Anglister, J. A. Freeman, R. Hildesheim, A. Manker; *Nature,* **1984**, *308*, 848.
102. I. Klodos, B. Forbush, III; *J. Gen. Physiol.,* **1988**, *92*, 46A (abstr.).
103. R. Bühler, W. Stürmer, H.-J. Apell, P. Läuger; *J. Membr. Biol.,* **1991**, *121*, 141.
104. W. Stürmer, R. Bühler, H.-J. Apell, P. Läuger; *J. Membr. Biol.,* **1991**, *121*, 163.
105. R. J. Clarke, P. Schrimpf, M. Schoneich; *Biochim. Biophys. Acta,* **1992**, *1112*, 142.
106. P. R. Pratap, J. D. Robinson; *Biochim. Biophys. Acta,* **1993**, *1151*, 89.
107. N. U. Fedosova, F. Cornelius, I. Klodos; *Biochemistry,* **1995**, *34*, 16806.
108. G. Bartolommei, N. Devaux, F. Tadini-Buoninsegni, M. R. Moncelli, H.-J. Apell; *Biophys. J.,* **2008**, *95*, 1813.
109. G. A. Figtree, C. C. Liu, S. Bibert, E. J. Hamilton, A. Garcia, C. N. White, K. K. Chia, F. Cornelius, K. Geering, H. H. Rasmussen; *Circ. Res.,* **2009**, *105*, 185.
110. S. E. Faraj, M. Centeno, R. C. Rossi, M. R. Montes; *Biochim. Biophys. Acta,* **2018**, *1861*, 355.
111. M. Habeck, E. Cirri, A. Katz, S. J. Karlish, H.-J. Apell; *Biochemistry,* **2009**, *48*, 9147.
112. J. C. Skou; *Biochim. Biophys. Acta,* **1960**, *42*, 6.
113. R. L. Post, A. S. Rosenthal; *J. Gen. Physiol.,* **1962**, *45*, 614A (abstr.).
114. J. S. Charnock, R. L. Post; *Nature,* **1963**, *199*, 910.
115. L. E. Hokin, P. S. Sastry, P. R. Galsworthy, A. Yoda; *Proc. Natl. Acad Sci U. S. A,* **1965**, *54*, 177.
116. K. Nagano, T. Kanazawa, N. Mizuno, Y. Tashima, T. Nakao, M. Nakao; *Biochem. Biophys. Res. Commun.,* **1965**, *19*, 759.
117. J. C. Skou; *Physiol Rev.,* **1965**, *45*, 596.
118. A. Schwartz, G. E. Lindenmayer, J. C. Allen; *Curr. Top. Membranes Transp.,* **1972**, *3*, 1.
119. R. Gibbs, P. M. Roddy, E. Titus; *J. Biol. Chem.,* **2018**, *240*, 2181.
120. R. Blostein; *Biochem. Biophys. Res. Commun.,* **1966**, *24*, 598.
121. S. Fahn, G. J. Koval, R. W. Albers; *J. Biol. Chem.,* **1966**, *241*, 1882.
122. T. Hexum, F. E. Samson, Jr., R. H. Himes; *Biochim. Biophys. Acta,* **1970**, *212*, 322.
123. F. Bastide, G. Meissner, S. Fleischer, R. L. Post; *J. Biol. Chem.,* **1973**, *248*, 8385.
124. I. Nishigaki, F. T. Chen, L. E. Hokin; *J. Biol. Chem.,* **1974**, *249*, 4911.
125. G. E. Shull, A. Schwartz, J. B. Lingrel; *Nature,* **1985**, *316*, 691.
126. S. Lutsenko, J. H. Kaplan; *Biochem.,* **1995**, *34*, 15607.
127. H. Matsui, H. Homareda; *J. Biochem. (Tokyo),* **1982**, *92*, 193.
128. M. Esmann, J. C. Skou; *Biochem. Biophys. Res. Comm.,* **1985**, *127*, 857.
129. I. Klodos, J. G. Norby, I. W. Plesner; *Biochim. Biophys. Acta,* **1981**, *643*, 463.
130. J. G. Nørby, I. Klodos, N. O. Christiansen; *J. Gen. Physiol,* **1983**, *82*, 725.
131. A. Yoda, S. Yoda; *Mol. Pharmacol.,* **1982**, *22*, 693.
132. J. G. Nørby, I. Klodos; *Progr. Clin. Biol. Res.,* **1988**, *268A*, 249.
133. J. D. Cavieres; *Curr. Top. Membr. Trans.,* **1983**, *19*, 677.
134. R. W. Albers, G. J. Koval, Siegel; *Mol. Pharmacol.,* **1968**, *4*, 324.
135. G. E. Lindenmayer, A. H. Laughter, A. Schwartz; *Arch. Biochem. Biophys.,* **1968**, *127*, 187.
136. G. J. Siegel, G. J. Koval, R. W. Albers; *J. Biol. Chem.,* **1969**, *244*, 3264.
137. A. K. Sen, T. Tobin, R. L. Post; *J. Biol. Chem.,* **1969**, *244*, 6596.
138. H. Matsui, A. Schwartz; *Biochim. Biophys. Acta,* **1966**, *128*, 380.
139. I. A. Svinukhova, A. A. Boldyrev; *FEBS Letters,* **1987**, *214*, 335.
140. J. D. Robinson; *Biochim. Biophys. Acta,* **1974**, *341*, 232.

141. J. J. Lacapère, N. Bennett, Y. Dupont, F. Guillain; *J. Biol. Chem.,* **1990**, *265*, 348.
142. A. M. Romani; *Arch. Biochem. Biophys.,* **2011**, *512*, 1.
143. M. D. Forgac; *J. Biol. Chem.,* **1980**, *255*, 1547.
144. Y. Fukushima, R. L. Post; *J. Biol. Chem.,* **1978**, *253*, 6853.
145. C. H. Pedemonte, L. Beauge; *Biochim. Biophys. Acta,* **1983**, *748*, 245.
146. P. A. Pedersen, J. R. Jørgensen, P. L. Jørgensen; *J. Biol. Chem.,* **2000**, *275*, 37588.
147. M. Palasis, T. A. Kuntzweiler, J. M. Argüello, J. B. Lingrel; *J. Biol. Chem.,* **1996**, *271*, 14176.
148. J. B. Lingrel, J. M. Argüello, J. van Huysse, T. A. Kuntzweiler; *Ann. N. Y. Acad. Sci.,* **1997**, *834*, 194.
149. E. Or, E. D. Goldshleger, D. M. Tal, S. J. Karlish; *Biochemistry,* **1996**, *35*, 6853.
150. H. Ogawa, T. Shinoda, F. Cornelius, C. Toyoshima; *Proc. Natl. Acad. Sci. U. S. A,* **2009**, *106*, 13742.
151. L. Yatime, M. Laursen, J. P. Morth, M. Esmann, P. Nissen, N. U. Fedosova; *J. Struct. Biol.,* **2011**, *174*, 296.
152. M. Laursen, L. Yatime, P. Nissen, N. U. Fedosova; *Proc. Natl. Acad. Sci. U. S. A,* **2013**, *110*, 10958.
153. M. Y. Abeywardena, E. J. McMurchie, G. R. Russell, J. S. Charnock; *Biochem. Pharmacol.,* **1984**, *33*, 3649.
154. J. M. Hamlyn, R. Ringel, J. Schaeffer, P. D. Levinson, B. P. Hamilton, A. A. Kowarski, M. P. Blaustein; *Nature,* **1982**, *300*, 650.
155. J. M. Hamlyn, M. P. Blaustein, S. Bova, D. W. DuCharme, D. W. Harris, F. Mandel, W. R. Mathews, J. H. Ludens; *Proc. Natl. Acad. Sci. U. S. A,* **1991**, *88*, 6259.
156. A. Kawamura, J. Guo, Y. Itagaki, C. Bell, Y. Wang, G. T. Haupert, Jr., S. Magil, R. T. Gallagher, N. Berova, K. Nakanishi; *Proc. Natl. Acad. Sci. U. S. A,* **1999**, *96*, 6654.
157. W. Schoner; *Exp. Clin. Endocrinol. Diabetes,* **2000**, *108*, 449.
158. J. M. Hamlyn, M. P. Blaustein; *Hypertension,* **2016**, *68*, 526.
159. J. G. Kaplan; *Annu. Rev. Physiol,* **1978**, *40*, 19.
160. Z. Xie, A. Askari; *Eur. J. Biochem.,* **2002**, *269*, 2434.
161. A. Zulian, C. I. Linde, M. V. Pulina, S. G. Baryshnikov, I. Papparella, J. M. Hamlyn, V. A. Golovina; *Am. J. Physiol Cell Physiol,* **2013**, *304*, C324-C333.
162. Z. Xie, T. Cai; *Mol. Interv.,* **2003**, *3*, 157.
163. W. Schoner, G. Scheiner-Bobis; *Am. J. Physiol Cell Physiol,* **2007**, *293*, C509-C536.
164. Y. Yan, J. I. Shapiro; *Curr. Opin. Pharmacol.,* **2016**, *27*, 43.
165. L. C. Cantley, L. Josephson, R. Warner, M. Yanagisawa, C. Lechene, G. Guidotti; *J. Biol. Chem.,* **1977**, *252*, 7421.
166. L. C. Cantley, L. G. Cantley, L. Josephson; *J. Biol. Chem.,* **1978**, *253*, 7361.
167. I. M. Glynn; *Biochem. J.,* **1962**, *84*, 75P (abstr.).
168. J. Järnefelt; *Biochim. Biophys. Acta,* **1962**, *59*, 643.
169. S. Fahn, M. R. Hurley, G. J. Koval, R. W. Albers; *J. Biol. Chem.,* **1966**, *241*, 1890.
170. H. Homareda, T. Ishii, K. Takeyasu; *Eur. J. Pharmacol.,* **2000**, *400*, 177.
171. R. Kanai, H. Ogawa, B. Vilsen, F. Cornelius, C. Toyoshima; *Nature,* **2013**, *502*, 201.
172. H.-J. Apell; *Bioelectrochemistry.,* **2004**, *63*, 149.
173. A. Schneeberger, H.-J. Apell; *J. Membr. Biol.,* **2001**, *179*, 263.
174. R. Goldshlegger, S. J. Karlish, A. Rephaeli, W. D. Stein; *J. Physiol (Lond),* **1987**, *387*, 331.
175. A. Schneeberger, H.-J. Apell; *J. Membr. Biol.,* **1999**, *168*, 221.
176. H.-J. Apell, A. Diller; *FEBS Lett.,* **2002**, *532*, 198.
177. V. S. Sokolov, A. A. Scherbakov, A. A. Lenz, Yu. A. Chizmadzhev, H.-J. Apell; *Biochemistry (Moscow) Supplement Series A,* **2008**, *2*, 161.
178. H. J. Apell, T. Hitzler, G. Schreiber; *Biochemistry,* **2017**, *56*, 1005.
179. J. V. Møller, C. Olesen, A. M. Winther, P. Nissen; *Q. Rev. Biophys.,* **2010**, *43*, 501.
180. H.-J. Apell; *Ann. N. Y. Acad. Sci.,* **1997**, *834*, 221.
181. P. De Weer, D. C. Gadsby, R. F. Rakowski; *Prog. Clin. Biol. Res.,* **1988**, *268A*, 421.
182. M. Nakao, D. C. Gadsby; *J. Gen. Physiol,* **1989**, *94*, 539.
183. A. Sagar, R. F. Rakowski; *J. Gen. Physiol,* **1994**, *103*, 869.
184. T. Friedrich, G. Nagel; *Biophys. J.,* **1997**, *73*, 186.
185. D. C. Gadsby, F. Bezanilla, R. F. Rakowski, P. De Weer, M. Holmgren; *Nat. Commun.,* **2012**, *3*, 669.
186. R. V. Gradinaru, H.-J. Apell; *Biochemistry,* **2015**, *54*, 2508.
187. B. Forbush, III; *Prog. Clin. Biol. Res.,* **1988**, *268A*, 229.
188. R. F. Rakowski, L. A. Vasilets, J. LaTona, W. Schwarz; *J. Membr. Biol.,* **1991**, *121*, 177.
189. F. Jaisser, P. Jaunin, K. Geering, B. C. Rossier, J. D. Horisberger; *J. Gen. Physiol,* **1994**, *103*, 605.
190. R. Bühler, H.-J. Apell; *J. Membr. Biol.,* **1995**, *145*, 165.
191. R. D. Peluffo, J. R. Berlin; *J. Physiol,* **1997**, *501*, 33.
192. R. D. Peluffo, Y. Hara, J. R. Berlin; *J. Gen. Physiol,* **2004**, *123*, 249.
193. R. D. Peluffo, R. M. Gonzalez-Lebrero, S. B. Kaufman, S. Kortagere, B. Orban, R. C. Rossi, J. R. Berlin; *Biochemistry,* **2009**, *48*, 8105.
194. R. D. Peluffo, J. R. Berlin; *Mol. Pharmacol.,* **2012**, *82*, 1.

195. L. J. Mares, A. Garcia, H. H. Rasmussen, F. Cornelius, Y. A. Mahmmoud, J. R. Berlin, B. Lev, T. W. Allen, R. J. Clarke; *Biophys. J.,* **2014**, *107*, 1352.
196. R. Goldshlegger, S. J. Karlish, A. Rephaeli, W. D. Stein; *J. Physiol,* **1987**, *387*, 331.
197. A. Bahinski, M. Nakao, D. C. Gadsby; *Proc. Natl. Acad. Sci. U. S. A,* **1988**, *85*, 3412.
198. B. Forbush, III; *J. Biol. Chem.,* **1988**, *263*, 7961.
199. H.-J. Apell, M. Roudna, J. E. Corrie, D. R. Trentham; *Biochemistry,* **1996**, *35*, 10922.
200. S. J. D. Karlish, D. W. Yates; *Biochim. Biophys. Acta,* **1978**, *527*, 115.
201. I. M. Glynn, D. E. Richards; *J. Physiol.,* **1982**, *330*, 17.
202. R. E. Moore, P. J. Scheuer; *Science,* **1971**, *172*, 495.
203. M. Ikeda, K. Mitani, K. Ito; *Naunyn Schmiedebergs Arch. Pharmacol.,* **1988**, *337*, 591.
204. M. T. Tosteson, J. A. Halperin, Y. Kishi, D. C. Tosteson; *J. Gen. Physiol,* **1991**, *98*, 969.
205. H. Ozaki, H. Nagase, N. Urakawa; *Eur. J. Biochem.,* **1985**, *152*, 475.
206. E. Habermann; *Toxicon,* **1989**, *27*, 1171.
207. X. Wang, J.-D. Horisberger; *FEBS Letters,* **1997**, *409*, 391.
208. M. T. Tosteson, G. S. Bignami, D. R. L. Scriven, A. K. Bharadwaj, D. C. Tosteson; *Biochim. Biophys. Acta,* **1994**, *1191*, 371.
209. P. Artigas, D. C. Gadsby; *Proc. Natl. Acad. Sci. U. S. A,* **2003**, *100*, 501.
210. P. Artigas, D. C. Gadsby; *J. Gen. Physiol,* **2004**, *123*, 357.
211. N. Harmel, H.-J. Apell; *J. Gen. Physiol,* **2006**, *128*, 103.
212. N. Reyes, D. C. Gadsby; *Nature,* **2006**, *443*, 470.
213. A. Takeuchi, N. Reyes, P. Artigas, D. C. Gadsby; *Nature,* **2008**, *456*, 413.
214. T. L. Hill, *Free Energy Transduction in Biology*, Academic Press, New York, **1977**, p. 229.
215. T. L. Hill, *Free Energy Transduction and Biochemical Cycle Kinetics*, Springer, New York, **1989**, p. 119.
216. H. J. Schatzmann; *Protoplasma,* **1967**, *63*, 136.
217. A. M. Rose, R. Valdes; *Clin. Chem.,* **1994**, *40*, 1674.
218. K. J. Swoboda, E. Kanavakis, A. Xaidara, J. E. Johnson, M. F. Leppert, M. B. Schlesinger-Massart, L. J. Ptacek, K. Silver, S. Youroukos; *Ann. Neurol.,* **2004**, *55*, 884.
219. T. Friedrich, N. N. Tavraz, C. Junghans; *Front Physiol,* **2016**, *7*, 239.
220. A. N. Shrivastava, V. Redeker, N. Fritz, L. Pieri, L. G. Almeida, M. Spolidoro, T. Liebmann, L. Bousset, M. Renner, C. Lena, A. Aperia, R. Melki, A. Triller; *EMBO J.,* **2015**, *34*, 2408.
221. T. Ohnishi, M. Yanazawa, T. Sasahara, Y. Kitamura, H. Hiroaki, Y. Fukazawa, I. Kii, T. Nishiyama, A. Kakita, H. Takeda, A. Takeuchi, Y. Arai, A. Ito, H. Komura, H. Hirao, K. Satomura, M. Inoue, S. Muramatsu, K. Matsui, M. Tada, M. Sato, E. Saijo, Y. Shigemitsu, S. Sakai, Y. Umetsu, N. Goda, N. Takino, H. Takahashi, M. Hagiwara, T. Sawasaki, G. Iwasaki, Y. Nakamura, Y. Nabeshima, D. B. Teplow, M. Hoshi; *Proc. Natl. Acad. Sci. U. S. A,* **2015**, *112*, E4465-E4474.
222. F. Beuschlein, S. Boulkroun, A. Osswald, T. Wieland, H. N. Nielsen, U. D. Lichtenauer, D. Penton, V. R. Schack, L. Amar, E. Fischer, A. Walther, P. Tauber, T. Schwarzmayr, S. Diener, E. Graf, B. Allolio, B. Samson-Couterie, A. Benecke, M. Quinkler, F. Fallo, P. F. Plouin, F. Mantero, T. Meitinger, P. Mulatero, X. Jeunemaitre, R. Warth, B. Vilsen, M. C. Zennaro, T. M. Strom, M. Reincke; *Nat. Genet.,* **2013**, *45*, 440.
223. E. A. Azizan, H. Poulsen, P. Tuluc, J. Zhou, M. V. Clausen, A. Lieb, C. Maniero, S. Garg, E. G. Bochukova, W. Zhao, L. H. Shaikh, C. A. Brighton, A. E. Teo, A. P. Davenport, T. Dekkers, B. Tops, B. Kusters, J. Ceral, G. S. Yeo, S. G. Neogi, I. McFarlane, N. Rosenfeld, F. Marass, J. Hadfield, W. Margas, K. Chaggar, M. Solar, J. Deinum, A. C. Dolphin, I. S. Farooqi, J. Striessnig, P. Nissen, M. J. Brown; *Nat. Genet.,* **2013**, *45*, 1055.
224. T. A. Williams, S. Monticone, V. R. Schack, J. Stindl, J. Burrello, F. Buffolo, L. Annaratone, I. Castellano, F. Beuschlein, M. Reincke, B. Lucatello, V. Ronconi, F. Fallo, G. Bernini, M. Maccario, G. Giacchetti, F. Veglio, R. Warth, B. Vilsen, P. Mulatero; *Hypertension,* **2014**, *63*, 188.
225. C. P. Roenn, M. Li, V. R. Schack, I. C. Forster, R. Holm, M. S. Toustrup-Jensen, J. P. Andersen, S. Petrou, B. Vilsen; *J. Biol. Chem.,* **2018**, doi: 10.1074/jbc.RA118.004591.
226. P. Lassuthova, A. P. Rebelo, G. Ravenscroft, P. J. Lamont, M. R. Davis, F. Manganelli, S. M. Feely, C. Bacon, D. S. Brozkova, J. Haberlova, R. Mazanec, F. Tao, C. Saghira, L. Abreu, S. Courel, E. Powell, E. Buglo, D. M. Bis, M. F. Baxter, R. W. Ong, L. Marns, Y. C. Lee, Y. Bai, D. G. Isom, R. Barro-Soria, K. W. Chung, S. S. Scherer, H. P. Larsson, N. G. Laing, B. O. Choi, P. Seeman, M. E. Shy, L. Santoro, S. Zuchner; *Am. J. Hum. Genet.,* **2018**, *102*, 505.
227. Castaneda M.S., E. Zanoteli, R. S. Scalco, V. Scaramuzzi, C. Marques, V, R. U. Conti, A. M. S. da Silva, B. O'Callaghan, R. Phadke, E. Bugiardini, R. Sud, S. McCall, M. G. Hanna, H. Poulsen, R. Mannikko, E. Matthews; *Brain,* **2018**, *141*, 3308.
228. M. V. Clausen, F. Hilbers, H. Poulsen; *Front Physiol,* **2017**, *8*, 371.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Research Article

# Range separation: the divide between local structures and field theories

David M. Rogers

*University of South Florida, 4202 E. Fowler Ave., CHE 205, Tampa, FL 33620, US*
E-mail: davidrogers@usf.edu

**Abstract.** This work presents parallel histories of the development of two modern theories of condensed matter: the theory of electron structure in quantum mechanics, and the theory of liquid structure in statistical mechanics. Comparison shows that key revelations in both are not only remarkably similar, but even follow along a common thread of controversy that marks progress from antiquity through to the present. This theme appears as a creative tension between two competing philosophies, that of short range structure (atomistic models) on the one hand, and long range structure (continuum or density functional models) on the other. The timeline and technical content are designed to build up a set of key relations as guideposts for using density functional theories together with atomistic simulation.

**Keywords**. Electronic structure, liquid state structure, density functional theory, Bayes' theorem, vapor interface, molecular dynamics.



**Citation:** D.M. Rogers (2019) Range separation: the divide between local structures and field theories. *Substantia* 3(1): 43-62. doi: 10.13128/Substantia-208



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Competing Interests:** The Author(s) declare(s) no conflict of interest.

Many of the most important scientific theories were forged out of controversy – like particles vs. waves, for which Democritus claimed (with his teacher, Leucippus of 5th century BC) that all things, including the soul, were made of particles, while Aristotle held to the Greek notion that there were continuous distributions of four or five elements.[1] It is telling to note that Aristotle's objection was strongly biased by his notion that the continuum theory was elegant and beautiful, and does not require any regions of vacuum. In addition, his conception of kinetic equations were first order – like Brownian motion, Navier-Stokes, or the Dirac equation, but not second order like Newton's or Schrödinger's. Newton sided with Democritus. In 1738, Daniel Bernoulli first explained thermodynamic pressure using a model of independent atomic collisions. That theory was not scheduled to be widely adopted until the caloric theory (which postulated conservation of heat) was overthrown by James Joule in the 1850s. Wilhelm Ostwald was famously stubborn for refusing to accept the atomic nature of matter until the early 1900s, after Einstein's theory of Brownian motion was confirmed by Jean Perrin's experiment.

The working out of gas dynamics by Maxwell and Boltzmann in the 1860s depended critically on switching between a physical picture of a 2-atom collision and a continuum picture of a probability distribution over

particle velocities and locations (Fig. 4a). Collision events drawn at random from a Boltzmann distribution were useful for predicting pressures and reaction rates. Whether that distribution represented a probability or an actual average over a well-enough defined physical system was left open to interpretation. Five decades later, Gibbs would argue with Ehrenfest[2] over this issue. Gibbs seemed to understand the continuous phase space density as *any* probability distribution that met the requirements of stationarity under time evolution. An observer with no means of gathering further information would have to accept it as representing reality. Ehrenfest argued that a well-defined physical system is exact, mechanical, and objective. The controversy was only resolved by the advent of the age of computation,[3] since we forgot about it. Three decades on, the physicist Jaynes championed the (subjective) maximum entropy viewpoint,[4] while mathematicians like Sinai and Ruelle[5-8] moved to do away with the whole subjectivity business by using only exact dynamical systems as starting assumptions.

Maxwell described light propagation by filling the continuum with 'idler wheels,' and the resulting partial differential equations inspired much of 20th century mathematics. Planck saw his own condition on quantized transfer of light energy as a regrettable, but necessary refinement of Maxwell's theory. Planck believed so strongly in that theory that he at first rejected Einstein's 1905 concept of the photon.[9] It was also five decades later, around 1955, when a field theory of the electron (quantum electrodynamics) was gaining acceptance from precise calculations of experimental details like the gyromagnetic ratio, radiation-field drag (spontaneous emission) and the Lamb shift. This quantum field theory is not a completely smooth continuum, since it incorporates particles using 'second quantization.' It understands particles as wavelike disturbances that pop in and out of existence in an otherwise continuous field. The technical foundations of that theory are derived by 'path-integrals' over all possible motions of Maxwell's idler wheels. As a consequence, infinities characterize the theory,[10] so that the mathematical status of many path integrals is still not settled[11] except in the Gaussian case,[12,13] and where time-sliced limits are well-behaved.[14]

This article discusses some well-known historical developments in the theory of electronic and liquid structure. As its topic is physical chemistry, this history vacillates without warning between experimental facts and technical details of the mathematical models conjured to describe them. The topics, outlined in Table 1, have been chosen specifically to highlight the debate between local structural and field theoretical models. Note that we have also presented the two topics in an idiosyncratic way to highlight their similarities. Differences between electronic and liquid structure theories are easy to find. By the nature of this type of article, we could not hope to be comprehensive. There has not been space to include many significant historical works, while it is likely several offshoots and recent developments have been unknowingly overlooked. Both histories trace their roots to the Herapath/Maxwell/Boltzmann concep-

**Table 1.** Contrasting long-range (LR) and short-range (SR) ideas showing stages of debate over atoms and electrons (top sections), along with concepts from hybrid theories (lower section).

| | SR/Discrete | LR/Continuous | |
|---|---|---|---|
| (Democritus) | atoms | elements | (Aristotle) |
| (Ehrenfest) | microstate | ensemble | (Gibbs) |
| (Einstein) | particle | wave | (Ostwald) |
| (Boltzmann) | distribution function | 1-body probability density | (Jaynes) |
| (Wein) | $n(\nu)$ | $\nu^2 d\nu$ | (Rayleigh-Jeans) |
| | $\hat{n}(r, p)$ | $n(r)$, $V^{\text{ext}}(r)$ | |
| | | Jellium | (Sommerfeld) |
| (Mott) | insulator | conductor | (Pauli) |
| (Hartree-Fock) | Slater determinant | Electron density | (Hohenberg-Kohn-Sham) |
| (Born- Oppenheimer) | nucleii | electrons | |
| | correlation hole | polarization response | |
| (Bohm-Pines) | ←−−−−Quasiparticle | | |
| | Phonon−−−−→ | | |
| | ←−−−−Cooper Pair | | |
| | Hybrid DFT−−−−→ | | |

tion of a continuous density (or probability distribution) of discrete molecules, and both remain active research areas that are even in communication on several points. We will find that, like Democritus and Aristotle, not only are there are strong opinions on both sides, but progress continues to be made by researchers regardless of whether they adopt discrete or continuum worldviews.

## ELECTRONIC STRUCTURE THEORIES

Between the lines of the history above, we find Bose's famous 1924 Z. Physik paper describing the statistics of bosons, which Einstein noted 'also yields the quantum theory of the ideal gas,' and the Thomas-Fermi theory of 1927-28 for a gas of electrons under a fixed applied voltage. Their basic conception was to model the 6-dimensional space of particle locations, $r$ and momenta, $p$ with the volume element,

$$g(p')dp' = dp' \int \delta(|p| - p')h^{-3}\, dr^3\, dp^3 = 4\pi V\, h^{-3}p'^2 dp' \qquad (1)$$

Using $p' = h\nu/c$ for photons of frequency $\nu$ provides $g(\nu)$, the number of available states for photons near frequency $\nu$. Applying Bose counting statistics to $n(\nu)$ photons occupying $2g(\nu)$ possible states for each frequency gives Bose's derivation of Planck's law. In the Thomas-Fermi (TF) model, $p'$ is electron momentum. Applying Fermi statistics to the occupancy number $N = 2\int_0^{hk_F} g(k\hbar)\, d(k\hbar)$ now gives a Fermi distribution for an ideal gas of electrons under a constant external potential (electrostatic voltage). In both cases the number of states is doubled – counting 2 polarizations for photons or 2 spin states for electrons. The result of the first procedure is a free energy expression for the vacuum. The result of the second is a free energy for electrons under a constant voltage.

This idea of a gas with uniform properties uses a long-range field to guess at local structure. Quantitatively, if the voltage at point $r$ is $\phi(r)$, then the theory predicts electrons will fill states up to maximum momentum of $k_F = \sqrt{(2m_e e_0 \phi(r))}/\hbar$, (where the kinetic energy is $E_F = \hbar^2 k_F^2/2m_e$ and $e_0$ is the electron charge) so the local density is,

$$n(r) = k_F^3/3\pi^2. \qquad (2)$$

The resulting model is then usually found to predict long-range properties of metals relatively well. Fig. 1a and b show plots of free energy *vs* number of electrons in an independent electron solution of the Schrödinger equation for a well of positive potential.[15] Panel b shows a simple adaptation of that model where electrons bind in pairs. The states of the electrons in these exact solutions still represent momentum levels, and are thus qualitatively very close to those of the Thomas-Fermi theory.

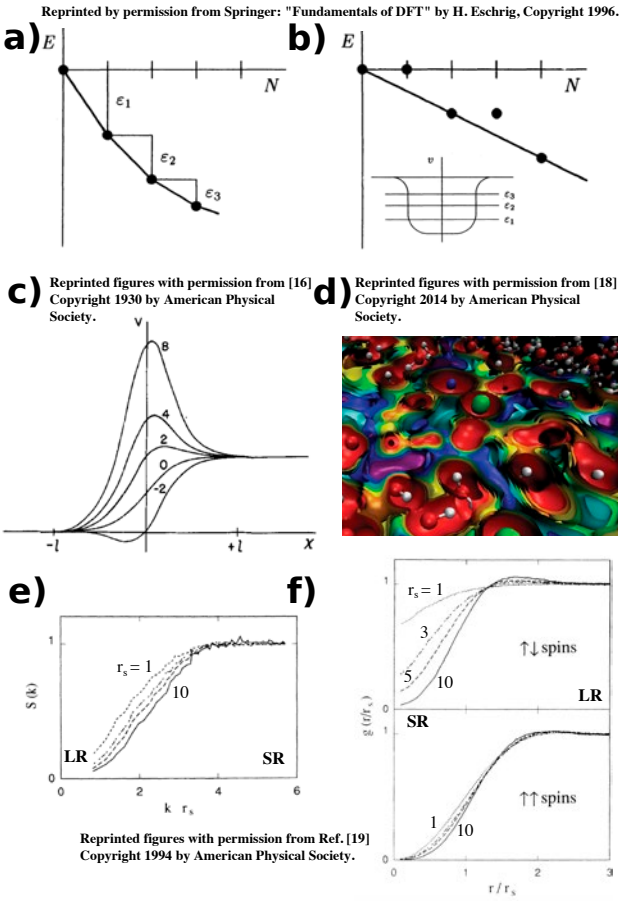



**Figure 1.** Long-range (left) and short-range (right) theories of electronic structure. (a) and (b) show free energy vs. electron number for a potential well.[15] (c) shows 'Epstein' profile of dielectric response[16,17] at a metal/vacuum interface. Numbers for each curve give the surface/bulk conductivity ratio. (d) shows surfaces of constant voltage at a water/vacuum interface, (e) and (f) show the correlation function of jellium from accurate calculations.[19]

The free electron gas evolved into the famous 'jellium' model of electron motion rather quickly, as can be seen by the earliest references in a discussion of that model from the late 20th Century.[20] The term jellium was coined by Conyers Herring in 1952 to describe the model of a metal used by Ewald[21] and others consisting of a uniform background density of positive charge. The electrons are therefore free to move about in gas-like motion. At high density, the electrons actually do act like a free gas, so it was possible to use the Thom-

as-Fermi theory to qualitatively describe the electronic contribution to specific heat, $C_v = \pi^2 k_B^2 T/2E_F$, as well as the spin susceptibility and width of the conduction band (after re-scaling the electron mass).[22]

These are long-range properties from the collective motion of many electrons. The predictions become poor for semi-metals and transition metals. It also rather poorly described the cohesive energy of the metal itself. Those cases fail because of the importance of short-range interactions that a free electron theory just doesn't have.[23]

The contrast becomes important at interfaces, as is visible when comparing Fig. 1c,d. On the left is an early model of local charge density response due to placing an external voltage at a point near a metal surface. On the right is a map of the local voltage for one surface configuration of an electrolyte solution computed using an accurate quantum density functional theory. Chloride ions are green, and sodium ions are blue. Treating one of the sodium ions as a test charge, the material response comes from rearrangement of waters (red and white spheres) and $Cl^-$ ions within a nuanced voltage field (colored surfaces).

It turns out that the electron gas in 'real' jellium behaves rather differently at low and high density. At low density, the electron positions are dominated by pairwise repulsion, and organize themselves into a lattice (of plane waves) with low conductivity.[24] This low-density state is named the 'Wigner lattice' after E. P. Wigner, who computed energetics of an electron distribution based on the lattice symmetry of its host metal.[25] At higher densities, collective motions of electrons screen out the pairwise repulsion at long range. This gives rise to a nearly 'free,' continuous distribution of electrons with higher conductivity more like we would picture for a metal. Fig. 2a, from a well-known particle-based simulation of Ceperly and Alder,[26] shows the Wigner lattice as well as both spin-polarized and unpolarized high-density states.

Taking the opposing side, early applications of self-consistent field (Hartree-Fock or HF) theory to molecules and oxides noticed that the long-range, collective 'correlated' behavior of the electrons was usually irrelevant to the short-range structure of electronic orbitals. Getting the short-range orbital structures right allowed HF theory to do well describing the shapes of molecules and the cohesive energy of metal oxides,[27] as well as magnetic properties.[28] More recent work has shown explicitly that a model that altogether omits the long-range tail of the $1/r$ potential still allows accurate calculations of the lattice energy of salt crystals.[29]

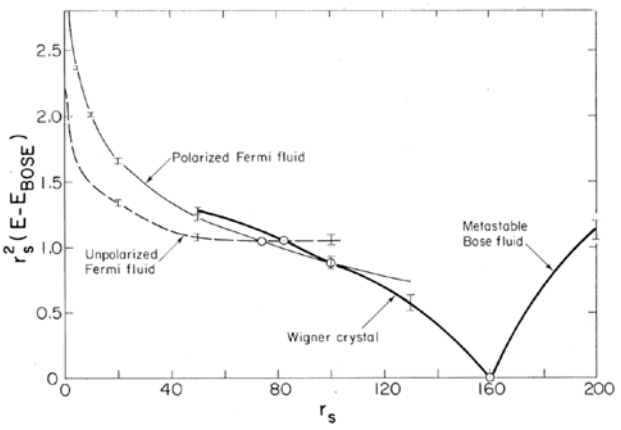
Although both theories worked well for their respective problems, the transition from insulating to conducting metals (as electron density increases) also proved to be difficult because it involved a cross-over between both short- and long-range effects. Because of this mixture of size scales required, relying exclusively on a theory appropriate for either short- or long-range produces results that increasingly depend on cancellation of errors. This sort of error cancellation is illustrated by the phenomenology of 'overdelocalization'.

Well known to density functional theorists, 'over-delocalization' is the tendency of continuum models for electron densities (having their roots in the long-range TF theory) to spread electrons out too far away from the nucleus of atoms. The result is that electron clouds appear 'softer' in these theories, and polarization of the charge cloud by the charge density of a far molecules contributes too much energy. On the other hand, induced-dipole induced-dipole dispersion forces are not modeled by simple density functionals, and so their stabilizing effect is not present. It has been found that the over-delocalization can be fixed by making a physical distinction between short and long-range forces. However, the resulting binding energies are not strong enough. After the correction, they need a separate addition of a dispersion energy to bring them back into agreement with more accurate calculations.[30] Thus, a bit of sloppiness on modeling short-range structure can compensate for the missing, collective long-range effects.

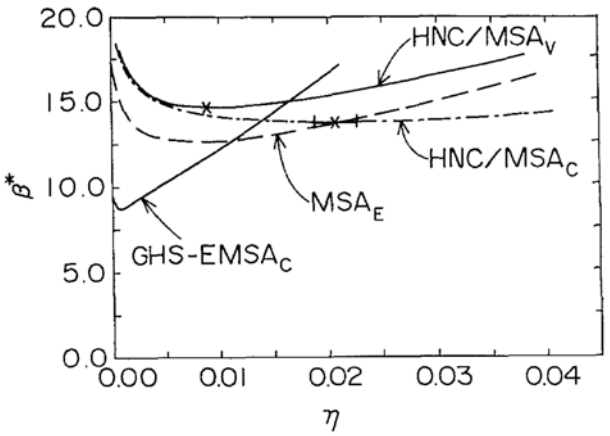
## HYBRID THEORIES IN ELECTRONIC STRUCTURE

When looking at properties like the cross-over between conducting and insulating behavior of electrons, it's not surprising that successful theories strike a balance between short-range, discrete structure and long-range continuum effects. Even in the venerable Born-Oppenheimer approximation from 1927, we see that atomic nucleii are treated as atoms (immovable point charges), while electrons are described using the wave theory. The separation in time-scales of their motion makes this work. By the time the atoms in a molecule have even slightly moved, the electrons have zipped back and forth between them many times over.

Correlation functions are a central physical concept in the debate between long and short range ideas. The distance-dependent correlation function, $g(r)$, measures the relative likelihood of finding an electron at the point, $r$, given that one sits at the origin. One of the first attempts at accounting for electron-electron interaction was to use perturbation theory to add electron interactions back into the uniform gas model ($g(r) = 1$). The first order perturbation modifies this by looking at




(a) Ground state energy vs. density for the uniform electron gas.[26] Four separate phases were observed (at zero temperature). Note that the density axis is reversed by the transformation $1/n = 4\pi r_s^3/3$. Reprinted figure with permission from Ref. 26. Copyright 1980 by the American Physical Society.




(b) Phase diagram of a z:z electrolyte like NaCl where $n$ is the cation concentration. Lines show the position of the spinodal using methods appropriate for each theory, and the minimum indicates a critical point for fluctuation in ionic concentration. Note the temperature axis is reversed by $\beta = z^2/dk_BT$ and $\eta = \pi nd^3/6$, $d$ is the ion diameter. Reprinted from Ref. 35, with the permission of AIP publishing.

**Figure 2.** Comparing phase diagrams of the electron gas dissolved ions. Both show an insulating phase at low density (labeled Wigner crystal in (a)) and a conducting phase at high density separated by a minimum. The corresponding transition in an electron gas has not been well studied, but critical temperatures feature in the phase diagram of superconducting cuprates (where $n$ is percent of solid impurities).[36]

interactions between electrons of the same spin. This interaction is termed the exchange energy, since it comes from pairs of electrons with the same spin exchanging momentum.[22] After the correction, electrons with parallel spin now have smaller density at contact, $g(r) = 1 - {}^9/_2(\sin(k_Fr) - k_Fr\cos(k_Fr))^2/(k_Fr)$.[6]

The correlation function between infinite periodic structures is $S(k)$, the long-range analogue of $g(r)$ (in fact its Fourier transform). The function $S(k)$ is called the structure factor by crystallographers. If the system consisted only of electrons, the structure factor could be measured directly by light or electron scattering experiments. There, $S(k)$ is the intensity scattered out at angle $\theta = 2\arcsin(\lambda k/4\pi)$ when the material is placed into a weak beam of photons or electrons of wavelength $\lambda$ pointed in the $\theta = 0$ direction. This function has been computed using an accurate particle simulation technique and shown in Fig. 1e,f.[19] The curves are labeled by $r_s = (3/4\pi n)^{1/3}$, measured in units of Bohr radii.

There is a duality between short and long range perspectives inherent in $g(r)$ and $S(k)$ as well. Long-range behavior appears at large $r$ when $g(r)$ approaches 1. At small $r$, the geometry of inter-particle interactions determines the shape of $g(r)$. Because particle dynamics is carried out in real-space, $g(r)$ tends to be used by its practitioners to characterize short and long-range structure. Analytical solutions of many models, and especially those aiding experimental measurements, are simpler in Fourier space. There, $S(0)$ is the integral of $g(r)$. It provides information on the total fluctuations in the number of particles, and is a long-range quantity from which the compressibility, partial molar volumes, and other properties can be computed.[31] Short-range structures that repeat with length $d$ show up as peaks in $S(k)$ at correspondingly large $k = 2\pi/d$.

Back to the metallic/insulator problem, between 1950 and 1953 Bohm and Pines pioneered the idea of explicitly splitting the energy function (Hamiltonian) governing electron motion into local and long-range degrees of freedom.[32-34] Using the intuition that long-range collective motions of electrons should look like the continuous plane-wave solutions to Maxwell's theory, they added and subtracted those terms and called them 'plasmons' (Fig. 4d). Just like photons, the plasmons are continuous waves when treated classically, but are quantized particles when understood quantum mechanically.

What remained after the subtraction was a Hamiltonian whose interactions were only short-ranged, but could not be treated with a continuum description. Instead, the short-range part describes interactions between effective discrete particles which Bohm and Pines dubbed 'quasiparticles'. The quasiparticles were like packs of electrons surrounded by empty space, 'holes.' The quasiparticles thus have larger mass and softer, screened, pair interactions (explaining why the mass has to be fixed when applying the free electron theory

to metals). These new 'renormalized' electron quasiparticles could even have effective pairwise attraction. This latter effect was a central component to the BCS model of superconductivity, where the quasiparticles are known as 'Cooper pairs.' Because of its dual representation, the Bohm-Pines model gave good answers for both cohesive energies and conductivities – and described the cross-over between insulating and metallic regimes as electron density is increased.[24]

For all its descriptive power, the Bohm-Pines approach was often lamented for its requirement for a specific set of approximations. Most damningly, it required inventing a continuum of plasmons to describe the long-range interactions of a finite set of electrons. This adds infinite degrees of freedom to a system with an initially finite number. It also required the plasmons to stop and the particles to commence at some cut-off wavelength. These troubles lead us into the problem of renormalization group theory, which is beyond the scope of the present article.

In fact, in 1954, just after the publication of the last article in the Bohm and Pines series above, Lindhard provided a model for collective electronic response of a metal that involved only the metal's correlation function (by means of its dielectric coefficient, $\varepsilon$).[34] Following a decade later in 1964-65 was Hohenberg, Kohn and Sham's density functional theory.[37-39] Both developments rephrased the description of electronic structure in terms of a continuous field of electron density. Linear response (perturbation) theory says that an initially homogeneous density $n_0$ responds to an applied field, $\phi$ as,

$$\Delta n(r) = n_0 \int \chi(r, r')\phi(r'), \tag{3}$$

where $\chi(r, r')$ is the Fourier transform of the structure factor above. Their defining characteristic is the focus on continuous response of that density to a continuous external field, $\rho(r) = \rho[\phi(r')](r)$.

The theory may be understood as a fully long-ranged point of view that includes short-range effects indirectly through $S(k)$. It shows how to use integration to calculate all thermodynamic quantities from structure factor. The only problem is that it does not broach the issue of how to predict the structure factor. One well-known method is to assume the probability of $n(r)$ is a Gaussian on function space (so the exponent depends on $\int n(k)^2/\chi(k)dk^3$, and $\chi(k)$ is just slightly different from $S(k)$). In that case, the inverse of the correlation function ($1/\chi(k)$) is a self-energy term plus the inter-particle energy function. This assumption is known as the random phase approximation (RPA), named because of its historical discovery by Bohm and Pines following from neglecting couplings between a set of linearly independent (Fourier) modes, $n(k)$. This ends up excluding all non-Gaussian fluctuations.

The 'dielectric' ideas encapsulated in the linear response theory of Eq. 3 can be combined with the free electron model of Eq. 2 ($T$ [$n$] proportional to $n^{5/3}$), or a wavefunction calculation of the kinetic energy, $T[n]$, to synthesize modern density functional theory (DFT).[20;40] It writes the electron configuration energy as,

$$A[\phi] = \inf_{n(r)} T[n] + E_{\mathrm{XC}}[n] + \int n(r) \left( \phi(r) + \frac{1}{2} \int dr'^3 \frac{n(r')}{4\pi\epsilon_0 |r - r'|} \right) dr^3 . \tag{4}$$

Now the (long-range) correlation function of the electron, $\chi$, is obtained from the curvature of $A[\phi]$. Mathematically, the unknown structure factor has been migrated into an unknown functional, $E_{\mathrm{XC}}[n]$. The initials stand for exchange and correlation, its two major components. The principle advantage gained by this rephrasing is that new, accurately known (usually short-range) terms like $T[n]$ can be added to $A[\phi]$ in order to decrease the burden on $E_{\mathrm{XC}}$ to model 'everything else.' The disconnect between short and long-range energies can be shoveled into some fitting parameters.

Again moving forward 40 years, the relative unimportance of long-range Coulomb interactions for local structuring noticed by Lang and Perdew[29,41] lead to the suggestion that the density functional method itself should also distinguish between short and long range structural effects. Implementation of this idea was perhaps first carried out by Toulouse, Colonna and Savin in 2004.[42] There, the local density approximation deriving its roots in the TF theory is applied to describe short-range interactions, while the HF theory is used to ensure proper electron-pair repulsion (exchange) energies at long-range. The association of HF with long-range and density functional (DF) with short-range apparently runs counter to our association between continuum, density-based, models for long-range interactions *vs.* discrete, particle-based models for short-range interactions. A major complication with our association is that it is known that the HF method describes the long-range (asymptotic) electronic interactions well, whereas the DF method does not. DF methods were historically used to describe the 'entire' energy function, and have thus been tailored to describe quasi-particles (the so-called exchange hole), rather than asymptotics. This association was put to the test shortly after by Vydrov and co. [43] using an earlier DF called LSDA that is not strongly tailored in this way. They separately averaged the short-

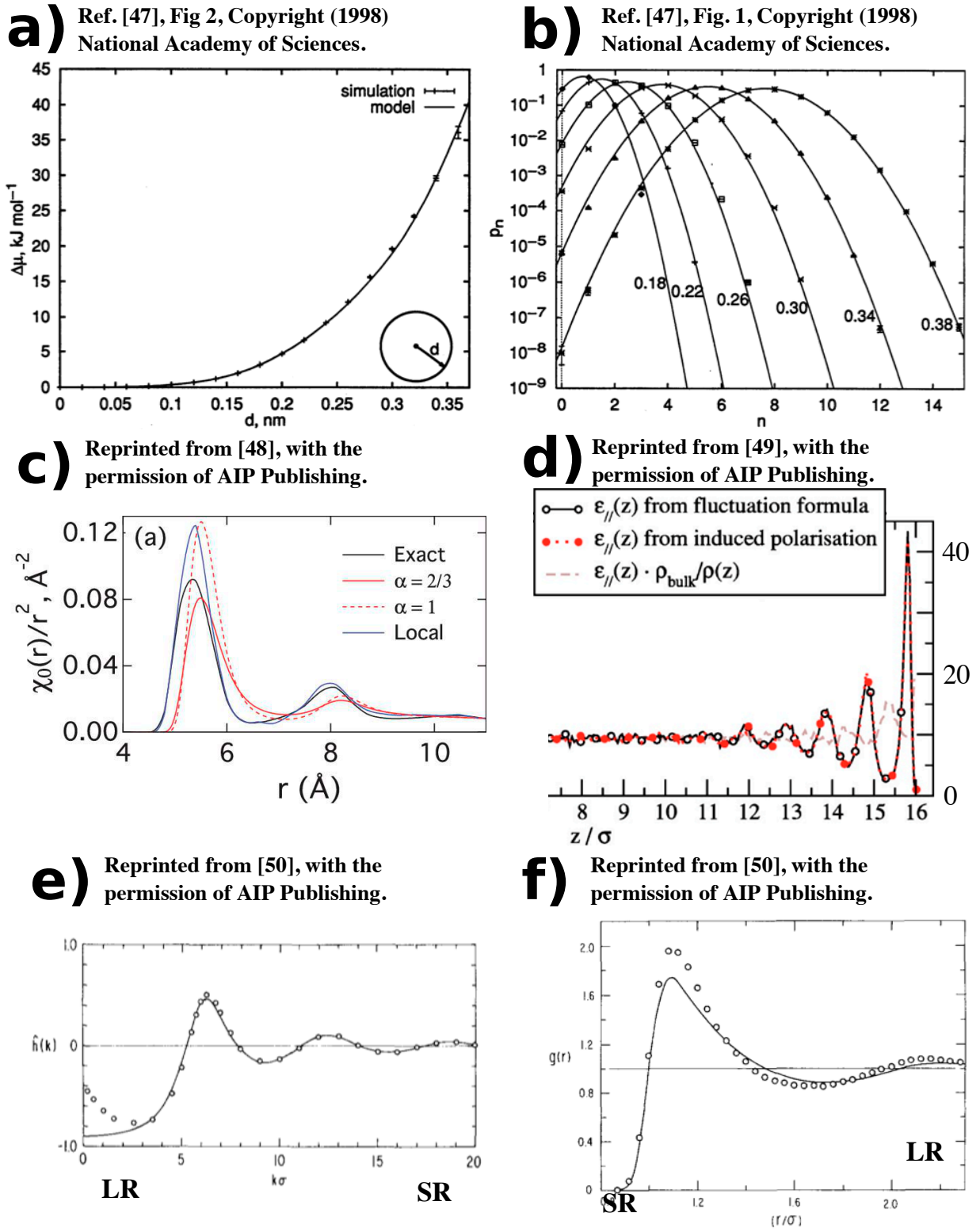



**Figure 3.** Short- and long-range theories of solvent dipole and electrolyte structure. (a) and (b) show free energies and number occupancy distribution for spherical cavities in water.[47] (c) shows the dielectric response in a spherical geometry[48] and (d) shows the dielectric permittivity computed in a slab geometry.[49] (e) and (f) show the correlation function of a supercritical Lennard-Jones fluid near $n = 0.52/\sigma^3$, $T = 1.34\varepsilon/k_B$.[50]

and long-range components of HF and DF and checked their ability to predict the cohesive, formation energies of small molecules. Doing so, they discovered that models with no HF at long range had similar descriptive power to those that used only DF at short range and only HF at long range. Split-range functionals are still an evolving research topic.

## LIQUID-STATE THEORIES

The divide between short and long-range, discrete, and continuous distributions also plays a key role in the development of thermodynamic theories for gasses and liquids. In the 1860s, Boltzmann proposed his transport equation for the motion of gas density over space and time. The model employed the famous stoßzahlansatz, which states that the initial positions of molecules *before* each collision is chosen 'at random.' (Fig. 4a) In the original theory, the probability distribution over such random positions was often confused with their statistical averages[44] – a point which lead to enormous confusion and controversy persisting even until 1960.[45]

This history very nearly parallels the development of electronic density theories. After electromagnetism and gas dynamics had been worked out at the end of the 19th century, Gibbs' treatise on statistical mechanics laid out the classical foundations of the relationship between statistics and dynamics of molecular systems. Nevertheless, there were contemporary arguments with Ehrenfest and others about the need for introducing statistical hypotheses into an exact dynamical theory.[2] Early on, it had been hoped that an exact study of the motion of the molecules themselves could predict the appropriate 'statistical ensemble' by finding long-time limiting distributions. However, that hope was spoiled by the notice that initial conditions must be described statistically. The idea persists even at present, though it has been tempered by the recognition that sustaining nonequilibrium situations requires an infinitely extended environment, which has to be represented in an essentially statistical way.[46]

The resolution, according to Jaynes,[4] is to understand the Boltzmann transport equation as governing the 1-particle probability distribution, $NP(r|C)$, rather than the average amount of mass, $n(r)$, at point $r$. It turns out that this switch in perspective from exact knowledge of all particle positions to probability distributions is one of the key ways of separating short and long-range effects. Two of the oldest and most widely known uses of this method are in the dielectric continuum theory dating from before Maxwell's 1870 treatise, even to Sommerfeld (Fig. 4c), and the Debye model of ionic screening from 1923. For both, a spatial field $E(r - r_0)$, emanating from a discrete molecule at $r_0$, is put to a bulk thermodynamic system whose average properties are well-defined using, for example, $P(r|E)$ for the dipole density $\mu(r)$ at point $r$, due to a field, $E$ or $n(r; \phi)$ for the ion density at point $r$ due to a voltage, $\phi$. Treating $\phi$ and $E$ as weak perturbations and looping $\mu(r)$ (or $n(r)$) back in as additional sources gives a self-consistent equation for the response of a continuum.

As was the case for electronic structure theory, the most concise description of this type of self-consistent loop is provided by a density functional equation for the Helmholtz free energy (with $\beta = 1/k_B T$),

$$\begin{aligned}\beta A[E,\beta] &= \inf_{\mu\in\{\mu\}} \left[-\log(g[\mu]) - \beta E\cdot\mu\right] \\ &\approx -\ln \sum_{\mu\in\{\mu\}} g(\mu) e^{\beta E\cdot\mu}.\end{aligned} \tag{5}$$

The curvature of $A$ with changing applied field, $E$, gives the response function which is related to the conventional dielectric. Consider first a case where μ contains enough information to exactly assign a dipole to every one of $N$ molecules. An example would be a single molecule with twice as many ways to create a small dipole as a large one, $g(4\text{ D}) = 2$ and $g(2\text{ D}) = 4$ (1D = 1 Debye). Then $g(\mu)$ is a product over counting factors. The free energy, $A$, will have jump discontinuities in its slope as the field, $E$ is varied because the solution jumps from one assignment ($\mu = 2$ D) to another ($\mu = 4$ D at $\beta E \geq (\ln 2)/(2\text{ D})$). Its graph is very much like Fig. 1a. In a discrete function space, density functional theory equations yield solutions exhibiting a discrete nature.

On the other hand, if $g(\mu)$ varies continuously with μ in some range of allowed average densities, then the solution will describe a smooth field free energy. Interestingly, starting from the first situation and computing

$$S[\bar{\mu},\beta] = -\beta \sup_E \left[A[E,\beta] + \int E(r)\cdot\bar{\mu}(r)dr\right] \tag{6}$$

leads to such a continuous version of $\log g(\mu) \approx S(\mu)$ (in fact its concave hull). This concave function allows densities that are intermediate between discrete possibilities for the system's state. Such intermediate densities could only be reached physically by averaging, so that $\bar{\mu}$ is an average polarization over possible absolute assignments of dipoles to molecules, μ.

After the theory of self-consistent response to a long-range field had been worked out, further development of liquid-state theory had to wait 40 years for developments in quantum-mechanical interpretation of

light absorption and scattering experiments. Some early history is given in Ref. 51 and Debye's 1936 lecture[52] in which he explains how electronic and dipole orientational polarization could be clearly distinguished from measurements of the dielectric capacitance of gasses along with the great advancements made in the 1920s (which Debye credits to von Lau in 1912) of using x-ray and electron scattering to confirm molecular structures already adduced by chemists from symmetry and chemical formulas alone. Thus, the long-range theory gave a comprehensive enough description of macroscopic electrical and density response that it could be used as a basis to experimentally determine local structure.

With statistical mechanics, quantum mechanics, and molecular structure in hand, liquid-state theories developed in the 1930s-50s through testing hypotheses about the partition function against experimental results for heat capacities. One of the earliest models was the 'free volume' (also known as cell model) theory, developed by Eyring and colleagues and independently by Lennard-Jones and Devonshire in 1937. The theory was put on a statistical mechanical basis by Kirkwood in 1950,[53] as essentially expressing the free energy of a fluid in terms of the free energy of a solid composed of freely moving molecules trapped, one each, in cages exactly the size of the molecular volume, plus the free energy cost for trapping all the molecules in those cages in the first place. It competed[54] with the 'significant structure' theory of liquids (also proffered by Eyring and colleagues[55,56]). In the significant structure theory (Fig. 4f), the partition function for the fluid is described as an average of gas-like and solid-like partition functions to account for the difference in properties between highly ordered and more disordered regions (which contain vacancies).

*Scaled Particle vs Integral Equations*

Also around that time, a competition emerged between the scaled particle theory[57] and the 'integral equation' approach based on (and now lumped together with) Percus and Yevick's[58,59] closure of a theory created by Ornstein and Zernike in 1914 to calculate the effect of correlated density fluctuations on the intensity of light scattered by critically opalescent fluids.[60] This connection was significant, since theories of the correlation function prior to 1958 applied the superposition approximation due to Kirkwood, Yvon, Born, and Green (ca. 1935).[61,62]

The scaled particle theory (SPT) approach takes the viewpoint that the number, sizes and shapes of molecules in a fluid are determined by integrating the work of 'growing' a new solute particle in the middle of a fluid. Its organizing idea is that the chemical potential of a hydrophobic solute is equal to the work of forming a nanobubble in solvent. For simple hard spheres, the work is $PdV$ , where $P = k_B T n_0 G(d)$, $n_0$ is the bulk solvent density, and $G(d)$ (Fig. 4b), the density of solvent molecules on the surface of the solute of diameter $d$. Hence, knowing the contact density for any shape of solute molecule provides complete information on the chemical potentials of those molecules. This very local idea can be related to counting principles at very small sizes,[63] and continued through to macroscopic ideas about surface tension at very large sizes – creating a way to interpolate between the two scales.

On the other hand, the integral equation approach expresses the idea that long-range fluctuations in density are well described by a multivariate Gaussian distribution. If the probability distribution of the density, $n(r)$, was actually Gaussian, its probability would be,[64]

$$P\,[n(r)] = P\,[n_0]\,\exp\left(-{}^{\beta}/_{2}\iint dr dr'(n(r)-n_0)G(r,r')(n(r')-n_0)\right)/Z[\beta G], \quad (7)$$

where $G(r, r') \equiv \text{const} \cdot \delta(r-r') - c(r-r')/\beta$. In the RPA, $-c(r)/\beta$ is energy for placing a pair of molecules at positions $r$ and $r'$. [65]

When they are not Gaussian distributed, the correlations in instantaneous densities, $n(r)$, provide a means of estimating $c$, the direct correlation function.[66] This long-range idea has been used to show that $G$ degenerates to the pairwise energy for very large separations ($G(r) \rightarrow U(r)$ as $r \rightarrow \infty$). For simple hard spheres, it can also be related to counting principles at short separations, since there the correlations must drop to -1, expressing perfect exclusion. Assuming both limits hold right up to the discrete boundary of a solute yields the mean spherical approximation (MSA, Fig. 4b).

These two theories thus express, in pure form, the divide between short-range and long-range viewpoints on molecular structure. Integral equation theories are most correct for describing continuum densities and smooth interactions. Theories that, like SPT, are based on occupancy probabilities of particles in well-defined local structures and geometries are most correct for describing short-range interactions that can contain large energies and discontinuous jumps.

Fig. 3b shows $P(n|d)$, the probability that a randomly chosen sphere of radius $d$ contains exactly $n$ discrete water molecules. Each curve is marked by its value of $d$ in nanometers. The free energy for creating an empty nanobubble of size $d$ in water is shown in its counterpart, Fig. 3a. Both computations are very closely related, and easiest to do from the local picture of scaled particle theory. The cavity formation free energy (Fig. 3a) is, in

principle, also able to be computed from a density functional based on relating the logarithm of Eq. 7 with the entropy.[64] However, when the calculation is done in the usual density functional way the cavity formation free energy is surprisingly difficult to reproduce.[67,68] This difficulty is related to the abrupt decrease in solvent density to zero at the cavity surface. In addition to mathematical difficulties,[69] this complicates creating a physically consistent functional from bulk properties alone. From scaled particle theory, we know the free energy should scale with the logarithm of the volume for small cavities, but later switch over to scale with the surface area. The transition distance is determined by the size of discrete solvent molecules.

*Perturbation Theories*

Slowly but surely during the same time period as integral equation theories were being developed the method of molecular dynamics emerged.[70] Its primary limitations of small, fixed, particle numbers, large numbers of parameters, finite sizes and short timescale simulations weigh heavy on the minds of its practitioners.[71] Early models of water needed several iterations before reproducing densities, vaporization enthalpies and radial distribution functions from experiment. Initial radial distributions from experiment were wrong, and the models had to be corrected and then un-corrected to chase after them.[72] Surprisingly, early calculations took the time and effort to calculate scattering functions and frequency-dependent dielectrics to compare to experiment.[73-75] By contrast, the bulk of 'modern' simulations report only the data that can be readily calculated without building new software.

By checking data from integral equations against molecular dynamics (MD) and scattering experiments it was clear by 1976 that many powerful and predictive methods had been created to describe the theory of liquids.[76,77] Nevertheless, there remained even then lingering questions about the applicability of integral methods to fluids where molecules contained dipole moments, and the treatment of long-range electrostatics in MD. Some difficulties in modeling phase transitions and interfaces were anticipated, but it was hardly expected that bulk molecular dynamics methods themselves would stall and eventually break down when simulating liquid/vapor and liquid/solid surfaces.

This trouble is illustrated by the simulation community's reception of the work leading to Fig. 3b,c. Both show the dielectric response function for water dipoles at the interface with a large spherical particle (left) or vacuum (right). The latter shows a correlation function

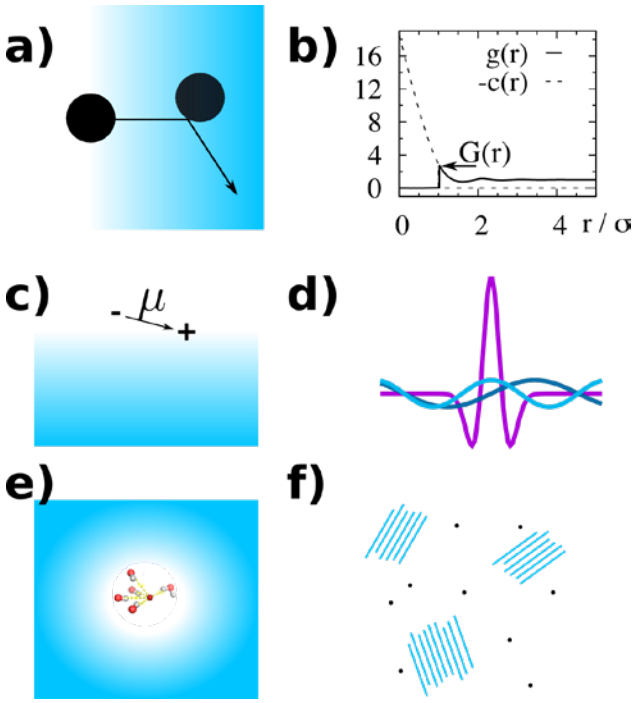



**Figure 4.** Hybrid discrete/continuum theories. (a) Boltzmann picture of scattering by one particle chosen 'at random' from the continuum. (b) Mean spherical approximation for the hard sphere fluid of diameter σ. $g(r)$ and $c(r)$ are known at $r << \sigma$ and $r >> \sigma$, but the central region is a guess. (c) Sommerfield conception of a dipole above a continuous polarizable medium. (d) Bohm-Pines conception of a quasiparticle (purple, central peak) and two long-range plasmons (blue). (e) Dressed ion, quasichemical, or Lorenz- Lorentz-Mossotti-Clausius[51] cavity models of a discrete molecule in a continuum solvent, (f) significant / inherent structure theory of a coexisting mixture of ordered and disordered regions making up an overall homogeneous phase.

computed from all-atom molecular dynamics by Ballenegger.[49] This full computation was preceded two years earlier by less well-cited theoretical work from the same author.[78] As of writing, the citations counts are 140 and 19, respectively. Even after its publication, the technical difficulties caused by simulating collective dipole correlations inside a finite size box cast a cloud over the interpretation that drove Ballenegger back into those fine details for the following nine years.[79,80] On the left (Fig. 3c) is a simulation of water's dipolar response next to a large sphere.[48] The finite-size effects are less severe, and a comparison (not common in contemporary literature) is made to analytical theories that apply to infinite systems. However, those analytical theories work best at long-range, and disagree on the short-range order. The disagreement is jarring because energetic contributions of long and short-range order are on the same order of magnitude.

It was also beginning to be recognized that there were two complementary approaches to the theory of fluid structure. The short-range viewpoint stated that the radial distribution function should be reproduced well at small intermolecular separations (small distance in real-space as in Fig. 3f). This leads to good agreement with interaction energies and pressures so that the virial and energy routes to the equation of state work well.[50] The long-range viewpoint instead emphasizes reproducing the structure factor at small wavevectors (as in Fig. 3e). Because of this, it favors using the compressibility route to the equation of state and leads to good agreement with fluctuation quantities.[81]

*Inherent structures*

Water proved to be a major challenge to molecular models because of its mixture of short-range hydrogen bonding and long-range dipole order.[82] One successful physical picture of water was provided by the Stillinger-Weber 'inherent structure' model introduced in the early 1980s.[83] It represented a cross between the 'significant structure' theory and the free volume theory. In it, molecules are fixed to volumes defined by their energetic basins, rather than by a rigid crystal lattice. Where the free volume theory had only one reference structure, the inherent structure (like the significant structure theory) had many. One for each basin. Each energetic basin looks, on an intermediate scale, like a distortion of one of the crystalline phases of ice. Thermodynamic quantities can be predicted using the energies and entropies associated to each basin – by virtue of the minimum energy structure and the number of thermal configurations mapping to that minimum.

## HYBRID THEORIES IN LIQUID-STATE STRUCTURE

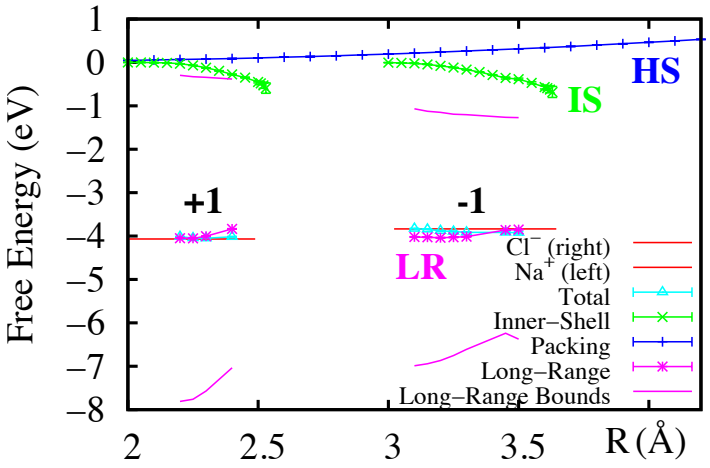
The Lennard-Jones fluid presented a challenge to the integral equation and scaled particle theories above because it contains both short-range repulsion and long-range attraction. At high densities, however, it was found that the radial distribution function was almost identical to the radial distribution for hard spheres (compare Fig. 3e and Fig. 4b). The transition from liquid to solid was also described fairly well using the hard-sphere model. On the other hand, at low densities the distribution function could be described by perturbation from the ideal gas. These two discoveries justify the use of a perturbation theory to calculate the effect of long-range interactions at very low and very high densities.[84] A comparison of molecular dynamics with integral equation plus correction theories is shown in Figs. 3e,f.[50]

At intermediate densities, however, a liquid-to-gas phase transition occurs that can be qualitatively understood, but not explained well as a perturbation from either limit. Instead, the integral equation method turns out to hold the best answer in the supercritical region.[85] It is often encountered in the form of a perturbation theory from the critical point.[86] It is no accident that the integral equation method works well here. Supercritical fluids are characterized by long-range correlations that can take maximum advantage of that theory. For the same reason, integral equations describe the compressibility well, but do poorly on the intermolecular energy.

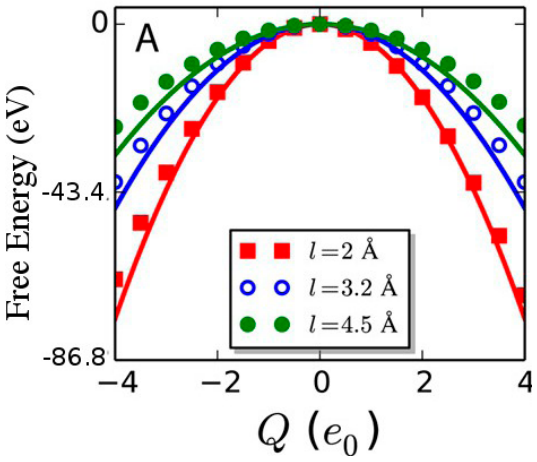
Comparing to developments in electronic structure raises the question of whether perturbation theory could fix the short-range correlations in high and low density fluids. This approach was popularized by Widom's potential distribution theory.[87] Its central idea is to drop a spherical void into a continuum of solvent, and then to drop a solute into its center. This divides the new molecule's chemical potential into a structural part (due to cavity formation) and a long-range part (due to response of solvent to the molecule). Originally, the former were based on a local density approximation from the hard sphere fluid and the latter from a pairwise term that amounted to a van der Waals theory.

Around 1999, this basic idea had been combined with older notions about working with clusters of molecules to create a new 'quasi-chemical' theory.[88] It refined the simple process of creating an empty sphere devoid of solvent into that of creating a locally well-defined cluster of solvent molecules. The free energy required for this process is still local and structural, but now the entire cluster of solute plus solvent can be regarded as one, local, chemical entity. In order to work with molecules that have 'loose' solvent clusters, a third step was also added. After pulling solvent molecules into a local structure and adding the long-range interactions between solute and solvent, the third step releases the solvent cluster, liberating any energy that might have been trapped by freezing them.[89]

The opposite of this short-range-first approach could be an inverse perturbation theory – first deciding on the long-range shape of correlation functions and second correcting them for packing interactions at short-range. This kind of correction would look like an adjustment to the solution of the Poisson-Boltzmann equation. Such an approach may first have been presented in Refs. 90;91, and followed with interesting modifications of the Debye theory.[92-94] Even more recently, the basic idea was rigorously applied to molecular simulation models by Remsing and Weeks. Their scheme eliminates a hard discon-




(a) Ion solvation free energy components for the short-range (empty cavity first) model computed from an MD model of NaCl in SPC/E water. $R$ is the cavity radius, 'HS' denotes the cavity formation cost, 'LR' is the full ion-SPC/E water interaction after a cavity is present, and 'IS' is the free energy of removing the cavity constraint.




(b) Interaction free energy of SPC/E water with a Gaussian charge distribution, $Q\exp(-r^2/l^2)/(l\sqrt{\pi})^3$. Points correspond to simulation data, while lines assume a constant dielectric model. Adapted with permission from Ref. 98, Copyright 2016 American Chemical Society.

**Figure 5.** Comparing components of the SR-first (left) and LR-first (right) calculations of the free energy gained on dissolving a charged ionic species in water.

tinuity between short and long-range in the first step by splitting the Coulomb pair potential into smooth, long-range and sharp, short-range parts. The long-range forces (from the smooth part of the potential) are used to compute a 'starting' density using RPA-like perturbation from a uniform fluid. Although it seems a lot like the molecular density functional method,[62,95,96] the density after the first step remains smooth at the origin, lacking any hard edges. It has previously been considered under the title 'ultrasoft restricted primitive model'.[97] Remsing and Weeks added a final step to this model to create a cavity at the origin and compared the results to MD simulations.

Detailed molecular simulations have been used to compare the two approaches with exact simulations by brute force calculation of all the energetic contributions. Focusing on the short-range structure leads to a model whose first step is to form an empty cavity in solution (blue curve in Fig. 5a, labeled 'Packing'). Fig. 5a shows the free energies of the next step ($Na^+$ and $Cl^-$ ions) divided into 'long-range' and 'inner-shell' parts of the re-structuring.[99] All points come from MD. If, instead, the long-range interaction between an ion and solvent occurs first, we are lead to couple the solvent to the smooth electric field of a Gaussian charge distribution. Fig. 5b shows the free energy of that first step as a function of charge for a variety of Gaussian (smoothing) widths. The lines show continuum predictions, and the points show MD.

Integral equation approaches to the dipolar solvation process have also continued independently. Matyushov developed a model for predicting the barrier to charge transfer reactions.[100] In that work, the dipole density response to the electric field of a dipole is worked out in linear approximation. A sharp cutoff is used to set the field to zero inside the solute, resulting in a hybrid short/long range theory. The approach succeeds because the linear response approximation (stating density changes are proportional to applied field) is correct at long range, where the largest contributions to the solvation energy of a dipole originate. Other authors have expanded on numerical and practical aspects of correlation functions.[101-103]

The theme of separating long-range, continuous *vs.* short-range, discrete interactions runs throughout numerous other molecular-scale models. Models in this category include the 'dressed' ion theory, which posits that ions in solution always go in clad with strongly bound, first shell, water molecules so that their radius is larger than would be suggested from a perfect crystal (Fig. 4e). These enlarged radii appear in the Stokes-Einstein equation to describe the effect of molecular shape on continuous water velocity fields when computing the diffusion coefficients for ions.[104] They should also appear to describe how excluded volume of ions will affect the continuous charge distribution predicted by the primitive model of electrolytes. This modification is not common, and so would yield some nonstandard

plots of hydration free energy as a function of ion concentration.[105] Solvent orientational order changes form again beyond about 1.5 micrometers due to the finite speed of light.[106] The Marcus theory of electron transport describes two separate, localized structural states of a charged molecule that interact with a continuously movable, long-range, Gaussian, field. Larger magnitude fluctuations in the solvent structure lead to broader Gaussians, which in turn are the cause of more frequent arrival at favorable conditions for the electron to jump. It is common practice in quantum calculations to explicitly model all atoms and electrons of a central molecule quantum-mechanically while representing the entirety of the solvent with a continuous dielectric field.[107-109]

The theories above are not perfect. They show issues precisely at the point where short- and long-range forces are crossing over. At high ionic concentrations, the dressed ion theory breaks down due to competition between ion-water and ion-ion pairing. When solvent molecules are strongly bound, the use of a continuous density field cannot fully capture their influence on thermodynamic properties. Even without strongly bound solvent, dielectric solvation models leave open the important question of whether electrons from the fully modeled molecule are more or less likely to 'spill out' into the surrounding solvent. Returning back to Aristotle's objection to discrete objects, it is known that density based models don't accurately capture the free energy of forming a empty cavity.[67,68] Thousands of years on, we are still vexed by the question of how to understand the interface between material objects and vacuums.

## THE FUTURE: A MIDDLE WAY

Early Eastern thought tends to place opposing ideas next to one another in an attempt to understand them as parts of a whole picture. Written around the beginning of the Middle ages, in 400 AD, the Lankavatara Sutra relates Buddha's view that this unity applies to atoms and 'the elements' (which refer to something like the classical Greek elements). Taking liberties, we can say he is discussing a process like instantaneous disappearance (annihilation) of a quantum particle in saying, "even when closely examined until atoms are reached, it is [only the destruction of] external forms whereby the elements assume different appearances as short or long; but, in fact, nothing is destroyed in the elemental atoms. What is seen as ceased to exist is the external formation of the elements." Bohr was well-known for his view on the 'complementarity' principle, stating in this context that the act of removing a particle makes its number more definite, while making the amount of energy it exchanged with an external observer undefined.[110] Perhaps inspiring to Bohr sixteen centuries later,[111] the quote concludes, "I am neither for permanency nor for impermanency … there is no rising of the elements, nor their disappearance, nor their continuation, nor their differentiation; there are no such things as the elements primary and secondary; because of discrimination there evolve the dualistic indications of perceived and perceiving; when it is recognised that because of discrimination there is a duality, the discussion concerning the existence and non-existence of the external world ceases because Mind-only is understood." Bohr's complementarity could be contrasted with physicist John Wheeler. He advocated, as a working hypothesis, that participants elicit yes/no answers from the universe. Replies come as discrete 'bits,' and are ultimately the reason that discrete structures emerge whenever continuum models try to become precise.[112] Wheeler, in turn, could be contrasted with Hugh Everett, whose working hypothesis was that the universe operates by pure wave mechanics.[113,114] A modern resolution of those debates invokes small random, gravitational forces to explain how quantum particles could become tied to definite locations.[115] It is does not appear that there will be a resolution allowing us to do away with either continuum or discrete notions.

Of course, it is impossible to deduce scientific principles if we include any elements of mysticism in a theory. Nevertheless, the debate on the separation between short and long-range seems to permeate history. This idea that a meaningful understanding of collective phenomena should be sought by combining physical models appropriate to atomic and macroscopic length scales was taken up even recently by Laughlin, Pines, and co-workers.[36] They state, "The search for the existence and universality of such rules, the proof or disproof of organizing principles appropriate to the mesoscopic domain, is called the middle way."

On one account it is clearly possible to set the record straight. There are well-known ways of converting local structural theories into macroscopic predictions and as vice-versa. Bayes' theorem states that, for three pieces of information, *A*, *B*, and *C*,

$$P\,(A|BC) = \frac{P\,(B|AC)P\,(A|C)}{P\,(B|C)}. \tag{8}$$

If 'C' represents a set of fixed conditions for an experiment, 'B' represents the outcome of a measurement, and 'A' represents a detailed description of the underlying physical mechanism (for example complete atomic coordinates), then Bayes' theorem explains how

to assign a probability to atomic coordinates for any given measurement, 'B'. Of course, in a reproducible experiment, $C$ will completely determine $B$, so $B = B(C)$. Thus, the probability distribution over the coordinates is a function only of the experimental conditions, $P(A|BC) = P(A|C)$. This summarizes the process of assigning a local structural theory from exactly reproducible experiments.

On the other hand, a local structural theory provides an obvious method for macroscopic prediction. Given a complete description, 'A,' simply follow the laws of motion when interacting with a macroscopic measuring device, 'B.' This would properly be expressed in the language above as $P(B|AC) = P(B|A)$, since the experimental conditions are irrelevant. Bayes' theorem then gives us a conundrum, $P(B|C) = P(B|A)$, stating that every microscopic realization of an experiment must yield an identical macroscopic outcome.

The solution to the puzzle is to realize that unless an experiment is exactly reproducible, $BC$ is always more informative than the conditions, $C$, alone and $P(A|BC) \neq P(A|C)$. This explains why studying exactly integrable dynamical systems is such a thorny issue, and is the central conceptual hurdle passed when transitioning from classical to quantum mechanics. Now identifying 'B' with a partial measurement that provides a coarse scale observation of some long-range properties, $P(A|BC)$ describes a distribution over the short-range, atomistic, and discrete degrees of freedom. Because of experimental uncertainty, the exact location of those atoms is evidently subjective and unknowable (since it is based on measurement of $B$). Nevertheless, it can in many cases be known to a high degree of accuracy.

Density functional theory traditionally focuses on $P(B|C)$, where 'B' is the average density of particles in a fluid and 'C' is the experiment where a bulk material is perturbed by placing an atom at the origin. However, with a minor shift in focus, $P(B\,A'C)$ can also be found, representing the average density under conditions where a particle is placed at the origin and some atomic information, $A'$ is also known. The objective of such a density functional theory would be to more accurately know the long-range structure by including some explicit information on the short-range structure. The dual problem is to predict $P(A|B'C)$, the distribution over coordinates when we are provided with some known information on the long-range structure. In a complete generalization, we might focus instead on $P(AB|A'B'C)$, representing the average density and particle distribution under conditions where density and particle positions are known only in part. Bayes' theorem shows us that such a generalization would just be the result of weaving the primal and dual problems together, since (given the redundancies, $B' = B'(B)$ and $A' = A'(A)$), $P(A|A'B'C) = P(A|B'C)/P(A'|B'C)$, and $P(B|A'B'C) = P(B|A'C)/P(B'|A'C)$.

The arguments above can be repeated for each of the elements in Table 1 – replacing SR with $A$ and LR with $B$. What emerges is a persistent pattern of logical controversy, where a problem can be apparently solved entirely from either perspective. In some areas, one or the other approach is more expedient. In every case, however, recognizing and using both sides has proved to be profitable. Comparing these two perspectives, we find that the discussion concerning the existence of long and short-range theories ceases, leaving only different ways to phrase probability distributions.

We have now arrived at a point in the history of molecular science where these two great foundations, short-range, discrete structures and long-range, continuum fields are at odds with one another. Molecular dynamical models are fundamentally limited by the world view that all forces must be computed from discrete particle locations. Computational methods treating continuum situations focus their attention on solving partial differential equations for situation-specific boundary conditions. Connecting the two, or even referring back to simple analytical models, requires time and effort that is seen as scientifically unproductive. What's worse, it reminds us that many, lucidly detailed, broad-ranging, and general answers were already presented in the lengthy manuscripts which set forth those older, unfashionable models.

Indeed, local and continuum theories are hardly on speaking terms. In molecular dynamics, the mathematics of the Ewald method for using a Fourier-space sum to compute long-range interactions are widely considered esoteric numerical details. Much effort has been wasted debating different schemes for avoiding it by truncating and neglecting the long-range terms.[116-118] On the positive side, the central issue of simulating charged particles in an infinite hall of mirrors has been addressed by a few works.[119-121] Much greater effort has been devoted to adding increasingly detailed parameters, such as polarizability and advanced functional forms for conformation and dispersion energies, to those atomic models. Apparently, automating the parameterization process [122] is unfundable. In the case of polarization and dispersion, the goal of these atomic parameters is, somewhat paradoxically, to more accurately model the long-range interactions. The problem of coupling molecular simulations to stochastic radiation fields has, apparently, never been considered as such. Instead, we can find comparisons of numerical time integration methods intended to enforce constant temperature on

computed correlation functions.[123] In continuum models based on partial differential equations, actual molecular information that should go into determining boundary conditions, like surface charge and slip length (or, more accurately, boundary friction[124]), are replaced by 'fitting parameters' that are, quite often, never compared with atomic models. Indeed, studies in the literature that even contain a model detailed enough to connect the two scales are few and far between.

We are also at a loss for combining models of different scales with one another. Of the many proposed methods for coupling quantum mechanical wavefunction calculations to continuous solvent, essentially all of them neglect explicit first-shell water structure that could be experimentally measured with neutron scattering, diffusion measurements, and IR and Raman spectroscopy. Jumping directly into applications is a disease infecting much of contemporary science. Rather than attempting to faithfully reproduce the underlying physics, many models are compared by directly checking against experimentally measured energies – and no clear winner has emerged (nor can it). To be correct, models must be checked for consistency with experiments at neighboring length scales. Similar remarks can be made for implicit solvent models coupling molecular mechanics to continuum. Even Marcus theory is not untouched. There is currently debate on the proper way to conceptualize its parameter that sets the 'stiffness' of the solvent linear response.[125]

In order to make progress, we must apparently work as if we had one hand tied behind our back. Used correctly, simulations provide a precise tool to answer a well-posed question within a known theory, or as a method of experimentation to discover ideas. However, when used absent a general theory, simply as a tool to reproduce or predict a benchmark set of experimental data, simulation is not capable of providing any detailed insight or understanding of molecular science.

## ACKNOWLEDGEMENTS

I thank the anonymous reviewers for their comments and suggestions.

## REFERENCES

1. George F. Bertsch and James Trefil et. al. Atom. In *Encyclopaedia Britannica*. Encyclopaedia Britannica, Inc., 2018.
2. Paul Ehrenfest and Tatiana Ehrenfest. *The Conceptual Foundations of the Statistical Approach in Mechanics*. Cornell Univ. Press, Ithaca, NY, 1959. Translation of Begriffliche Grundlagen der statistichen Auffassung in der Mechanik, 1912, by Michael J. Moravcsik.
3. Barry W. Ninham. The biological/physical sciences divide, and the age of unreason. *Substantia*, 1(1): 7–24, 2017. doi: 10.13128/Substantia-6.
4. E. T. Jaynes. Where do we stand on maximum entropy? In R. D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, page 498. M.I.T Press, Cambridge, 1979. ISBN 0262120801,9780262120807.
5. David Ruelle. Is there a unified theory of nonequilibrium statistical mechanics? In *Int. Conf. Theor. Phys.*, volume 4 of *Ann. Henri Poincaré*, pages S489–95. Birkhäuser Verlag, Basel, 2003.
6. G. Gallavotti. Heat and fluctuations from order to chaos. *Eur. Phys. J. B*, 61:1–24, 2008. doi: 10.1140/epjb/e2008-00041-1.
7. S. R. S. Varadhan. Large deviations. *Ann. Prob.*, 36(2):397–419, 2008. doi: 10.1214/07-AOP348.
8. Hugo Touchette. The large deviation approach to statistical mechanics. *Phys. Rep.*, 478(1–3):1–69, 2009. doi: 10.1016/j.physrep.2009.05.002.
9. S. Toulmin. The evolutionary development of natural science. *American Scientist*, 55(4):456–471, 1967. URL http://www.jstor.org/stable/27837039.
10. E. Jaynes. Probability in quantum theory. In W. H. Zurek, editor, *Complexity, Entropy, and the Physics of Information*. AddisonWesley, Reading MA, 1990.
11. Freeman J. Dyson. Missed opportunities. *Bull. Amer. Math. Soc.*, 78(5):635–652, 1972.
12. C. M. De Witt. Feynman's path integral: definition without limiting procedure. *Commun. Math. Phys.*, 28:47–67, 1972.
13. Maurice M. Mizrahi. Phase space path integrals, without limiting procedure. *J. Math. Phys.*, 19:298, 1978. doi: 10.1063/1.523504.
14. Hagen Kleinert. *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets*. World Scientific, Singapore, 2009. ISBN 981-270-008-0. 5th edition.
15. H. Eschrig. *The Fundamentals of Density Functional Theory*. B. G. Teubner Verlagsgesellschaft, Leipzig, 1996.
16. Carl Eckart. The penetration of a potential barrier by electrons. *Phys. Rev.*, 35:1303–1309, Jun 1930. doi: 10.1103/PhysRev.35.1303. URL https://link.aps.org/doi/10.1103/PhysRev.35.1303.
17. Stuart A. Rice, Daniel Guidotti, Howard L. Lemberg, William C. Murphy, and Aaron N. Bloch. Some

comments on the electronic properties of liquid metal surfaces. In *Aspects of The Study of Surfaces*, volume 27 of *Advances in Chemical Physics*, pages 543–634. John Wiley & Sons, New York, 1974.

18. Bernhard Sellner and Shawn M. Kathmann. A matter of quantum voltages. *J. Chem. Phys.*, 141(18): 18C534, 2014. doi: 10.1063/1.4898797.
19. G. Ortiz and P. Ballone. Correlation energy, structure factor, radial distribution function, and momentum distribution of the spin-polarized uniform electron gas. *Phys. Rev. B*, 50:1391–1405, Jul 1994. doi: 10.1103/PhysRevB.50.1391. URL https://link.aps.org/doi/10.1103/PhysRevB.50.1391.
20. N. D. Lang and W. Kohn. Theory of metal surfaces: Charge density and surface energy. *Phys. Rev. B*, 1:4555–4568, Jun 1970. doi: 10.1103/PhysRevB.1.4555. URL https://link.aps.org/doi/10.1103/PhysRevB.1.4555.
21. P. Ewald and H. Juretschke. Atomic theory of surface energy. In R. Gomer and C. Smith, editors, *Structure and Properties of Solid Surfaces: A Conference Arranged by the National Research Council*, page 117. U. Chicago Press, 1952.
22. David Pines. Electrons and plasmons. In *Elementary Excitations in Solids*, pages 56–167. CRC Press, Boca Raton, FL, 1999. ISBN 978-0-7382-0115-3.
23. J. C. Slater and H. M. Krutter. The Thomas-Fermi method for metals. *Phys. Rev.*, 47:559–568, 1935.
24. P. Nozières and D. Pines. Correlation energy of a free electron gas. *Phys. Rev.*, 111(2):442–454, 1958.
25. E. Wigner and F. Seitz. On the constitution of metallic sodium. II. *Phys. Rev.*, 46:509–524, 1934.
26. D. M. Ceperley and B. J. Alder. Ground state of the electron gas by a stochastic method. *Phys. Rev. Lett.*, 45:566–569, Aug 1980. doi: 10.1103/PhysRevLett.45.566. URL https://link.aps.org/doi/ 10.1103/PhysRevLett.45.566.
27. N. F. Mott. The basis of the electron theory of metals, with special reference to the transition metals. *Proceedings of the Physical Society. Section A*, 62(7):416, 1949. URL http://stacks.iop.org/0370-1298/62/i=7/a=303.
28. E. P. Wigner. Effects of the electron interaction on the energy levels of electrons in metals. *Trans. Faraday Soc.*, 34:678–685, 1938. doi: 10.1039/TF9383400678.
29. Peter M. W. Gill and Ross D. Adamson. A family of attenuated coulomb operators. *Chem. Phys. Lett.*, 261(1):105–110, 1996. ISSN 0009-2614. doi: 10.1016/0009-2614(96)00931-1. URL http://www.sciencedirect.com/science/article/pii/0009261496009311.
30. Marielle Soniat, David M. Rogers, and Susan Rempe. Dispersion- and exchange-corrected density functional theory for sodium ion hydration. *J. Chem. Theory. Comput.*, 142:074101, 2015.
31. Pablo G. Debenedetti. The statistical mechanical theory of concentration fluctuations in mixtures. *J. Chem. Phys.*, 87(2):1256–1260, 1987.
32. David Bohm and David Pines. Screening of electronic interactions in a metal. *Phys. Rev.*, 80:903–904, Dec 1950. doi: 10.1103/PhysRev.80.903.2. URL https://link.aps.org/doi/10.1103/PhysRev.80. 903.2.
33. David Bohm and David Pines. A collective description of electron interactions. I-III. *Phys. Rev.*, 82: 625, 1951. **85** (1952), 338; **92** (1953), 609.
34. R. I. G. Hughes. Theoretical practice: the Bohm-Pines quartet. *Perspectives on Science*, 14:457–524, 2006.
35. E. González-Tovar, M. Lozada-Cassou, L. Mier y Terán, and M. Medina-Noyola. Thermodynamics and structure of the primitive model near its gasliquid transition. *J. Chem. Phys.*, 95:6784, 1991. doi: 10.1063/1.461516.
36. R. B. Laughlin, David Pines, Joerg Schmalian, Branko P. Stojković, and Peter Wolynes. The middle way. *Proc. Nat. Acad. Sci. USA*, 97(1):32–37, 2000. ISSN 0027-8424. doi: 10.1073/pnas.97.1.32. URL http://www.pnas.org/content/97/1/32.
37. P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964. doi: 10.1103/PhysRev.136.B864. URL https://link.aps.org/doi/10.1103/PhysRev.136.B864.
38. W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965. doi: 10.1103/PhysRev.140.A1133. URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133.
39. P. C. Hohenberg, Walter Kohn, and L. J. Sham. The beginnings and some thoughts on the future. In Samuel B. Trickey, editor, *Advances in Quantum Chemistry*, volume 21, pages 7–26. Academic Press, San Diego California, 1990.
40. von Helmut Eschrig. Legendre transformation. In *The Fundamentals of Density Functional Theory*, volume 32 of *Teubner-Texte zur Physik*, pages 99–126. B. G. Teubner Verlagsgesellschaft, Leipzig, 1996.
41. David C. Langreth and John P. Perdew. Exchange-correlation energy of a metallic surface: Wavevector analysis. *Phys. Rev. B*, 15:2884–2901, Mar 1977. doi: 10.1103/PhysRevB.15.2884. URL https://link.aps.org/doi/10.1103/PhysRevB.15.2884.
42. Julien Toulouse, Francois Colonna, and Andreas Savin. Long-range–short-range separation of the

electron-electron interaction in density-functional theory. *Phys. Rev. A*, 70:062505, 2004.
43. Oleg A. Vydrov, Jochen Heyd, Aliaksandr V. Krukau, and Gustavo E. Scuseria. Importance of short-range versus long-range Hartree-Fock exchange for the performance of hybrid density functionals. *J. Chem. Phys.*, 125(7):074106, 2006. doi: 10.1063/1.2244560. URL http://link.aip.org/link/?JCP/125/074106/1.
44. P. and T. Ehrenfest. Begriffliche grundlagen der statistischen auffassung in der mechanik. *Encykl. Math. Wiss.*, (IV 2, II, Heft 6):90 S, 1912. Reprinted in 'Paul Ehrenfest, Collected Scientific Papers' (M. J. Klein, ed.), North-Holland, Amsterdam, 1959. (English translation by M. J. Moravcsik, Cornell Univ. Press, Ithaca, New York).
45. E. Jaynes. Gibbs vs Boltzmann entropies. *American J. Phys.*, 33:391, 1965. doi: 10.1119/1.1971557.
46. Giovanni Gallavotti. Ergodicity: a historical perspective. equilibrium and nonequilibrium. *Eur. Phys. J. H*, 41(3):181–259, 2016.
47. G Hummer, S Garde, A E Garca, A Pohorille, and L R Pratt. An information theory model of hydrophobic interactions. *Proc. Nat. Acad. Sci. USA*, 93(17):8951–8955, 1996. URL http://www. pnas.org/content/93/17/8951.abstract.
48. Mohammadhasan Dinpajooh and Dmitry V. Matyushov. Free energy of ion hydration: Interface susceptibility and scaling with the ion size. *J. Chem. Phys.*, 143(4):044511, 2015. doi: 10.1063/1.4927570.
49. V. Ballenegger and J.-P. Hansen. Dielectric permittivity profiles of confined polar fluids. *J. Chem. Phys.*, 122:114711, 2005. doi: 10.1063/1.1845431.
50. J. D. Weeks, D. Chandler, and J. C. Andersen. Role of repulsive forces in determining the equilibrium structure of simple liquids. *J. Phys. Chem.*, 54:5237–5247, 1971.
51. H. Kragh. The Lorenz-Lorentz formula: Origin and early history. *Substantia*, 2(2):7–18, 2018. doi: 10.13128/substantia-56.
52. Peter Debye. Methods to determine the electrical and geometrical structure of molecules. In *Nobel Lectures in Chemistry*, Dec 1936.
53. John G. Kirkwood. Critique of the free volume theory of the liquid state. *J. Chem. Phys.*, 18(3), 1950.
54. Donald A. McQuarrie. Theory of fused salts. *J. Phys. Chem.*, 66(8):1508–13, 1962. doi: 10.1021/j100814a030.
55. H. Eyring, T. Ree, and N. Hirai. Significant structures in the liquid state. *Proc. Nat. Acad. Sci. USA*, 44(7):683–91, 1958.
56. Henry Eyring and R. P. Marchi. Significant structure theory of liquids. *J. Chem. Educ.*, 40(11):562, 1963. doi: 10.1021/ed040p562.
57. H. Reiss, H. L. Frisch, and J. L. Lebowitz. Statistical mechanics of rigid spheres. *J. Chem. Phys*, 31: 369, 1959. doi: 10.1063/1.1730361.
58. Jerome K. Percus and George J. Yevick. Analysis of classical statistical mechanics by means of collective coordinates. *Phys. Rev.*, 110(1):1–13, 1958. doi: 10.1103/PhysRev.110.1. URL https://link.aps.org/doi/10.1103/PhysRev.110.1.
59. J. K. Percus. Approximation methods in classical statistical mechanics. *Phys. Rev. Lett.*, 8(11):462–3, 1962.
60. L. S. Ornstein and F. Zernike. Accidental deviations of density and opalescence at the critical point of a single substance. *Proc. R. Neth. Acad. Arts Sci.*, 17:793–806, 1914. URL http://www.dwc.knaw.nl/DL/publications/PU00012643.pdf.
61. H. Ted Davis. *Statistical Mechanics of Phases*. VCH Publishers, New York, 1996.
62. Arieh Ben-Naim. *Molecular Theory of Solutions*. Oxford Univ. Press, Oxford, 2006. ISBN 0199299692.
63. Henry S. Ashbaugh and Lawrence R. Pratt. Colloquium: Scaled particle theory and the length scales of hydrophobicity. *Rev. Mod. Phys.*, 78:159–178, Jan 2006. doi: 10.1103/RevModPhys.78.159. URL https://link.aps.org/doi/10.1103/RevModPhys.78.159.
64. David J. E. Calloway. Surface tension, hydrophobicity, and black holes: The entropic connection. *Phys. Rev. E*, 53(4):3738–3744, 1996.
65. Derek Frydel and Manman Ma. Density functional formulation of the random-phase approximation for inhomogeneous fluids: Application to the gaussian core and coulomb particles. *Phys. Rev. E*, 93: 062112, Jun 2016. doi: 10.1103/PhysRevE.93.062112. URL https://link.aps.org/doi/10.1103/ PhysRevE.93.062112.
66. Kenneth E. Newman. Kirkwood–Buff solution theory: derivation and applications. *Chem. Soc. Rev.*, 23:31–40, 1994. doi: 10.1039/CS9942300031.
67. Guillaume Jeanmairet, Maximilien Levesque, and Daniel Borgis. Molecular density functional theory of water describing hydrophobicity at short and long length scales. *J. Chem. Phys.*, 139(15):154101, 2013. doi: 10.1063/1.4824737.
68. Guillaume Jeanmairet, Maximilien Levesque, Volodymyr Sergiievskyi, and Daniel Borgis. Molecular density functional theory for water with liquid-gas coexistence and correct pressure. *J. Chem. Phys.*, 142(15):154112, 2015. doi: 10.1063/1.4917485.

69. J. T. Chayes, L. Chayes I, and Elliott H. Lieb. The inverse problem in classical statistical mechanics. *Commun. Math. Phys.*, 93:57–121, 1984.
70. Loup Verlet. Computer "Experiments" on classical fluids. I. thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, 159(1):98–103, 1967.
71. J. Karl Johnson, John A. Zollweg, and Keith E. Gubbins. The Lennard-Jones equation of state revisited. *Molecular Physics*, 78(3):591–618, 1993. doi: 10.1080/00268979300100411.
72. A. K. Soper. The radial distribution functions of water as derived from radiation total scattering experiments: Is there anything we can say for sure. *ISRN Physical Chemistry*, 2013:279463, 2013. doi: 10.1155/2013/279463.
73. A. Rahman and F. H. Stillinger. Molecular dynamics study of liquid water. *J. Chem. Phys.*, 55: 3336–3359, 1971. doi: 10.1063/1.1676585.
74. F. H. Stillinger. Low frequency dielectric properties of liquid and solid water. In E. W. Montroll and J. L. Lebowitz, editors, *The Liquid State of Matter: Fluids Simple and Complex*, pages 341–431. North-Holland, New York, 1982.
75. H. L. Friedman and C. V. Krishnan. Thermodynamics of ion hydration. In F. Franks, editor, *Water: A Comprehensive Treatise*. Plenum Press, New York, 1973.
76. J. A. Barker and D. Henderson. What is "liquid"? understanding the states of matter. *Rev. Mod. Phys.*, 48(4):587–671, 1976.
77. H. B. Singh and A. Holz. Structure factor of liquid alkali metals. *Phys. Rev. A*, 28:1108–1113, Aug 1983. doi: 10.1103/PhysRevA.28.1108. URL https://link.aps.org/doi/10.1103/PhysRevA.28.1108.
78. V. Ballenegger and J.-P. Hansen. Local dielectric permittivity near an interface. *Europhys. Lett.*, 63: 381–387, 2003.
79. V. Ballenegger, A. Arnold, and J. J. Cerdá. Simulations of non-neutral slab systems with long-range electrostatic interactions in two-dimensional periodic boundary conditions. *J. Chem. Phys.*, 131:094107, 2009. doi: 10.1063/1.3216473.
80. V. Ballenegger. Communication: On the origin of the surface term in the Ewald formula. *J. Chem. Phys.*, 140:161102, 2014. doi: 10.1063/1.4872019.
81. R. L. Perry, J. D. Massie, and P. T. Cummings. An analytic model for aqueous electrolyte solutions based on fluctuation solution theory. *Fluid Phase Equil.*, 39:227–266, 1988.
82. Jhon Mu Shik and Henry Eyring. Liquid theory and the structure of water. *Ann. Rev. Phys. Chem.*, 27:45–57, 1976.
83. F. H. Stillinger and T. A. Weber. Inherent structure in water. *J. Phys. Chem.*, 87:2833–40, 1983.
84. L. Verlet and J. Weis. Perturbation theory for the thermodynamic properties of simple liquids. *Mol. Phys.*, 24(5):1013–1024, 1972.
85. C. Caccamo. Integral equation theory description of phase equilibria in classical fluids. *Physics Reports*, 274:1–105, 1996.
86. L. Reatto. Phase separation and critical phenomena in simple fluids and in binary mixtures. In Carlo Caccamo, Jean-Pierre Hansen, and George Stell, editors, *New Approaches to Problems in Liquid State Theory*, volume 529 of *NATO Science Series C: Math. and Phys. Sci.*, pages 31–46. Kluwer Academic, 1998. ISBN 978-0-7923-5671-4. doi: 10.1007/978-94-011-4564-0.
87. B. Widom. Potential-distribution theory and the statistical mechanics of fluids. *J. Phys. Chem.*, 86: 869–872, 1982.
88. L. R. Pratt and S. B. Rempe. Quasi-chemical theory and implicit solvent models for simulations. In L. R. Pratt and G. Hummer, editors, *Simulation and theory of electrostatic interactions in solution*, pages 172–201. ALP, New York, 1999.
89. David M. Rogers and Susan B. Rempe. Probing the thermodynamics of competitive ion binding using minimum energy structures. *J. Phys. Chem. B*, 115(29):9116–9129, 2011.
90. Benjamin P. Lee and Michael E. Fisher. Density fluctuations in an electrolyte from generalized Debye-Hückel theory. *Phys. Rev. Lett.*, 76(16):2906–2909, 1996.
91. Itamar Borukhov, David Andelman, and Henri Orland. Steric effects in electrolytes: A modified Poisson-Boltzmann equation. *Phys. Rev. Lett.*, 79:435–438, Jul 1997. doi: 10.1103/PhysRevLett.79.435. URL https://link.aps.org/doi/10.1103/PhysRevLett.79.435.
92. Phil Attard. Asymptotic analysis of primitive model electrolytes and the electrical double layer. *Phys. Rev. E*, 48:3604–3621, Nov 1993. doi: 10.1103/PhysRevE.48.3604. URL https://link.aps.org/doi/10.1103/PhysRevE.48.3604.
93. Roland Kjellander. Decay behavior of screened electrostatic surface forces in ionic liquids: the vital role of non-local electrostatics. *Phys. Chem. Chem. Phys.*, 18:18985–19000, 2016. doi: 10.1039/C6CP02418A.
94. Roland Kjellander. Focus article: Oscillatory and long-range monotonic exponential decays of electrostatic interactions in ionic liquids and other electrolytes: The significance of dielectric permit-

tivity and renormalized charges. *J. Chem. Phys.*, 148(19):193701, 2018. doi: 10.1063/1.5010024.

95. T. J. Sluckin. Density functional theory for simple molecular fluids. *Mol. Phys.*, 43(4):817–849, 1981. doi: 10.1080/00268978100101711.
96. Shuangliang Zhao, Rosa Ramirez, Rodolphe Vuilleumier, and Daniel Borgis. Molecular density functional theory of solvation: From polar solvents to water. *J. Chem. Phys.*, 134(19):194102, 2011. doi: 10.1063/1.3589142.
97. Arash Nikoubashman, Jean-Pierre Hansen, and Gerhard Kahl. Mean-field theory of the phase diagram of ultrasoft, oppositely charged polyions in solution. *J. Chem. Phys.*, 137:094905, 2012. doi: 10.1063/ 1.4748378.
98. Richard C. Remsing and John D. Weeks. Role of local response in ion solvation: Born theory and beyond. *J. Phys. Chem. B*, 120(26):6238–6249, 2016. doi: 10.1021/acs.jpcb.6b02238.
99. David M. Rogers and Thomas L. Beck. Modeling molecular and ionic absolute solvation free energies with quasichemical theory bounds. *J. Chem. Phys.*, 129(13):134505, 2008. doi: 10.1063/1.2985613.
100. Dmitry V. Matyushov. Solvent reorganization energy of electron-transfer reactions in polar solvents. *J. Chem. Phys.*, 120:7532, 2004. doi: 10.1063/1.1676122.
101. Lu Ding, Maximilien Levesque, Daniel Borgis, and Luc Belloni. Efficient molecular density functional theory using generalized spherical harmonics expansions. *J. Chem. Phys.*, 147:094107, 2017. doi: 10.1063/1.4994281.
102. David M. Rogers. Extension of Kirkwood-Buff theory to the canonical ensemble. *J. Chem. Phys.*, 148: 054102, 2018. doi: 10.1063/1.5011696.
103. Mikhail A. Vorotyntsev and Andrey A. Rubashkin. Uniformity ansatz for inverse dielectric function of spatially restricted nonlocal polar medium as a novel approach for calculation of electric characteristics of ion-solvent system. *Chemical Physics*, 521:14–24, 2019. ISSN 0301–0104. doi: 10.1016/j.chemphys.2019.01.003. URL http://www.sciencedirect.com/science/article/pii/S0301010418309108.
104. S. Koneshan, R. M. Lynden-Bell, and Jayendran C. Rasaiah. Friction coefficients of ions in aqueous solution at 25°C. *J. Amer. Chem. Soc.*, 120(46):12041–12050, 1998. doi: 10.1021/ja981997x.
105. L. Blum and Yaakov Rosenfeld. Relation between the free energy and the direct correlation function in the mean spherical approximation. *J. Stat. Phys.*, 63(5–6):1177–1190, 1991.
106. Gunnar Karlström and Per Linse. Retardation effects breaking long-range orientational ordering in dipolar fluids. *J. Chem. Phys.*, 132(5):054505, 2010. doi: 10.1063/1.3305325.
107. Pengyu Ren, Jaehun Chun, Dennis G. Thomas, Michael J. Schnieders, Marcelo Marucho, Jiajing Zhang, and Nathan A. Baker. Biomolecular electrostatics and solvation: a computational perspective. *Quart. Rev. Biophys.*, 45(4):427–491, 2012. doi: 10.1017/S003358351200011X.
108. B. Mennucci, E. Cancès, and J. Tomasi. Evaluation of solvent effects in isotropic and anisotropic dielectrics and in ionic solutions with a unified integral equation method: theoretical bases, computational implementation, and numerical applications. *J. Phys. Chem. B*, 101(49):10506–10517, 1997. doi: 10.1021/jp971959k.
109. Timothy T. Duignan, Drew F. Parsons, and Barry W. Ninham. A continuum solvent model of the multipolar dispersion solvation energy. *J. Phys. Chem. B*, 117(32):9412–9420, 2013. doi: 10.1021/jp403595x.
110. David M. Rogers. The Einstein-Podolsky-Rosen paradox implies a minimum achievable temperature. *Phys. Rev. E*, 95:012149, 2017. doi: 10.1103/PhysRevE.95.012149.
111. Niels Bohr. *Atomic Physics & Human Knowledge*. Chapman & Hall, London, 1958. page 20.
112. John A. Wheeler. Information, physics, quantum: The search for links. In Shun'ichi Kobaysahi and Nihon Butsuri Gakkai, editors, *Proc. 3rd International Symposium on Foundations of Quantum Mechanics*, pages 354–368. Physical Society of Japan, Tokyo, Japan, 1989.
113. H. Everett. *The Many-Worlds Interpretation of Quantum Mechanics*. Princeton University Press, Princeton NJ, 1973.
114. Lev Vaidman. Many-worlds interpretation of quantum mechanics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018.
115. Lajos Diósi. How to teach and think about spontaneous wave function collapse theories: Not like before. In S. Gao, editor, *Collapse of the Wave Function: Models, Ontology, Origin, and Implications*, pages 3–11. Cambridge Univ. Press, Cambridge UK, 2018. doi: 10.1017/9781316995457.002.
116. Robert H. Wood. Continuum electrostatics in a computational universe with finite cutoff radii and periodic boundary conditions: Correction to computed free energies of ionic solva-

tion. *J. Chem. Phys.*, 103(14):6177–6187, 1995. doi: 10.1063/1.470445.

117. Henry S. Ashbaugh and Robert H. Wood. Effects of long-range electrostatic potential truncation on the free energy of ionic hydration. *J. Chem. Phys.*, 106(19):8135–8139, 1997. doi: 10.1063/1.473800.
118. Billy W. McCann and Orlando Acevedo. Pairwise alternatives to ewald summation for calculating long-range electrostatics in ionic liquids. *J. Chem. Theory Comput.*, 9(2):944–950, 2013. doi: 10.1021/ct300961e.
119. Shinichi Sakane, Henry S. Ashbaugh, and Robert H. Wood. Continuum corrections to the polarization and thermodynamic properties of Ewald sum simulations for ions and ion pairs at infinite dilution. *J. Phys. Chem. B*, 102(29):5673–5682, 1998.
120. Philippe H. Hünenberger and J. Andrew McCammon. Ewald artifacts in computer simulations of ionic solvation and ionion interaction: A continuum electrostatics study. *J. Chem. Phys.*, 110(4):1856–1872, 1999. doi: 10.1063/1.477873.
121. Luc Belloni and Joel Puibasset. Finite-size corrections in simulation of dipolar fluids. *J. Chem. Phys.*, 147(22):224110, 2017.
122. Phillip S. Hudson, Stefan Boresch, David M. Rogers, and H. Lee Woodcock. Accelerating QM/MM free energy computations via intramolecular force matching. *JCTC*, 14(12):6327–6335, 2018. doi: 10.1021/acs.jctc.8b00517.
123. Joseph E. Basconi and Michael R. Shirts. Effects of temperature control algorithms on transport properties and kinetics in molecular dynamics simulations. *J. Chem. Theory Comput.*, 9(7):2887–2899, 2013. doi: 10.1021/ct400109a. URL http://pubs.acs.org/doi/abs/10.1021/ct400109a.
124. Benjamin Cross, Chloé Barraud, Cyril Picard, Liliane Léger, Frédéric Restagno, and Élisabeth Charlaix. Wall slip of complex fluids: Interfacial friction versus slip length. *Phys. Rev. Fluids*, 3: 062001, Jun 2018. doi: 10.1103/PhysRevFluids.3.062001.
125. Richard C. Remsing, Ian G. McKendry, Daniel R. Strongin, Michael L. Klein, and Michael J. Zdilla. Frustrated solvation structures can enhance electron transfer rates. *J. Phys. Chem. Lett.*, 6(23): 4804–4808, 2015. doi: 10.1021/acs.jpclett.5b02277.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Research Article

# Hydration of silica and its role in the formation of quartz veins - Part 2

John Elliston

*Elliston Research Associates Pty Ltd, 10B The Bulwark, CASTLECRAG 2068, New South Wales, Australia*
E-mail: john.elliston@ellistonresearch.com.au

OPEN ACCESS

**Citation:** J. Elliston (2019) Hydration of silica and its role in the formation of quartz veins - Part 2. *Substantia* 3(1): 63-94. doi: 10.13128/Substantia-209



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Competing Interests:** The Author(s) declare(s) no conflict of interest.

**Absract**. The aqueous chemistry of silica and the formation of abundant polymeric silica species in natural sediment accumulations is set out in Part 1. Part 2 continues to describe the amorphous silica in sediments and presents more evidence for the formation of quartz veins from silica gels. The physical chemistry for the formation of oolites in quartz veins, ptygmatic folding of quartz veins, the enhanced growth of crystals in quartz veins, the precipitation of gold in quartz and the mechanism for replacement or metasomatism by quartz (silicification) are detailed. Current physical chemistry applies not only to precursor silica gel but also to the hydrous precursors of other crystalline rock minerals such as clays, mica, hydrous ferromagnesian minerals and mineral deposits. "The Origin of Rocks and Mineral Deposits - using current physical chemistry of small particle systems" provides a new basis for understanding geological phenomena. This has now been published in Elliston[1] 2017 (Connor Court, Brisbane, ISBN 978-1-925501-36-0).

Part 2 continues the illustrations and diagrams numbered in sequence from Part 1 and the conclusions relate to both parts of this article.

**Keywords**. Accretion, concretion, charged particle, ptygmatic folds, precursor mineral, enhanced crystal growth, metasomatism, porphyroid, granite.

## THE NATURE OF SILICA IN VEINS

Dispersions of monomer are at first stable but as polymer formation develops the monomer is more rapidly consumed. The movement of monomeric silica dispersion tends, over time, to transfer significant quantities of quartz. The old concept of quartz "sweating out of" the host granites or sediments finds a complete explanation in what is now known of the properties and behaviour of polymeric silicic acids.

The dispersions or gels (accumulating in the precursor quartz veins) of poly silicic acids are unstable in the sense that the particles grow in size and aggregate into denser precipitate, or gels (Figure 48). It should be noted that, when substantial salt concentrations and alkaline conditions prevail, the gelling process does not produce a homogeneous gel. The silica usually appears

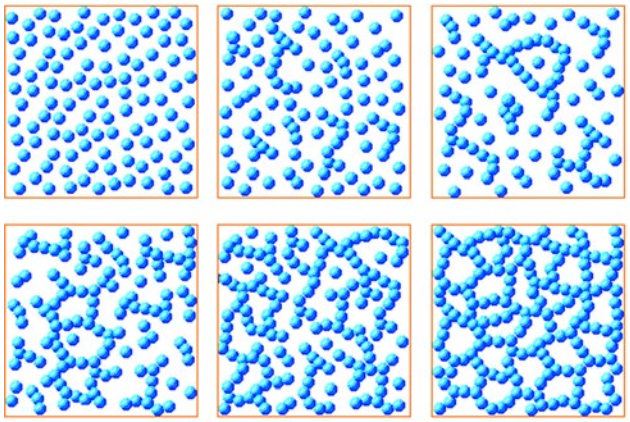


**Figure 48.** The gelling process involves formation of an increasing amount of the sol being converted to microgel with increasing viscosity until it solidifies. This "solidification" in the first instance is achieved when there is sufficient three-dimensional cross linking to entrap the solvent in pockets. Initially it is a wet weak watery gel which becomes essentially a dispersion of solvent in "spongy" solid. A gel is by definition a visco-elastic solid with all linked parts of the meshwork immobilised in relation to all other linked parts. It is "sensitive" to disruption by shock or shear (thixotropic). [From Iler, 1979, p. 232]

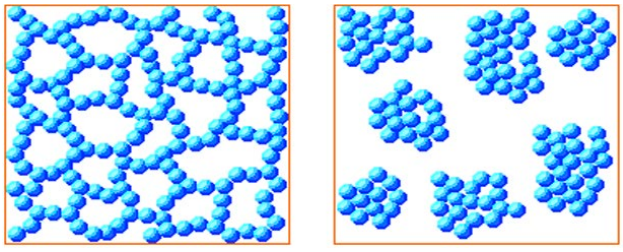


**Figure 49.** This diagram is a two-dimensional representation of the difference between a three-dimensional gel meshwork with its cross-linked chains of silica spheres and a precipitate of three-dimensional clusters of colloidal silica globules.

as a white precipitate or a white opaque gel, owing to partial precipitation before gel formation.

Figure 49 indicates the difference between gel and precipitate of the primary particles which are about 3 to 6 nm in natural silica gels. Due to the presence of salt and higher pH (most pore fluids have a pH about 7.5 to 8.2) the white silica gels of the quartz vein precursors are thought to be partly precipitated to aggregates of around 30 nm which "gel" to form chains of these spheres.

Maximum thixotropic sensitivity occurs when such small three-dimensional aggregates link together into a larger network extending through the liquid medium (Figure 50). Aggregates link together through hydrophobic bonding and such linkages are readily broken by

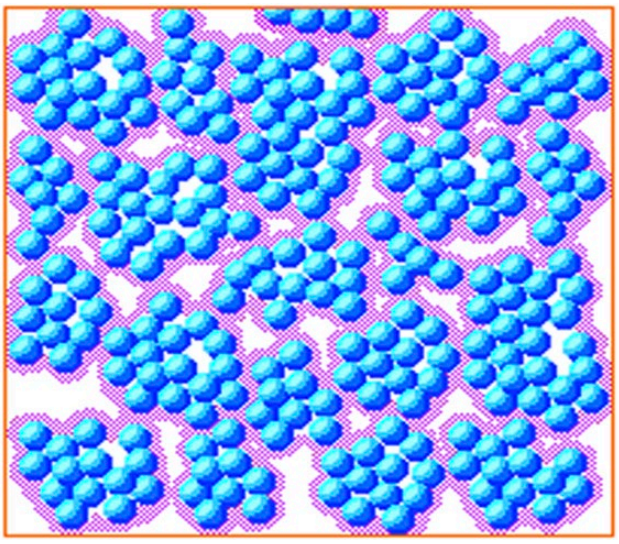


**Figure 50.** Maximum thickening effect and thixotropy are produced by small aggregates linking together into an extensive three-dimensional gel network throughout the liquid medium. Natural silica gel occurring as the precursor to vein quartz is thought to be partly precipitated (white colour), thixotropic, and dense enough to support mineral and wall rock fragments. In most cases it is probably of this type. [From Iler, 1979, p. 591]

shearing forces and readily re-established when the mass is at rest. This would apply to the precursor gel stage of quartz vein development where the vein material is highly thixotropic.

### *The amorphous silica in sediments*

The gel structure of amorphous silica in sediments is rather uncertain. Under the conditions of most newly deposited marine sediments where pH is around 7.8 to 8.2 with salts present, chaining of silica spheres would probably involve particles about 1 nm in size. While these conditions would favour stable polymeric silica, the chained particles are too small to be seen with an electron microscope.

It is also not clear how the small spherical silica particles would interact, or have their normal 'chained meshwork' structure modified by the presence of other charged sediment particles. Just as the behaviour of ionic species in the pores of a mud are severely modified by their interaction with charged surfaces, so also is the behaviour of charged particles.

Silica spheres under natural conditions are negatively charged and at about 1 nm in size they would be expected to interact with the positively charged edges of clay platelets. For montmorillonite these platelets are

about 30 to 50 nm in size but for illites the platelets can be 1000 to 1500 nm.

The 'gel structure' of silica in natural muds cannot be considered as separate from that of the sediment particles as a whole unless the sediment is essentially silica such as chert, jasper, or one of the deep marine siliceous oozes. A further complication is the fact that much of the hydrated silica in sediments is derived from the hydrolytic degradation of clays. Progression of the 'zip fastener' reaction to complete hydrolysis yields sheet-like networks of silica gel, residual tetrahedral layers or fragments of them, from the hydrolysis of clays, ferromagnesian silicates and quartz (in slightly alkaline sea water).

These are surfactants and as polyelectrolytes, that is polymers whose repeating units are ionised or ionisable in water, they are strongly surface active. This polymeric form of silica usually adsorbs on oppositely charged surfaces but it may also adsorb on neutral or negatively charged surfaces. An entropy driving force can enable one adsorbing polymer to free to solution many previously bound water-molecules. This yields a net increase in the number of kinetic units in the system. Much of the amorphous silica in natural muds can therefore be regarded as 'coating' the other particles.

*The chemical equilibria of silica in sediments*

Silica is also in chemical equilibrium with the other constituents of the sediment. The hydrolysis of clays, the 'zip fastener' reaction, is reversible and depends on the concentration of one of the reactants, namely water. If water is in excess, under most diagenetic conditions the clays will continue to slowly hydrolyse. Without shear, when water is removed from the system, the tetrahedral and octahedral sheets of hydrolysed clay platelets re-combine to restore the original clay structure. Iler[2] (1979, p.193) points out that over a long period of time monomeric silica, $Si(OH)_4$, reacts with $Al^{3+}$ ions at 25°C to form colloidal aluminium silicate of halloysite composition:

$$2Si(OH)_4 + 2Al^{3+} + H_2O \rightarrow Al_2Si_2O_5(OH)_4 + 6H^+$$

There is no question that where the amorphous silica constitutes all or most of the sediment such as in a chert bed, quartzite, or siliceous ooze, the gel structure is a chained meshwork of spherical particles. Figure 51 from Iler[2] (1979, p. 235) indicates the type of meshwork packing of spheres when they are large enough to be observed by electron microscopy. In a sediment they are probably more like those illustrated in Part 1 that show the way they aggregate through micro-accretions

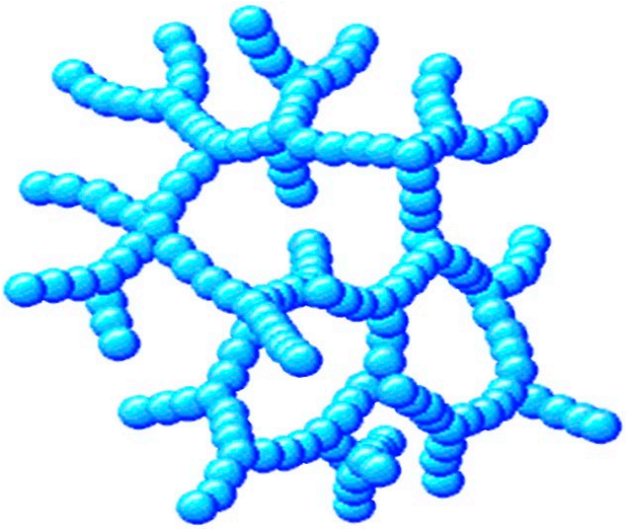


**Figure 51.** In natural silica gels the density and structure vary by a number of different kinds of packing in the meshwork arrangement of the particle chains. In a more open regular meshwork such as that illustrated here, the number of particles linked to a given particle can vary. In this example each joining sphere is touched by 3 or 4 surrounding spheres but denser linking arrangements are common.

to more stable 'close-packed' macro-aggregates when the sediment is re-liquefied and involved in flow. The extremely small size of the ultimate particles in this form of silica must be kept in mind.

*The 'aging' of silica gels in sediments*

Iler[2] (1979, pp. 224-230) points out that the silicic acid "balls" or particles chaining together to form a gel can be linked by inter-particle siloxane bonds when catalysed by a base (Figure 52). Also, the laws of solubility apply, whereby according to the Otswald-Freundlich equation, solubility at the negative radius of curvature (Figure 53), in the neck between contacting spheres, is less than elsewhere on the spheres. At equilibrium conditions with the monomer, and as the gel ages, the polymerisation tends to increase in the crevice at the point of contact between the spheres while solution would occur from the protuberant portions of the sphere as in Figure 54. Electron microscope studies of the fine porous silica gel (Sugar and Guba[3], 1954) have shown that the structure is indeed made up of a thread-like fibre network, but the fibres were made up of chains of spheres as in Figure 55.

The "chaining" arises because new charged particles from dispersion are added to a doublet in alignment or to the end of a chain. When added in this position a new

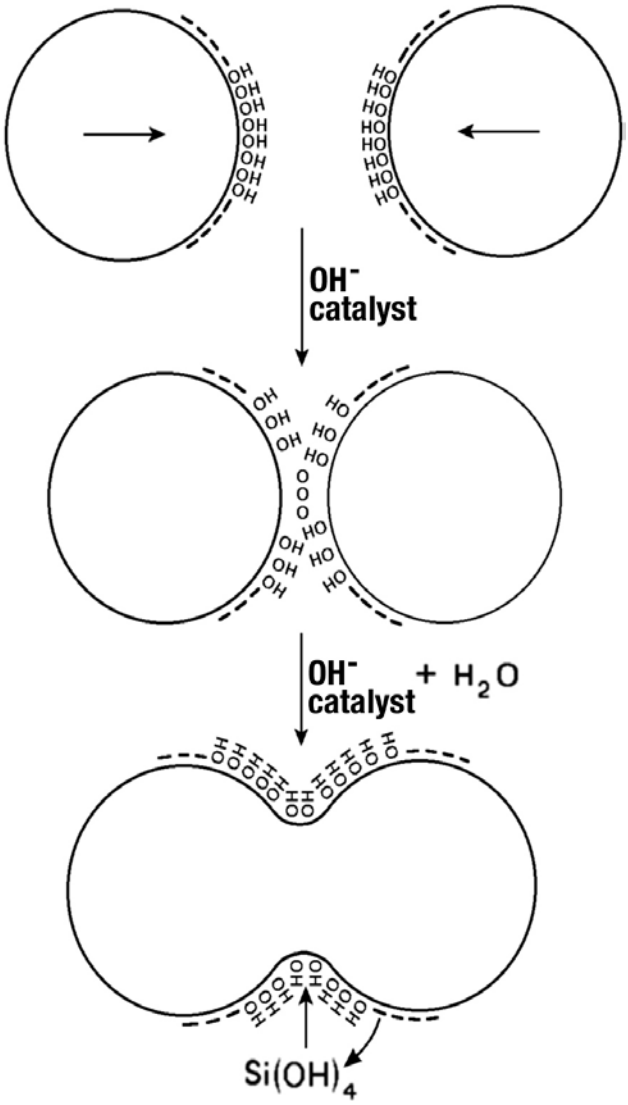



**Figure 52.** When there is little or no charge repulsion between silica particles such that they can come together to form chains (linked by van der Waals' attraction), chemical siloxane bonds, catalysed by the presence of hydroxyl ions, also develop as indicated. Once bonded the particles grow together by diffusion and condensation of additional $Si(OH)_4$. The resilience and strength of some silica gels is greatly increased when the chains of silica spheres are also chemically linked. [From Iler[2], 1979, p. 224]

particle has only to overcome the repulsive force due to one of the particles it is joining. If it approaches from the side position, the other similarly charged particle or particles would contribute to the repulsion. It is therefore at the end position that the joining particle can first approach close enough for van der Waals attraction to exceed the coulombic repulsion.

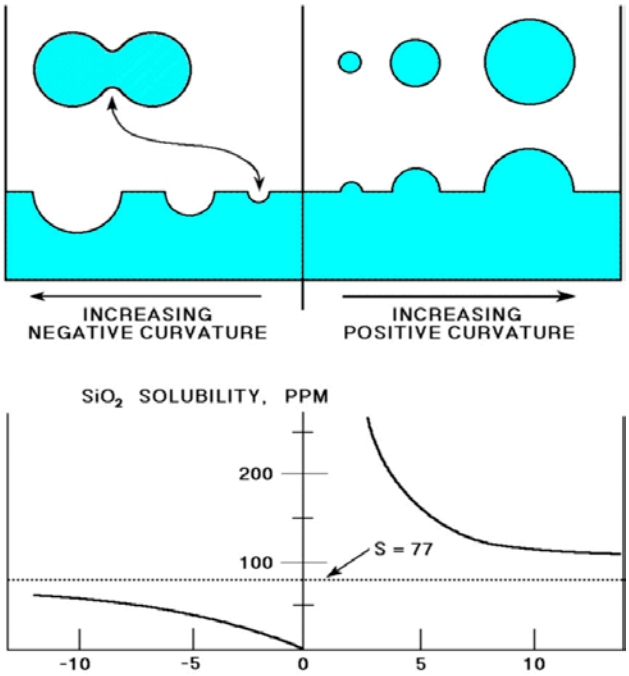



**Figure 53,** The solubility of silica varies inversely with the radius of curvature of the surface according to the Ostwald-Freundlich equation. The positive radii of curvature are shown as cross-sections of particles and as projections from a silica surface. Negative curvatures are shown as depressions or holes in a silica surface or in the crevice between two adhering particles. The solubility in ppm is indicated in the lower diagram. Two important practical effects are: (a) When very small particles are brought into the same suspension as larger ones, especially at high pH where OH- ions catalyse solution and deposition, small particles dissolve and larger ones grow. (b) At the point of contact between two flocculated colloidal silica particles in a chain where the radius of curvature is negative and small, the solubility is low and silica dissolves from the particle surfaces to be deposited at the point of contact. (From Iler[2], 1979, p50)

## MORE EVIDENCE FOR QUARTZ VEINS FORMING FROM SILICA GELS

### *Orbicular structures and oolites in quartz veins*

The wetter parts of jaspoidal lodes, quartz magnetite bodies, and quartz veins which crystallise from gelatinous precursors, may develop oolitic concretions during their gel stage. Spherical and globular structures are not very commonly found in quartz veins and siliceous lodes because, like colloform banding, injection or episodic re-mobilisation of the precursor lode materials usually destroy primary precipitation structures.

However, when siliceous oolites and orbicular structures are found in vein or mineral lode material they constitute an unequivocal indication of their development in a gelatinous medium by diffusion of hydrous

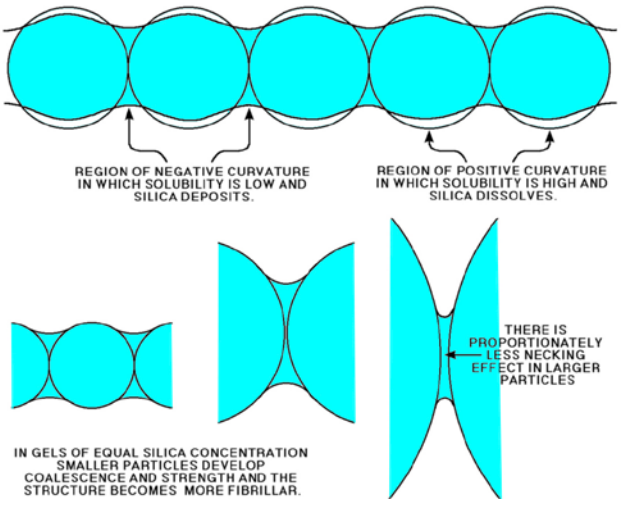



**Figure 54.** A chain of small particles is converted to a fibre or rod by the laws of solubility. Solubility at the negative radius of curvature of the neck between the spheres is less than elsewhere on the spheres so that, as the system tends to equilibrium, silica builds up in the region of the contacting spheres and "rounds off" protuberances. The effect is greater for smaller particle chains. [Adapted from Iler[2], 1979, p. 230]

**Figure 55.** Stages in the aging of a gel are illustrated from the chaining of spheres and their greater coalescence, to the coarsened structure which results from drying or heating where the structure disintegrates to irregular rounded particles. [Adapted from Iler[2], 1979, p. 530]

polymeric silica particles. The diffusion processes which give rise to oolites depend on the structure of precursor stage silica gel in the veins and diffusion of mobile particulate species within them. Having regard to the nature of the vein gels an occasional occurrence of oolites resulting from these simple diffusion processes would certainly be expected.

The precursor vein host and mobile species may be identical chemically yet quite different structurally (eg. $Si(OH)_4$ and silicic acid polymers). This is why it is quite possible to develop silica concretions in silica gel (as Lougheed's[4] microspheres in Figure 56), ferric hydroxide concretions or oolites in ferric hydroxide matrix (Lindgren[5], 1933, p. 277: Carozzi[6], 1960, p. 352), or calcite oolites in limestone or an essentially calcitic matrix as in Figure 57 of Lindgren's Short Creek oolites.

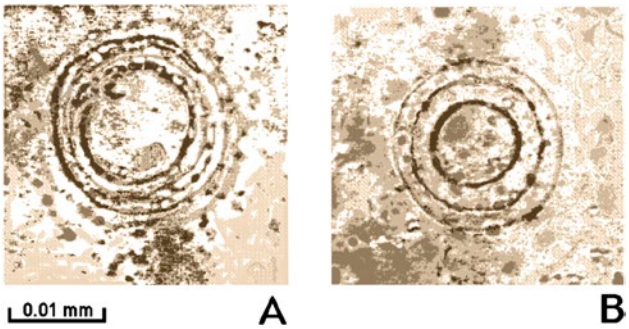



**Figure 56.** Examples of silica microspheres or concretions of silica in chert from the siderite-chert facies of the Gunflint Iron Formation, Kakabeka Falls, Ontario. These silica in silica concretions illustrated by Lougheed[4] (1983, p. 327) may be coated sparsely with micron-sized carbonate crystals (white) and carbon flecks (spots and dusting).

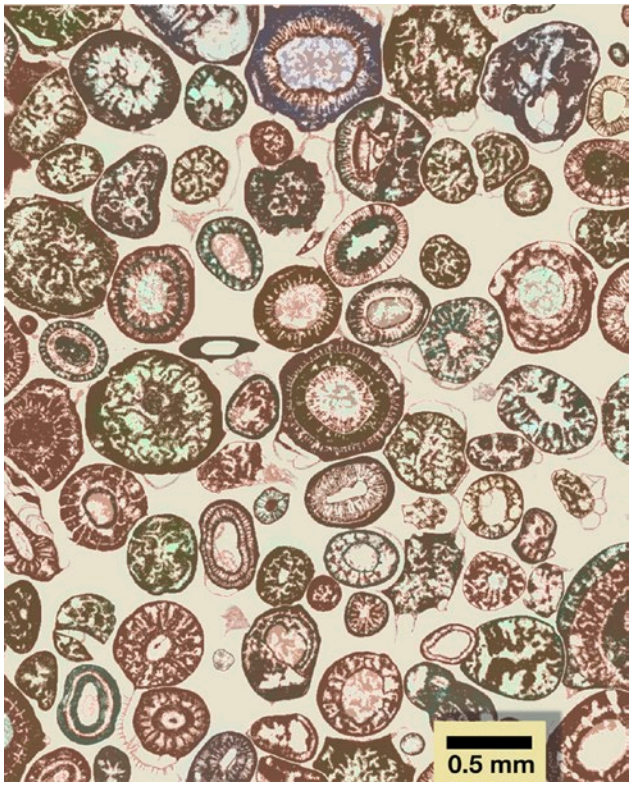



**Figure 57.** Calcite oolites in a calcitic matrix further emphasise the point that gelatinous oolite precursors do not have to be different in composition from the gelatinous lime muds in which they form. They differ in precursor gel structure in that the oolitic nuclei must be 'close packed' and synerectic while the host mud is an open-meshwork porous gel through which very small particles can diffuse. These oolites from Short Creek, Missouri, (Lindgren[5], 1933, p. 267) show the evidence of their soft gelatinous precursors in the mutual indentation and distortion, fracture fragments, composites, overgrowths, and syneresis shrinkage patterns.

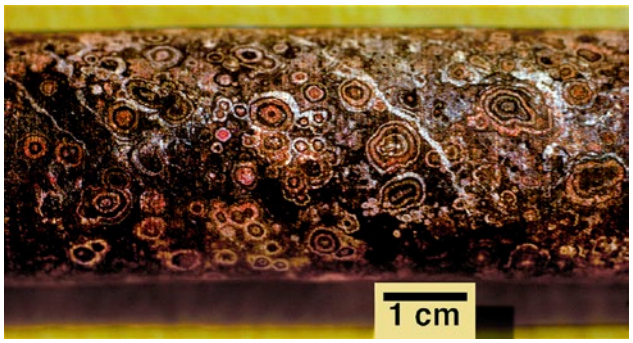



**Figure 58.** Orbicular quartz-magnetite from drill core through the lower part of the Peko diapiric ore pipe, Tennant Creek. The oolitic texture indicates that precipitation of precursor lepidocrocite and polymeric silica round synerectic nuclei is a sol-gel transition rather than a chemical reaction or from supersaturated solution. Clearer quartz veins which cut the oolites reflect the soft particulate nature of the precursors.

Similarly, because the 'colloidal processes' by which concretions develop are not dependent on the chemistry, mixed oolites with successive rims of different substances are commonly developed as in Figure 58.

*Some examples of siliceous oolites in veins and lodes*

The concretions of silica in chert illustrated by Lougheed[4] in Figure 56 are significant in showing that the development of the concentric spheres is a physical or structural change in the polymeric precursor particles and not chemical. The hydrous silica particles ("little balls") do not initially react with each other chemically but they deposit and build up round a synerectic nucleus from very small sol particles dispersed in the surrounding chert precursor to the 'close packed' dense gel of the concretion. It is a sol to gel transition at the precursor stage whereas the chemical reaction is the subsequent loss of water on crystallisation.

After condensation and lithification of the whole formation, these concretions could not be seen at all if it were not for the small carbonate crystals and carbon flecks which coated the former surfaces of these denser globules within the gel.

Radiating quartz crystals have grown radially round dawsonite needles in a large impure quartz vein at Mt. Gee in the Flinders Ranges in South Australia as shown in Figure 59. The dawsonite [$Na_3Al(CO_3)_3{\cdot}2Al(OH)_3$] apparently grew as slender needles supported in the soft precursor silica gel of the vein so that subsequent quartz crystallisation was nucleated from their rod-like surfaces. Concentric banding or tubular zones equidistant from dawsonite needles (Figure 60) reflect the concretionary banding similar to the carbon dusted zones in Figure 56 or the chlorite bands in Part 1, Figure 45.


**Figure 59.** At the precursor stage, denser concretionary polymeric silica precipitated around dawsonite needles grown in the pre-crystalline silica gel. The denser concretionary silica subsequently crystallised to form a radiating spherulitic pattern of crystallisation which now forms tubular shaped clusters of radiating fine crystals round the nucleating dawsonite needles. The concentric banding or zones probably represent concretionary desorption of impurities (Liesegang-type) prior to crystallisation. The outcrop is at Mt. Gee in the Flinders Ranges, SA.

Much of the large pyritic gold-copper lode at Mt. Morgan in Queensland was highly siliceous. Some of this silica was ordinary white vein quartz but most of it was a dull grey earthy quartzite-looking silica. Apart from the abundant sulphides it was discoloured with traces of impurities such as kaolinite, sericite, chlorite, or haematite.

One of the early phases of this mineralisation in the lode system is a massive injection of rather impure cherty grey silica. This early emplacement of siliceous material could be regarded as a "dirty early" quartz vein although,


**Figure 60.** Concretionary polymeric silica at the precursor stage has formed round dawsonite needles grown in silica gel. It has subsequently crystallised in a radiating spherulite-like pattern to form interfering coarse tube-like clusters of fine radiating quartz crystals along the slender dawsonite 'needles' from which they nucleated. This vein quartz is from Mt. Gee in the Flinders Ranges, SA.

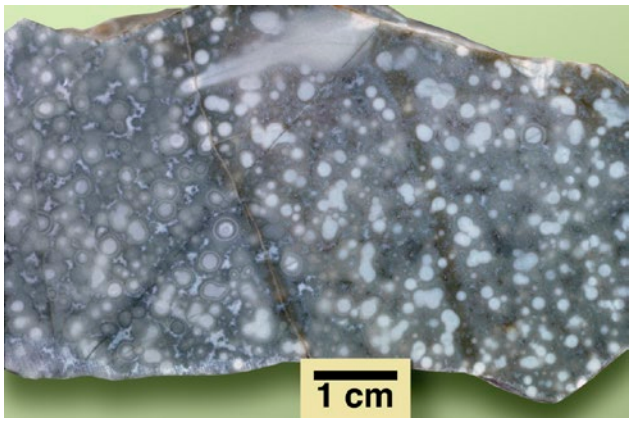



**Figure 61.** An irregular bulbous intrusion of impure cherty grey silica is part of the Mt. Morgan Mine lode system in Queensland. It has an "orbicular" or oolitic texture formed by concretion round small accretionary (denser synerectic) nuclei which were developed during intrusion and flow of the precursor polymeric silica. The texture positively indicates the silica gel precursor of this siliceous lode or 'elvan'.

like many such early emplacements in highly hydrous lode systems, it is rather bulbous and irregular in form and it is re-invaded by several phases of later mineralisation and later vein networks. However, the early impure grey silica intrusion clearly had a silica gel precursor because it is oolitic (Figure 61). The overlapping and merging concretions which are abundantly developed, resulted in the nickname "bird's eye" porphyry.

Its concretionary texture reflects the fact that this impure silica gel precursor was re-mobilised into the vein or body in which it is found. The flow of the liquefied precursor generated small denser random nuclei. In the subsequent 'static' stage, when the permeable gelatinous matrix material had again 'gelled', these synerectic close-packed micro-accretions form the nuclei for concretionary overgrowth. Concretion is by diffusion of particles individually to the surface of the nuclei where they precipitate to become part of each close-packed concretionary mass. Similar 'dirty' grey quartz to this concretionary 'elvan' is also illustrated as the matrix of the chloritic fragment breccia in Part 1, Figure 37.

In another more jaspoidal variety of this early mobilised silica gel at Mount Morgan, the oolites are rather larger and some of the concentric banding where it was outgrown in a scalloped pattern, was called locally "hurgledurgelite" (because at the mine there was some uncertainty as to its origin).

Some of the larger concretions in this marginal jaspoidal chert body (Figure 62) have grown over and incorporated smaller ones. This is common in colloidal gels where the concretion is essentially the same composition as the less dense gel in which it is developing. These large round 'blobs' of synerectic silica show little or no concentric internal banding indicating that their growth involved a degree of infill or incorporation of their silica gel matrix rather than its displacement. Displacive concretion occurs when the growing 'close-packed' gel mass is stronger than its enclosing less dense matrix and this is usually the case for mixed layer oolites or when other colloids are involved.

A siliceous lode cored in drilling at the Abra prospect at Jillawara, Western Australia, contains jasper concretions (Figure 63). Rather complex small oolites and their fracture fragments occur in early quartz-jasper vein material and are cut by later stylolitic quartz veinlets. These jasper oolites have all the features of full-scale larger concretions which show successive rims, mutual indentation, complex cores, plastic and fractured ovoids, regrowth on fragments, folding and unconformities in the rim patterns, etc. Their development clearly indicates that the matrix vein material was originally a particulate polymeric silica gel precursor.

The Smithsonian Institute displays a piece of a massive dark siliceous lode from near Gilroy in California which contains abundant jasper oolites (Figure 64). The highly hydrated lode in which the oolites developed has

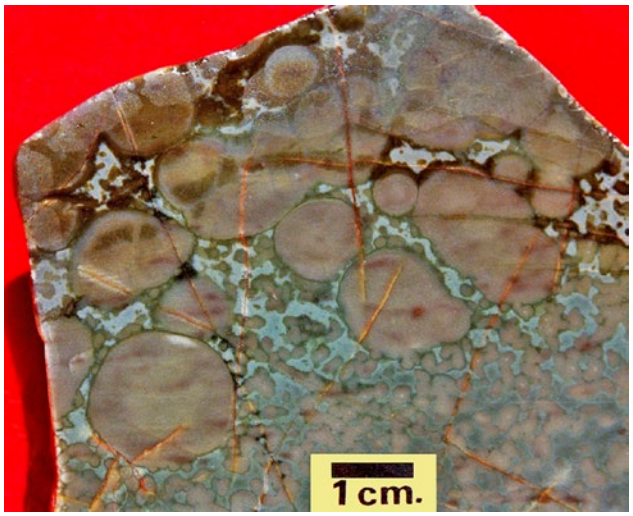



**Figure 62.** Some of the larger concretions in the Mt. Morgan siliceous 'elvan' grow over and incorporate smaller ones. Where the concretion has essentially the same composition as the less dense gel medium in which it is developing, the growth of the 'close-packed' nodule is often as much by infill and incorporation of the host silica gel matrix rather as by its displacement (like bedding preserved through concretions in sediments). Syneresis exudes impurities to the rims.

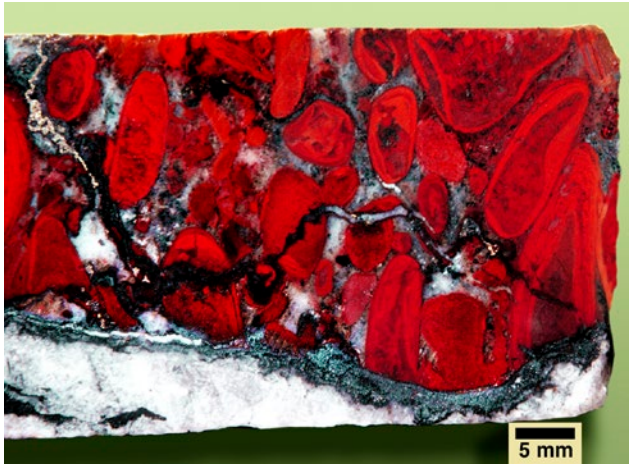



**Figure 63.** Brilliant jasper oolites occur in core through a quartz lode at the Abra Prospect, Jillawarra, W.A. Jasper is metacolloidal and a soft precursor stage of these overgrowths on accretionary siliceous nuclei is indicated by their plastic deformation, mutual indentation, and fragmentation. They are cut by later stylolitic quartz veinlets which meandered through the soft gelatinous silica precursors.

subsequently contracted and further white polymeric silica partly fills the syneresis cavities. Earlier khaki chert rims these cavities and some of it fills veinlets which cut the soft precursor oolites.

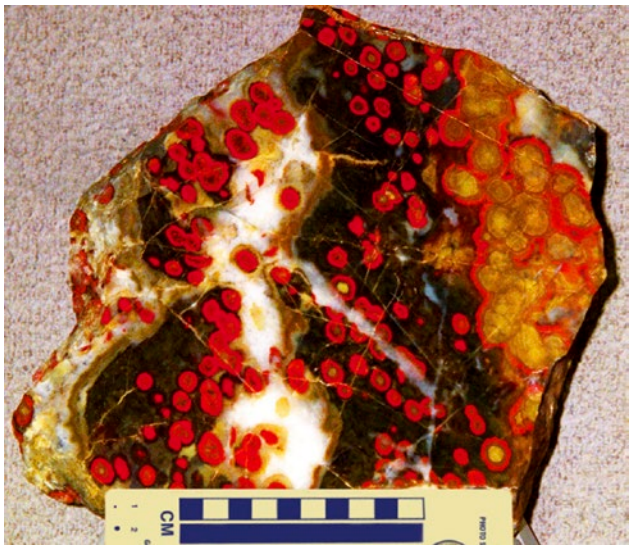



**Figure 64.** Later diagenetic colloidal silica branching into veinlets, fills syneresis cavities in an oolitic jasper lode. Earlier khaki chert rims the syneresis cavities in a similar manner to its occurrence in geodes and it fills veinlets which cut the soft precursor oolites. The jaspers, the chert, and the oolites undeniably indicate precursor silica gel. This Smithsonian Institute specimen is from near Gilroy, California, USA.

Studies of the Peko diapiric quartz-magnetite ore pipe some fifty years ago established conclusively that the quartz and magnetite of the main breccia body could not have been deposited from solution. The quartz magnetite is an intrusive, brecciated, and flow sheared mass which incorporates many fragments of the greywacke-shale wall rocks into which it intrudes. Some of these are several metres in size, rotated and unaltered with bedding still preserved. The massive quartz-magnetite must have been fluid at some stage in order to allow the blocks of included sediments to rotate and yet it had to 'set' quickly in order to suspend these blocks and the pieces of magnetite it contains like those in the Warrego ore pipe (Part 1, Figure 38) or the magnetite breccia pieces in the quartz lode at New Cobar Mine (Part 1, Figure 39).

The intrusive quartz-magnetite body with its central core of rich sulphide and gold mineralisation is itself auto-brecciated. It is too heavy (density 4.5) to have intruded as actual quartz-magnetite into the greywacke-shale sequence (density 2.7) and if the silica and iron had in fact been in hydrothermal solution together to form this deposit it should have contained or consisted of iron silicates rather than the respective oxides.

Evidence for colloidal precursors of the precipitated quartz, magnetite, and ore sulphides came from textures preserved in some of the auto-brecciated blocks and

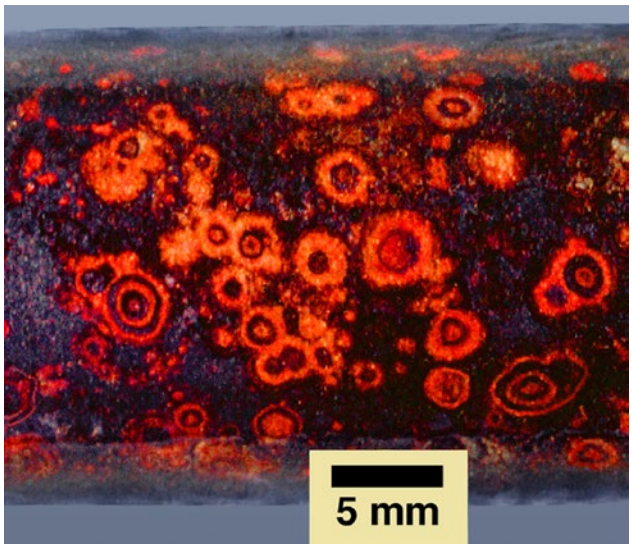



**Figure 65.** Red jasper-magnetite oolites are found in the lower parts of the Peko diapiric quartz-magnetite ore breccia pipe at Tennant Creek, N.T. Composite ooolites and some disruption of the rims reflect the soft nature of the precursor magnetite (lepidocrocite) and particulate colloidal silica. The texture and the metacolloidal jasper leave no doubt that this lode was emplaced as a precursor colloidal mass.

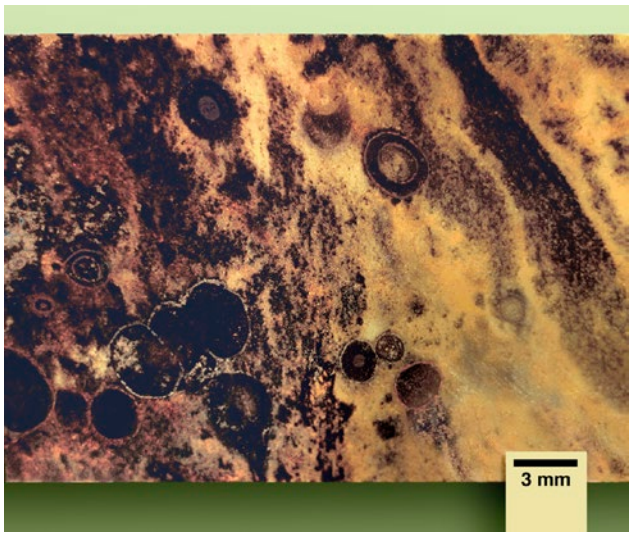



**Figure 66.** Jaspoidal chert-magnetite oolites are preserved in part of the Peko Mine main lode at Tennant Creek, NT. This section of the ore breccia pipe has not been disrupted by later re-mobilisation. Original colloform textures are often preserved in large rafts, or near the lode wall where later thixotropic re-liquefaction and intrusion of the ore pipe slurry has not affected the whole semi-consolidated mineral mass.

zones which had been less disturbed in positions deeper and peripheral to the main ore zone.

These blocks and lower zones in the Peko ore pipe contain quartz-magnetite oolites (Figure 58), quartz-chlorite oolites (Part 1, Figure 36), jasper-magnetite oolites (Figure 65), and jaspoidal chert-magnetite oolites (Figure 66). There are other concretionary overgrowths and colloform banded structures in the ore sulphides and gangue minerals in this pipe but the siliceous oolites are clear evidence of the gelatinous nature of the silica in the lode at the precursor stage and the ability of smaller polymeric silica particles to diffuse within it.

*Particle interactions involved in the formation of oolites*

The mechanism of oolite formation requires the specific adsorptive, synerectic, and concretionary properties of the parent gel system to the extent that any alternative genesis is precluded.

A poor understanding of the 'colloidal processes' which give rise to concretions is reflected in the geological literature. This is sometimes acknowledged (eg. Heinrichs[7], 1984, p. 239) and therefore the main steps in the auto-precipitation of metastable colloidal suspensoids in sediment pore fluids or within a gel meshwork are summarised as follows.

The process of concretion within a small particle system such as a fine-grained sediment, is characterised by the diffusion of macromolecules or colloidal particles from the surrounding disperse phase, particle by particle, to accumulate about a central nucleus. Such concretion produces spherically or elliptically symmetric accumulations of higher particle density and compaction than the medium from which it is diffusing. The process is 'static' during diffusion and the two gel phases are not liquid or the gel meshwork of either disrupted. It is usual for the main matrix to be thixotropic and have been mobilised to form denser accretions that then form synerectic nuclei for the concretionary overgrowth.

In most systems the disperse particles diffuse down a concentration gradient in the interstices of the surrounding matrix gel of the system. This gradient is created by the removal of particles from dispersion as they precipitate at the nucleus or internal surface on which the concretion is developing. The nuclei constitute a macroscopic aggregate of particles or an internal 'surface' which is the boundary of the much denser gel within the more permeable and 'open framework' gel of the surrounding matrix (Figure 67). When Brownian motion moves the charged particles in the direction of the precipitating surface, they "see" the deficiency or absence of similar charge and are thus less repelled in this direction. Diffusion acts to keep the concentra-

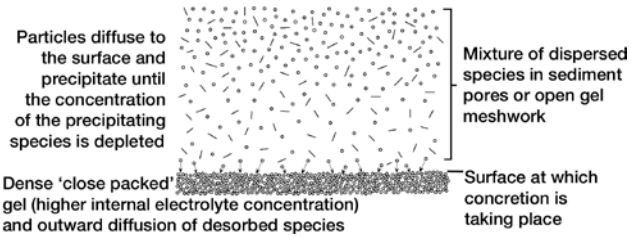



**Figure 67.** Diagram of the surface of a synerectic nucleus to which colloidal particles are diffusing to build up a concretionary overgrowth. This internal surface within a sediment such as a mud, ooze, or slime is merely a denser 'close packed' gel. The active concretionary nucleus is synerectic, condensing under van der Waals' attractive forces such that surface energy, internal surfaces, and adsorptive capacity are reduced. Synerectic liberation of sorbed electrolytes within the nucleus and the new overgrowth. This perpetuates the precipitation of new species.

tion constant and thus is established the gradient along which a continuing supply of particles move to the precipitating surface.

The mechanism of precipitation at the surface of the developing concretion is generally dependent on the nature of this surface and its charge but in all cases, it results in a reduction of the total surface and of surface energy.

The formation of oolites or any form of concretion depends on a nucleus or internal surface at which there is a change in physical conditions. The precipitation is not a purely chemical reaction but more usually a change in the parameters which determine the nature of the surface charge on the colloidal particles. As pointed out (Part 1 pages 47 to 50) substances with points of zero charge (pzc) near the middle of the pH range (like FeO·OH, 6.8; $Fe_2O_3$, 4.8-8.6; AlO·OH, 7.8-8.8) show marked variation in their surface charge dependent properties as a consequence of relatively minor variations in parameters such as pH, electrolyte concentration, etc.

Thermodynamically, the most stable state for the total system is that of minimum free energy. Since there is free energy associated with the particle surfaces, in proportion to their area, stability requires that the total surface area be minimised, which can only be achieved by their coagulation into larger particles. This process is favoured by interparticle attractive forces (van der Waals), but opposed by the electrostatic repulsion of similarly-charged particles. Therefore, a colloidal suspension, or sol, can exist in metastable condition, with no observed tendency to coagulate, as long as the electrostatic repulsion effect predominates. However, coagulation of a stable sol can be induced by changing the system in any of several ways. These include:

(a) Changing parameters, such as temperature or pH, that determine the nature of the surface charge on the dispersed particles.
(b) Addition of oppositely-charged particles (eg. electrolyte ions, or particles of another colloidal system with different pzc) to the system, to neutralise the residual surface charge.
(c) Reduction of the mean interparticle distance to the point where attractive van der Waals forces predominate over the electrostatic repulsive forces. van der Waals attractive forces vary approximately proportionately as the inverse of the cube of the interparticle distance and therefore strengthen rapidly as the interparticle distance diminishes. Electrostatic repulsive forces vary inversely as the square of the interparticle distance, and therefore strengthen less rapidly at decreasing interparticle distances. Reduction of the interparticle distance can be a consequence of various processes such as load compression, removal of interparticle fluid, increased kinetic motion at increasing temperatures, etc.
(d) Addition of "bridging" particles, such as long-chain polymers, which suffer no electrostatic repulsion and can link the sol particles into loosely bound flocs, which then become large enough to settle into more compact aggregates.

The main physical condition which changes at the surfaces of the nuclei on which oolites develop, is the concentration of the electrolyte. The 'close packed' nuclei where the particles of the denser gel are more cohesive, semi-ordered, and spontaneously drawing together (synerectic) under the influence of van der Waals strong forces of attraction (Part 1, Figure 6) are, as a cluster, reducing internal surfaces, surface energy and thereby reducing adsorptive capacity.

This means that the electrolytes adsorbed on the particle surfaces within the nucleating cluster are desorbed into the residual interparticle fluid within the cluster. This fluid at the higher electrolyte concentration is exuding to the surface of the cluster as the denser gel condenses and ages. Mielenz and King[8], in their 1953 paper on the physical-chemical properties and engineering performance of clays, discussed syneresis and recognised that (p. 237) in more concentrated systems van der Waals attractive forces operate to draw the clay particle clusters together in spite of the repulsion due to their diffuse cationic hulls. They say (p. 238) that in the more rigid gel that may thus form, syneresis will cause a contraction of the diffuse cationic layer of each particle which can only be accomplished by expulsion of water with a different electrolyte content to that of the original suspension.

Any dispersed particle in the surrounding medium which approaches or contacts the internal surface therefore encounters a higher electrolyte concentration which reduces the thickness of its Helmholtz double layer thereby reducing the force of repulsion between particle and surface or particle and other similarly charged dispersed particles.

Particles therefore precipitate at the surface of the synerectic nucleus and having passed through this transition from the metastable dispersed phase to become part of the 'close packed' dense gel nucleus, they also condense or 'age' under the influence of van der Waals strong attractive forces further reducing surface energy and internal surfaces. As part of the synerectic nucleus, the particles accumulating on the perimeter in turn desorb their adsorbed electrolyte species which they brought to the surface by diffusion as particles. This then exudes to the new surface perpetuating the growth of the concretion (Figure 68).

Some concretions can therefore grow to enormous size because the growth is only limited by maintenance of an open water-rich gel meshwork through which a continuing supply of particulate species can diffuse (Figure 69, Moeraki boulder).

These non-chemical sol-gel transitions, the diffusion of ions and small charged particles within fine grained sediments or open meshwork gels can be confusing for geologists if the essential nature of these particulate systems is not kept in mind. The colloidal size range spans three orders of magnitude from 1 to 1000 nanometres and there is no precise "cut-off" at these limits. Both smaller and larger particles retain certain colloidal properties, the smaller ones behaving increasingly like molecules and ions whereas the larger ones behave increasingly like ultra-fine granular material. Within the colloidal size range the material is semi-solid (visco-elastic with a Bingham yield point) with an almost infinite variety of structures, flocs, packing arrangements, particle clusters and 'chains', etc.

These 'colloidal processes' which develop concretions are not dependent on chemistry. They require a diffusive gradient in a 'weak wet' open meshwork gel in which there are pre-existing small clusters or larger nuclei of 'close packed' particles where van der Waals attractive forces are drawing them strongly together (syneresis) so that the mass exudes electrolyte which will precipitate additional diffusing particles at their surface.

The conclusion is that the concretionary process can only occur in aqueous particle systems.

## PTYGMATIC QUARTZ VEINS

### *Examples of ptygmatic quartz veins*

One of the clearest indications that the formation of ptygmatic veins involves colloidal precursors and a soft yielding host material is the occurrence of ptygmatic chert micro-veins in a laminated chert matrix as shown in Figure 70. This structure is due to physical differences (like viscosity) between the vein precursors and that of its matrix which is almost identical chemically. Both the ptygmatic micro-veins and their host chert are metacolloidal and have obviously been derived by the dehydration and crystallisation of sedimentary polymeric silicic acids.

The 'loopy' nature of the several successive injections of vein chert into its laminated or bedded host precursor ooze, is due to an inherent property of colloidal silicic acid. This has a viscosity dependant on the rate of flow during its remobilisation and injection into the veins. The successive thixotropic liquefactions and rapid re-gelling to preserve the 'loops', which is indicated by later ptygmatic veins cutting several earlier generations, are also a characteristic of colloidal systems.

Figure 71 shows a ptygmatic quartz vein in drill core from the New Cobar Mine at Cobar, N.S.W. This vein cuts across the cleavage in its host shale yet in some places it has been offset by further movement on the cleavage planes. The precursor polymeric silica in the ptygmatic vein appears to have been still mobile at the stage of this later minor movement as small 'spurs' and offshoots from it project into cleavage planes similar to those in Figure 72. However, shrinkage cracks or breaks in the internal quartz of this ptygmatic vein itself show no relation to the cleavage in its host shale.

A soft shale host is indicated by the double ptygmatic quartz vein intersected in the drilling at Orlando Mine, Tennant Creek, NT. These veins are not in parallel straight fissures subsequently folded because the 'loops' in the folding are quite discordant and the thickness of the shale material between these adjacent veins is quite variable. The veins also contain abundant fragments of their chloritic host sediment and at the time of intrusion they were clearly able to flow into the soft precursor paste in any direction. The whole system appears to have been under hydraulic pressure which was carrying the load as a water-rich diagenetic gel and it is inferred that only a small differential pressure was required to intrude these veins (Figure 73, Orlando).

The detail of most ptygmatic quartz veins shows that they are post-cleavage as in Figure 71 and have micro-veinlets or apophyses which extend out into the fabric of their soft precursor host rocks. The bedding planes

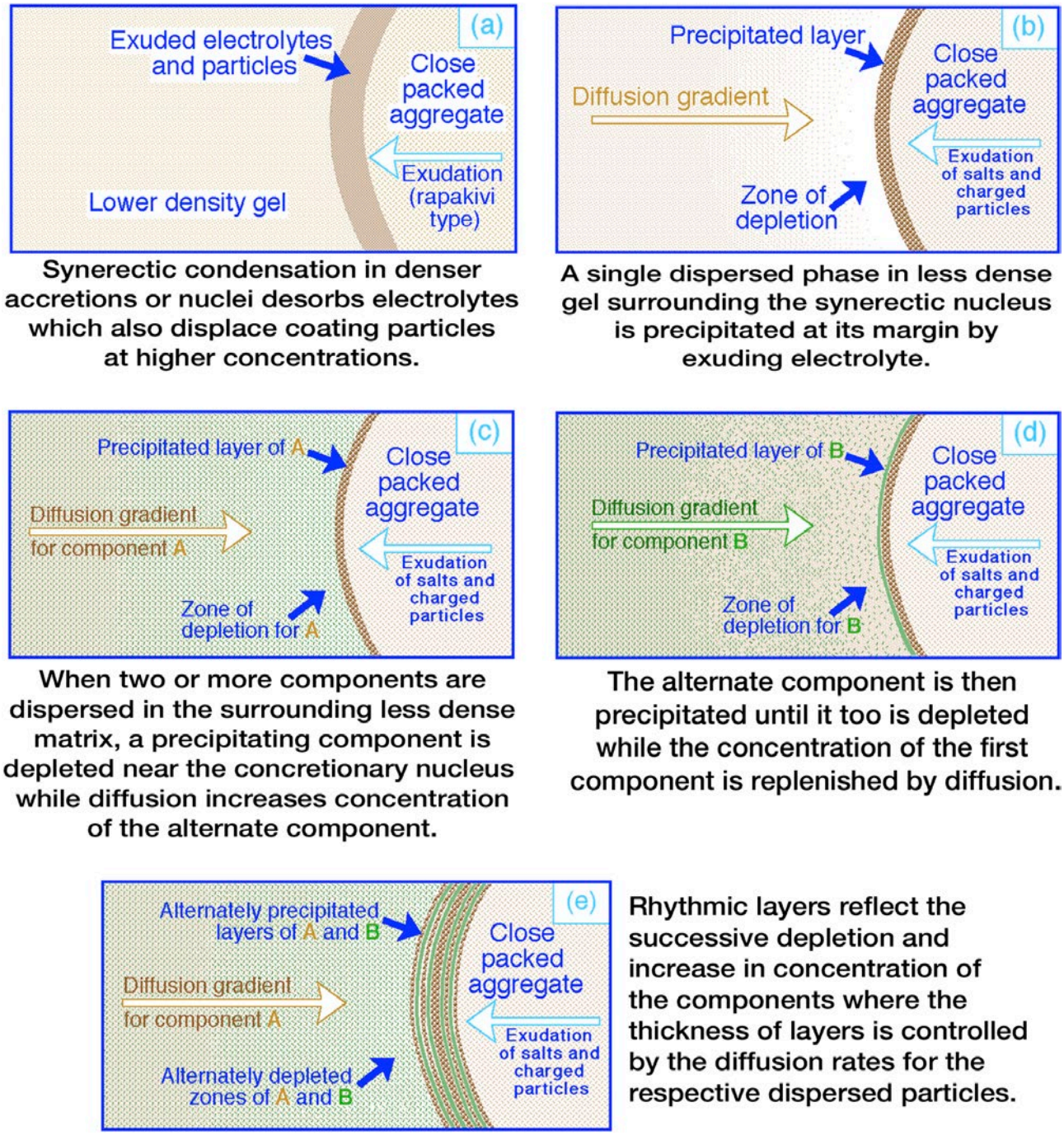



**Figure 68.** A schematic diagram indicates the exudation from synerectic nuclei of desorbed electrolyte and 'surface coating' charged particles which precipitate concretionary overgrowths or rimming bands. The principal stages in the saturation - precipitation - depletion - diffusion and re-saturation cycle that gives rise to rhythmic layering are indicated. Essentially the mechanism by which the diffusing particles are precipitated at the surface of the growing nucleus, is by coagulation. They encounter an increase in the concentration of electrolyte diffusing outward from the nucleus. This reduces the thickness of the Helmholtz double layer so that they can 'plate out' on the surface.

in this shaly dololutite from drill core at Woodcutters Mine, NT., are 'wavy' having responded to the soft diagenetic deformation at the time the ptygmatic quartz vein was intruded.

Slobodskoy[9] (1970, p. 449) describes monomineralic ptygmatic quartz veins which merge along their strike

**Figure 69.** A very large calcareous concretion has weathered out of soft shales. It is at Moeraki, south of Timaru, New Zealand. Rapid shoreline erosion of the bank of soft shale is exposing a number of these concretions which are left on the beach as the "Moeraki Boulders". The size of the concretion indicates that the process is self-perpetuating and can continue as long as hydroxy-carbonate particles can diffuse to its surface.


**Figure 70.** Ptygmatic micro-chert veins in a laminated matrix chert show the 'loopy' nature of the fluid precursor silica gel injected into a precursor ooze. Both matrix and ptygmatic veinlets are metacolloidal chert so the difference at the time the precursor was injected is in the gel structure or consistency - plastic gel and non-Newtonian fluid. The shear thinning property of the fluid silica polymer causes the looping and bulging. This specimen from Richie Lake[17], Sokoman Iron-formation is illustrated by Dimroth and Chauvel (1973, p. 122).

into ptygmatic quartz-oligoclase aplitic veins. He points out that crossing veins do not necessarily displace each other, and that the folding is such that the walls would not fit back together if the vein material were removed. The host rock has had to yield.

**Figure 71.** The detail of a ptygmatic quartz vein in the core of the New Cobar Mine at Cobar NSW, shows that it cuts across the cleavage yet it has been offset in places by further movement on the cleavage planes. Precursor polymeric vein silica also appears to have been mobile at this stage as small 'spurs' and offshoots of it project along the cleavage planes.

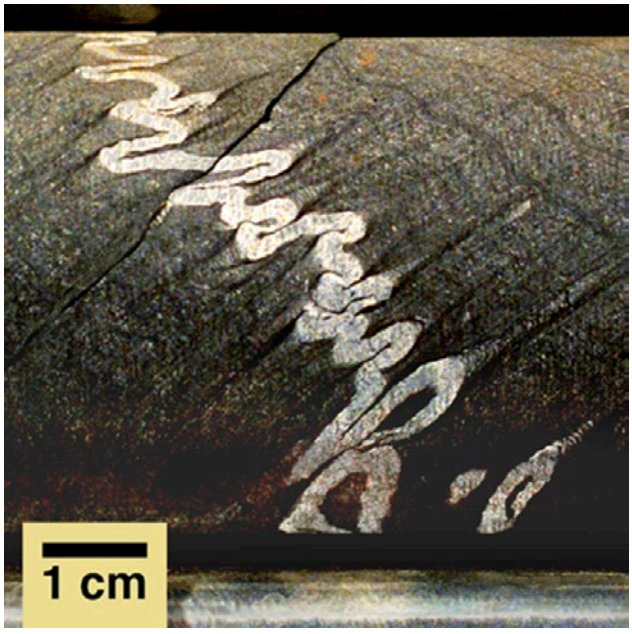



**Figure 72.** A ptygmatic quartz vein intrudes approximately parallel to the bedding in soft mobile dololutite. The ptygmatic loops have diffuse "tails" of quartz extending from them into the de-watering cleavage indicating that the vein was emplaced in the soft sediment before compaction and de-watering was completed. Woodcutters Mine, east of Batchelor, Northern Territory.

*The rheology of ptygmatic vein material*

This rheology, that is thixotropy, rheopexy, and viscosity dependant on the rate of flow, is quite specific to particle systems which have a Bingham yield point and where their cohesion depends on glue-like particles "sticking" together (interparticle attraction without chemical linkages).

The 'loopy' discordant folding itself arises where flow round the outside of the loops or through the

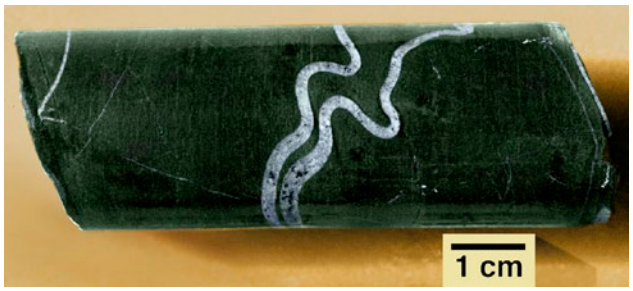



**Figure 73.** These two diverging and 'wandering' ptygmatic quartz veins in chloritic shales were intersected in drilling at Orlando Mine Tennant Creek, NT. The veinlets were apparently intruded when the diagenetic sediment was water-rich, soft and carrying at least some of the overburden load by hydraulic pressure. Small differential pressures appear sufficient to intrude the fluid precursor silica gel.

'necks' or constricted parts of the vein is more rapid and therefore more fluid (lower viscosity) than on the inside of the loops or in wider parts of the vein.

Such thixotropic particle systems re-liquefy as non-Newtonian fluids which have a viscosity dependant on the rate of shear (Mysels[10], 1959, p. 269; Van Olphen[11], 1963, p. 131; Yariv and Cross[12], 1979, p. 388). That is the viscosity diminishes as the shear or flow rate increases. This is subject to hysteresis which is the time taken for the viscosity to reach equilibrium at a particular rate of shear, but the overall effect during pasty flow is one of shear thinning behaviour. The paste is more fluid and mobile wherever it is moving faster. This is the reason for the in-welling and bulbous pod-like intrusions. Once re-mobilised, the precursor material pours in until it is slowed by equalisation of the small pressure differentials. When flow rates decline to the critical rate at which rheopectic re-setting re-establishes the gel condition, flow stops abruptly and the whole structure is preserved for eventual lithification.

The intrusion of a vein or a paste of polymeric silica particles as a fluid into a homogeneous plastic yielding host is indicated in Figure 74. It is the variable viscosity across the intruding sheet as it flows round a curve which causes it to loop sinuously back and forth down the slight pressure gradient. The amplitude of the loops is related to the thickness of the pasty sheet intruding and the differential viscosity which varies with the rate of flow causes the vein to 'loop' or zig-zag continuously as it is injected into a uniform soft medium. It behaves in the same way as a thin stream of cream does as it is poured slowly from a jug. For non-Newtonian fluids the phenomenon is called 'shear thinning'.

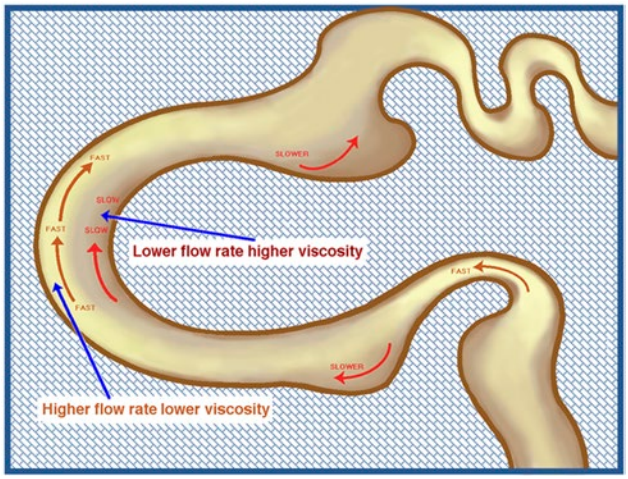



**Figure 74.** Diagram of an intrusive ptygmatic vein comprising a pasty mixture of particulate sediment components like clay, polymeric silica, chlorite, etc. More hydrous and lower viscosity components tend to separate from the mixtures when they are mobilised. At higher rates of flow these more fluid materials are even less viscous. This causes their intruding veins to bend tortuously as they flow along pressure gradients, "jet" through thin connecting necks, well into bulbous forms, or diffuse in cloud-like masses into the plastic host rocks. The viscosity is dependent on the rate of flow.

*The stress fields indicated by ptygmatic folding*

The ptygmatic structures exhibit highly disharmonic, extremely tortuous folds which are distinguished from shear folds by the fact that no relation exists between the thickness of the folded material with respect to crest and limb of the folds. Their limbs can be thickened and the crests thinned whereas the reverse is true of shear folds. There are also indications of blastic crystal growth where the borders of ptygmatic veins are blurred by subsequent crystal growth across their margins.

Where two or more ptygmatic veins cross each other, they do so without any tectonic interdependence; that is, forces which might have been responsible for the intense fold pattern in one vein, are inconsistent with those needed to fold the next crossing ptygmatic vein. These in turn are not consistent with the folding in the third, or fourth, etc. (Figure 70). In this type of random discordant folding any analysis of the stress fields which may have caused the folds can only be interpreted by isotropic or fluid deformation because it is not possible to establish solid phase transmission patterns of mechanical vectors. Any deformation pattern in a solid phase must be anisotropic. The deformation in this case arises within the pasty semi-solid vein material itself as it is injected into its soft and yielding host.


**Figure 75.** A ptygmatic sediment dike invades and transects the Greta coal seam apparently prior to its consolidation. The material of the dike is sandy with abundant clay, and close examination has shown that much of the rounded granular fine quartz is accretionary. Wet sediments trapped under gelatinous semi-consolidated coal apparently gave rise to this ptygmatic dike. Pelton Colliery, Newcastle, New South Wales.

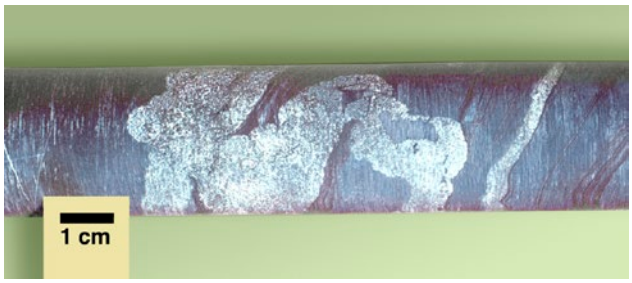



**Figure 76.** A ptygmatic-type clastic mud dyke or "dirty early quartz vein" has been intruded into soft diagenetic banded hematite shales. These early "out-wellings" of entrapped fluid through fine oozy sediment are episodic with each successive reactivation becoming more quartz rich until the filling material resembles vein quartz. Explorer 8 drill core, Tennant Creek.

These observations, and the fact that many tightly folded ptygmatic veins "go straight" for a segment (Figure 75), or have each branch folded independently when they fork into two or more branches, make it difficult to accept any hypothesis that the veins were injected as straight veins and then heaped into the highly tortuous fold pattern as the host rock and vein collapsed together.

Many others, such as Sederholm[13], 1907, p.89; Holmquist[14], 1920, p.212; Niggli[15], 1925, p.14, have suggested the veins were folded during their "injection", that is, in a liquid state through flow movements relative to a plastic host rock. In this view the veins were never straight and their curved rather meandering structure resulted from the tendency of the fluid phase to follow the paths of least resistance within the soft and yielding host rock. They often show complex over-folded structures which could never be formed by the opening of fissures. It is also observed that thick layers produce folds of larger amplitude than thin ones and it has been pointed out that during re-mobilisation a mass transport of material from the host rock into the ptygmatic veins is indicated.

In reality there is no question of the random isotropic nature of ptygmatic folds, Part 1 Figure 30, Figure 70, and Figure 76. The pasty vein material is "heaped in" on top of, and transecting or adding to, earlier vein material. Completely random folding of the type described by renowned Australian professor S. W. Carey as "chicken guts" folding can be developed. There is very clear evidence that the previously injected ptygmatic vein retained the same plastic character as its host material. It responds similarly when re-injected by a later ptygmatic vein as in Figure 70. This is a conspicuous example of ptygmatic injection of chert veins into a soft laminated chert host.

*Ptygmatic veins develop in soft pliable pre-crystalline materials as their hosts*

In many cases there is clear evidence that ptygmatic veins have been injected into soft precursor host materials. Most of them are into sediments at the diagenetic stage (syn-diagenetic) or they intrude into soft walls of newly deposited mineral matter associated with hydrothermal zones. In some cases, there is quite definite evidence that the host materials were at the soft precursor stage when the veins were injected.

Figure 75 is a ptygmatic clay-rich sediment dyke in the Greta coal seam photographed in the workings of the Pelton Colliery at Cessnock. The coal was obviously soft and unconsolidated at the time of its intrusion. There is no question that ptygmatic veins occur in soft sediments. This is shown by the intrusion of mud dykes and 'muddy' quartz veins through haematite shales near Juno Mine in Tennant Creek, Part 1, Figures 30, and Figure 76.

A ptygmatic chlorite vein meandering across slate cleavage and merging with chlorite in its host Appalachian shale-slates is illustrated by Weaver[16], 1984, p. 49. It is reproduced here as Figure 77.

In Figure 78, Dimroth and Chauvel[17] (1973, p. 122 record a small ptygmatic chert vein injected into greenalite in the Sokoman Iron Formation at Lac Helluva, Labrador. Needle-like minnesotaite crystals have nucleated on the vein margins and developed by growth of the greenalite out into its surrounding host. These spiky crystals clearly developed in the soft host after the ptyg-


**Figure 77.** During diagenesis the highly fluid pale green chlorite content of the Conasauga shales was able to flow under a small differential pressure into a ptygmatic veinlet as illustrated by Weaver (1984, fig. 38). Chlorite in the veinlet, possibly only a little 'cleaner', is the same as that in the fabric of the diagenetic mud from which it merges and oozes out into the vein from a series of interlayer offshoots.

ment with model substances, but in the case of natural sediments (Part 1, Figure 30) it is water rich mud trapped under the gelatinous ferruginous shale that is light enough to well upward into the plastic host materials under a very low-pressure differential.

These enigmatic ptygmatic veins are not an enigma if the material of the vein is intruded into soft host material as a thixotropic paste.

*Ptygmatic veins of other metal oxides*

The ptygmatic vein form is entirely dependent on the rheology of the intruding material and its soft precursor host. Similar ptygmatic veins and vein hosts


**Figure 78.** Round a stylolitic quartz vein from the Sokoman Iron Formation, Lac Helluva, Labrador, Dimroth and Chauvel[17] (1973, p. 122) show minnesotaite needles nucleated at right angles to the sinuous convolutions of the vein. The vein, by facilitating water release, has nucleated the crystallisation and it was clearly emplaced prior to the time the minnesotaite crystallised from its surrounding hydrous ferromagnesian precursor.

include mica-schists (Figure 77), slates (Figure 73), cherts (Figure 70), banded iron formations (Figure 79), coal seams (Figure 75), and haematite shales (Figure 76). Ptygmatic veins of single minerals or where it appears that involved eutectics could not be invoked to lower 'melting points', include quartz veins (Figure 72), calcite veins, chlorite veins (Figure 77) haematite veins (Figure 80), magnetite veins (Figures 79 and 81), and mud (Part 1, Figure 30).

In Figure 80 a highly sinuous and 'loopy' vein of haematite is illustrated where its precursor ferric hydroxide has injected into previously soft chlorite near the margin of the One-Oh-One lode near the Orlando Mine at Tennant Creek. This vein is also complexly branched, thickened, and apparently remobilised several times. The bladed haematite crystals are in random clusters which have no relation to vein walls, its flow during intrusion, or any regional stress field.

Micro stylolite-like ptygmatic veinlets of magnetite (Figures 79 and 81) occur in the Mary Ellen Mine in the Biwabik Iron Formation, Mesabi District, of the Lake Superior iron ore province (van Hise and Leith[19], 1911). The ptygmatic magnetite veinlet in Figure 81 occurs among haematite oolites in cherts which clearly indicate a very soft diagenetic environment because the ptygmatic magnetite veinlet cuts their margins. Some precursor internal amorphous silica has also been re-mobilised

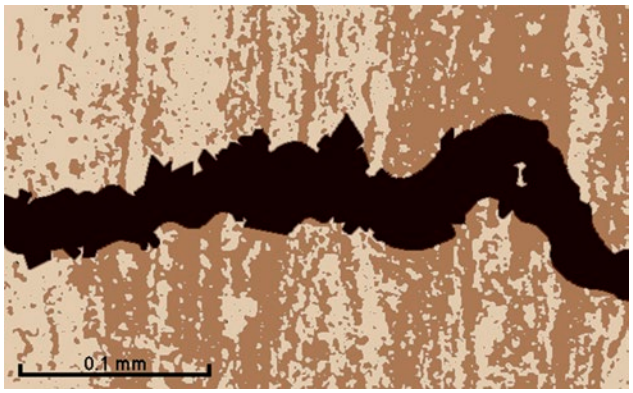



**Figure 79.** A ptygmatic micro-veinlet of magnetite (black) cutting chert (white) and hematite (grey) bands is from the Mary Ellen Mine, Biwabik, Minnesota. The magnetite octahedra projecting across the veinlet boundary clearly indicate that crystallization is later and that the magnetite had an intrusive non-crystalline precursor with shear thinning properties as indicated by the ptygmatic folding. A thixotropic ferric hydroxide vein precursor is unmistakably indicated. (From Lougheed[4], 1983, p. 339.)

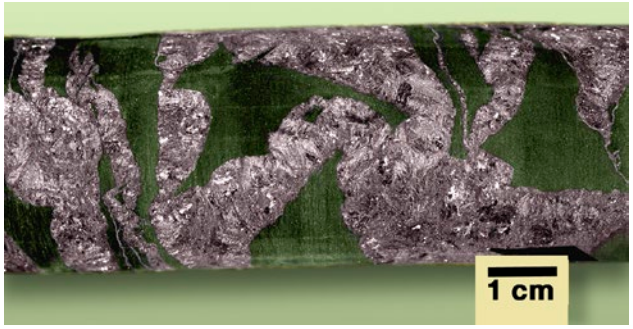



**Figure 80.** A complex ptygmatic haematite vein system has intruded into the marginal chloritic sediment at the One-Oh-One prospect near the Orlando Mine, Tennant Creek. The massive crystalline specularite contains slivers of chlorite wall-rock and the ptygmatic folding indicates that the precursor ferric-hydroxide was colloidal at the time it was intruded (viscosity dependant on the rate of flow).

within them, probably in syneresis shrinkage cracks, and this intrusive infilling silica itself has assumed a slightly ptygmatic form.

The obviously once fluid micro-veinlet of magnetite in Figure 79 also cuts banded chert - haematite unit from a banded iron formation at the same location. These occurrences indicate clearly that the mobile iron hydroxide precursors do not react chemically with their polymeric silica (chert) hosts.

As Lougheed[4] (1983, p. 339) claims, they are evidence for the mobility of the iron oxide now crystallised as magnetite. It could not be suggested that the ptygmatic micro-veinlet amongst the oolites or crossing the chert laminae was injected as fluid molten magnetite, as a precipitate from meteoric weathering, or as magnetite in some solid crystalline state. The magnetite crystal growth across the margin of the veinlet in Figure 79 indicates that the crystallisation is later. What then are the alternatives? Clearly the rheology and shear-thinning properties of the ferric hydroxide precursor gels are needed to interpret these observations of vein development and to account for the diagenetic mobility of the magnetite precursors.

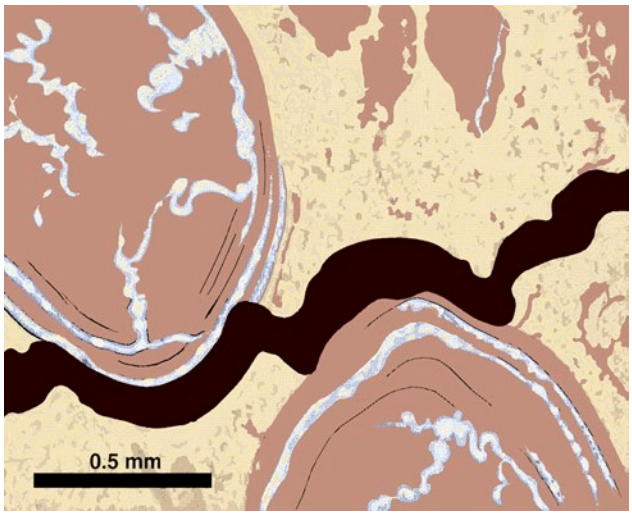



**Figure 81.** A stylolitic veinlet of magnetite cuts the margins of hematite oolites in a chert matrix at the Mary Ellen Mine, Biwabik, Lake Superior province. These hematite oolites have well developed syneresis shrinkage cracks filled with quartz which also extends into 'saddle reefs' or layers within the rim zones. The remobilised silica gel precursor has been intruded (ptygmatically in places). A ptygmatic magnetite veinlet cutting soft precursor chert and hematite oolites is very clear evidence for viscosity dependent on the rate of shear (flow). Recorded by Lougheed[4] (1983, p. 339).

## THE ENHANCED GROWTH OF CRYSTALS IN PRE-CRYSTALLINE VEINS

*The crystallisation of vein minerals is after vein emplacement*

It has always been held that the crystalline minerals observed in veins were formed after the vein had been emplaced and established as a hydraulic fracture or opening in the rock. The clear evidence for this, particularly in sinuous veins, is that the crystal growth is often nucleated on the vein walls and the orientation of the crystals follow round the sinuous curvature of the vein wall always projecting towards its centre. The vein walls

were there and the vein opening established before the crystals grew inward from them as in Part 1, Figure 47.

Whatever mechanism of vein emplacement could be envisaged, this unequivocally implies that the vein was filled with some precursor fluid or substance from which the crystals subsequently grew. The crystal growth was sometimes insufficient to fill the vein space leaving vughs or chains of crystal cavities in the centre of the vein. These crystal cavities are like those in condensing and crystallising chert geodes as illustrated in Part 1, Figure 26.

The evidence for the growth of crystals inward from condensing colloidal precursors as in geodes, is also often seen in veins as in Part 1, Figure 43 where the crystals grow inward from banded colloform silica near the margin of the vein. Occasionally colloform silica fills the whole vein space as in Part 1, Figure 46.

*Enhanced crystal growth in precursor vein gels*

The colloidal nature of vein precursors is further indicated by enhanced or surface catalysed crystal growth. Crystals growing in gelatinous media, such as soft vein precursors like polymeric silica gels, are supported within this soft material. Crystals can develop directly within the material itself or by diffusion of soluble ions or small particles through the media to the growing crystal face. For example, quartz crystals grow within vein precursor silica gels by diffusion of the very small and mobile $Si(OH)_4$ species to 'infill' the condensing gel meshwork by the simple exothermic chemical reaction:


Frequently the gel coating on a particular crystal face of a growing crystal will catalyse growth on that face to the detriment of other crystal faces of differing crystallographic orientation. When this occurs on very small newly nucleated 'seed crystals', it can result in spectacular needle-like crystals as one of the very small faces grows to the exclusion of all others. Fine acicular rutile needles which are often found in clear vein quartz are a good example (Part 1, Figure 27). Fine needle like or bladed crystals of dawsonite, chlorite, haematite, or magnetite also commonly reflect the enhanced crystal growth of the included mineral in the gel of the vein precursor. Chlorite crystals grown from wall rock material included in vein quartz from the Wagga Tank Prospect drill core are illustrated in Part 1, Figure 42.


**Figure 82.** Radiating acicular hematite needles have grown out like spikes into a gelatinous dolomite precursor of the lode at Juno Mine, Tennant Creek. This 'wild growth' from a central nucleation point is typical of bladed and spherulitic crystal patterns developed in gelatinous media where enhanced (catalytic) crystal growth on one of the small seed crystal faces proceeds almost to the exclusion of growth on other faces.

Where many crystals are nucleated together such as on the surface of a small synerectic 'gel ball' like an accretion, framboid, or oolite set within gelatinous media, spherulitic growth (Part 1, Figure 45), 'wheat sheaf' structures, or axiolitic outgrowths can develop. A good example of this type of enhanced crystal growth in gels producing a radiating acicular texture is the slender haematite needles developed in the gelatinous dolomite precursor of the Juno Mine lode at Tennant Creek (Figure 82).

Because of the catalysis on a single specific face of the 'seed crystal', needle-like crystals of many different minerals and chemical compounds have been grown experimentally in polymeric silica and other gels to demonstrate this phenomenon (Henisch[20], 1970).

Sometimes the catalytic crystal growth enhancement induced by the host gel for the developing crystal occurs on two of the 'seed crystal' faces. In this case bladed or platy crystals develop. Bladed magnetite crystals in quartz from Peko Mine are shown in Figure 83.

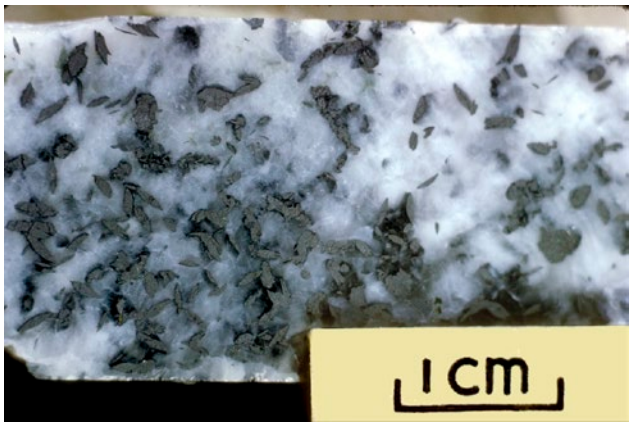



**Figure 83.** Catalytic crystal growth enhancement sometimes occurs on two faces of the original seed crystal resulting in growth of a bladed or platy crystal form. This magnetite in quartz from Peko Mine in Tennant Creek has probably grown from dispersed lepidocrocite to developed this bladed form instead of the more usual compact octahedra.

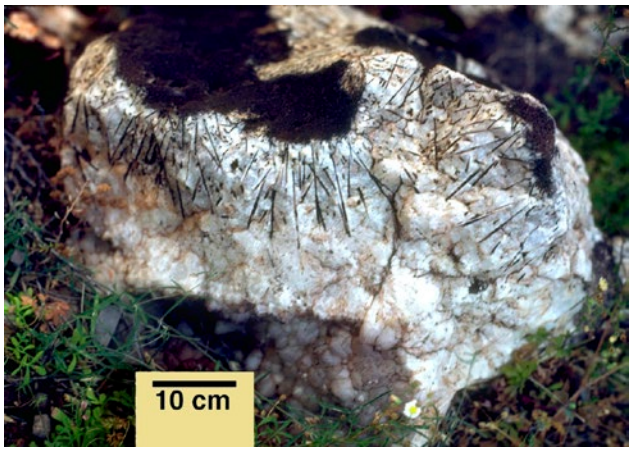



**Figure 84.** Colloform vein quartz from Mt. Gee, Flinders Ranges, South Australia, contains delicate acicular dawsonite needles. A precursor gel stage of the quartz is indicated by the enhanced growth of these needle-like crystals. They were supported in a yielding medium through which material for crystal growth could diffuse. Subsequent quartz crystallisation nucleates at, and radiates from, the pre-existing dawsonite needles.

These crystals most probably grew from a dispersion of fine iron hydroxide or hydroxy oxide particles like lepidocrocite [$FeO(OH).Fe_2O_3.H_2O$].

The silica host for the bladed magnetite must have been a polymeric gel in order to support the heavy magnetite crystals and to provide the catalytic coating which favoured crystal growth on two faces of the magnetite crystals. These are normally compact octahedra.

Magnetite and quartz rimmed oolites (Figure 58) occur in this same Peko main lode. These definitely indicate the dispersion of particulate hydrated iron oxide species.

Crystal growth in the gelatinous medium of the early vein material is enhanced because the gel allows diffusion of the crystallising species to the growing crystal face; it acts as a crystallising catalyst to the face it is coating; and the gel yields to support the delicate needle like crystal as it grows.

Figure 59 illustrates vein quartz from Mt. Gee, Flinders Ranges, South Australia, with scalloped crustiform and subspherical colloform concretionary textures see also Figure 60 and Figure 84. Here the same polymeric silica precursor gel has supported the delicate acicular dawsonite needle growth.

*Precipitation of gold in quartz*

The catalytic surface coating which enhances crystal growth in gels and facilitates the growth and development of well faceted crystals in gelatinous media directly from dispersed particulate species is quite significant. The phenomenon has been recognised for many years following the experiments of Hatschek and Simon[21] (1912) and Boydell[22] (1925) who demonstrated that gold sols diffused into silica gel would crystallise directly to metallic gold. In fact, silica gel appears to be a 'scavenger' for insoluble gold particles dispersed in aqueous fluid. The irregular and sometimes high concentrations of metallic gold in quartz veins is thought to be due to the 'quality', that is the gel structure and degree of hydration, of the polymeric precursor silica in the vein and to its maintenance for a long period during hydrothermal fluid seepage so that very low trace levels of gold particles are concentrated.

Hatschek and Simon's[21] experiments have since been repeated and further investigated by Henisch[20] (1970, p. 19). Henisch points out that the gel structure has some degree of long-range ordering which is why crystal faces of different orientations and having a different relationship to this ordering, are "catalysed" to different extents. Stumm[23] (1992, p. 218) confirms that a surface catalytic effect is observed when the surface of the substrate or gel coating "matches well" with the crystal to be formed.

*Crystal growth requires particles to move to the surface*

Removal of a mobile species from dispersion in the gel by its crystallisation creates a diffusion gradient towards the crystal surface. Henisch[20] (p. 51) notes that convection currents are suppressed in gelatinous media and therefore movement must be essentially by diffusion but a very important function of the gel media is also to suppress nucleation. Without growth on many closely

spaced competing nuclei, the faces of crystals supported in the gel medium are supplied by a steady diffusion of particles or ions so that large and well-formed crystals are able to develop.

Since enhanced crystal growth occurs in a number of different types of gelatinous materials and involves many different particulate and ionic dispersed species, it is very doubtful that the actual catalytic action is fully understood. Healy[24] (1969) grew galena crystals from PbS sols in silica gel and from these experiments he was able to discern that in many cases small acicular tubules, not necessarily of PbS, first developed and that along the surfaces of these minute hollow rod-like structures, small perfect cubes of PbS developed. This type of surface nucleation and growth of the crystals in gels appears to be similar to the radiating quartz spherulites developed round dawsonite needles at Mt. Gee in South Australia.

Any substance precipitating from solution must first form molecular species which must then come together with other molecules to develop a crystallite nucleus. Stumm[23] (1992, p. 217) points out that if this nucleus is smaller than a single unit cell, the growing crystallites are most likely to be amorphous. This is because these nascent or part formed lattice units of a crystal must have 'broken' or 'unsatisfied' lattice bonds. Such a cluster of crystalline molecules less than a unit cell in size has enormous surface energy and is highly charged by virtue of its unsatisfied lattice points. Such exceedingly small charged particles in an aqueous system are immediately hydrated with dissociated water or "coated" with polar water as an adsorbed water monolayer. It is this water of hydration, and any adsorbed water monolayer which creates the metastable 'gel' condition of the material.

The very small particles if they are formed slowly enough or at very dilute concentration form a metastable hydrate (critical cluster) which then impedes further crystallisation until the particles aggregate into synerectic clusters which can spontaneously desorb water by drawing together under van der Waals strong interparticle forces of attraction.

Silica differs from the other metal oxides, hydroxides, sulphides, etc. in that in its dispersed or 'soluble' form the neutral monomeric $Si(OH)_4$ species predominates. Because the ionisation constants of the polymers are greater than that of the monomer it reacts more rapidly with the dimers and higher polymers than with another monomer. Quartz crystal growth within precursor silica gels in quartz veins is therefore by diffusion of the monomer to react at the face of the growing crystal with surrounding polymer so that siloxane linkages then form the developing lattice.

## REPLACEMENT

### *Replacement criteria*

'Replacement' or 'metasomatism' are often routinely used with little regard to the physico-chemical process by which the 'replacement' of one mineral with another is actually accomplished.

To accomplish such a 'replacement', certain requirements are obvious. New matter in the form of ions, radicals, molecules, or very small particles must be moved in to the replacement site. There must be a mechanism for the exchange of material at this site with the new material arriving and the 'replaced' matter must be moved out.

One of the first papers to logically address these requirements and to consider processes involved in effecting 'replacement' was Holser[25] (1947). He points out that the replacing substance must have access to all parts of the material being replaced and says that most replacement must therefore be accomplished in openings of the very smallest size if it is to be complete. The types of openings he considers are:

1) Supercapilliary - that is cracks, fissures, openings, intergranular pores, etc. in which fluids can flow and effect transport of dissolved or suspended matter over relatively long distances.
2) Capilliary - (a) Open highly hydrated gel meshwork with particles in 'loose' or open structures such that fluid flow is restricted and solutes and sols move by diffusion.
   (b) Dense strongly crosslinked gel meshwork in which fluid movement is mainly by diffusion and movement of sols and solutes is restricted to vary markedly according to particle size, shape, and charge (gel chromatography).
3) Grain boundaries, hydrated crystallites, or incomplete lattices where some diffusion of fluids is possible and fluids and ions can be transferred by reversible "shuffling" reactions like adsorption-desorption reactions.
4) Intra lattice spacings where very limited movement of ions occurs within the lattice itself such as the slow migration of entrapped silanol terminations within a strained natural quartz lattice (White[26], 1971).

No substantive 'replacement' would be possible within a crystal lattice because of the inability to diffuse many larger ions or particles either in or out. Therefore, if observations suggest replacement of a whole mineral or rock volume it must have been porous and gel-like so that the replacing ions, molecules, or particles could gain access to all parts of it. Similarly, when replacement occurs, the material substituted must remain gel-like at least for some time so that the replaced ions, molecules,

or particles can continue to get out.

Considering the various aggregates, flocs, tactoids, or curd-like particle associations and the three orders of magnitude in the size range of colloidal particles themselves, the size of pore spaces probably extends over four orders of magnitude or more. There is clearly a gradient from open highly hydrated sediment gel meshwork to dense strongly crosslinked and semi-ordered 'close packed' particle aggregates. Openings the size of those in densely crosslinked gels would severely limit or preclude fluid flow and therefore restrict fluid transport of any sort. As Liesegang phenomena so clearly indicate, ion or small particle migration in such aging or compact gels is controlled by diffusion.

Holser[25] (1947) points out that matter may be transported in solution or suspension and that such fluids may hydrolyse, disperse, disaggregate, or mechanically disintegrate particles and particle clusters. The surfaces in turn may precipitate, adsorb, or settle material into the spaces left by these processes.

If the requirement of metasomatism is to "change the body" of the material being replaced, then the "openings of the smallest size" must give access to all parts of the body being replaced. It must be like a sponge with very small pores. Consequently, the replacement can only be achieved by ions, molecules, or very small particles. A gel or the gelatinous precursors of a wide range of ore and rock minerals which "soak" for very long periods of time in ore forming fluids, are indeed susceptible to 'replacement'.

Transport by fluids is both by stream flow or seepage and by diffusion. Ions and charged particles diffuse down a concentration gradient. Thus, if an ion or charged particle is precipitated or adsorbed from the fluid, further ions or particles will move along the gradient so created towards the place of precipitation. In the case where the diffusing ion or particle is exchanged for one then released by substitution, the concentration gradient in respect of these newly released ions or particles will be in the reverse direction and they will diffuse outwards. Diffusion is therefore particularly appropriate for replacement. It allows movement of ions or particles to the replacement site and removal of those displaced. This was recognised by Lindgren[5] (1933, p.177): -

*"The great importance of diffusion is probably in the mechanism of replacement. In the ultimate small spaces available for metasomatism, there is constant change of concentration; and diffusion attends to the moving up of the new molecules and the removal of the by-products of replacement. Diffusion acts easily in a gel."*

Chemical reactions affect the rate of diffusion and may stop or impede it (as in the case of Liesegang rings) but Holser[25] (1947) concludes that diffusion is much more effective than fluid flow in metasomatism. Four types of replacement are recognised:

1) Filling the spaces occupied by fluid within the original pore space or gel meshwork.
2) Displacing or pushing aside the existing gel structure by the formation of a precipitated or denser structured gel as ions or charged particles diffuse to and expand the nucleus around which they are precipitating (like oolite growth, etc.).
3) Infilling the original gel meshwork by diffusion of a disperse phase to the surface of this "infilling volume" or replacement front (like the diffusion of monomeric or oligomeric silicic acid into its polymers to make denser gels).
4) Replacement of units in the original gel meshwork by a chemical reaction or by radicle or particle exchange which results in formation of a new gel.

These types of 'replacement' correspond to the normal behaviour of dispersed species within colloidal materials:

a) Coagulation or auto-precipitation of the metastable sols in open spaces.
b) Concretion.
c) Infill concretion.
d) Chemisorption.

In each case the particles of the system are interacting with ions, other particles, and surfaces, to condense, reduce surface energy, and to progress toward the lower energy crystalline state.

## EXAMPLES OF REPLACEMENT

### *Silicification*

Solutions or gels of polysilicic acids are somewhat unstable in that these molecules or particles grow in size and aggregate into denser precipitates or gels. They accumulate in precursor quartz veins, in the interparticle spaces in mineralising systems and stock-works, in the pores of sediment gels, or in natural openings like the syneresis cavities in the centres of geodes. A well-known example is the 'secondary' quartz surrounding sand grains and filling the pores of sandstones so that quartzites result from their eventual crystallisation.

Significant volumes of sediment near veins and mineral channels (elvans) are "silicified", that is, infilled with 'secondary' or polymeric silica which attracts more of the mobile monomeric and oligomeric species. It then becomes denser and hardens due to the crosslinking and 'infilling' of the readily diffusible neutral monomer [$Si(OH)_4$].

This silicification and infilling of expanded pore spaces, veins, geodic cavities, etc. is actually achieved largely by chemisorption. After the "little balls" of colloidal silica have established a meshwork, the mobile silicic acids or the monomer 'add on' to the pre-existing polymers or disordered solvated crystal surfaces by chemical condensation reactions. However, in many cases the silica infilling is first achieved simply by the physico-chemical sol-gel transition of the very small colloidal or "little ball" species of polymeric silica precipitating as 'infill' concretions or building up on surfaces.

*The example of silica replacing wood*

While it is easy to envisage the infilling and creation of progressively denser silica gels in existing fluid filled pores and hydraulically opened veins, lode systems, etc., actual 'replacement' or 'metasomatism' which has also been described as "molecule by molecule replacement", requires the removal of the material being replaced.

The commonly observed silicification of wood with preservation of its complete cellular and structural detail is perhaps one of the simplest examples of this exchange as in Figure 85. Iler[2] (1979, pp. 88-91) in his discussion of the silicification of biogenetic materials points out that chemical degradation must occur before the resulting space can be filled by silica deposition. That is, the cellulose of the wood or log being petrified must decompose by hydrolysis (gelatinise or simply "rot") before the replacing silica could diffuse to all parts of it. Only the short chained oligomeric or monomeric silicic acids can diffuse within the gelatinous hydrocellulose mass and larger colloidal silica polymers cannot pass through the cell walls. For this reason silicification for the most part involves the monomeric $Si(OH)_4$ which must precipitate as a gel to allow continuing diffusion within the specimen.

Cellulose or polycellulose $(C_6H_{10}O_5)x$, the main constituent of wood, is well suited to such replacement by the mobile silicic acids. The repeating basic sugar-like chain of six carbon atoms has a an oxygen atom forming a cyclic link between the first and fifth carbon in the chain and this repeating group contains four hydroxyls. It can hydrolyse or 'decay' to several intermediate acids like humic, tannic, talonic, idonic, mannoic acid, etc. and the simpler tartaric and acetic acids also probably occur among the breakdown products. The basic chemisorption or condensation reaction with silica is:

$$(C_6H_{10}O_5)(C_5H_9O_4) - O\boxed{H + HO} - Si(OH)_3 \longrightarrow (OH)_3SiO(C_5H_9O_4)(C_6H_{10}O_5)_x + H_2O$$

**Figure 85.** The trunk of a giant redwood tree felled and buried by the ash of a nuée ardente some 3.4 Ma ago is now completely replaced by silica and almost re-exposed by erosion (some excavation). Original cellular structures are largely preserved in this 'petrified forest' some 12 km due south of Mount St. Helena in the Napa valley of California.

The natural decay of cellulose also results in an efflux of $CO_2$ as well as the various organic acids. Finally, the silica also displaces the carbon forming -Si-O-Si-O-Si- chains together with organic and hydroxyl radicals. However, as Iler[2] (1979, p. 89) points out the actual chemistry of this displacement of the complex organic radicals by chemisorption has not yet been properly determined. It appears to depend largely on their steric configuration and catalysis at the exchange sites. It is recognised that natural silicification of wood proceeds to various stages. In many cases the silicified wood contains unusual amounts of organic residues which can sometimes be demineralised, embedded, sectioned, and stained in a similar manner to living tissue.

Cellulose decomposes by hydrolysis and oxidation much faster than lignin and because of these differenc-

es in chemical stability, the cellulose can be replaced by the silica before the lignin is attacked. In all cases of this type of replacement by chemisorption, the removal of dispersed mobile silicic acid species is by its chemical bonding to the decayed wood structure. This lowers its concentration and creates a diffusion gradient towards the site at which it reacts. Similarly, the organic acids or radicals displaced increase their concentration within the petrifying wood. This creates diffusion gradients outwards for these organic species. There are many places where water coloured like tea with the tannic and humic acids can be seen seeping out from under sand-dunes containing buried logs and wooden debris.

Fossilised wood has been produced reasonably successfully in the laboratory by Leo and Barghoorn[27] (1976). Wood specimens were boiled in water to free them of gas and then subjected to a succession of immersions for long periods in ethyl orthosilicate at 70°C. Silicified wood closely resembling natural petrifaction of a geologically young age was produced. After treatment, residual organic matter in the silica filled specimens was removed with oxidising acid and the silica lithomorph was found to faithfully reproduce the original organic woody structures. Since the organic matter can be removed by oxidising agents, it is clear that the siliceous lithomorph is quite porous which, in the diagenetic environment, would make it an ideal substrate for further deposition of silica.

Laboratory duplication of natural chert which contains micro-organisms is very much easier because diffusion distances are much smaller. Oehler and Schopf[28] (1971) silicified specimens of filamentous algae by embedding the algae in silica gel and then autoclaving it at 2-4 Kb for 2-4 weeks at 150°C. Under these conditions the gel undergoes syneresis until completely solid then it crystallises to a cryptocrystalline cherty form preserving the algal filaments.

*The replacement of shells by opal*

The significance of the replacement of calcareous shells by opal is that this is direct and positive evidence of replacement by the "little balls", the actual charged particles of polymeric silica, which exchange for gelatinous hydroxy carbonate species. Particle exchange by hydroxide and sulphide charged particles can be just as important as ion exchange or 'molecule by molecule' exchange in effecting replacement.

The complete diagenetic hydrolysis or "gelation" of fossil shells is usually indicated by plastic distortion of the entire calcareous fossil (Figure 86), their 'welding' together and re-crystallisation, dolomitisation, or replacement by galena, pyrite, glauconite, or silica.

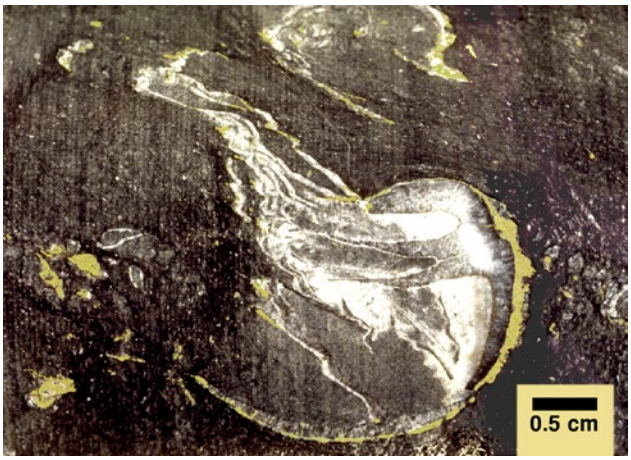



**Figure 86.** A distorted carbonate fossil is from pyritic black shales at Mt. Bulga N.S.W. The plastic or semi-fluid nature of the distortion indicates the way in which the original carbonate had been hydrolysed to hydroxy carbonate at the time of distortion. This is confirmed by its partial replacement by sulphide which requires hydrolysis (or gelation) before the sulphide particles can diffuse into the material to substitute for the hydroxycarbonate they are displacing.

The calcium ion, $Ca^{++}$, does not hydrolyse appreciably in aqueous solution because of its extraordinarily high hydration energy. However, in neutral or slightly alkaline solution it readily exchanges to reach an equilibrium with $Na^{+}$, $Mg^{++}$, and other ions so that a range of stable gelatinous mixed Ca-Na, Ca-Mg, and Ca-Al hydroxy carbonates are formed. Actual species in these amorphous mixed hydroxy-carbonates are not easily defined but a few of those which have been identified and named are listed in Table 2.

Occurrences of oolitic limestones (Figure 57) and large calcareous concretions in shales (Figure 69) clearly indicate the extensive hydrolysis of carbonates in the diagenetic environment and confirm the fact that equilibria in the pore fluid chemistry favours the hydrolysed species and stable gels for a long period of time during diagenesis.

It is important to recognise that substantial fossil shells like brachiopods and molluscs can hydrolyse completely in this environment and can therefore be replaced by silica as in Figure 87. In fact, Newell et al.[29] (1953), point out that in many cases the macro-fossils are selectively replaced by silica in preference to their limy matrix. Most readily silicified are bryozoans, tetracorals, tabulate corals, punctuate brachiopods, followed by non-punctuate brachiopods, molluscs, where replacement is usually spongy and imperfect, and finally echinoderms,

**TABLE 2. Hydrous Carbonate Minerals**

| HYDROXYCARBONATE (Hydrolyses to gel under aqueous conditions) | COMPOSITION (Formulation is usually indefinite due to highly variable hydration) |
|---|---|
| GAYLUSSITE | $CaCO_3.Na_2CO.2H_2O$ |
| PROTODOLOMITE | Gelatinous partly ordered hydrous Ca - Mg Carbonate. |
| PIRSSONITE | $CaCO_3.NaCO_3.2H_2O$ |
| ALUMINOHYDROCALCITE | Hydrated carbonate of calcium and aluminium. |
| GAJITE | Hydrous calcium-magnesium carbonate. |
| NESQUEHONITE | $MgCO_3.3H_2O$ |
| ARTINITE | $MgCO_3.Mg(OH)_2.3H_2O$ |
| HYDROGIOBERTITE | $MgCO_3.Mg(OH)_2.2H_2O$ |
| HYDROMAGNESITE | $3MgCO_3.Mg(OH)_2.3H_2O$ |
| DAWSONITE | $Na_3Al(CO_3)_3.2Al(OH)_3$ |
| TRONA | $Na_2CO_3.HNaCO_3.2H_2O$ |
| THERMONATRITE | $Na_2CO_3.H_2O$ |
| NATRON | $Na2CO3.10H2O$ |

foraminifers, and calcareous sponges where silicification is often limited only to the surface.

The hydrolysis which in most cases appears to be extensive and often complete, proceeds whether or not the fossils are replaced, or partly replaced by silica. This complete or extensive hydrolysis of fossil shells, calcareous remains, etc., 'gelatinises' the original microcrystalline carbonates which then coarsely re-crystallise when the sediments finally dehydrate and lithify.

Such fossils and debris distort or disintegrate if the sediments are disturbed while they are plastic. They 'weld' together and some develop rims like 'sooty' pyrite, glauconite, or haematite, and partial replacement by small patches of chalcedony in their centres or marginal to such fragments has also been observed (Chilingar et al.[30] 1967, p. 243). Remains like crinoid ossicles re-crystallise in optical continuity with patches of their limy matrix (Adams et al.[31] 1984, p. 44), or where the surrounding matrix is extensively re-crystallised in optical continuity with several crinoid plates or fossil remains, it is called "syntaxial overgrowths" (Adams et al.[31] 1984, p. 57).

The main point is that for replacement to occur, any fossil, buried log, shale layer, bedding laminae, or mineral precursor must be porous and gel-like so that the replacing ions, molecules, or silica particles can gain access to all parts of it by simple diffusion.

*Other examples of replacement*

A great many gelatinous mineral precursors can be 'replaced' at the diagenetic stage but actual replacement

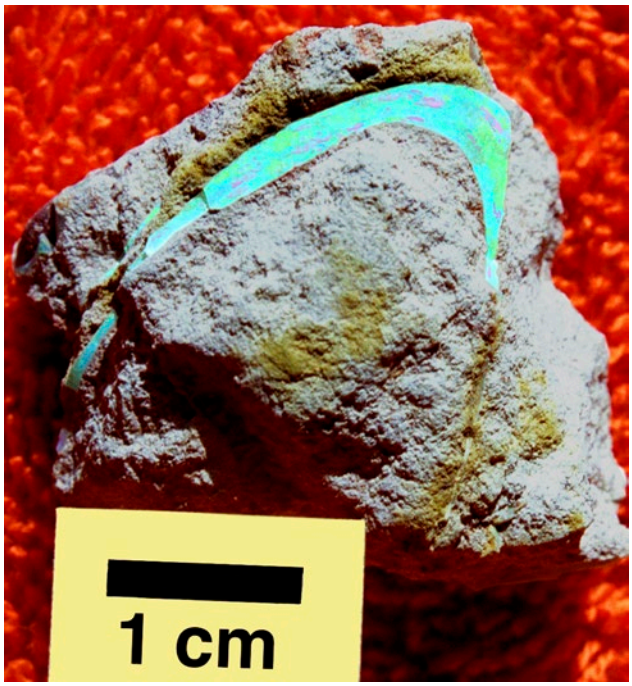



**Figure 87.** A calcareous fossil shell (brachiopod) has been completely hydrolysed to a gelatinous hydroxy carbonate into which silicic acid [probably $Si(OH)_4$] has been able to diffuse to entirely replace the original shell with opal. This specimen is from Coober Pedy in South Australia.

within such permeable media involves exchange of a mobile ion for one adsorbed; of an ion for a charged particle (macromolecule); or a charged particle for an ion or another charged particle; or of a chemical radicle or group for a similar chemical radicle or macromolecular group freed by the reaction at the exchange site.

This could be confused with infill concretion where material infilling the gel meshwork or particle interstices is merely added to available spaces. Replacement actually substitutes for, and liberates to the dispersed phase, some part of such meshwork or adsorbed particle system. Many minerals like the common precursor silica gels in quartz veins become denser by simple "one-way traffic" diffusion (only water comes out in the reverse direction) of additional molecular species into the consolidating mineral mass. The diffusion gradient is maintained by release of surface adsorbed ionic species or by chemisorption which simply releases water as a reaction product.

The most obvious examples of silica 'replacement' are those where biogenic material or fossils have been replaced. The two examples illustrated are:

- 5 km SW of Calistoga in the Napa Valley in California, or some 12 km due south of Mount St Helena, is a petrified forest where a number of large red-

wood trees (Sequoia sempervirens) were felled by an ancient (3.4 Ma) nuée ardente. The very large tree trunks buried in the volcanic ash have been completely silicified (Figure 85).

- Calcareous fossils have been selectively replaced by silica with more extensive replacements recorded in bryozoans, tetracorals, tabulate corals, punctuate brachiopods (Newell et al.[29], 1953; and Figure 87).

'Replacement' or silicification is a very real phenomenon which has been widely recognised for many years. The physico-chemical interactions by which it is achieved have remained somewhat obscure and ill-defined but replacement processes become clear and greatly simplified when the intrinsic properties of polymeric silica particle systems and the gelatinous mineral precursors of vein quartz are recognised.

## CONCLUSIONS

There is clear evidence for the conclusion that quartz veins are emplaced, not by 'hydrothermal solutions', but by gelatinous polymeric silica carried in the normal outflow of aqueous fluids and brines during diagenesis of sediments and to a greater extent from those that have been reliquefied as pastes and slurries. Fluids released during diagenesis carry suspensoids and sols of amorphous polymeric silica gel. A very limited amount of the transported silica is actually dissociated or ionic $Si(OH)_4$ at any stage.

Early in diagenesis fluids which escape from basin sediments tend to be more dilute suspensoids. These episodically 'break across' formations or form dykes and veins injected by hydraulic fracture from overpressured zones as the entrapped fluid-rich suspensoids containing the soluble and suspended matter thixotropically re-liquefy and break out.

The rapid re-setting or solidification which suspends wall rock fragments or contained mineral matter and which preserves the shape and form of the veins, is due to rheopexy of the hydrous polymeric silica. Once established, the crosscutting precursor veins, 'elvans', ore pipes, silicified zones, stockworks, lodes, stringer zones, etc. are permeable and continue to augment the general diffusion of fluids upwards and out of de-watering sedimentary materials.

Much, if not most of the silica accumulates in the quartz veins, pipes, or lodes, during this 'static' or 'slow seepage' stage. Pore fluids tend to be saturated with monomeric and oligomeric species of silicic acid and since these react preferentially with larger polymers rather than with other monomers or oligomeric species, the open meshwork or more fluid-rich gels infill and become denser. Each successive mobilisation of precursor quartz vein material is a thicker or denser gel. It becomes whiter and 'cleaner' due to rejection during fluid flow of other differently shaped colloidal particles.

The aqueous chemistry of silica involves not only the dissolution of mono silicic acid but a sensitive equilibrium over a wide range of parameters to the oligomers and polymerisation to colloidal silica.

The crystallisation of quartz in veins and lodes is from dense polymers during final dehydration and long after the veins were emplaced. Crystallisation of these dense polymers entraps fluid inclusions (including some oil and organic matter), Boehm lamellae, and helicitic structures. Residual hydroxyl groups cause displacements and 'strain' in quartz crystal lattices so that virtually all naturally occurring quartz is 'strained' and has an undulatory extinction under the petrological microscope. The extinction in deformed and folded quartz is unrelated to any schistosity or regional cleavage.

All the observations relating to the occurrence and behaviour of quartz and silicification in nature are consistent with its aqueous chemistry and the formation of a wide range of particulate polymeric species from which quartz most commonly crystallises.

That quartz veins are actually emplaced as successively mobilised fluid particulate suspensoids of polymeric silica is indicated by:

1) The fluid nature of the injected vein quartz where the mobility and plasticity of the pre-crystalline quartz was clearly not due to melting or the exceedingly low solubility of quartz in water.
2) The actual occurrence of silica gels in some veins.
3) The occurrence of biogenic and carbonaceous matter, brines, and oil in small inclusions within the vein quartz.
4) The occurrence of vein quartz as breccia matricies for rotated angular fragments of heavier wall rock and other minerals such as magnetite, sulphides, etc. The suspension of these fragments in their quartz matrix is due to rheopexy of the pre-crystalline polymeric silica.
5) The re-brecciation, episodic re-liquefaction and re-intrusion of vein quartz is due to the thixotropy of the pre-crystalline polymeric silica.
6) The occurrence of some jasper, opal, and chalcedony veins confirms the mobility of their hydrated meta-colloidal silica precursors.
7) The occurrence of colloform and Liesegang-type banding in some vein quartz or sections of quartz veins clearly indicates the metacolloidal and diffusive nature of the silica precursors.

8) The common occurrence of druse and miarolitic cavities in vein quartz and lodes is undoubtedly due to the syneresis and contraction of the original precursor polymeric silica of the veins.
9) The occurrence of siliceous oolites in cherts, jaspers, and vein quartz is due to syneresis of the oolite nuclei. Such concretionary rim growth can only occur in aqueous particle systems.
10) Surface catalysis or the enhanced crystal growth on a specific single face of very small crystallite nuclei results in the growth of slender acicular needles, rosettes, blades, spherulites, wheat sheaf structures, etc. where mineral crystals like rutile tourmaline, chlorite, dawsonite, and magnetite are developed in vein quartz. This type of enhanced crystal growth is due to the gelatinous nature of the polymeric silica in the precursor host vein material.
11) Occurrences of ptygmatic quartz veins clearly indicate viscosity related to fluid flow rates during intrusion of the non-Newtonian vein precursors.
12) Parallel striations on some faces of large quartz vein crystals reflect "crystal growth in steps" from a gelatinous surface coating.
13) Silica like potch, opal, or chert which replaces hydrolysed brachiopod shells clearly indicates a diffusive media in the host and a dispersed particulate phase of the replacing precursor silica.
14) Replacement of large tree trunks, originally largely cellulose, by silica clearly indicate the diffusive media of the decaying cellulose logs and the dispersion of the replacing particulate silica species.
15) Pervasive silicification of wall rocks, lode zones, stockworks, etc. and of the initially intruded polymeric vein quartz precursors is accomplished by minute dispersed monomeric and oligomeric silicic acid species which can diffuse within the silicifying host media.

The interactions and behaviour of the particulate and hydrous species of polymeric silica are more widely known and better documented than for any other colloid. This can therefore be applied to the mobilisation of silica into veins and lodes, the silicification of wall rocks, the formation of opal, the formation of ptygmatic veins, the replacement of shells and tree trunks, etc. This behaviour of particulate silica species will then provide a better basis for understanding how sulphide particles might similarly be mobilised into veins and lodes, permeate shales, form framboids, replace fossils and plant fragments or selectively replace fine shale bands in syndiagenetic orebodies, etc.

## GLOSSARY

**Accretion**: is rapid formation of clusters of similar shaped particles to form 'close packed' and pre-ordered aggregations at net lower surface energy in any remobilized concentrated fluid paste containing colloids. Crystallisation of these pre-ordered aggregates occurs subsequently to then form a 'porphyroblastic' texture where the large crystals are set in a finer grained matrix of crystallised sedimentary material.

**Acicular:** describes a crystal that is needle-like in form. It is also used to describe rod-shaped sedimentary particles when their length is more than three times their width.

**Adsorption**: is the adherence or fixation on a surface (usually but not necessarily a colloid because of the enormous area, surface energy, and charge) of an ion or charged particle. The uptake by a surface of a solute or dispersion can occur by electrostatic, dipolar, quadrapolar, linkages or hydrogen bonding, etc. Where a chemical linkage is involved, the surface-controlled reaction is called chemisorption. The dispersed ions and charged particles compete for adsorption sites on all available surfaces. Changes in concentration, pH, in the availability of surfaces, and in the permeability (spacing of the meshwork through which the ions and particles can diffuse) often have quite marked effects in exchanging and replacing surface adsorbed species.

**Aggregate**: refers to a mass or body of any sub-units such as smaller gelatinous accretions or concretions. These can crystallise as a mosaic of small interlocking crystals, as a composite of complexly intergrown crystals, or in optical continuity as a single ovoidal crystal. Rounded or irregular zones and patches of granular sediment or matrix cemented by infill concretion have also been referred to as aggregates.

**Bingham Yield Point:** Charged particle systems "gelled" as cohesive, fractural, thixotropic, and viscoelastic solids may be envisaged as having many weak links between particles and fewer strong ones. Stress (application of force) disrupts weak linkages continuously and stronger linkages at an increasing rate until uniform viscous flow is achieved (the rate of shear is proportional to the shearing stress). In systems like natural sediments, viscous flow begins gradually through a plastic deformation stage. The theoretical point at which stress would be sufficient to initiate uniform viscous flow is called the Bingham yield point

**Boehm lamellae:** are chains or bands of fluid inclusions within quartz crystals that are usually deformed or folded. They are a feature of the precursor polymeric silica and unrelated to the later crystal lattice structure.

**Clay hydrolysis:** clay minerals are created by the reaction of water (hydrolysis) with more structured silicates to form sandwich-structured platelets that are fully hydrated with silanol terminations on their external surfaces. Hydrolysis of existing clay minerals therefore refers to the slow progression of hydrolysis inward along the octahedral (brucite or gibbsite) layers from the edge of the platelets. This further hydrolysis occurs during diagenesis when clay-rich sediments are 'soaking' for long periods in slightly alkaline seawater or exposed to pore fluids in thick sediment accumulations. The progression of the hydrolysis reaction between the tetrahedral and octahedral layers in clay platelets has been referred to as the "zip fastener reaction".

**Colloform texture:** describes the finely banded semi-circular or spheroidal layering of minerals that are precipitated from colloidal sols and crystallise from these gelatinous rhythmically layered precursors.

**Concretion**: this is the slow or step–wise accumulation of material about a central nucleus to produce a banded–textured spherical or elliptical accumulation of higher particle density and compaction than the medium in which the particles are diffusing. A concretion may be homogeneous, being self–nucleated, homogeneous but nucleated on a foreign body, or heterogeneous (i.e. banded) with or without a specific nucleus. The active process of concretion depends on colloidal particles individually diffusing towards the precipitating surface represented by the boundary of a higher density gel aggregate with the less dense surrounding medium through which the particles are diffusing. Concretion could be considered to represent "adsorption" of ions or colloidal sol particles onto a growing nucleus, and finally onto a growing macroscopic aggregate of particles or a denser gel surface. Removal of such particles from dispersion by precipitation at a nucleus or at an interface between random open meshwork gel and denser ordered gel creates a diffusion gradient (fewer particles in that vicinity). To equalise the concentration, other particles under Brownian motion arrive in turn to precipitate (adsorb) and accumulate on the surface.

**Crustiform:** describes a vein in which the mineral filling is deposited in layers on the wall rock.

**Crystal growth**: requires particles to move to the surface Removal of a mobile species from dispersion in the gel by its crystallisation creates a diffusion gradient towards the crystal surface. Henisch[20] (p. 51) notes that convection currents are suppressed in gelatinous media and therefore movement must be essentially by diffusion but a very important function of the gel media is also to suppress nucleation. Without growth on many closely spaced competing nuclei, the faces of crystals supported in the gel medium are supplied by a steady diffusion of particles or ions so that large and well formed-crystals are able to develop. Crystal growth in a solid is usually by chemical reaction or ordered arrangement of molecules by heating or change of physical shape. Crystallisation from gas (sublimation) is by addition of molecules to active lattice sites (Pt. 1, Figure 230).

**Crystal lattice**: describes the stable meshwork of chemical bonds that hold the atoms of a crystal together in an ordered repetitive pattern of unit cells so that the compound that has crystallised achieves a low energy state.

**Crystallisation of feldspar**: a number of natural clays in close packed aggregates react spontaneously with alkali metal ions and monomeric silica to feldspar and water with the liberation of heat. Feldspathoids are sometimes formed as an intermediate product. Reactions are described on page 214 of Elliston[1], 2017.

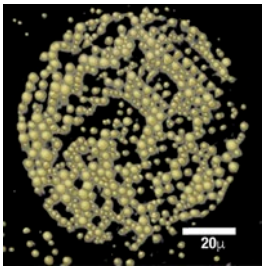
**Crystallisation of quartz**: most natural quartz has crystallised from compact polymeric species to which a further and continuing supply of the monomer is available. Some details are set out on pages 10-13, pages 21-26, pages 126-130, and pages 215-216 of Elliston[1], 2017.

**Displacive concretion:** is where the growth of the denser gel rimming layers round the synerectic nucleus displaces the surrounding weaker gel matrix.

**Enhanced crystal growth:** occurs in gelatinous media because of the support for the initially delicate skeletal or needle-like crystals and the catalytic effect of certain gel coatings on solid crystal faces. The catalytic surface coating that enhances crystal growth in gels and facilitates growth and development of well faceted crystals directly from dispersed particulate species is very significant. Where the catalytic surface coating applies particularly to one facet of an initial small 'seed crystal', it can result in acicular or needle-like growth. Freedom of reactants to diffuse to and from a growing crystal face, and suppression of competing crystal nuclei also enhance the growth of crystals in gelatinous media. Yield of the medium to the developing crystal shape and conversion from the high-energy state of dispersed gel or sol particles to the low energy crystalline state also enhance the growth of crystals. For the crystallisation of synerectic accretions or concretions in deposits of gelatinous pastes or slurries, the pre-ordering due to close packing of the particles strongly accelerates crystal growth.

**Elvan:** A now largely disused Cornish term (from Celtic, "white or pale rock") for intruded rocks having the composition of quartz, quartz porphyroid, chert, etc.

and usually associated with intruded mineral deposits. An elvan may contain chlorite, tourmaline, fluorite, or topaz as accessory minerals.




**Framboids:** are globular clusters of iron hydroxy mono-sulphide in various stages of crystallisation to tiny pyrite cubes and grains that make up small raspberry-like spheroidal aggregates usually some 15 to 25 microns in size. Framboidal clusters are synerectic and they often nucleate infill concretion of additional hydroxy-sulphide mineral or form the nuclei for concretionary overgrowth. Some framboidal clusters display long range ordering patterns. Framboids are usually composed of iron sulphide but framboids of brunckite (ZnS gel), sphalerite, native copper, digenite (CuS), chalcocite, covellite, native arsenic, and magnesioferrite or haematite have been recorded.

**Gel**: a gel is essentially a semi-solid meshwork of fine particles coagulated or flocculated by the inter-tangling of long chained polymers where the particles are linked to form a visco-elastic permeable solid by interaction between electric charges on their surfaces. Synthetic gels of pure clay, silica gel, gelatin, agar, etc. contain similar particles or macromolecules but natural gelatinous sediments are complex mixtures of charged particles having a wide range of sizes, shapes, and compositions. Hydrolysis reduces most particles to the colloidal size range but these can form a matrix to larger residual grains. Gelatinous ferric hydroxide, hydroxy carbonates, hydrated organic matter and silica gel occur in most sediment and sometimes as major constituents. However, the most common particles in pelitic sediments are clays, amorphous silica, and hydrous ferromagnesian minerals. These common particles are shaped as platelets, spheres, and rods respectively and in the "gelled" or coagulated condition they link together to form 'house of cards' or 'book-house' structures, 'strings of beads', and 'scaffold-like' structures of rods. Natural sediment can be thought of as these several types of structures randomly intermeshed and securely cross-linked together by the mutual satisfaction of coulombic charge sites and by van der Waal's attractive forces where particles are appropriately packed or close enough. It is not surprising that wet sediments have shear strength!

Semi-solid gelatinous sediments are thixotropic and have a definite yield value (Bingham yield point). The strength of the sediment fabric is very sensitive to water content and to the presence of flocculating or deflocculating agents. Liquefaction is isothermal and mechanically induced but the linkages between particles tend to re-form during viscous flow of the mud. The systems are "self-healing" but there is a time delay in reverting to the original gel strength called hysteresis. Cross-links are more readily broken at higher temperatures. Transition from an elastic gel to a liquid of relatively lower viscosity occurs revertably over a narrow temperature range. The more concentrated gels require higher temperatures but the thermal energy "softens" the paste and makes it easier to disrupt the fabric of particle linkages. Gels melt!

**Helmholtz double layer**: in an aqueous electrolyte solution in the vicinity of a charged surface the aqueous phase is divided into four regions of distinct dielectric behaviour. The innermost region consists of preferentially oriented water molecules in contact with the solid surface and where specific ions are adsorbed without their hydration shells. This is called the inner Helmholtz layer. The region further from the surface (β in Figure 1.7) contains both free water molecules and molecules attached to hydrated ions. This is called the outer Helmholtz layer and is defined by the by the closest approach that a fully hydrated charged ion can make to the solid – liquid interface. Further out from the surface the concentration of counter-ions (having a charge of opposite sign to the surface) decreases with increasing distance in the Gouy-Chapman diffuse layer. The outer and inner Helmholtz layers are referred to as the Helmholtz double layer.

**Hydrated silica polymers**: a diagrammatic representation of the polymerisation behavior of silicic acid is shown in Part 1, Figure 5.

**Hydrophobic bonding:** hydrophobic (water rejecting) colloids such as emulsions can involve the adsorption of dispersible organic particles or molecules on solid surfaces such as silica (or sulphide minerals in the important example of the floatation process). This adsorption means that the adsorbed hydrophobic molecule or particle must displace the solvent (adsorbed water monolayer) from the surface. This binding of hydrophobic molecules or particles to surfaces by short-range chemical forces or longer range (electrostatic and van der Waal's attraction) is called hydrophobic bonding.

**Illite:** this is a general name for a group of three-layer mica-like clay minerals intermediate between muscovite and montmorillonite. Illite flakes are generally much larger than montmorillonite platelets but they do not have the expanding lattice characteristics of smectites. However, the specific adsorption of potassium ions on the hydrolysed margins of illite clay platelets is some

500 times greater than their affinity for competing ions at the same molar concentration (equivalent solution strength). Illite accretions therefore readily crystallise to potassium feldspar. Illite is also called hydromica.

**Infill concretion**: growth of gelatinous precipitate on small synerectic nuclei within intergranular pore spaces can either displace the sediment grains or simply fill the pore spaces without displacement. Infill concretions are rounded "blobs" or nodules of relatively coarse sediment in which the pore spaces have been filed and the grains 'cemented' together by concretionary growth of the intergranular gel. Infill concretions may contain concentric or rhythmic bands like displacing concretions but the bedding and granularity of the original sediment are preserved within the concretion.

**Lepidocrocite:** is a ferroso-ferric hydroxide of iron [$FeO(OH).Fe_2O_3.H_2O$] that occurs as a precursor mineral in the mixture of iron hydroxides that form the intrusive lodes and veins of haematite and magnetite.

**Liesegang banding:** concentric rings or bands are developed during syneresis of pre-crystalline natural sediment colloids (such as clay, chert, hydroxy-carbonates, or siliceous shale). The Liesegang banding is a response to rhythmic changes in adsorption equilibria for pigment particles coating gel surfaces. The release/resorption of pigment particles is caused by different anion and cation diffusion rates as electrolytes are exuded from synerectic material. The bands parallel surfaces from which the fluid is lost.

**Lithomorph:** is the skeletal shape or pattern of rock mineral such as silica, that is left when cell walls or organic structures have been replaced and the original decayed organic matter removed.

**Macromolecule, macromolecular:** refers to very large molecules or very small crystallites such as clay platelets that by the continuity of their chemical linkages are essentially large molecules.

**Meshwork:** in relation to particle systems 'meshwork' is used to describe the static situation where particles of different shapes and sizes are randomly linked to each other by coulombic and van der Waal's forces to form a diffusible, plastic, fractural, and thixotropic visco-elastic solid.

**Metasomatism:** means the changed body of the mineral or rock. See 'Replacement' for further discussion of the process in terms of current chemistry.

**Minnesotaite:** is a green to brown ferromagnesian silicate mineral commonly found in chloritic or iron-rich sediments.

**Mobilisation**: means the liquefaction of a body of semi-consolidated sediment or other particulate material usually by earthquake shock or gravity sliding downslope.

**Montmorillonite:** this is a group of clay minerals that 'swell' by further reaction with water (hydrolyse). Typical montmorillonites have a three-layer crystal lattice with one sheet of aluminium-magnesium hydroxide (octahedral layers) between two sheets of hydrated silica (tetrahedral layers). A synonym for montmorillonite is smectite.

**Newtonian Fluids:** melts like that of ice and most other crystalline solids behave as Newtonian fluids. The rate of flow resulting from a shearing stress (application of a force) is proportional to the stress applied.

**Non-Newtonian Fluids:** the viscosity of non-Newtonian fluids varies with the rate of flow. This fluid behaviour is typical of thixotropic 'gels' (pastes and slurries) where the shear strength depends on interparticle linkages. As an approximation, it can be thought of as "the faster it flows – the more fluid and mobile it becomes". In many systems involving the flow of pastes and slurries, it is referred to as "shear thinning".

**Ossicle:** is a calcified individual element or piece of a skeleton such as an echinoderm shell. The term is usually used for larger pieces but has also been used for tiny bones, etc.

**Polymeric silicic acid:** the condensation of monomeric silica [$Si(OH)_4$] forms oligomeric silicic acids and many varieties of silica gel including the naturally occurring 'little balls' of amorphous silica that are adsorbed on the surfaces of most sediment substrates.

**Ptygmatic:** this word is used to describe the appearance of 'loopy' disharmonic folds that are obviously not due to any uniform stress field or constant direction of deformation. Internal forces within the precursor pastes of the intruded substances develop this pattern of fluidal folding as they flow into soft visco-elastic host materials. Ptygmatic folding is due to the shear thinning properties of the intruded pastes (viscosity dependent on the rate of shear) and the fold patterns are preserved by rheopectic re-setting (see Figure 74).

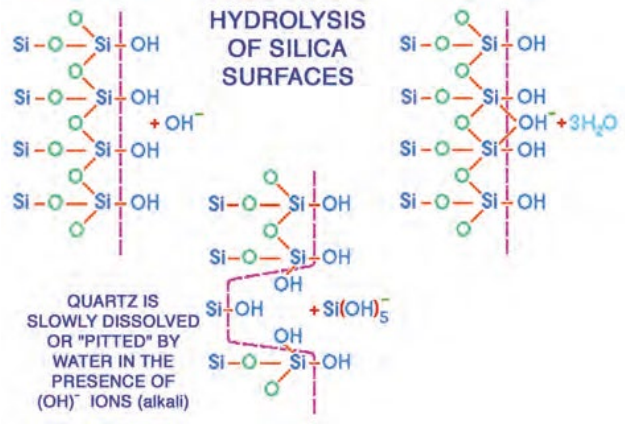
**Rheological separation:** is an important principle by which components of gel meshworks such as semi-consolidated sediments separate by differences in their fluid properties when the meshwork is physically disrupted or disturbed. The more mobile (less viscous or water-rich) components simply "flow out" of the agitated or disturbed pastes at the stage where its major compo-

nents would re-assume a non-fluid or more viscous gel condition.

**Rheopexy**: is the accelerated resetting to a gel condition in a flowing colloidal dispersion subjected to shear by laminar flow of a thick paste. Thixotropic liquids may rapidly revert to a higher viscosity condition when linkages establishing between particles throughout the flowing mass overcome the momentum of the moving mass. This "instant re-freezing" preserves flow foliation, the shape and form of intrusions, suspends fragments, etc.

**Silica**: Ordinary sand or crystalline quartz ($SiO_2$). Other polymorphs of silica such as cristobalite or tridimite are less common.

**Silica Polymers, where do all the natural silica polymers come from?** Quartz is not soluble in water including normal ground water and stream water in the cycle of erosion. It is transported by the streams and rivers as gravel and sand grains and generated by coastal erosion as sea sand. Quartz does not dissolve in seawater by dispersion of anions and cations as a solution but it does hydrolyse (react with water) in slightly alkaline conditions (seawater pH 7.9 to 8.3) by a process called "proton promoted dissolution" (Iler[2], 1979, fig. 1.11.) This is shown diagramatically as: -




Molecular dispersion from crystalline silica in water as $Si(OH)_4$ is catalysed by hydroxyl ions of an alkali or base. Seawater is slightly alkaline and therefore silica (and most silicate surfaces) "disperse" by these surface reactions. In sea water and within marine sediments the small neutral $Si(OH)_4$ molecules polymerise to short chain polymeric silicic acids called "oligomers".

**Silicic Acid**: See Part 1, Figures 1 and 2 and captions.

Where do all the natural silica polymers come from? The answer to this is that there are immense quantities of sand and silicates soaking for thousands of years in even greater quantities of slightly alkaline sea water. Reversible chemical reactions are driven by the quantities of reacting substances. In the late stages of diagenesis when natural sediments are losing water, the hydration reactions reverse and siloxane linkages predominate.

**Silanol**: Is a fully hydrated form of silicic acid that can condense to a direct chemical silicon-oxygen-silicon linkage by loss of a water molecule.


**Siloxane**: the direct silicon-oxygen-silicon chemical linkage is called siloxane.

**Sol**: is a homogeneous suspension or dispersion of colloidal particles in a liquid or gas. In the glossary of geology, a sol is also defined as a completely mobile mud that is in a more fluid form than a gel.

**Spherulite:** is a rounded or spherical body of acicular (needle-like) crystals radiating from a central point or small nucleus. Spherulites are a fairly common arrangement of feldspar crystals but are formed by many other minerals crystallising in diffusive media. Spherulites range in size from microscopic to several centimeters in diameter. Part 1, Figure 45 illustrates an example.

**Small particle systems**: are materials or substances made up of small particles. They are independent of the chemical composition of the particles but the very small size of the component particles means that the surface charge enables their interaction with other charged particles and ions in the pore fluids surrounding them. Mud is "sticky" because the particles cling to each other and to surfaces they come in contact with. Examples of small particle systems are mud, clay, silica gel, thick paints, food colloids (like yoghurt, cream, soup, etc.).

**Surface charge**: See Figures 3 and 4 and captions and description in associated text.

**Surface chemistry**: is the study of the special chemistry that is related to the solid-water interface. Surface chemistry and colloid chemistry are closely interrelated because the behaviour of colloidal particles is dependent on the properties of the very large surfaces they present to the solvent in relation to their very small volume. The solvent, ions, complexes, and other charged particles interact with all surfaces but are especially important in their interactions with colloidal particles.

**Surface energy**: is the difference in energy per unit area between the surface of a given crystal lattice or substance and the energy of the same number of atoms

(comprising the unit area) situated within the bulk of the crystal or substance. Surface energy is clearly dependent on the atomic geometry of the atoms exposed within the unit area of surface. The atoms exposed at the surface are able to interact with particles, ions, solvent, or other substances. They have 'dangling bonds' or charge that can compensate each other, hydrate, adsorb surface species, or form new chemical compounds. Bonds or linkages of atoms comprising a comparable area within the crystal or substance are in equilibrium with those surrounding them.

**Surfactant:** a particular class of solutes that show dramatic effects on surface tension are highly active in relation to surface adsorption and are called surfactants. They are dispersions or solutes such as soaps, detergents, long chain alcohols, and polymeric silica

**Syneresis**: is the spontaneous aging or contraction of a gel meshwork within itself by the establishment of a greater density of cross–linkages and elimination of water. The particles or particle–chains achieve greater co–ordination. The total surface energy is lowered, and the internal surface and adsorptive capacity are reduced. The contraction and greater gel density causes shrinkage cracks or a pattern of holes or channels (like those in cheese) which is independent of whether or not the gel is immersed in water. In syneresis the particles move closer together under the influence of van der Waal's attractive forces so that the less dense, sparse, weak "watery" gels tend to be less or non–synerectic. The crystalline state is the low energy state of matter.

**Tactoid:** is like a floc but distinguished by a high internal ordering of the particles within the cluster. In the process of manipulating the repulsive forces between particles to promote coagulation or flocculation, it is possible to achieve a structured or partly structured aggregate such as stacks of platelets, interlocking arrays of chains of particles, or bundles of aligned rods.

**Thixotropy**: is the isothermal reversible re-liquefaction of a gel or coagulated sol. Thixotropy is due to mechanical shock or shear which disrupts the gel particle linkages allowing the colloids of the system to revert to a dispersed sol or more fluid gel at the same fluid content. This isothermal gel to sol or to more-fluid-gel transformation is reversible and repeatable. Thixotropy is mainly induced by shock. A short sharp oscillation throughout the gelatinous mass is more effective in destroying all, or sufficient of, the interparticle linkages at the one time so that the meshwork structure will collapse and the material revert to a fluid. Differential liquefaction depending on hydration (differing Bingham yield points) of different colloidal components allows separation of more mobile hydrous materials which can then simply flow out of the disturbed mixture.

**van der Waals' forces:** the strong attraction due to interaction between dipoles when small particles or molecules are in very close proximity to each other is called van der Waals' forces of attraction. These forces exist between all matter in very close proximity.

**"Zip fastener reaction":** see 'Clay hydrolysis'

ACKNOWLEDGEMENTS

The permission of CRA Exploration Pty Ltd to publish the content of this company report on the hydration of silica and the formation of quartz veins is gratefully acknowledged. Permission from Connor Court to publish photographs and information from the treatise, Elliston[1], 2017, is also thankfully acknowledged. The author is also appreciative and thankful for the interest and financial contribution of AusIndustry for monitoring this research since 1984 and ensuring that the conclusions were reached by the prescribed scientific method.

REFERENCES

1. Elliston, John, **2017**. *The origin of rocks and Mineral Deposits - using current physical chemistry of small particle systems*. Connor Court Publishing Pty Ltd, Brisbane. 706 pp. ISBN 978-1-925501-36-0.
2. Iler, R.K., **1979**. *The Chemistry of Silica*. John Wiley and Sons, New York. 483 p.
3. Sugar, I., and Guba, F., **1954**. Proc. 3rd Int. Congr. Electron Microsc., by Royal Microscopical Society, London, p. 530.
4. Lougheed, M.S., **1983**. Origin of Precambrian iron-formations in the Lake Superior region. *Geol. Soc. Amer. Bull*., 94: 325-340.
5. Lindgren, W., **1933**. *Mineral Deposits*. McGraw-Hill, New York, 930 pp.
6. Carozzi, A.V., **1960**. *Microscopic Sedimentary Petrology*. John Wiley and Sons, New York, 485 pp.
7. Heinrichs, T., **1984**. The Umsoli chert, turbidite testament for a major phreatoplinian event at the Onverwacht/Fig Tree transition (Swaziland suoergroup, Archean, South Africa), *Precambrian Research*, 24: 237-283
8. Mielenz, R.C. and King, M.E., **1955**. Physical-Chemical Properties and Engineering Performance of Clays. *Calif. Dept. Nat. Resources, Bull.* 169, 196-294.
9. Slobodskoy, R.M., **1970**. Origin of ptygmatic veins in the contact aureole of granitoid batholiths (Altai Region, U.S.S.R.), *Tectonophysics*, 9: 447-457.

10. Mysels, K.J., **1959**. *Introduction to Colloid Chemistry*. Interscience, New York, N.Y., 270 pp.
11. van Olphen, H., 1963. *An Introduction to Clay Colloid Chemistry*. Interscience Publishers, New York, 301 pp.
12. Yariv, S., and Cross, H., **1979**. *Geochemistry of Colloidal Systems*. Springer, Berlin-New York, 450 pp.
13. Sederholm, J.J., **1907**. On Granite and Gneiss. *Bull. Comm. Geol. Finlande*, 23: 1-110.
14. Holmquist, P.J., **1920**. Om pegamatitpalingenes och ptygmatisk veckning. *Geol. Fören. Stockholm Forh*., 42: 191-213.
15. Niggli, P., **1925**. Über das Grundgebirge des Schwarzwaldes, *Mitt. Aarguer Naturforsch. Ges*., 17: 1-35.
16. Weaver, C.E., **1984**. *Shale-slate Metamorphism in Southern Appalachians*. Elsevier, Amsterdam. Developments in Petrology 10. 239 pp.
17. Dimroth, E., and Chauvel, J.J., **1973**. Petrography of the Sokoman iron formation in part of the central Labrador trough, Quebec, Canada. *Geol Soc. Amer. Bull*., 84: 111-134.
18. Erdmannsdörfer, O.H., **1938a**. Über Versuche zur Nachbildung ptygmatischer Falten. *Zentr-Mineral*., A: 257-261.
19. van Hise, C.R., and Leith, C.K., **1911**. *The Geology of the Lake Superior Region*. US Geol. Survey, Monogr. 52.
20. Henisch, H.K., **1970**. *Crystal Growth in Gels*. Penn. State Uni. Press, University Park, Pennsylvania. 111 pp.
21. Hatschek, E., and Simon, A.L., **1912**. Gels in Relation to Ore Deposition, *Trans. Inst. Min. Met*., 21: 451-479.
22. Boydell, H.C., **1925**. The Role of Colloidal Solutions in the Formation of Mineral Deposits. *Bull. Inst. Min. Met*., 243: 1-103.
23. Stumm, W., **1992**. *Chemistry of the solid-water interface: Processes at the mineral-water and particle-water interface in natural systems*. A Wiley-Interscience publication, New York, 428 pp.
24. Healy, T.W., **1969**. *The effect of polyelectrolytes on PbS crystal growth in silica gel*. Report on Research at the University of Melbourne, Geopeko Technical Seminar, Mount Morgan, February 1969. (Unpub.)
25. Holser, W.T., **1947**. Metasomatic Processes. *Econ. Geol*., 42: 384-395.
26. White, S., **1971**. Hydroxyl ion diffusion in quartz, *Nature*, 230: 192
27. Leo, R.F., and Barghoorn, E.S., **1976**. (Duplicated the fossilization of wood in the laboratory) *Bot. Mus. Leafl. Harv. Univ*., 25 (1), 1. (See Iler, 1979, pp. 91 and 113.)
28. Oehler, J.H., and Schopf, J.W., **1971**. (Silicified filamentous algae in silica gel) *Science*, 174: 1229.
29. Newell, N.D., Rigby, J.K., Fischer, A.G., Whiteman, A.J., Hickox, J.E., and Bradley, J.S., **1953**. *The Permian reef complex of the Guadalupe Mountains Region, Texas and New Mexico*. W.H. Freeman and Company, San Francisco, 236 pp.
30. Chilingar, G.V., Bissel, H.J., and Wolf, K.H., **1967**. *Diagenesis of carbonate rocks*. In Diagenesis in Sediments, (G. Larsen and G.V. Chilingar, Eds.), Developments in Sedimentology 8, Elsevier, Amsterdam, pp. 179-322.
31. Adams, A.E., Mackenzie, W.S., and Guilford, C., **1984**. *Atlas of sedimentary rocks under the microscope*. John Wiley & Sons, New York, 104pp.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Feature Article

# Chuckles and Wacky Ideas

Carl Safina

*Stony Brook University, 100 Nicolls Rd, Stony Brook, NY 11794, US*
*The Safina Center, 80 North Country Road, Setauket, NY 11733, US*
E-mail: csafina@safinacenter.org

OPEN ACCESS

**Citation:** C. Safina (2019) Chuckles and Wacky Ideas. *Substantia* 3(1): 95-99. doi: 10.13128/Substantia-210



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Competing Interests:** The Author(s) declare(s) no conflict of interest.

Another big group of dolphins had just surfaced alongside our moving vessel —leaping and splashing and calling mysteriously back and forth in their squeally, whistly way, with many babies swift alongside their mothers. And this time, confined to just the surface of such deep and lovely lives, I was becoming unsatisfied. I wanted to know what they were experiencing, and why to us they feel so compelling and so—close. This time I allowed myself to ask them the question that for a scientist is forbidden fruit: Who are you? Scientists usually steer firmly from questions about the inner lives of animals. Surely they have inner lives of some sort. But like a child who is admonished that what they really want to ask is impolite, a young scientist is taught that the animal mind—if there is such—is unknowable. Permissible questions are "it" questions: about where it lives, what it eats, what it does when danger threatens, how it breeds. But always forbidden—always forbidden—is the one question that might open the door to the interior: Who? There are good reasons to avoid so fraught an inquiry and the cans of worms such a door could open. But the barrier between humans and animals is artificial, because humans are animals. And now, watching these dolphins, I was tired of being so artificially polite; I wanted more intimacy. I felt time slipping for both of us, and I did not want to risk having to say good-bye and realizing that I'd never really said hello. During the cruise I'd been reading about elephants, and elephant minds were on my own mind as I wondered about the dolphins and watched them pacing fluidly and freely in their ocean realm. When a poacher kills an elephant, he doesn't just kill the elephant who dies. The family may lose the crucial memory of their elder matriarch, who knew where to travel during the very toughest years of drought to reach the food and water that would allow them to continue living. Thus one bullet may, years later, bring more deaths. Watching dolphins while thinking of elephants, what I realized is: when others recognize and depend on certain individuals, when a death makes the difference for individuals who survive, when relationships define us, we have traveled across a certain blurry boundary in the history of life on Earth—"it" has become "who." "Who" animals know who they are; they know who their family and friends are. They know their enemies. They make strategic alliances and cope with chronic rivalries. They aspire to higher rank and wait for their chance to challenge the existing order. Their status affects their offspring's prospects. Their life follows the arc of a career. Personal relationships define them. Sound familiar? Of course. "They" includes us. But a vivid, familiar life is not the domain of humans alone. We look at the world through our own eyes, naturally. But by looking from the inside out, we see an inside-out world. This book takes the perspective of the world outside us—a world in which humans are not the measure of all things, a human race among other races. To understand anything, really, one must go deep, to the roots. In our estrangement from nature we have severed our sense of the community of life and lost touch with the experience of other animals. So while I went in search of particular "who" animals,

I delved into new findings about thought, emotion, and consciousness that apply to many animals. And because everything about life occurs along a sliding scale, understanding the human animal becomes easier in context, seeing our human thread woven into the living web among the strands of so many others. This project differs from other "animal thinking" books in one fundamental way. I'd intended to take a bit of a break from my usual writing about conservation issues, to circle back to my first love: simply seeing what animals do, and asking why they do it. I traveled to observe some of the most protected creatures in the world—elephants of Amboseli in Kenya, wolves of Yellowstone in the United States, and killer whales in the waters of the Pacific Northwest—yet in each place I found the animals feeling human pressures that directly affect what they do, where they go, how long they live, and how their families fare. So in this book we encounter the minds of other animals and we listen—to what they need us to hear. The story that tells itself is not just what's at stake but who is at stake. The greatest realization is that all life is one. I was seven years old when my father and I fixed up a small shed in our Brooklyn yard and got some homing pigeons. Watching how they built nests in their cubbyholes, seeing them courting, arguing, caring for their babies, flying off and faithfully returning, how they needed food, water, a home, and one another, I realized that they lived in their apartments just as we lived in ours. Just like us, but in a different way. Over my lifetime, living with, studying, and working with many other animals in their world and ours has only broadened and deepened —and reaffirmed—my impression of our shared life. That's the impression I'll endeavor to share with you in the pages that follow.

**Keywords**. Ethology, theory of mind, behaviorism, progress of science.

This chapter is an excerpt from a book of the author (C. Safina, *Beyond Words: What Animals Think and Feel*, Henry Holt And Company, New York, 2015). The same book is available in Italian as: "*Al di là delle parole*" by Adelphi Press (2015). For more contents please visit the website: CarlSafina.org.

I'd never deny that formal scientific research in controlled conditions has been exceptionally helpful. I'll also never lose sight of the fact that the real lives of animals are too expansive for laboratories to adequately reflect. Yet many behaviorists work only in labs (or, far worse, philosophy departments). Now we'll see how, by slicing reality salami-thin and marinating it in jargon, researchers who confuse sometimes amuse.

The search for intelligent life on Earth produces a few chuckles along the way. One dog-loving researcher videotaped dogs in a neighborhood park during two years before arriving at the following conclusions: If a dog wanted to play with another dog it was facing, it would usually perform the "play invitation" (that familiar bow: front end crouched low, rear end high). But if the dog the play seeker wanted to romp with was facing away, the play seeker would first get the other dog's attention – with a paw, for instance, or by barking. In one of those science-marches-on moments, the researcher tells us, "They seem to be reacting to distinct cognitive states." In everyday terms: from two years of video analysis she discovered that a dog can distinguish another dog's face from its butt. May I please say this: a dog's behind is not a "distinct cognitive state." Why not just say that dogs get other dogs' attention before inviting play? Too obvious to seem like science?

Just minutes after I started searching the formal academic literature for "theory of mind," a typical recent study popped up. Titled "On the Lack of Evidence That Non-Human Animals Possess Anything Remotely Resembling a 'Theory of Mind,'" it was published in the *Philosophical Transactions of the Royal Society*. The authors begin, "Theory of mind entails the capacity to make lawful inferences about the behaviour of other agents on the basis of abstract, theory-like representations of the causal relation between unobservable mental states and observable states of affairs." (Translation: by watching another's behavior, we can guess at what they may be thinking.) They continue:"We are entirely agnostic (for our present purposes anyway) about whether an organism's states are modal or amodal, discrete or distributed, symbolic or connectionist or even about how they come to have their representational or informational qualities to begin with… Of course, there are innumerable other factors that also contribute to shaping a biological organism's behavior."

I can probably understand that study—I just don't want to.

Two guys from Rutgers University (where I got my own PhD, so I am favorably inclined) have published a review called "Reading One's Own Mind: A Cognitive Theory of Self-Awareness." Here we go: "We'll start by examining what is probably the most widely held account of self-awareness, the 'Theory Theory' (TT). The basic idea of the TT of self-awareness is that one's access to one's own mind depends on the same cognitive mechanism that plays a central role in attributing mental states to others… Theory Theorists argue that the TT is supported by evidence about psychological development and psychopathologies… After making our case against

the TT and in favor of our theory, we will consider two other theories of self-awareness to be found in the recent literature."

No, thanks! Theorizing about theorizing seems a very poor substitute for actually watching living beings do their thing.

"Theory of mind" is probably the most oversold concept in human psychology, as well as the most underappreciated, oft-denied aspect of non-human minds. We've all been in relationships where we thought, "I don't know where I stand with her" or "I don't know what to expect of him."

As John Locke said in the 1600s, "one man's mind could not pass into another man's body." The painter Paul Gauguin wrote of his thirteen-year-old Tahitian wife, "I strive to see and think through this child." Joni Mitchell sang, "There's no comprehending, / Just how close to the bone and the skin and the eyes / And the lips you can get / And still feel so alone." The Roman poet Lucretius—in what W. B. Yeats called "the finest description of sexual intercourse ever written" (not to mention a good translation)—observed bleakly,

*They gripe, they squeeze, their humid tongues they dart, As each would force their way to t'others heart:*
*In vain: they only cruise about the coast,*
*For bodies cannot pierce, nor be in bodies lost.... All ways they try, successless all they prove,*
*To cure the secret sore of ling'ring love.*

"The tragedy of sexual intercourse," Yeats howled, "is the perpetual virginity of the soul." Paul Valéry, another poet, noted that "the interchange of human things between men requires that brains be impenetrable." Praise the poets for being good scientists. The scientist Nicholas Humphrey, says, "There are no doors between one consciousness and another. Everyone knows directly only of his or her own consciousness and not anyone else's!"

If I want to sneak up on you, or fantasize while flirting, or steal from you, it is crucial that my mind remain unreadable. The more we *could* open into each other's minds, the more our brains would need a way to get up and lock the door. So yes, we observe, we resonate, but ultimately we guess. That's the most we can do. We can choose to reveal ourselves or hide our cards. But the choice is ours.

Chimps have mainly a theory of chimp mind, if we might put it that way; dolphins, mainly of dolphin mind. Humans often experience difficulty understanding even human needs and predicting other people's actions. And humans who assume that other animals are not even conscious—or who ignore their capacity for conscious experience—show how faulty our theory-of-mind talents are.

People in Japan and the Faeroe Islands kill dolphins and pilot whales by running steel rods into their spinal columns while they squeal in pain and terror and thrash in agony. (In Japan, it's illegal to kill cows and pigs as painfully and inhumanely as they kill dolphins.) The lack of compassion for dolphins and whales indicates that humans' "theory of mind" is incomplete. We have an empathy shortfall, a compassion deficit. And human-on-human violence, abuse, and ethnic and religious genocide are all too pervasive in our world. No elephant will ever pilot a jetliner. And no elephant will ever pilot a jetliner into the World Trade Center. We have the capacity for wider compassion, but we don't fully live up to ourselves. Why do human egos seem so threatened by the thought that other animals think and feel? Is it because acknowledging the mind of another makes it harder to abuse them? We seem so unfinished and so defensive. Maybe incompleteness is one of the things that "makes us human."

While some people seem unable to sense the minds of non-human animals, other people see human-like minds in everything. Our minds automatically discern human-like faces in things like clouds, the moon, even in food. Many believe that rocks, trees, streams, volcanoes, fire, and other things have thoughts, that *everything* has a mind and is inhabited by spirits that might act for or against us. That's called *panpsychism*. The religion that follows from this primal human assumption is *pantheism*. It is common among tribal hunter-gatherer peoples, and it's also alive and well in modern life. On the summit of Mount Kilauea, in Hawaii, I've seen offerings of money and liquor, put there by people who think that volcanoes have a god within who watches, tallies favors, and sometimes acts vindictively. Don't get the volcano mad by ignoring it. A little more booze and a few more bills, some flowers and some food and a roast pig occasionally, and the volcano's fiery goddess, Pele, will perhaps be mollified. And this is in the United States, where anyone can just stroll into the visitors' center and learn some volcano geology. (Park rangers have asked visitors to stop leaving offerings of food, money, flowers,incense, and liquor on Kilauea because in sum the offerings are more clearly appreciated by rats, flies, and roaches than by the goddess.) It appears that deep belief in the supernatural comes naturally to us.

"Nonhuman animals may arrive at beliefs based on evidence," writes philosopher Christine M. Korsgaard,

"but it is a further step to be the sort of animal that can ask oneself whether the evidence really justifies the belief, and can adjust one's conclusions accordingly. "Yet it is many *humans* who are demonstrably incapable of asking whether evidence justifies their beliefs, then adjusting their conclusions. Other animals are great and consummate realists. Only humans cling unshakably to dogmas and ideologies that enjoy complete freedom from evidence, despite all evidence to the contrary. The great divide between rationality and faith depends on some people choosing faith over rationality, and viceversa.

Other animals' actions and beliefs are evidence-based; they don't believe anything *unless* the evidence justifies it. Other animals attribute awareness only to things that are actually aware. While a dog might bark to rouse someone sleeping on the living room couch, they never seek assistance from the sofa itself. Or from volcanoes. They easily discriminate living things from inanimate objects and even from impostors. True, skilled duck hunters' decoys and calls fool passing ducks enough to get them to swerve into gun range, but the ruse must be elaborate or it won't work. Fish can be hard to fool even with artificial lures painstakingly designed to look and act like the real thing.

Years ago, while doing research that involved tagging migrating falcons, I lured the falcons to my net with tethered live starlings. The frightened starlings did not enjoy this; nor did I. So I put a stuffed starling on a string, wings in flight position, behind the net. Of course, in nature absolutely everything that looks like a bird and is covered with feathers and has a gleaming eye and moves up and down *is* a bird. Yet the stuffed bird never fooled one single falcon. They all sized it up, at a glance, as somehow "not real," and ignored it. *That* is impressive. Other animals are exceptionally good at identifying and reacting to predators, rivals, and friends. They never act as if they believe that rivers or trees are inhabited by spirits who are watching. In all these ways other animals continually demonstrate their working knowledge that they live in a world brimming with other minds, as well as their knowledge of those minds' boundaries. Their understanding seems more acute, pragmatic, and, frankly, better than ours at distinguishing real from fake.

So I wonder: Do humans really have a better-developed theory of mind than other animals? People watching a cartoon of nothing more than a circle and a triangle moving around and interacting almost always infer a story, involving motives and personalities and genders. Children talk to dolls for years, half-believing – or firmly believing – that the doll hears and feels and is a worthwhile confidant. Many adults pray to statues, fervently believing that they're listening. When I was a teenager, our next-door neighbors (Americans who'd been born and raised in New York) kept religious statues in every room except their bedroom, lest the Virgin witness human lust. All of this indicates a common human inability to distinguish conscious minds from inanimate objects and evidence from nonsense.

Children often talk to a fully *imaginary* friend whom they believe listens and has thoughts. Monotheism might be the adult version. We populate our world with imaginary conscious forces and beings – good and evil. Most present-day people believe they're helped or hindered by deceased relatives, angels, saints, spirit guides, demons, and gods. In the world's most technologically advanced, most informed societies, a majority of people take it for granted that disembodied spirits are watching, judging, and acting on them. Most leaders of modern nations trust that a sky god can be asked to protect their nation during disasters and conflicts with other nations.

All of this is "theory of mind" gone wild, like an unguided fire hose, spraying the whole universe with presumed consciousness. Humans' "superior" theory of mind is in part pathology. The oft-repeated line "Humans are rational beings" is probably our most half-true assertion about ourselves. There is in nature an overriding sanity and often, in humankind, an undermining insanity. We, among all animals, are also frequently irrational, distortional, delusional, worried.

Yet I also wonder: Is our pathological ability to generate false beliefs, to elaborate upon what does not exist, also the very root of human creativity? Is our tendency to imagine and even cling to what is false the foundation of all our inventive genius?

Perhaps believing false things comes bundled with our peculiar, oddly brilliant ability to envision what is not yet, and to imagine a better world. No one has explained where creativity arises, but some human minds lurch along sparking new ideas like a train with a stuck wheel. It's not rationality that's uniquely human; it's irrationality. It's the crucial ability to envision what is not, and to pursue unreasonable ideas.

Perhaps other animals don't need to manipulate logic because their actions are logical. They don't need tools because they are self-sufficient in their special abilities. Perhaps humans need logic and tools because without them we cannot survive, in a sense unable to succeed just as we are. Perhaps this is intuited in the story of the Fall, the trade-off in going from self-contained creatures like all the others to creatures needing a new way to access new knowledge so that, with much craft and

effort, our distinctly human abilities might compensate for our distinctly human frailties.

Insight, shared to various degrees by other apes, wolves and dogs, dolphins, ravens, and a few other creatures, relies on an ability to see what is not there. As does turning homeward, or waiting for the mate who happens to be gone at the moment. Perhaps the depth of human insight comes with genes that give us a capacity not just to imagine what isn't there but to insist on it, to fervently hold and pursue unmoored beliefs. What is more irrational than a nonexistent melody, or the dream of human flight, or holding fixed the light of an image, or capturing a musical performance so that it may be heard again and again, or diving deep into the sea and breathing underwater? Who could have imagined such things? Who else.

Along for the ride on that singular ability to imagine comes sheer brilliance and utter madness. And maybe more than anything, what "makes us human" is our ability to generate wacky ideas.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Feature Article

# The increased anthropogenic gas emissions in the atmosphere and the rising of the Earth's temperature: are there actions to mitigate the global warming?

Francesco Barzagli[1,2], Fabrizio Mani[2]

[1] *University of Florence, Department of Chemistry, via della Lastruccia 3, 50019 Sesto Fiorentino, Italy*
[2] *ICCOM CNR, via Madonna del Piano 10, 50019 Sesto Fiorentino, Italy*
E-mail: fabrizio.mani@iccom.cnr.it

**Abstract.** Some frozen bodies have been recently discovered in the Alp glaciers because the global warming is forcing the ice to retreat. Many years have passed since the first perception of a strong link between the temperature of the Earth and the amount of some gases in the atmosphere, the so called greenhouse gases. Today there is a general consensus among the governments, the scientists and industrial organizations of most countries in recognizing the relationship between the increase of the atmospheric $CO_2$ concentration resulting from over a century of combustion of fossil fuels and the observed global warming. The development of technologies to reduce the anthropogenic emissions should not be further delayed, in accordance with the Paris Agreement that recommended keeping the global mean temperature well below 2 °C above pre-industrial levels to reduce the risks and impacts of climate change.
This paper gives an overview of the different greenhouse gases, their emissions by economic sectors and the international treaties that require the most developed countries to pursue the objective of reducing their greenhouse gas emissions. Amongst the different actions directed towards a low-carbon economy, the chemical capture of $CO_2$ from large stationary emission points is the most efficient and widespread option. Additionally, new technologies are currently exploited to capture $CO_2$ directly from air and to convert $CO_2$ into fuels and valuable chemicals.

**Keywords.** Global warming, climate changes, greenhouse gas emissions, $CO_2$ capture, $CO_2$ utilization.



**Citation:** F. Barzagli, F. Mani (2019) The increased anthropogenic gas emissions in the atmosphere and the rising of the Earth's temperature: are there actions to mitigate the global warming?. *Substantia* 3(1): 101-111. doi: 10.13128/Substantia-69



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Competing Interests:** The Author(s) declare(s) no conflict of interest.

## THE GLOBAL WARMING AND THE POLICIES FOR ITS MITIGATION

It is very likely the relationship between the Earth's temperature, climate and the concentration of some gases, the so called greenhouse gases (GHGs), in the atmosphere. As a matter of fact, the greenhouse effect made our planet habitable with an average temperature of 18 °C, otherwise it would be – 19 °C.

If we look back to hundreds of thousands years ago, cooler glacial and warmer interglacial cycles occurred with periods of about 100,000 years (Figure 1). They are related to the variation of the amount of solar radiation with time, caused by the precession of the equinoxes (the rotation of the Earth's direction axis), the variation of the obliquity of the Earth's axis with respect to the perpendicular to the plane of the orbit around the sun, and the variation of the eccentricity of the orbit that varies the Earth-Sun distance. It must be noted that the variation of $CO_2$ concentration over time was a *consequence* of the variation of the temperature: the increasing temperature released more dissolved $CO_2$ from the oceans and permafrost, thus increasing the greenhouse effect that accelerated the global warming. The opposite effect occurred when the temperature decreased.

The last glacial period ended about 21,000 years ago, and currently we are in an interglacial period of very low increasing Earth's temperature that has been accelerated in the last century, most likely by the increasing GHG emissions from human activities. The anthropogenic GHG emissions, predominantly carbon dioxide, add to the "natural" greenhouse effect and could result in Earth's temperature rising and subsequent climate change.

The "greenhouse effect" and the global warming have a long history, that started two centuries ago. The famous French mathematician and natural philosopher Jean-Baptiste Fourier (Auxerre, 1768 – Paris, 1830), suggested in the late 1820 that the atmosphere limits the heat loss from the Earth's surface, that is warmer than it would be in the absence of this effect. In 1860 John Tyndall (Leighlinbridge, 1820 – Haslemere (UK), 1893), an Irish scientist, measured the absorptive power of some gases and discovered that water vapour and "carbonic acid" (carbon dioxide) absorb the re-emitted heat from the Earth's surface that cools overnight. He realised that climate changes could be related to the concentration of these gases. Svante Arrhenius (Vik, 1859 – Stockholm, 1927), a Swedish physicist and chemist, Nobel laureate for Chemistry in 1903, in 1896 calculated that 50% increase of $CO_2$ concentration in the atmosphere would take thousands of years and would increase the Earth's temperature of 2.5-3 °C. Arrhenius concluded that the world population would benefit in the future from a warmer climate that would prevent new glacial ages, thus affording more land for harvesting. Contrary to the Arrhenius' belief, the 50% of $CO_2$ concentration has increased in the last two centuries, because of the fossil fuel combustion to sustain the continuously increasing demand of energy of the industrial revolution and the economic growth of the population that, additionally, rose from about 1 billion in 1800 to today 7.6 billion.

Now, there is a general consensus among the governments, the scientists, and industrial organisations of most countries about the correlation (95-100% probability) between the GHG emissions in the atmosphere originating from the human activities, the rise of the Earth's temperature and the climate change (Figure 2).[3-5] It has become a worldwide priority to reduce the anthropogenic GHG emissions, particularly those of $CO_2$, the main component of GHGs, together with the techniques for the adaptation to climate change.

Afterwards the first stations at South Pole and Mauna Loa, Hawaii, in 1950 began measuring the $CO_2$ concentration in the atmosphere, accurate data were available. In 1988 the *Intergovernmental Panel on Climate Change* (IPCC) was established by the *World Meteorological Organization* (WMO) and the *United Nations Environmental Programme* (UNEP) to provide "*policy-*

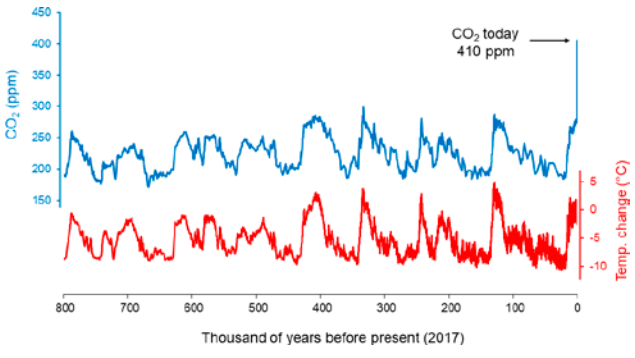



**Figure 1.** Correlation between $CO_2$ concentration[1] in the atmosphere and Earth's temperature[2] over the last 800,000 years. Temperature change is the difference from the average of the last 1000 years.

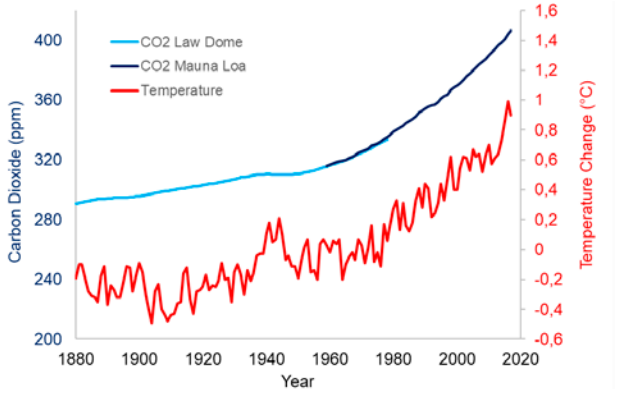



**Figure 2.** Correlation between the change in the mean annual temperature records and the $CO_2$ concentration. Temperature data from NASA/GISS;[3] $CO_2$ concentration data from Mauna Loa, Hawaii,[4] and from ice cores from Law Dome, Antarctica.[5]

*makers with regular assessments of the scientific basis of climate change, its impacts and future risks, and options for adaptation and mitigation*".

Until now IPCC has released five *Assessment Reports*,[6,7] and the sixth will be completed in 2021. The fifth Assessment Report (IPCC AR5)[8] is referred to 2014 and is based on the work of 831 worldwide experts on physics, engineering, chemistry, meteorology, oceanography, ecology, economics. The scenarios provided on the GHG emissions by human activities and global warming are indisputable.

The $CO_2$ concentration in the atmosphere, largely the main component of GHGs, increased from 280 ppm (0,028 % v/v) of the pre-industrial level (the beginning of the industrial society is conventionally fixed to 1750) to today 410 ppm (0,041%; April 2017). In the same time the Earth's temperature increased approximately of 1.0÷1.2 °C, most of which in the last century. Before 1750 the mean temperature, even if with ± 0.3 °C variations, and GHG concentration remained roughly constant for hundreds of years.

Carbon dioxide emissions from fossil fuel combustion and industrial processes account for about 76% of the current total GHG emissions. The percentage of the other GHGs is reported in Table 1 as $CO_2$-equivalent ($CO_2$eq), that takes into account for the relative amount of emissions and for the global warming potential (GWP) relative to $CO_2$.[9] $GWP_{100}$ measures the warming effect of a mass of a GHG relative to that of the same mass of $CO_2$, over a period of 100 years. The lifetime of each GHG in the atmosphere, and consequently its GWP, is different to each other, because of the different reactivity with the other components of the atmosphere and with solar radiation.

About 75% of overall anthropogenic $CO_2$ emissions between 1750 and 2010 occurred in the last 60 years, because of the unrestrainable growth of the population (from 2.5 billion in 1950 to 7.6 billion in 2018), the energy intensive lifestyle of the population and the economic activities of the developed countries, and the socio-economic growth of rapidly developing countries (currently, China, India, Brazil), that require more and more energy production. Total anthropogenic GHG emissions increased over 1970 to 2012 of 91% from 24 to 47 Gtonne $CO_2$eq/y, the highest in human history. Also the rate of warming of the atmosphere and ocean since 1950 is the greatest ever recorded.[10]

The Kyoto Protocol (December 1997) is an international treaty that commits the 39 most industrialised countries to tackle the global warming by reducing their GHG emissions in the atmosphere to a level that "*would prevent dangerous anthropogenic interference with the climate system*". The six greenhouse gases taken into consideration by the Kyoto Protocol were carbon dioxide ($CO_2$), methane ($CH_4$), dinitrogen oxide ($N_2O$), sulfur hexafluoride ($SF_6$), hydrofluorocarbons (HFCs) and perfluorocarbons (PFCs) (Table 1). The treaty was signed and ratified by 187 countries and entered into effect on 2005, after being ratified by at least 55 of the most industrialised countries which accounted in total for at least 55% of the total $CO_2$ emissions for 1990 ("55%" clause). USA and Australia did not ratify the treaty; China, India and Brazil had no targets of reduction. By 2012 the signatory countries should have fulfilled the cut of GHG emissions of 5.2% below the 1990 level (–8% for European Union); the reduction target 2013-2020 should be –18%. European Union met the objective of Kyoto Protocol by 2011.

In the 21st Paris Climate Conference (COP21, 2015), an agreement was signed by 195 countries and entered in force in 2016. For the first time the countries signatories agree to carry out actions to limit the increase of the Earth's temperature in the range 1.5 - 2 °C above pre-industrial levels; the increase of temperature from today should be comprised between 0.65 °C and 1.15 °C. Each country is committed to provides the GHG inventories every five years, starting from 2023. However, it must be pointed out that the Paris Protocol is not a legally binding treaty, and, additionally, a country that did not accomplish its reduction target may purchase carbon credits (GHG certificates) from other countries that have no reduction obligation or are below their reduction target. In 2017 Donald Trump declared he is going to withdraw US from the Paris Agreement, which was previously signed by the former US President Barack Obama.

To keep the temperature increase below 2 °C relative to pre-industrial level, the $CO_2$ concentration in the atmosphere by 2100 should be about 450 ppm, compared to current 410 ppm. The fulfilment of that objective relies on some strategies, namely reducing fossil fuel

**Table 1.** Contribution of each gas to global GHG emissions, relative to $CO_2$, based on the amount of gas emitted and on the relative global warming potential ($GWP_{100}$).

| | $GWP_{100}$ | emissions ($CO_2$eq) |
|---|---|---|
| $CO_2$ | 1 | 76% |
| $CH_4$ | 21 | 16% |
| $N_2O$ | 310 | 6% |
| HFC/PFC[a] | 650 ÷ 11,700 | 2%[b] |
| $SF_6$ | 23,900 | |

[a] hydrofluorocarbons (HFCs) and perfluorocarbons (PFCs); [b] summed fluorides.

combustion increasingly substituted by renewable energy sources, improving the efficiency of energy production and use, enhancing the $CO_2$ capture from large-point sources, the so called Carbon Capture and Sequestration (CCS) technology. Without mitigation scenarios, by 2100 the $CO_2$ concentration in the atmosphere is expected to increase up to 750 ppm and the Earth's surface temperature between 3.7 to 4.8 °C. Obviously, the mitigation objectives cannot be an obstacle to the increasing food production and to the socio-economic development of the world population that is expected to grow to at least 9 billion over the next 35 years.

From the data reported in Table 1 it is clear that the greatest contribution to the overall GHG effect comes from $CO_2$ emissions, mainly originating from fossil fuel combustion in power plants, transportation and building heating. Livestock farming, agricultural and other land use, waste management, account for most of non-$CO_2$ ($CH_4$ and $N_2O$) GHG emissions. Due to their sparse point sources, most of the non-$CO_2$ emissions cannot be abated. Consequently, the strategies aimed at reducing the overall GHG emissions should be focused on the abatement and capture of $CO_2$ emissions from the energy sectors (fossil fuel power generation without CCS technology should phase out by 2100),[8] industry and transport. In summary, most of the sectors of the human activities must be redirected towards a sustainable low–carbon economy. Replacing coal and oil by less carbon containing fuels in all of the sectors of energy production, are feasible objectives. For instance, an immediate great contribution to the $CO_2$ emission abatement from combustion (between 11% and 25%) should be gained by replacing carbon rich fossil fuels with natural gas ($CH_4$).

The global GHG emissions by economic sector are reported in Table 2.

Low carbon electricity must play a crucial role in accelerating the global transformation to a low-carbon society, by substantially increasing the use of renewable technologies: photovoltaic cells, wind farm, solar energy, will continue to grow and to become cheaper and more competitive compared to fossil fuel combustion. However, it must be pointed out that wind and solar are intermittent energy sources, and their transformation and storage in the form of chemical energy would be a feasible solution. Nuclear energy also cannot be omitted, even though in Europe its contribution is decreasing; however, contrary to popular belief and mass media information, 59 new nuclear reactors are under construction around the world.

**Table 2.** The global GHG emissions percentage by economic sector.[a]

| | |
|---|---|
| Power plants | 38% |
| Agriculture and forestry | 22% |
| Transport | 20% |
| Buildings | 10%[b] |
| Industry | 10%[b] |

[a] there is poor agreement amongst different sources on the share of individual sector: the data are adapted from references 8 and 11; [b] doesn't comprise the consumption of electricity.

In Europe, the production of electricity by renewable sources (wind, solar, biomass) should increase from the current 32% to 80% by 2050. Afforestation, reduced deforestation and bioenergy production are natural sinks of $CO_2$. To date, the decarbonisation of energy generation occurs at a greater rate than in industry, building and transport sectors. As worldwide transportation sector accounts for about 20% of $CO_2$ emissions from fossil fuel combustion, it is expected a substantially reduction of the $CO_2$ emissions attained from technological innovations that include more efficient thermal engines, cleaner fuels (natural gas, biofuels produced by biomass and regenerated fuels), light materials and electric propulsion systems. Hybrid, plug-in-hybrid and full electric vehicles (powered by improved batteries or fuel cells) should eventually replace those equipped with thermal engines. By 2025, it is expected that the electric cars equipped with more efficient batteries will have cruising range over 600 km and substantial reduction of charge time. Sustainable biofuels should replace kerosene in aviation and diesel fuel in heavy duty trucks.

The building conditioning should reduce their $CO_2$ emissions by about 90% by 2050: this objective can be achieved by the new zero-energy buildings, and by refurbishing as much as possible the yet existing buildings, in particular the commercial and tertiary ones.

The industrial sector, especially the cement and steel production, could reduce their GHG emissions (mostly $CO_2$) by about 80% with more energy efficient processes and increased recycling of the wastes and by-products. Also, CCS technology should be applied to reduce $CO_2$ emissions of the industrial sector.

The agricultural sector is expected to have a less impact in the GHG reduction with a non-$CO_2$ GHG emission ($CH_4$ and $N_2O$) reduced by 45-50%, thanks to an improved land and fertiliser use, improved livestock farming, and bio-gas recovery from organic manure. Moreover, improved agricultural and forestry activities can increase $CO_2$ sink and can provide feedstock for energy and industry. It must be considered that the biosphere (land and oceans) takes part to the global cycle of $CO_2$ through photosynthesis of green plants and phytoplankton, that represent a natural 50% sink of the global anthropogenic emissions of $CO_2$.

It is hard to believe that hydrogen could be a substitute of fossil fuels in a short-term (hydrogen economy), even if it could have a crucial role in the conversion of $CO_2$ into liquid fuels. Hydrogen is currently produced by fossil fuels (mostly from methane), because its production from water electrolysis is not cost-effective.

Finally, an appreciable contribution to the mitigation scenarios could be given by a less energy consuming lifestyle of the population of the most developed countries, for instance less mobility demand, less energy use in households, choice of longer-lasting products, less disposable items, reduction in food wastes; moreover, it would be highly beneficial recycling wastes into industrial new products (Italy is a European leader in this field).

Noticeable, thanks to the policies of low-carbon technologies, yet worldwide adopted mostly for energy production, the global emissions of $CO_2$ remained stable to 35.8 Gton $CO_2$/year in the last three years (2014-2016). On the contrary, the GHG emissions increased in 2017 because of the growing industrial emissions that weren't compensated by the increased energy production by renewables and by the reduction of coal use.

Benefiting from low-carbon energy sources and energetic efficiency, European Union has set up the ambitious objective of the following reductions compared to 1990, to be completed before 2020 (before 2030 in parenthesis):[11]

1) 20% (40%) reduction of GHG emissions;
2) 20% (27%) of the overall energy from renewable sources;
3) 20% (27%) of the increase of energetic efficiency.

Currently the 26% reduction of $CO_2$ emissions has been attained in Italy. By 2050 the reductions of GHGs in the 28 countries of the European Union should be: $CO_2$ – 63%; $CH_4$ – 60%; $N_2O$ – 26%. Nevertheless, it must be pointed out that Europe accounts for only 9.6% of the worldwide $CO_2$ emissions (compared to 14.0% of US and 29.2% of China; 2016 data),[10] and once these objectives were reached, they would not sufficient for the 2 °C target.

Last but not least scenarios are the technologies of $CO_2$ capture from large point sources (the $CO_2$ concentration in the exhaust gases may be comprised between 5% and 40% v/v), such as fossil fuelled power plants and some industrial processes, and the safe $CO_2$ storage underground (CCS technology). Notwithstanding its low concentration (0,04% v/v), $CO_2$ can be also captured directly from air (DAC technology). Contrary to the $CO_2$ storage, in the carbon capture and utilization option (CCU technology), pure $CO_2$ could be used as a feedstock for producing chemicals and fuels.

## TECHNOLOGIES OF $CO_2$ SEPARATION FROM GAS MIXTURES

The $CO_2$ separation from gas mixtures is a technology applied at industrial scale in hydrogen and ammonia production, natural gas processing and sweetening. These methodologies can be also applied to large fixed-point sources, such as cement and steel production, and to post combustion gases from fossil fuelled power plants, the main sources of GHG emissions (Table 2). Chemical capture of $CO_2$ by a liquid alkaline solution (the absorbent) is recognized as the most efficient technology for dilute $CO_2$ (low partial pressure) removal from a gas mixture.

Different technologies for $CO_2$ capture have been also proposed, based on physical methods, cryogenic and membrane separation processes, biological fixation, but none of them went into application to large scale separation of $CO_2$ from exhaust gases because of the low efficiency or high costs.

In this section, an overview of the chemical capture of $CO_2$ with possible application to power plants is presented.[12,13]

Combustion of fossil fuels with air produces exhaust gases containing 4-15% (v/v) $CO_2$, $N_2$ (from air), with residual $O_2$, water vapour, and variable amount of sulfur and nitrogen oxides as well as particulate matter. The $CO_2$ percentage depends on the carbon content of the fossil fuel and the technology employed: the lowest value refers to a gas turbine combined cycle, where the combustion is accomplished with a large excess of air.

A typical coal-fired power plant of 1000 MW can emit about $3 \cdot 10^6$ $m^3/h$ of exhaust gases containing 15% (v/v) of $CO_2$.[14] The storage underground of that huge amount of combustion gases is not a feasible option, because of the high compression costs and of the very large geologic reservoirs where the gas mixture should be stored. On the other hand, the storage in the deep sea is not safe, and would increase the water acidity which is harmful for sea life. Therefore, it is firstly necessary to remove $CO_2$ from the gas mixture, afterwards the nearly pure $CO_2$ is compressed and injected underground (carbon capture and storage, CCS technology). An accurate geological investigation must be performed to select the site of $CO_2$ storage, that should reduce as most as possible leakage in time of sequestered $CO_2$ from the reservoirs.[15]

The employed technologies for the chemical capture of $CO_2$ are substantially similar to each other and differ, at most, in the liquid absorbents. To be a cost-effective process and to avoid millions of tons of wastes per year (the carbonated absorbent), the $CO_2$-loaded absorbent

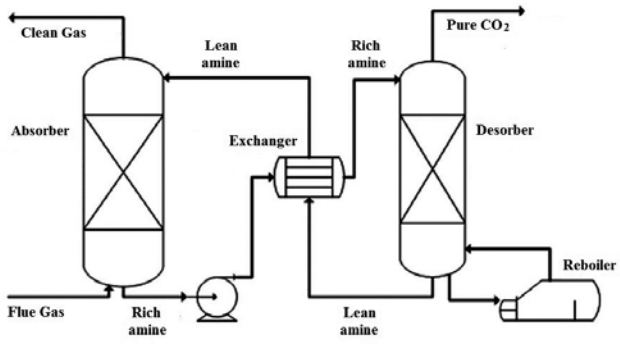



**Figure 3.** A simplified flow sheet for the $CO_2$ removal process

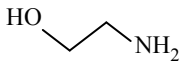
must be regenerated and recycled: the reactions of $CO_2$ with the absorbent must be reversible.

The equipment for $CO_2$ capture comprises the stainless-steel absorber (the scrubber) and desorber (the stripper) units connected to each other through a heat exchanger (Figure 3). The absorber and the desorber are packed columns that maximize the gas-liquid exchange surface, thereby enhancing the reaction rate. The absorbent circulates continuously between the two devices in a continuous cyclic process. The gas stream (12-15 % $CO_2$ v/v) is injected to the absorber (kept at about 40-50 °C) and the carbonated solution exiting from the absorber is preheated by the cross-heat exchanger and sent to the desorber where it is heated to 110-130 °C (at pressure of 1-2 bar) by steam. The regenerated solution is cooled and then it is circulated back to the absorber and reused for further $CO_2$ capture. Finally, the nearly pure $CO_2$ released from the top of the stripper can be compressed at 100-200 bar and transported to the storage site by a pipeline.

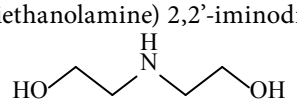
The size of the equipment to be fitted in a power plant is proportional to the flow rate of the exhaust gas *i.e.* to the amount of $CO_2$ to be captured. The height/diameter of the packed columns may be 15 m/7 m for the absorber and 10 m/4.5 m for the desorber; the plants have a capacity of $CO_2$ capture in the range of $3\text{-}4{\cdot}10^6$ tonne/year.

Most of the absorbents for $CO_2$ removal from gas mixtures are based on aqueous solutions of primary and secondary alkanolamines;[12-17] a few examples are:

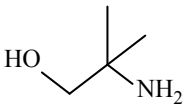
MEA (monoethanolamine) 2-aminoethanol




DEA (diethanolamine) 2,2'-iminodiethanol




AMP (aminomethylpropanol) 2-amino-2-methyl-1-propanol




The hydroxyl functionality of the amines provides their sufficient solubility in water and substantially lowers their vapour pressure, to reduce as much as possible the amine loss by evaporation. In the continuous search of more efficient absorbents, blends of amines and non-aqueous absorbents have been also investigated.[18-21] The concentration of the aqueous absorbents is usually limited to 30% (wt/wt), to reduce corrosion of the equipment and amine loss by heating, yet pursuing the target of 90% (v/v) of $CO_2$ removal from the gas stream.

The main reactions of $CO_2$ with aqueous primary and secondary alkanolamines are:

$$AmH + CO_2 + H_2O \rightleftarrows HCO_3^- + AmH_2^+ \quad (1)$$

$$2AmH + CO_2 \rightleftarrows AmCO_2^- + AmH_2^+ \quad (2)$$

where AmH denotes the free amine; $AmCO_2^-$ and $AmH_2^+$ indicate, respectively, the amine carbamate and the protonated amine. Equation (2) doesn't apply to tertiary amines that are unable to form carbamate, as well as to amines featuring steric hindrance around the amine functionality (AMP) because the carbamate is less stable than bicarbonate in aqueous solution.

The forward reactions (1) and (2) are exothermic and the reverse endothermic reactions account for $CO_2$ release and amine regeneration in the desorber.

Whatever the technology and absorbent may be used, the overall process of $CO_2$ separation from gas mixtures is energy intensive, therefore the $CO_2$ capture from a fossil fuelled power plants reduces the output electric power by 20% up to 40%, depending on the process configuration and fuel used; the cost of $CO_2$ capture from a power plant can be as high as 50-60 $/tonne $CO_2$. As a result, more fuel is consumed (additional 15-45%), more $CO_2$ is emitted that must be captured, for a given output of electric power.[22-26] The main operating cost of any process of $CO_2$ removal is the heat for absorbent regeneration, namely to reverse the exothermic absorption reactions (1) and (2). Additional energy is required to pump the absorbent within the entire apparatus and for final $CO_2$ compression. Moreover, the thermal and oxidative degradation of the alkanolamines may be another serious concern in the CCS technology.[27]

Compared to organic absorbents, very few inorganic solvents have been investigated, mainly aqueous $Na_2CO_3$, $K_2CO_3$ and $NH_3$.

Aqueous alkali carbonates do not suffer of thermal degradation and loss of the absorbent, have low regeneration energy and high absorption capacity (mass $CO_2$/ mass absorbent), but have low rate of reaction with $CO_2$.[28]

$$CO_3^{2-} + CO_2 + H_2O \rightleftarrows 2HCO_3^- \quad (3)$$

Absorbents based on aqueous $NH_3$ display fast absorption rate, significantly lower regeneration energy and thermal and oxidative stability compared to alkanolamines, but entail a major concern related to its high volatility.[29-31]

The reactions of aqueous ammonia with $CO_2$ are:

$$NH_3 + CO_2 + H_2O \rightleftarrows HCO_3^- + NH_4^+ \quad (4)$$

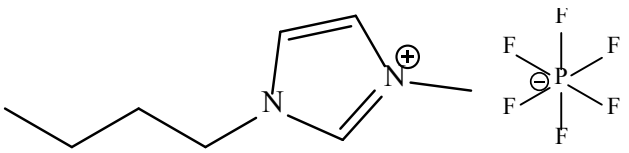
$$NH_3 + HCO_3^- \rightleftarrows CO_3^{2-} + NH_4^+ \quad (5)$$

In the absence of water, ammonium carbamate is the sole reaction product

$$2NH_3 + CO_2 \rightleftarrows NH_2CO_2^- + NH_4^+ \quad (6)$$

With the purpose of substantially reducing the energy penalty of absorbent regeneration, new absorbents based on "ionic liquids" and "demixing solvents" have been recently developed. Both methodologies avoid the heat wasted to bring the diluent to the desorption temperature (sensible heat), a significant share of the overall desorption energy; it must be pointed out that water account for 70 wt% of the aqueous absorbents. Additional cost saving and advantages come from the reduced size of the equipment and from the negligibly vapour pressure and high thermal stability of ionic liquids.

Ionic liquids are organic salts in the liquid phase at room temperature (RTILs): as an example of a common ionic liquid, the chemical structure of 1-butyl-3-methylimidazolium hexafluorophosphate ([BMIM]$PF_6$), a common ionic liquid is reported.




One-component RTILs containing an amine functionality or mixtures of RTILs and alkanolamines have been exploited for the $CO_2$ capture.[32-34] Because those absorbents are liquid before and after the $CO_2$ capture, no added diluent is necessary. To overcome the intractable viscosity of most of the carbonated absorbents based on RTILs, commercially available and inexpensive secondary amines (2-(butylamino)ethanol, for example) have been recently formulated[35,36] that reversibly react with $CO_2$ at room temperature and pressure to form liquid carbonated species without any aqueous or organic diluent.

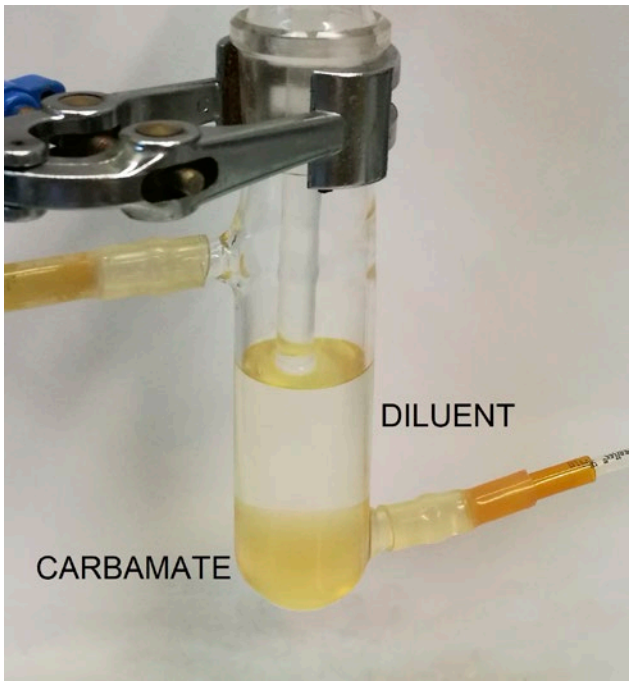



**Figure 4.** Two liquid phase recovered from $CO_2$ capture: the lower phase is the carbonated absorbent and the upper phase is predominantly the diluent with a small amount of the amine carbamate.

Demixing solvents are based on two liquid-liquid phase separation. Upon $CO_2$ capture, some aqueous or non-aqueous amines split into two separate, immiscible, liquid phases (Figure 4) which separate by virtue of their different density.[37,38] Only the lower phase that contains the carbamate and the protonated amine must be thermally regenerated, thus avoiding to heat the diluent in the upper phase.

## DIRECT $CO_2$ CAPTURE FROM THE ATMOSPHERE

The objective of zero-emission energy should be fulfilled by 2100 in most of the developed countries. Meanwhile, the lifetime of $CO_2$ in the atmosphere and the inertia of the climate change, strongly suggest to reduce

**Figure 5.** Proposed design to capture 1 million tonnes of $CO_2$ per year. Photo-illustration: courtesy of Carbon Engineering Ltd.

the $CO_2$ concentration in the atmosphere. Moreover, the direct $CO_2$ capture from air (DAC technology) is the only method to contrast the dispersed emissions from transport, heating systems of buildings and biomass burning, that cannot be captured at their sparse sources. A comprehensive overview of DAC is provided by the American Physical Society report (June 2011).[39]

The DAC method is at the early stage of investigation and no proposed process is today suitable for large scale application because of the low efficiency and high costs. Because of the very low concentration of $CO_2$ in the air (0,04% v/v), large air-absorbent contactors are necessary equipped with many fans to blow air to the absorber (Figure 5).

The absorbents so far used are concentrated aqueous solutions of NaOH or KOH (2–3 mol $dm^{-3}$) which capture $CO_2$ as soluble $Na_2CO_3$ or $K_2CO_3$; the efficiency of $CO_2$ capture is usually no more than 50%.[40] To be a feasible process, the hydroxide regeneration is accomplished with lime

$$2Na^+ + CO_3^{2-} + Ca(OH)_2 \rightarrow CaCO_3 + 2Na^+ + 2OH^- \quad (7)$$

Once separated from the solution, calcium carbonate is calcinated at 900-1000 °C to restore quicklime (CaO) and to release $CO_2$

$$CaCO_3 \rightarrow CaO + CO_2 \quad (8)$$

The entire energy requirement of the process has been estimated 17 GJ/tonne $CO_2$ captured ($4.7 \cdot 10^6$ kWh/tonne $CO_2$ captured) and about half is due to the calcium carbonate calcination.[41] The production of the same amount of energy (thermal and electric) from coal combustion, releases in the atmosphere 1.89 ton of $CO_2$: more $CO_2$ is emitted than captured! The $CH_4$ combustion produces less $CO_2$ but it doesn't compensate the investment, maintenance and overall operational costs. To make the DAC technology attractive, it is mandatory to produce the energy to run the process (thermal and electrical) with photovoltaic cells and solar heat concentration. Benefiting of the advantage of the DAC technology that can be placed everywhere, areas with higher solar radiation should be preferred. Moreover, the aqueous NaOH or KOH solutions must be replaced by new absorbents that require less regeneration energy, yet maintaining sustainable efficiency. If that method will be successfully implemented at a pilot-scale, $CO_2$ will be captured from air by using the solar radiation, as green plants are used to do.

## FROM $CO_2$ TO VALUABLE PRODUCTS

At present, the carbon capture and utilisation (CCU) technologies are non-profit options, because of their high costs. Notwithstanding, the CCU technology is more and more studied, because it has the potential of converting $CO_2$ into value-added chemicals and synthetic fuels, combined with the mitigation of $CO_2$ emissions, yet at a low extent.[42-48] In other words, the energy depleted $CO_2$ is captured and converted into reusable chemical energy, contrary to the $CO_2$ storage underground of CCS technology. It must be pointed out that CCS technology can store underground billions of tonnes $CO_2$ per year (about six million per year from a single 1000 MW power plant), whereas CCU relies on different products that overall could capture millions of tonnes of $CO_2$ per year.

The very high stability of $CO_2$ ($\Delta G° = -395$ kJ $mol^{-1}$) is a great advantage in the energy production from the combustion of carbon containing fuels [equation (9), for example], but has an adverse effect on its reactivity. For instance, the reverse of reaction (9) is thermodynamically disfavoured, whereas the reduction of $CO_2$ with hydrogen, [reaction (10)], features a severe kinetic obstacle; much energy together with catalysts therefore are necessary to convert $CO_2$ into useful chemicals.

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O(g) \quad (9) \quad \Delta H° = -803 \text{ kJ mol}^{-1}, \ \Delta G° = -801 \text{ kJ mol}^{-1}$$

$$CO_2 + 3H_2 \rightarrow CH_3OH + H_2O \quad (10) \quad \Delta H° = -131 \text{ kJ mol}^{-1}, \ \Delta G° = -9 \text{ kJ mol}^{-1}$$

Europe is leader in the study of the CCU technology, in particular Germany, thanks to its long-lasting traditional leadership in the chemical industry. The first company that has demonstrated (2015) the feasibility of the production of a liquid fuel from $CO_2$, $H_2O$ and renewable energy is based in Dresden.

Without any doubt, the most challenging option of CCU is the conversion of $CO_2$ into liquid fuels (power to liquid technology, PtL), to reduce the dependence from the fossil fuels and to address the progressive decarbonisation of the fuels for the transportation sector (an example of the so called circular economy). The most promising PtL technology is the methanol production,[49] obtained by reacting $CO_2$ with hydrogen [equation (10)]. To increase its rate, the reaction is accomplished at 200 °C with copper-based catalysts; notwithstanding, the yield of reaction is no more than 40%, based on today technologies. The cost, mainly due to the cost of electricity, is estimated to be about 600-700 euro/tonne $CH_3OH$, which is not competitive with the standard production of methanol from methane, and with the methane itself as a fuel. To be sustainable, the reaction (10) must be accomplished with solar and wind energy, so that intermittent and fluctuating energy is stored as disposable chemical energy of methanol. Methanol, directly or in blends, can be used as fuel for thermal engines in transportation, or converted into gasoline (methanol to gasoline, MtG, process) or into dimethyl ether, a possible substitute of propane, a liquefied petroleum gas (LPG). Liquified DME has been also proposed as an alternative fuel to diesel for compression ignition engines. Combustion of DME eliminates particulate and greatly reduces nitrogen oxides from exhaust emissions, compared to conventional diesel fuel, but at the expense of about half energy density.[50]

Biofuels as alternative to the fossil fuels are currently produced at industrial scale (millions of tonnes every year), mainly in Brazil and USA. Gasoline blended with 25% up to 85% of ethanol is delivered in USA, and ten million of vehicles in Brazil are fuelled by 100% ethanol.[51]

All the efforts to imitate the photosynthesis of the green plants that converts sunlight into chemical energy are failed because the energy costs to produce useful chemicals from artificial photosynthesis by far overcome the energy output of the combustion of those artificial fuels. Consequently, it is much more advantageous to allow the nature make most of the work. Based on that strategy, ethanol is produced in Brazil from sugarcane, whereas corn is the main feedstock in USA. Biodiesel as alternative fuel for diesel engines is produced with the alkaly-catalyzed transesterification process which converts vegetal oils into methyl or ethyl esters, featuring a reduced viscosity compared to the natural sources.[52]

The production of biofuels points out some problems.[51] The cost of the raw material (planting, irrigation, fertilization, harvesting and transportation) accounts for 60% to 75% of the cost of biodiesel producing. If the life cycle assessment of the process is taken into account, the biofuels are still not a viable alternative to fossil fuels, in the absence of the government support. As a final consideration, it should be a better option to use farmland for food production instead of crop-based biofuels.

In the search of nonedible sources of biofuels, any form of biomass can be converted into a liquid fuel by means of a thermochemical process, but at unsustainable costs. In that contest, algae-based biodiesel has emerged as a promising option, because it doesn't entail a reduction of food production and features a substantially higher photosynthetic efficiency compared to land crops.[53,54]

Using $CO_2$ for the manufacture of plastics and speciality chemicals is a further option to store and re-use $CO_2$. However, the estimated worldwide production of such products is about 180 million tonnes every year, that corresponds to less than 1% of the anthropogenic $CO_2$ emissions. Compared to the production of fuels, the production of chemicals doesn't have an appreciable impact on the reduction of $CO_2$ emissions.

Taking the advantage of the thermodynamically favoured and fast acid-base reactions between $CO_2$ and $NH_3$, it has been recently developed an innovative process that integrates the $CO_2$ capture with the production of urea, the most worldwide used nitrogen fertilizer, more than $10^8$ tonne/year. The $CO_2$ capture (15% v/v in air) in water-ethanol produces solid mixtures of ammonium bicarbonate and carbamate [reactions (4), (6)]. By heating the solid mixtures at 165 °C in a closed vessel without any external pressure, both ammonium carbamate, and bicarbonate are converted into urea.[55,56]

$$NH_2CO_2NH_4 \rightleftarrows NH_2CONH_2 + H_2O \quad (11)$$

$$2NH_4HCO_3 \rightleftarrows NH_2CONH_2 + CO_2 + 3H_2O \quad (12)$$

The industrial production of urea is carried out with $NH_3$ and purified $CO_2$ in the gas phase at high temperature (180 –230 °C) and pressure (150 – 250 bar). Pure $CO_2$ is obtained by the conventional aqueous amine scrubbing and thermal stripping. The advantage of process based on the solid ammonium salts compared to the industrial process, is the potential energy saving because both the $CO_2$ purification step with aqueous amine scrubbing and the high pressure working are avoided, yet with efficiency (about 47% with respect to $NH_3$) and reaction time (60 min at most) comparable with the industrial process.

As a final consideration, 60 million tonnes of $CO_2$ are employed in different commercial sectors every year, and are currently extracted from natural sources under-

ground. A cheap capture technology from exhaust gases yet recovering high purity $CO_2$, could replace the current $CO_2$ production that is re-emitted in the atmosphere and the end of its utilization cycle.

## CONCLUSIONS

The increased greenhouse effect originating from human activities is most likely responsible of the increase of Earth's temperature in the last century, and possibly of the climate change. The climate change has, and will have to a greater extent in the future, adverse impacts on the society development and world economy, because of the increasing extreme weather events such as storms, floods, drought and heat waves. The frequency of snowfall and rain is reduced in the recent years, but they are heavier. The objective of mitigation of climate change cannot be further delayed, and many possible actions have been proposed to reduce the GHG anthropogenic emissions. As most of the GHG emissions is due to combustion of fossil fuels, the reduction of dependence from fossil fuels would provide further benefits to the economy of most countries, whereas the improved air quality will have noticeable beneficial effects on human health.

The world economy will be more and more dependent from solar and wind energy; this form of energy is intermittent, and its storage as chemical energy (renewable fuels) and chemicals (fertilizer, plastic) by using the CCU and DAC technologies should be possible options.

Innovative solutions in all of the sectors of the human activities that include both the reduction of combustion of fossil fuel and the CCS, CCU and DAC technologies can contribute to the objective of a progressive decarbonisation of the world economy in the sectors of energy generation, transport and industry. However, that objective appears doubtful in the absence of governmental obligations and of the carbon tax. Meanwhile, adaptation strategies to the foreseen extreme events of the climate change should be adopted.

## REFERENCES


1. D. Luthi, M. Le Floch, B. Bereiter, T. Blunier, J.-M. Barnola, U. Siegenthaler, D. Raynaud, J. Jouzel, H. Fischer, K. Kawamura, T.F. Stocker, *Nature* **2008**, *453*, 379.
2. J. Jouzel, V. Masson-Delmotte, O. Cattani, G. Dreyfus, S. Falourd, G. Hoffmann, B. Minster, J. Nouet, J.M. Barnola, J. Chappellaz, H. Fischer, J.C. Gallet, S. Johnsen, M. Leuenberger, L. Loulergue, D. Luethi, H. Oerter, F. Parrenin, G. Raisbeck, D. Raynaud, A. Schilt, J. Schwander, E. Selmo, R. Souchez, R. Spahni, B. Stauffer, J.P. Steffensen, B. Stenni, T.F. Stocker, J.L. Tison, M. Werner, E.W. Wolff, *Science* **2007**, *317*, 793.
3. Data from NASA's Goddard Institute for Space Studies (GISS)
4. Dr. Pieter Tans, NOAA/ESRL (www.esrl.noaa.gov/gmd/ccgg/trends/) and Dr. Ralph Keeling, Scripps Institution of Oceanography (scrippsco2.ucsd.edu/).
5. D.M. Etheridge, L.P. Steele, R.L. Langenfelds, R.J. Francey, J.-M. Barnola and V.I. Morgan. *Carbon Dioxide Information Analysis Center*, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A, **1998**.
6. IPCC, 2001: *Climate Change 2001, Synthesis Report. Contribution to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, UK, 398 pp.
7. IPCC, 2007: *Climate Change 2007, Synthesis Report. Contribution to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, UK, 939 pp.
8. IPCC, 2014: *Climate Change 2014, Fifth Assessment Report*; www.jrc.nl.
9. *US Environmental Protection Agency*. www.epa.gov
10. *Fossil $CO_2$ and GHG emissions of all world countries*; EUR 28766, Publication Office of the European Union, Luxemburg, **2017**.
11. *European Environmental Agency*, www.eea.europa.eu/clima/policies /strategies /2020.
12. *IPCC Special Report on $CO_2$ Capture and Storage* (2005); B. Metz, O Davidson, H. de Coninck, M. Loos, and L Meyer , *Cambridge University Press*, Cambridge UK, **2005**, p.442. See :http//www.ipcc.ch/ipccreports/srccs.htm
13. *$CO_2$ capture from Existing Coal-fired Power Plants*; DOE/NETL-401/110907
14. H.H. Rao, R.B.H. Tan, *Energy Fuels*, **2006**, *20*, 1914
15. F.M. Orr Jr, *Energy Environ. Sci*., **2009**, *2*, 449
16. G. Astarita, D.W. Savage, A. Bisio, *Gas Treating with Chemical Solvents*, Wiley, New York, **1994**.
17. G.T. Rochelle, *Science* **2009**, *325*, 1652.
18. F. Barzagli, F. Mani, M. Peruzzini, *Energy Environ. Sci*., **2010**, *3*, 772.
19. F. Barzagli, M. Di Vaira, F. Mani, M. Peruzzini, *ChemSusChem*, **2012**, *5*, 1724.
20. D. Bonenfant, M. Mimeault, R. Hausler, *Ind. Eng. Chem. Res*., **2005**, *44*, 3720.
21. W.J. Choi, K.C. Cho, S.S. Lee, J.G Shim, H.R. Hwang, S.W. Park, K.J. Oh, *Green Chem*. **2007**, *9*, 594.
22. K.Z. House, C.F. Harvey, M.J. Aziz, D.P. Schray, *Energy Environ. Sci*., **2009**, *2*, 193.

23. A.B. Rao, E.S. Rubin, *Ind. Eng. Chem. Res*. **2006**, *45*, 2421.
24. J. Oexmann, A. Kather, *Int. J. Greenhouse Gas Control* **2010**, *4*, 36.
25. E.J. Stone, J.A. Lowe, K.P. Shine, *Energy Environ. Sci*. **2009**, *2*, 81.
26. J. Davison, *Energy* **2007**, *32*,1163.
27. A.J. Reynolds, T.V. Verheyen, S.B. Adeloju, E. Meuleman, P. Feron, *Environ. Sci. Technol*. **2012**, *46*, 3643.
28. B. A. Oyenekan, G.T. Rochelle, *Int. J. Greenhouse Gas Contr*. **2009**, *3*, 121.
29. V. Darde, K. Thomsen, W.J.M. van Well, E.H. Stenby, *Int. J. Greenhouse Gas Control* **2010**, *2*, 131.
30. H. Huang, S-G. Chang, T. Dorchak, *Energy Fuels* **2002**, *16*, 904.
31. A. Pérez-Salado Kamps, R. Sing, B. Rumpf, G. Maurer, *J. Chem. Eng. Data* **2000**, *45*, 769.
32. D. Camper, J.E. Bara, D.L. Gin, R.D. Noble, *Ind. Eng. Chem. Res*. **2008**, *47*, 8496.
33. C. Wang, X. Luo, H. Luo, D. Jiang, H. Li, S. Dai, *Angew. Chem., Int. Ed*. **2011**, *50*, 4918.
34. D.J. Heldebrant, C.R. Jonker, P.G. Jessop, L. Phan, *Energy Environ. Sci*. 2**008**, *1,* 487.
35. F. Barzagli, S. Lai, F. Mani, *ChemSusChem* **2015**, *8,* 184.
36. F. Barzagli, F. Mani, M. Peruzzini, *Environ. Sci. Technol*. **2016**, *50*, 7239.
37. Q. Zhuang, B. Clements, J. Dai, L. Carrigan, *Int. J. Greenhouse Gas Control* **2016**, *52*, 449.
38. F. Barzagli, F. Mani and M. Peruzzini, *Int. J. Greenhouse Gas Control* **2017**, *60*, 100.
39. R. Socolow, M. Desmond, R. Aines, J. Blackstock, O. Bolland, T. Kaarsberg, N. Lewis, M. Mazzotti, A. Pfeffer, K. Sawyer, J. Siirola, B. Smit, J. Wilcox, *Direct Air Capture of $CO_2$ with Chemicals: A Technology Assessment for the APS Panel on Public Affairs*, American Physical Society, **2011**.
40. M. Mahmoudkhani, K.R. Heidel, J.C. Ferreira, D.W. Keith, R.S. Cherry, *Energy Procedia* **2009**, *1*, 1535.
41. R. Baciocchi, G. Storti, M. Mazzotti, *Chem. Eng. Process*. **2006**, *45*, 1047.
42. X. Yin, J. R. Moss, *Coord. Chem. Rev*. **1999**, *181*, 27.
43. A.S. Bhown, B.C. Freeman, *Environ. Sci. Technol*. **2011**, *45*, 8624.
44. P. Markewitz, W. Kuckshinrichs, W. Leitner, J. Lissen, P. Zapp, R. Bongartz, A. Schreiber, T.E. Müller, *Energy Environ. Sci*. **2012**, *5*, 7281.
45. M. Aresta, A. Dibenedetto, A. Angelini, *J. $CO_2$ Utilization* **2013**, *3-4*, 65.
46. G. Centi, E.A. Quadrelli, S. Perathoner, *Energy Environ. Sci*. **2013**, *6*, 1711.
47. V. Barbarossa, G. Vanga, R. Viscardi, D.M. Gattia, *Energy Procedia* **2014**, *45*, 1325.
48. A. Goeppert, M. Czaun, J.P. Jones, G.K.S. Prakash, G.A. Olah, *Chem. Soc. Rev*. **2014**, *43*, 7995.
49. G. A. Olah, A. Goeppert, G. K. S. Prakash, *Beyond Oil and Gas: The Methanol Economy, Second updated and enlarged edition*, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, **2009**.
50. A.M. Namasivayam, T. Kovakianitis, R.J. Crookes, K.D.H. Bob-Manuel, J. Olsen, *Applied Energy* **2010**, *87*, 769.
51. J. Hill, E.Nelson, D. Tilman, S. Polaski, D. Tiffant, *Proc. Natl. Academy Sci. USA* **2006**, *103(30)*, 11206.
52. L.C. Meher, D. Vidya Sagar, N.S. Naik, *Renewable and Sustainable Energy Reviews* **2006**, *10*, 248.
53. G. Huang, F.Chen, D. Wie, X.W. Zhang, G. Chen, *Applied Energy* **2010**, *87*, 38.
54. P.T. Vasudevan, M. Briggs, *J. Ind. Microb. Biotech.* **2008**, *35*, 421.
55. F. Barzagli, F. Mani, M. Peruzzini, *Green Chem*. **2011**, *13*, 1267.
56. F. Barzagli, F. Mani, M. Peruzzini, *J. $CO_2$ Utilization* **2016**, *13*, 81.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Feature Article

# The 'Consciousness-Brain' relationship

Jean-Pierre Gerbaulet[1], Pr. Marc Henry[2]

[1] *N-LIGHT Endowment Fund, 30 rue de Cronstadt, Paris*
[2] *Université de Strasbourg, UMR 7140, 4 Rue Blaise Pascal, 67000 Strasbourg*
E-mail: jpg@n-light.org

**Abstract**. From a thought experiment on the observation of a human intellect by itself, we will attempt to demonstrate that, unlike what many neuroscientists postulate, assemblies of neurons do not generate consciousness: Consciousness pre-exists any material system.

**Keywords**. Consciousness, meaning, information, activity, neurons.



**Citation:** J.-P. Gerbaulet, Pr. Marc Henry (2019) The 'Consciousness-Brain' relationship. *Substantia* 3(1): 113-118. doi: 10.13128/Substantia-161



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Competing Interests:** The Author(s) declare(s) no conflict of interest.

## INTRODUCTION

Understanding that the nature of consciousness is a real challenge for Western cultures which heavily focus on the scientific method for understanding natural phenomena, one usually refers in this case to the "hard problem".[1]

By contrast, Eastern cultures traditionally adopt philosophical approaches to the problem, such as Hinduism[2] or Buddhism[3], with a notable Western exception (Eckhart Tolle).[4]

In a nutshell, three main visions are fighting each other over tackling the "hard problem" from the Western side.[5]

A first position is *physicalism*, a kind of monism stating that physical laws are perfectly valid for explaining the existence of both mind and body. Such a vision (Thales, Leucippus, Democritus, Epicurus) is a broader version of materialism taking for granted that there exists in the universe, in addition to matter, energetic phenomena such as electromagnetism, that are physical and real. In this view, physical states (size, mass, shape, energy, etc.) and mental states (beliefs, desire, emotions, etc.) are made of the same "stuff".

A second position is *dualism* (Plato, Descartes), stating that mental and physical states are both real and made of two different materials that cannot be assimilated to one another.

Finally, a third position is *illusionism,* stating that consciousness simply does not exist and involves some sort of introspective illusion. According to D.J. Chalmers, this illusion is a close relative to the meta-problem of consciousness, i.e. the problem of explaining why we think that there is a problem of consciousness. In fact, illusionism states that distinguishing between

easy problems and the hard problem distracts our attention from the hard question which is: "And then what happens"?[6-8]

In contrast with the above approaches, we would like to draw your attention to an Einstein's remark made in the context of "how to deal with the threat of the atom bomb":

"*A new type of thinking is essential if mankind is to survive and move toward higher levels*".[9]

Owing to the generality of this statement, such a remark has been widely diffused out of its context in several versions, among which we shall retain this one:

"*No problem can be solved from the level of consciousness that created it*".

Such a formulation is quite reminiscent of Gödel's incompleteness theorems.[10] Applied to the 'hard problem' or the 'hard question' of consciousness, it means that the bottom-up logic, typical of western thinking, in which consciousness is the result of long-range coherence in neural activity[11], may be considered as a dead-end. If a theoretical model has recently been proposed for decoding brain wave information,[12] it remains that it does not address subtle aspects of consciousness.

It is thus our deep conviction that a "new" approach (as far as Western minds are concerned) is to consider a top-down logical process inspired by Eastern thinking where consciousness pre-exists any material system such as neurons or brain.

In other words, we plan to demonstrate that consciousness cannot be an emergent property of neural activity. Owing to the importance of such an assertion for Western minds, the demonstration proposed in this paper is concise and readable by non-scientists. A more technical and scientific demonstration is published as a separate paper showing how this top-down approach fits into current scientific knowledge.[13]

## DEFINITIONS

Our aim in this paper is to give a wide audience access to the concise demonstration of the logical necessity to consider consciousness as the source of reality. The presentation has thus necessarily many gaps that will be addressed in a forthcoming article.

Among the gaps, the very first one is a *good definition of consciousness. We sincerely think that the best way of handling the consciousness concept is to assign it an "identity card" in order to recognize it by its manifestations in space and time.*[14] On this ground, we state that consciousness is the tool that allows us to find a meaning in information, either analyzed by intelligence or coming directly from feelings and intuition (qualia). It *a priori* applies to most living beings.

Our demonstration below thus necessarily implies the existence of two other dimensions (one space-like, the other one time-like) located outside a 4D space-time framework.[15] With such two extra-dimensions, consciousness would acquire an extra-human value and it would then be designated by Consciousness. It has as real (probably more real) features than our so-called "objectivity" attached to our manifest 4-D space-time horizon. We will assume that the extra time dimension is the ordering element that generates different attributes within itself as illusions[16] as developed elsewhere[13].

Our line of thought in this matter is inspired by chemistry, a science where thermodynamics uses static general concepts putting constraints on dynamical aspects, which allows selecting among all possible paths the most favorable to evolution. Consequently, we shall now focus on the framework rather than on what may happen within the framework, a problem which will be addressed later.[13] Concerning dynamics, we will be considering time as an emanation of consciousness, the question of its topology (linear, curved or fractal) being thus irrelevant to our demonstration.

Similarly, we have introduced the conceptof **activity**, which is generally used in thermodynamics to combine energy and entropy within a single entity. We therefore recommend reading "energy/entropy" whenever you come across the word "activity", unless you are familiar with thermodynamics.[17] And if you are reluctant to the concept of entropy, just think "energy". It is close enough to make laypeople understand the idea.

## SCIENTIFIC BASES, POSTULATE

To demonstrate the priority of consciousness over neurons, we will use principles of computer science,[18,19] information theory,[20] Gödel's incompleteness theorems[10] and the laws of thermodynamics.[17]

And we will refer to the following postulate: ***any phenomenon preexisting another one is able to participate in the creation of the latter, whereas the contrary is impossible***.

This postulate, which conditions the possibility to create a principle from another one, should clearly explain how a space-time-matter framework used by conventional science is able to emerge from a non-local Consciousness following a hierarchical cascade, hereafter named "the thought experiment", where a person observes the functioning of his/her own intellect.

## CONTEXT AND DESCRIPTION OF THE THOUGHT EXPERIMENT

The thought experiment that we will propose relates to what is called in psychology: metacognition. Some evolutionary psychologists hypothesize that humans use metacognition as a survival tool, which would make metacognition the same across cultures. Writings on metacognition date back as far as two works by the Greek philosopher Aristotle (384-322 BC): "*On the Soul*"[21] and the "*Parva Naturalia*".[22] Today, metacognition is studied in the domain of artificial intelligence and modelling. Therefore, it is the main domain of interest of emergent systemics.

In such an experiment, the Subject and the Object are the same since the person observes his/her intellect by means of the latter. Although all the parameters of the Subject and the Object are identical, they operate in the self-observation process at different chronological and hierarchical levels. The result is that the situation can be summarized by the relationship between five protagonists: **consciousness, meaning, information, activity and neurons**.

Organized in couples, their specific relationship allows for the proper functioning of the whole:

- **Consciousness and meaning,**
- **Meaning and information,**
- **Information and activity,**
- **Activity and neurons**.

### *Consciousness-meaning*

The intellect is a system comprising, by analogy with a computer[23], a *hardware* (material device) and several types of *software* (immaterial devices). The difference with a computer is that the physical entity is able to repair itself by creating *de novo* material components (cells) necessary to its proper functioning.

In the software-hardware couple composing a computer, hardware without software would only be a set of 'dumb' electronic circuits: central unit, memories, I/O interfaces, peripherals. Even if Artificial Intelligence equipped computers are able to write software, to self-educate and self-duplicate themselves, even to self-improve their level of performance, they have initially been fitted with software designed by conscious beings, without which they would be unable to operate.

Moreover, electronic components are designed and manufactured by conscious beings, not by the computers themselves using 3D-printers for instance, owing to difficulties in implementing evolutionary processes and to the "salt contingency problem" raised by Alex Ellery in 2017.[24]

In a computer, *since software gives life to hardware, it has a functional anteriority over hardware.*

Now, in a computer, the process of cognition and memorization is based on the manipulation of binary digits, the so-called "bits" (with just two possible values 0 and 1), a succession of such bits being called "information". An important aspect is that, at computer level, such information has no meaning, even if bits are combined and manipulated according to logical rules inferred from the existence of consciousness. Meaning only appears as soon as information is combined with consciousness.[25]

***Thus, it is consciousness that gives a meaning to information, and thereby possesses a functional anteriority over meaning.***

### *Meaning-information*

The way consciousness gives meaning to information is by considering pieces of information which, once compared to memorized other pieces of information, are placed in a context which gives them a meaning.

We typically find ourselves in the framework of the information theory, where *meaning is defined as information in a context*.[26] Although of a similar nature to the point to be often confused in everyday's language, information and meaning are not identical.

At the end of his life, the great physicist John Wheeler considered that, in the universe, all could be made of information.[27] In our thought experiment this basically means that, within a field of information, consciousness has the ability to select pools of information of varying sizes thus defining "objects" or "things" that could be differentiated by their respective information content. Obviously, as evidenced by the fluidity of thought, such pools of information should not be considered as static entities, but rather as dynamic systems exchanging information.

***Since it is the meaning that gives its value to a given amount of information it chronologically anteriorizes information and is, therefore, hierarchically superior to it.***

### *Information-activity*

Based on the above considerations, it follows that, in our thought experiment, characterizing pools of information solely by their number of bits is not enough. One may assume that within a given pool of information, some groups of bits that are considered by consciousness as having a high meaning will not be easily transferred

to another pool of information, since such groups of bits give an identity to the information pool. Thus, transferring them, would inevitably make the pool lose its identity. Here appears, in a logical way, the **conscious "I"** which holds a number of bits sufficient to give itself an identity within the whole information field.

This means that besides the information content, one should also introduce an information **availability** that could be low or high depending on its importance for the definition of the identity of the pool. As soon as two pools have not the same information availability, information is expected to flow from the pool having the higher availability towards the pool having the lower availability. By such information transfers, the information availability of the emitter decreases, whereas the information ability of the receiver increases, allowing pools of information to undergo evolution on two levels. At a first level, pools may just change their information content by exchanging non-meaningful bits that are readily available. At a second level, pools may also change their identity by exchanging meaningful bits that are not readily available.

It suggests introducing a new concept, **information activity**, defined as ***the product of information content by information availability***.[13] Consequently, one may meet pools having small information content that are not readily available, corresponding to a low activity pool. Conversely, pools characterized by high information content that is readily available for information transfers would be qualified as high activity pools. Such a definition of information activity has also the consequence to make duality appear within a non-dual information field. Accordingly, a given activity value may be associated either to a low information availability within a large pool of information or to a highly available information coming from a small pool of information. In the first case, activity may be associated to "moving" information allowing evolution and change in "time", while in the second case it becomes associated to "structural" information defining conservation and identity in "space". A space-time frame thus emerges quite naturally by the action of consciousness giving meaning to various pools of the information field.

From this analysis, it follows that ***information is unique*** *in the information field, whereas* ***activity*** *characterizing the intensity of information transfers* ***has a dual character*** responsible for an *energy/entropy duality* in the physical world.

Such a duality is reflected by the existence of two universal constants:

- *Boltzmann's constant* $\boldsymbol{k_B}$ *ruling the minimum information content viewed as an entropy (statistical physics)*
- *Planck's constant* ***h*** *ruling the minimum information activity for observing an energy change viewed as a frequency (quantum physics) or as a temperature (thermodynamics).*

***Consequently, one can assert that information chronologically anteriorizes activity, and is therefore, hierarchically superior to it****.*

*Activity-neurons*

Having given birth to concepts of entropy S and energy W through the concept of vibration f ($W = h \cdot f$) and temperature T ($W = k_B \cdot T$), it remains introducing the "matter" concept through a third universal constant intimately associating space to time. The reason for it clearly stems from the fact that it is the same consciousness acting on a unique information field that creates time as moving information, and space as structural information. The two concepts referring to the same amount of information should thus necessarily be linked as two different viewpoints about the same parameter depending on information availability. The basic postulate of equivalence between space and time stemming from the theory of relativity, another most important physical theory in science, is thus logically introduced.

By this definition, the third universal constant should be a ***speed c*** *i*mposing an upper limit to the transfer of moving structural information between information pools.

From the above considerations, it follows that *two kinds of elements should exist in a physical universe*: those able to propagate with the *maximum allowed speed c, known as "photons"*, and those that propagate at *speeds* $v < c$*, known as "matter"*. In the second case, one may assign to a material object with an energy E, an inertial coefficient m or "mass", linked to it by $m = E/c^2$.

Adding the two other universal constants, we may write *the fundamental identity of our physical world*:

$$\mathbf{E = m \cdot c^2 = h \cdot f = k_B \cdot T},$$

meaning that our reality is made of a combination of *inertia (mass m), spontaneous vibration (frequency f) and spontaneous movement (temperature T).*

Going back to our computer analogy, it should now be clear that neurons are likened to hardware since they are the cells dedicated to information processing. Each neuron is an information-processing unit linked to other neurons to form a network with various crucial physical nodes at the levels of brain, heart and intestines. The nodes of the network are linked together to form an **intranet-like physical body** which behaves in an

autonomous way and can be likened to a set of circuits: network nodes (brain, heart, intestines), Input interfaces (the five senses plus a sixth one relaying feelings and intuitions), Output interfaces (limbs, voice, ...), associated to neuronal, and possibly non-local, memories. The physiological complexity of the whole allows it to perform processing functions, but not interpretations.

In a nutshell, even if the **intranet-body** possesses a certain processing autonomy, the directions of its actions are given, at each stage of the process, by the meaning of the intermediate results interpreted by our **consciousness**.

It then appears that neurons, which are in the physical world the material interface for manipulating information, are located at the very end of the hierarchy described in our thought experiment.

This analysis shows that ***activity plays a role chronologically anterior hence hierarchically superior to the one of matter, making it impossible to state that consciousness emerges from the physical activity of neurons***. **It is the opposite**.

SYNTHESIS:

***We have hereby demonstrated that consciousness anteriorizes meaning, which anteriorizes information, which anteriorizes activity, which anteriorizes neurons.***

Consequently, the relationship between consciousness acting on a unique information field, and brain acting in a four-dimensional space-time, acquires in this environment the status of a law:

***LAW:***

***Consciousness preexists neurons and cannot be an emergent property of them.***

We shall deduct from it 5 corollaries, some of them remaining to be confirmed.

**Corollary 1: C*onsciousness exists independently from the neurons.***

**Corollary 2:** ***Matter originates in consciousness (spirit).***

We posit that consciousness preexists not only neurons but matter in general. By likening consciousness to spirit, one could, subject to further confirmation, deduct that matter originates in spirit.

**Corollary 3:** ***Extension to non-local consciousness:***

Subject to similar confirmation, matter, activity, information, meaning and consciousness would be states of decreasing vibratory levels of a same principle, ***the ground state of which would be pure Consciousness***, and the forms closer to this fundamental level would be subtler or less material.

We might then postulate that this fundamental state being without precursor, it would be at the origin of all that exists. There may then be a high probability that the Primordial Consciousness be located outside space-time, since being at the origin of it, it could hardly belong to it (Gödel's theorem).

This Primordial Consciousness could be named ***Non-local Consciousness***.

**Corollary 4:** ***Generalization***

- Non-local Consciousness would preexist all that exists in the observable universe or *manifest world*.
- Its expressions would be of a decreasing level when *coherence diminishes*: meaning, information, activity, and finally inert matter.
- They would be of an increasing level when *coherence grows*: structured matter (crystals), unconscious life, life conscious of the world, then of itself, and, at last, of the fact to be conscious of being conscious, this most advanced state being the one of Humanity.

By analogy with the ***geometrical fractalisation***, this cascade of levels could be named ***conceptual fractalisation***.

**Corollary 5:** ***Practical consequences***

In our thought experiment, the energy considered is a mix of chemical energy, well known by biologists (mass m and temperature T), and of electromagnetic energy (frequency f), tolerated by them. Since these energies are the ones concerning the object-oriented language, as defined in our companion paper,[13] nothing prevents from having more subtle energies working at the meta-language level, such as vital energy or Psi energy, largely ignored by mainstream neuroscientists. Using the brain computer metaphor, it may be time to update our own software.[28]

By contrast, traditional medicines commonly use these energies and the '*informational function*' of consciousness to cure patients, with track records of several millennia.

This contradiction is the main subject of interest of the experiments, underway or in project, by the N-LIGHT Research Institute, its members and its partners.

REFERENCES

1. D. J. Chalmers, *J. Consciousness Stud.* **1995**, *2*, 200.
2. G. Sri Aurobindo Gose, *The Life Divine Vol. 21 & 22*, Sri Aurobindo Ashram Publication Department, Pondicherry, **2005**.
3. H. H. Dalaï Lama, *Consciousness at the crossroad, Conversation with the Dalaï Lama on Brain Science and*

*Buddhism*, Snow Lion Publications, Ithaca, New York, **1999**.
4. E. Tolle, *Oneness With All Life: Inspirational Selections from A New Earth*, Dutton, Penguin Group, New York, **2008**.
5. D. J. Chalmers, *J. Consciousness Stud.* **2018**, *25*, 6.
6. D. C. Dennett, *Phil. Trans. R. Soc. B* **2018**, *373*, 20170342.
7. N. Block, *Behavorial and Brain Sci.* **1995**, *18*, 227.
8. J. E. Bogen, S. Bringsjord, D. Brown, D. J. Chalmers, D. Gamble, D. Gilman, G. Güseldere, A. Murat, B. Mangan, A. Alva; E. Pöppel, D. M. Rosenthal, A. H. C. van der Hejiden ; P. T. W. Hudson, A. G. Kurvink, N. Block, Ned, *Behavorial and Brain Sci.***1997**, *20*, 144.
9. A. Einstein, M. Amrine, *The New York Times Magazine* **1946**, June, 23, p.7.
10. K. Gödel, Kurt, *Monatsh. Math. Phys.* **1931**, *38*, 173.
11. S. Dehaene, J.-P. Changeux, L. Naccache, Lionel (2011), in *Research and Perspectives in Neurosciences*, (Eds.: S. Dehaene, Y. Christen), Springer-Verlag, Berlin, **2011**, pp. 55-84.
12. S. Sen, Siddhartha, *J. Consciousness Stud.* **2018**, *25*, 228.
13. M. Henry, J.-P. Gerbaulet, *Substantia (submitted).*
14. J.-F. Houssais, MD, *Les 3 niveaux de la conscience,* Guy Trédaniel, Paris, **2016**.
15. M. Henry, *Inference : Int. Rev. Sci.*, Vol. 2, issue 4; http://inference-review.com/article/super-saturated-chemistry
16. Many thanks to referee #1 for drawing our attention to this point.
17. W. J. Gibbs, *The scientific papers of J. Willard J. Gibbs, Vol. 1*, Longmans Green and Co, London, **1906**.
18. A. M. Turing, *Proc. London Math. Soc.* **1937**, *42*, 230.
19. A. M. Turing, *Proc. London Math. Soc.* **1937**, *43*, 544.
20. C. E. Shannon, *Bell Syst. Tech. J.* **1948**, *27*, 379.
21. Aristotle, "*De Anima (On the Soul)*", translated by Hugh Lawson-Tancred, Penguin Classics, **1986**.
22. Aristotle, "*Parva Naturalia (short treatises on nature)*", translated by J. I. Beare & G. R. T. Ross, Oxford, **1931**.
23. R. M. Biron, *J. Comput. Higher Educ.* **1993**, *5*, 111.
24. A. Ellery, in *Proceedings of the ECAL 2017*, (Eds.: C. Knibbe et al.), The MIT Press, Cambridge MA, **2017**, pp. 146-153
25. R. Mukhopadhyay, J. *Consciousness Stud.* **2018**, *25*, 184.
26. M. Burgin, R. Feistel, *Information* **2017**, *8*, 139.
27. J. A. Wheeler, K. Ford, *Geons, Black Holes and Quantum Foam. A Life in Physics*, W.W. Norton & Co., New York, **1998**, pp. 63-64.
28. G. Weber, *Evolving beyond thought: Updating your own software*, CreateSpace Independent Publishing Platform, **2018**.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Historical Article

# Dmitry I. Mendeleev and his time

Dmitry Pushcharovsky

*Lomonosov Moscow State University, Department of Geology, Vorob'evy gori, 1, 119899 Moscow, Russia*
E-mail: dmitp@geol.msu.ru

**Abstract.** The history of the creation of Periodic table and of the Mendeleev's discovery of Periodic Law is considered. The different approaches used by Mendeleev's colleagues are discussed. The contribution of the Periodic system to the extension of the scientific ideas in geology and best of all in geochemistry and mineralogy is illustrated by the discovery of new chemical elements and by the isomorphic replacements in minerals. The details of uneasy history of Mendeleev's nomination to the St. Petersburg Academy and for the Nobel Prize are given.

**Keywords**. Periodic table, isomorphism, Nobel Prize, electronic structure of atom.



**Citation:** D. Pushcharovsky (2019) Dmitry I. Mendeleev and his time. *Substantia* 3(1): 119-129. doi: 10.13128/Substantia-173



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Competing Interests:** The Author(s) declare(s) no conflict of interest.


Periodic table of chemical elements on the front of the main building of the Central Board of Weights and Measures in St. Petersburg; height – 9 m, area – 69 m$^2$; red colour - elements, known in the Mendeleev lifetime, blue colour – elements discovered after 1907 (Public domain)

## INTRODUCTION

The United Nations declared 2019 as the International Year of the Periodic Table. This decision is related to the 150th anniversary since its first version elaborated by the prominent Russian chemist Dmitry Ivanovich Mendeleev (1834-1907, Fig. 1) was published on the 17th of February 1869. On this date he sent his table to the publisher and simultaneously distributed it among his colleagues in Russia and abroad.

In connection with UN resolution it is necessary to address the question whether it is really urgent to discuss the events related to Mendeleev's discovery. Researchers all over the world consider that as before it contributes the further development of many scientific branches. On the basis of the Periodic Table they search the answers to the many mysteries which Nature still hides. Besides that the history of its creation clearly justifies the absolutely non-linear process which usually accompanies the scientific progress [1]. These aspects are the focus of the present paper which is devoted to some applications of the Periodic Table, to its author and to the time when he made his historical discovery.

**Figure 1.** D.I Mendeleev (photo from public domain).

## BIOGRAPHY

Mendeleev was born on the 27th of January (8th of February) 1834 in Tobolsk – the first Siberian town established in 1587, located between Ural and Western Siberia (Fig. 2). He was the last among 17 children in the family of his father Ivan Mendeleev, the director of local gymnasium, and his mother Maria Kornil'eva, a daughter of the "middle class" landowner. In the gymnasium Dmitry was not a brilliant student and had very modest marks in Latin and Scripture, however showing an evident interest in mathematics and physics. He was 10 years old when his father passed away. His mother inherited a small glass factory and she managed it until Dmitry finished gymnasium in 1849. The same year the factory was burned down and the family moved first to Moscow and then to St. Petersburg.

Mendeleev was unable to continue his education immediately. Finally one year later, in 1850, he was admitted to the faculty of mathematics and physics of the Main Teacher's Training Institute in St. Petersburg. Here he also had some problems with his studies. When he was a first year student he failed all the exams except for mathematics. However the turning point occurred at the end in 1855, when he graduated in the Institute with the excellent certificate and with the golden medal. As a result he obtained the position of senior teacher in the Crimean town Simferopol. It was the critical period of the Crimean war and it was the reason why Mendeleev moved to Odessa where he continued to teach in the Richelieu gymnasium.

In 1856 Mendeleev returned to St. Petersburg where he defended his thesis for the Master degree in Chemistry. At that time he began to deliver lectures in organic chemistry. In 1864 he was elected professor of chemistry in the Petersburg Technological University and one year later, in 1865, he defended his thesis for the Doctor's degree. Two years later he became the chair of Inorganic Chemistry in St. Petersburg University.

## PRIVATE LIFE

In the spring of 1862 in St. Petersburg Mendeleev married Feozva Leshcheva, who was 6 years older. She was a stepdaughter of the Russian poet Piotr Ershov,

**Figure 2.** Tobolsk at the end of XIX century: left - the Bogoyavlensky church were Mendeleev was baptized; right - the gymnasium where Mendeleev studied (permission of Museum and archives of Dmitri Mendeleev in St. Petersburg).

who was Mendeleev's teacher of Russian literature in the Tobolsk gymnasium (Fig. 3a). However the relations within the family didn't get on and in 1881 the spouses divorced. Mendeleev's second wife, Anna I. Popova, was 26 years younger than him (Fig. 3b). During 1876-1880 she studied at the Academy of Art in St. Petersburg. Omitting many details of their love story, I can only mention that in December 1880 her father sent Anna to Italy to put distance between her and Mendeleev. She stayed in Rome for 4 months. At that time her main supervisor was Alessandro Rizzoni, a Russian painter of portraits and genre scenes. She also attended the classes in the Academy Gigi (L'Accademia Gigi – L'Accademia Libera del Nudo). On the 14th of March 1881 Mendeleev arrived to Rome to meet Anna and on the 5th of May they came back to St. Petersburg. The same year the Orthodox Church accepted Mendeleev's divorce. However he was condemned to penance for the following six years and during that period he could not be married. However in April 1882 in spite of this verdict the priest of the Admiralty Church in St. Petersburg, whose name was Kuntsevich, received 10 thousand rubles and married Mendeleev with his sweetheart Anna Popova. As a result the breach of inhibit led to the deprivation of Kuntsevich's holy orders.

Mendeleev had seven children with his wives. His and Anna Popova's eldest daughter Lyubov (Lyuba – Fig. 3c) was married to Alexander Blok, the prominent Rus-


**Figure 3.** Ladies of Mendeleev's family: Mendeleev with his first wife Feozva Leshcheva (a); Anna Popova – his second wife (b); c – Mendeleev's daughter Lyubov (permission of the Museum and archives of Dmitri Mendeleev in St. Petersburg).

sian poet of Silver Age (period from the last decade of the 19th century up to first two or three decades of the 20th century). He dedicated to Lyuba his first cycle of poetry *Stikhi o prekrasnoi Dame* (*Verses About the Beautiful Lady*, 1905).

## WORK ON THE PERIODIC TABLE

Mendeleev worked in St. Petersburg University until 1890, and it is just here he made his most significant discovery – the creation of Periodic Table of chemical elements. He began to give a lecture course "Fundamentals of Chemistry" in October 1867. During 1868-1871 he summarized it in 5 issues with the same name. During the composition of this edition Mendeleev noticed that the properties of chemical elements definitively obey some periodicity. This regularity became specifically clear when he arranged the elements according to their atomic weights, even though some of their values needed a correction. Later on this approach justified the prediction of some chemical elements which were unknown at that time.

The history does not give an unambiguous answer to some questions related to the events when the first version of the Periodic Table was completed. It is known [2] that on Monday 17th of February 1869 Mendeleev prepared the manuscript with the title written by him in French: "Essai d'une systeme des elements d'après leur poids atomiques et fonctions chimiques". It is curious, because Mendeleev's gymnasium mark for foreign languages was far from excellent. In the last decade of February he also finished the work on the corresponding paper with the additional information which was published the same year in the Journal of the Russian Chemical Society – the first chemical journal in Russia [3] (Fig. 4).

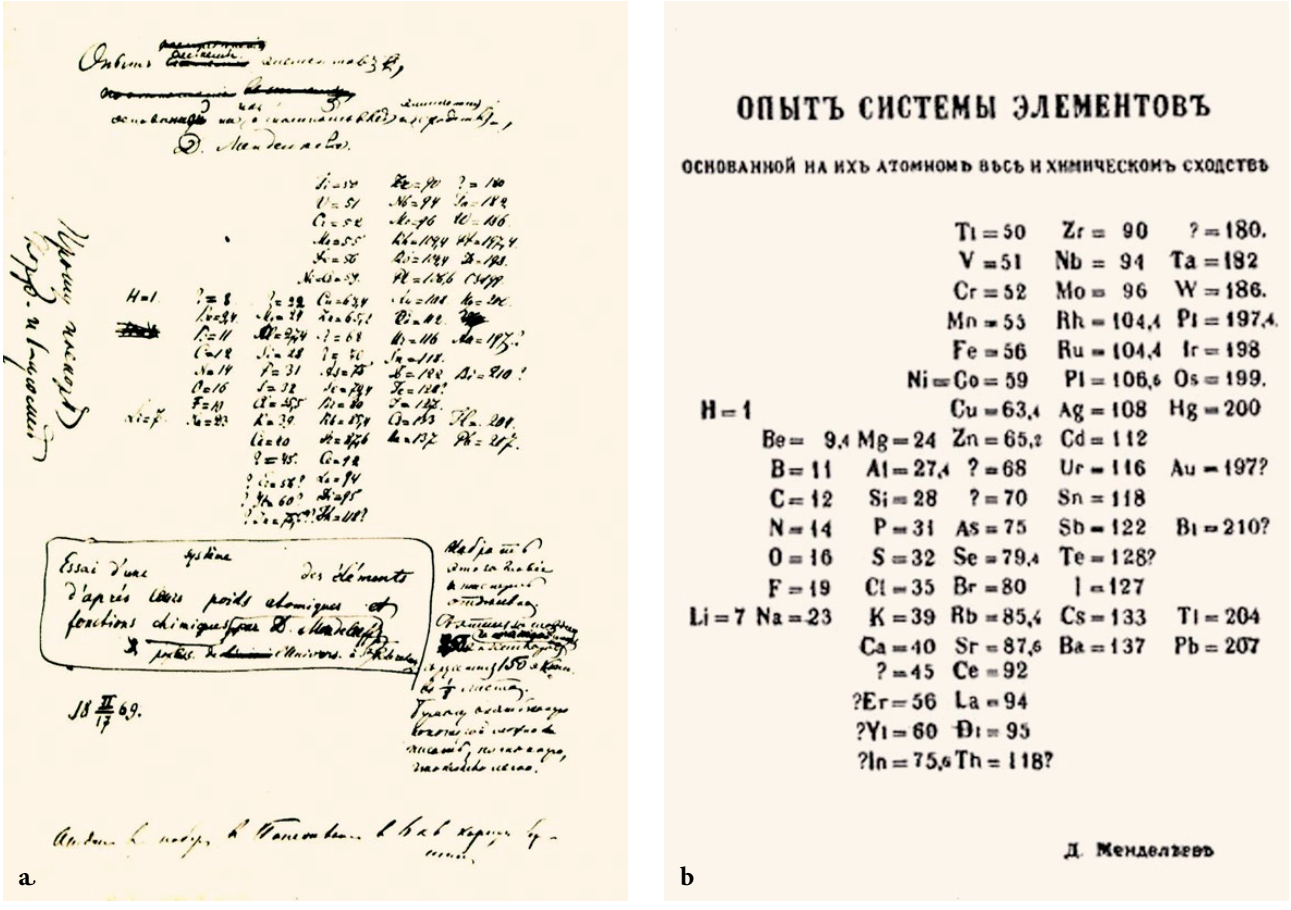



**Figure 4.** Mendeleev's manuscript "Essay of the system of elements according to their atomic weights and chemical properties", 17th of February 1869 (a). The first version of the Mendeleev's Periodic system distributed before his report among the members of Russian Chemical Society and published in the beginning of the first two issues of "Fundamentals of Chemistry" in March 1869 (b) (permission of Museum and archives of Dmitri Mendeleev in St, Petersburg).


**Figure 5.** Two of the world's oldest Periodic Table Charts: a - printed in 1876 and exposed in St. Petersburg University (Public domain); b - found at University of St. Andrews in Scotland by Dr. A.Aitken and printed in 1885 in Vienna (Public domain, photo St. Andres University)

From the very beginning Mendeleev understood that his discovery needed international recognition. Therefore immediately, already in February, he sent his table to his colleagues in Western Europe. Apart from that, on the 6th of March his famous report with the same title of his paper was presented by professor N.A.Menshutkin – the first editor of the Journal RCS – during the meeting of the Russian Chemical Society. In 1906 Mendeleev remembered these events [4]: "In 1869 I sent to many chemists the separate page "Essai d'une systeme des elements d'après leur poids atomiques et fonctions chimiques" – "Essay of the system of elements according to their atomic weights and chemical properties" and provided this information to the Russian Chemical Society during its meeting in March 1869 "On the correlation between properties and atomic weights of the elements". From that it is unclear whether the author gave the presentation or not. According to some data just on the 17th of February he had to leave St. Petersburg for an inspection of the cooperative cheese dairy in Tver province. But because this day became the day of discovery of Periodic Table the departure was postponed until the beginning of March. During this trip Mendeleev also planned to visit his homestead Boblovo, where his house had been restored at that time. However, other records of that time show that Mendeleev personally gave a presentation during the meeting of Chemical Society on the 6th of March. Anyway all these details deviate back in comparison with the very essence of Mendeleev's discovery.

Step by step improving the first version of the Periodic system Mendeleev continued his work until 1871, when the table gained the perfect well-known form [5] (Fig. 5). That year he visited several well-known chemical centers where he gave lectures devoted to his Periodic Table of chemical elements and the same year he presented his famous article "Periodic validity for chemical elements". According to [6], perhaps this discovery inspired US physicist Eugene Wigner, the Nobel laureate in 1963, who in his lecture on this occasion at the Stockholm City Hall, formulated the philosophy of scientific research work: "… science begins when a body of phenomena is available which shows some coherence and regularities, that science consists in assimilating these regularities and in creating concepts which permit expressing these regularities in a natural way" [7].

## MENDELEEV'S COLLEAGUES

As it often happens with important discoveries, which correspond to the challenges related to the scientific ideas about Nature, several researchers in different countries at the same time were thinking about the periodicity in the system of the chemical elements. Julius Lothar Meyer (1830-1895), who worked in Germany, and British chemist John Alexander Newlands (1837-1898), contributed in a significant way to the development of the ideas concerning the periodicity of elements [6]. Their main results will be reviewed below, however initially in connection with Mendeleev's discovery it is worth mentioning the Italian chemist Stanislao Cannizzaro (1826 –1910, Fig. 6a), whose fate had been complicated. He studied medicine and chemistry at the universities of Palermo, Naples and Pisa. In 1849 he took an active role in the popular revolt in Sicily. It was suppressed and Cannizzaro was condemned to death. He fled to Paris and since 1855 he began to work in different Italian Universities. In 1871 he was elected as a member of the Italian Senate and later on he became its vice-chairman. As a member of Senate, Cannizzaro supervised the scientific education in Italy.

Cannizzaro brought the attention to the concepts already present in literature between atom and molecule. In this respect it is worthy to mention the fundamental paper by A.Avogadro [8], published approximately half a century earlier. Moreover, Cannizzaro elaborated and revised the system of the crucial chemical notions: definition of chemical formula, differences between atom and molecule, atomic and molecular

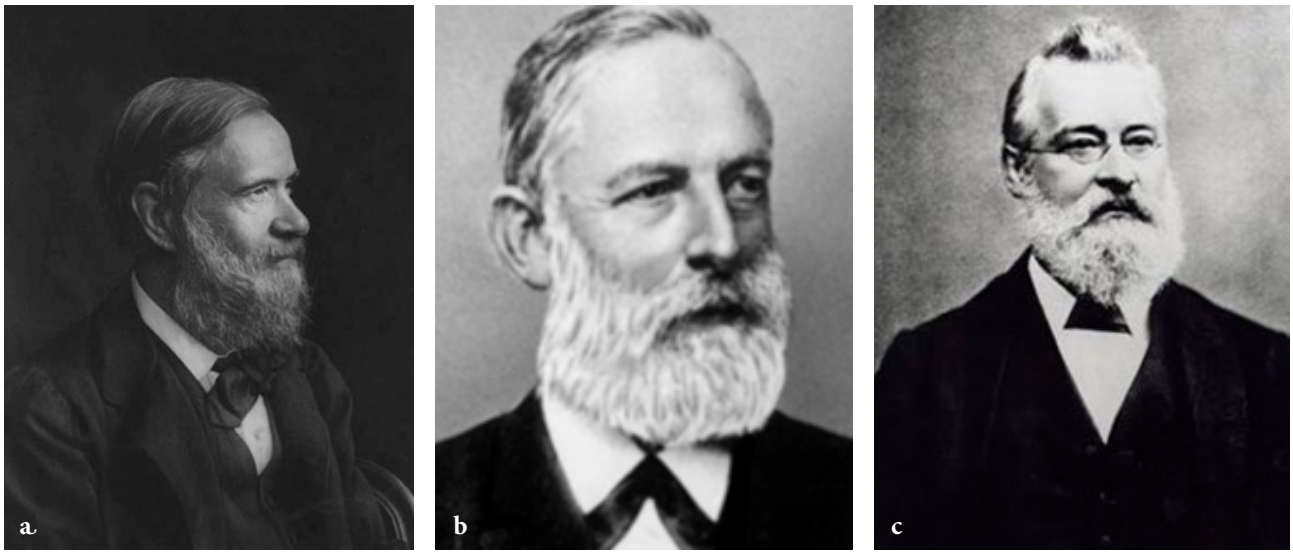



**Figure 6.** Mendeleev's colleagues: Stanislao Cannizzaro (a); Julius Lothar Meyer (b); John Alexander Newlands (c) (photos from public domain).

weights. J. Berzelius published the first data of atomic weights (consequent to the definition of isomorphism by Mitscherlich) as early as 1828 [9]. However Cannizzaro provided their most accurate values. His historical significance is connected primarily to these results. He expressed his theory and the distinction between atomic and molecular weights in the pamphlet [10,11] which he distributed among the participants of the International Chemistry Congress in Karlsruhe in September 1860. Mendeleev and Julius Lothar Meyer were among the attendees and together with the leading European chemists they highly appreciated Cannizzaro's contribution to general chemistry. Many years later Mendeleev said: "I consider him (Cannizzaro) as my real predecessor", because he determined by himself the values of atomic weights and created a necessary fulcrum".

Lothar Meyer (Fig. 6b), who never used his first given name, was a German chemist and a foreign member of St. Petersburg Academy (1890). In the beginning of their careers both Mendeleev and Meyer worked in Heidelberg with R.Bunsen, who elaborated the spectral analysis. Meyer is one of the pioneers in developing the first periodic table of chemical elements. In Meyer's birthplace Varel (Lower Saxony, Germany) there is a memorial with three sculptural portraits of Meyer, Mendeleev, and Cannizzaro.

In 1864 Meyer composed a table with 28 elements allocated in six columns according to their valences. Obviously such arrangement of limited number of chemical elements revealed the similarity of their chemical properties within the same vertical column. In connection to this approach Mendeleev argued that this system is just a simple comparison of some elements on the basis of their valences.

Such values are even not constant for the same element and therefore should not be considered as its crucial characteristic. Consequently, Meyer's table could not pretend for the full description of elements and did not reflect their inherent Periodic Law.

Only half a year after the first version of Mendeleev's Periodic Table was printed in 1869, Meyer published a revised and expanded version of his 1864 table, which was similar to that published by Mendeleev. This paper "Die Natur der chemischen Elemente als Function ihrer Atomgewichte" ("The Nature of the Chemical Elements as a Function of their Atomic Weight" - *Annalen der Chemie* (1870)) [12], contained the table and the plot with the correlation between atomic volumes and atomic weights for the known chemical elements at that time. It is worthy to recall that Meyer unjustly reproached Mendeleev for the correction of some atomic weights in the Periodic Table. However several years later he wrote: "I confess frankly that I lacked the courage for far-sighted assumptions which Mendeleev expressed with certitude" [2, 13].

Approximately at the same time the British chemist Newlands (Fig. 6c) suggested his own version of the Periodic system of the chemical elements. In the beginning of 1864 Newlands was impressed by the paper, which claimed that for most of the chemical elements the values of atomic weights are multiple of 8. Obviously the author's opinion was erroneous, however Newlands

decided to continue his research in this direction. He composed the table where the elements were ordered according their atomic weights. In his paper dated 20th of August 1864 he emphasized the periodicity in the arrangement of chemical elements [14]. After he numbered the elements and compared their properties he noticed the repeating pattern of elements where every 8 each element had similar chemical properties as the first one in common with the eighth note in musical octave. This mysterious musical harmony finally compromised the whole concept which exhibited similarity with Mendeleev's Periodic Table only externally.

One year later, on the 18th of August 1865, Newlands published the new table which he called "Law of Octaves" [15]. On the 1st of March 1866 in the Chemical Society he gave a talk "Law *of* Octaves and the Causes of Numerical Relations among the Atomic Weights", which received the hostile reception on behalf of the audience. In particular, G. C. Foster, professor of physics at the University College of London, humorously inquired whether the speaker had ever examined the elements according to the order of their initial letters [16]. According to [6], in 1884 Newlands collected his various papers on the discovery of the Periodic Law in [17].

In 1887 the London Royal Society awarded Newlands with the Davy Medal "For his discovery of the periodic law of the chemical elements". This medal is given annually since 1877 to an outstanding researcher in the field of chemistry. Five years earlier Dmitry Mendeleev and Lothar Meyer received the Davy Medal from Royal Society "For their discovery of the periodic relations of the atomic weights". Newlands rewarding seemed rather ambiguous, however he primarily revealed the periodic variation of the chemical properties of the elements which is reflected in his Law of octaves, and it is obviously his merit. Mendeleev emphasized that "…due to his works it was possible to perceive Periodic law in its first stages" [18, 19].

## PERIODIC TABLE AND MINERALOGY

The Periodic system contributed to the progress in many natural sciences. It significantly extended the scientific ideas in geology and best of all in geochemistry and mineralogy [20]. The discovery of new minerals and consequently of the chemical elements in their composition contributed to the creation of the Periodic table. At the same time the Periodic table indicated some shortcomings in the scientific ideas about these elements. One of the first results of its use was the revision of the atomic weights of uranium and rare earth elements as well as the transfer of the latter from the divalent calcium analogues to the group of trivalent elements. The significance of this correction becomes more important nowadays when the use of the rare earth elements is estimated at 2000 tons per year only in Russia [21]. Electronics and photonics use about 70% of this quantity and thus the hunt for rare earth elements is expanding all over the world.

Besides atomic weights Mendeleev composed his Periodic table on the basis of the chemical properties of the elements. Thanks to that he predicted the analogues of aluminum (gallium) and of silicon (germanium). Both elements were discovered in 1876 [22] and in 1886 [23], respectively. They are widely used in semiconductor technology and thus, the industrial demand for them is growing up. Finally it is worthy to note that when Mendeleev was still alive, the noble gas group was discovered. This discovery definitively indicated that the periods include octets of chemical elements where the 9th element is similar to the 1st one, and have no analogy with the musical octaves. These elements are also of geochemical interest, namely He and Ne are important constituents of the Gas Giants - Jupiter and Saturn.

During several decades after the publication of the Periodic table researchers in different countries continued to think over the question whether a more fundamental property of the chemical element than its atomic weight exists. Thus in 1913, six years after D.Mendeleev passed away, the young British physist Henry Moseley introduced a new characteristic - "atomic number" which is equal to the number of positive charges in the atomic nucleus and consequently to the number of electrons in the neutral atom [24].

The electronic model of atoms enlarged the ideas related to their behavior in the geochemical processes. In particular, in 1958 the German mineralogist Hugo Strunz discovered gallite $CuGaS_2$ - the first Ga-mineral with a crystal structure identical to the widespread chalcopyrite $CuFeS_2$. Thus everybody began to think that gallium, which is a rare chemical element, can be hidden in chalcopyrite. However all attempts to find gallium in chalcopyrite failed because it and iron have different electronic structure: there are 18 electrons in the outer shell of Ga whereas Fe contains only 13 electrons and thus there is no isomorphic replacement between these minerals.

Professor Vladimir Vernadsky at Moscow University highly appreciated the important contribution of the Periodic law to mineralogy [20]. In the end of XIX century he composed the table of isomorphic elements with emphasis on so-called Vernadsky's rows. The atomic radii were not known at that time and thus the isomor-

phic replacements were examined only within the vertical groups of the Periodic table. Therefore Vernadsky's rows did not receive an acknowledgement on behalf of mineralogists and geochemists and as a result for some time the Periodic table was shifted back in their mind.

This situation radically changed when in 1926 Victor Goldsmith, a Norwegian mineralogist, on the basis of the interatomic distances and the experimentally determined values of radii for $O^{2-}$ = 1.32 Å and $F^{-}$ = 1.33 Å, composed the system of ionic radii and formulated the rule for isomorphic replacements [25]. He indicated that the size difference for the ions involved in such substitution cannot exceed 10-15%. Thus three parameters, namely atomic weight, atomic number and ionic radius were used to characterize each element in the Periodic table. After that the diagonal rows of elements which correspond to the directions of possible isomorphic replacements were revealed within the Periodic table. The following examples illustrate Goldsmith's rule: $Li^{+}$ - $Mg^{2+}$ - $Sc^{3+}$; $Na^{+}$ - $Ca^{2+}$ - $Y^{3+}$ - $Th^{4+}$; $Al^{3+}$ - $Ti^{4+}$ - $Nb^{5+}$ - $W^{6+}$. This idea allowed to explain the complete substitution between $Na^{+}$ and $Ca^{2+}$ in feldspars – the main rock forming minerals in the Earth crust, according the scheme $Na^{+} + Si^{4+} = Ca^{2+} + Al^{3+}$ [26]. This diagonal also contains yttrium and in association with it the whole group of rare earth elements. They always replace calcium in the minerals and that's the reason why these elements were considered primarily as bivalent.

The recent theoretical calculations and experimental results indicate a dramatic transformation of electronic structure in some atoms at high pressures. It leads to some changes in their chemical properties and consequently the formation of several new materials with the unexpected stoichiometry. For example, the cubic $NaCl_3$ was synthesized at the pressure 55-60 GPa and at the temperature >2000 K [27]. Similarly such exotic compounds were found in some other systems, namely Mg – O and Al – O. Obviously these phenomena still require an appropriate explication, however the Periodic table is a starting point for such works.

In general the mineralogical observations and conclusions extend the ideas related to the periodic variations of electronic structure at no ambient conditions, of ionic radii, ionization potential and some other notion of energetic crystal chemistry.

## MENDELEEV AFTER HIS DISCOVERY

Mendeleev's lifeline shows that he had many interests and hobbies. He was friends of many artists (Fig. 7), knew painting, he liked to play chess and producing

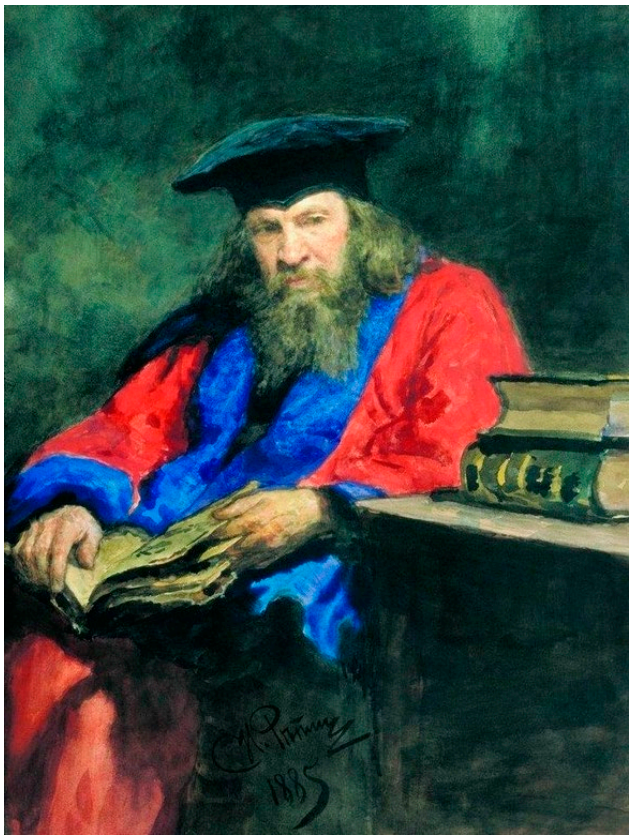


**Figure 7.** Repin I.E (1885). Mendeleev in the mantle of Edinbourgh University honorable professor (permission of Tretyakov Gallery, Moscow).

suitcases was among his unusual hobbies. These items were of exceptional quality because Mendeleev invented a completely unique glue. Therefore all the merchants in St. Petersburg tried to get just these suitcases directly "from Mendeleev".

During his last years Mendeleev promoted the establishment of the first Siberian University in Tomsk and the Polytechnic Institute in Kiev. In 1866 he initiated the foundation of the first Chemical Society in Russia. In 1890 Mendeleev had to leave St. Petersburg University due to his support to the student's movement related to the displeasure of life and studies conditions. In 1892 the minister of finance S.J. Vitte suggested Mendeleev to be the head of the new Central Board of Weights and Measures in Russia. Being on this position Mendeleev insisted on the implementation in Russia of the metric system which was essentially accepted in 1899 (Fig. 8). In the beginning of January 1907 he fell ill with pneumonia and on the 20$^{th}$ of January he passed away. His tomb is in Volkov's cemetery in St. Petersburg (Fig. 9). At his funeral in St. Petersburg, his students carried a large

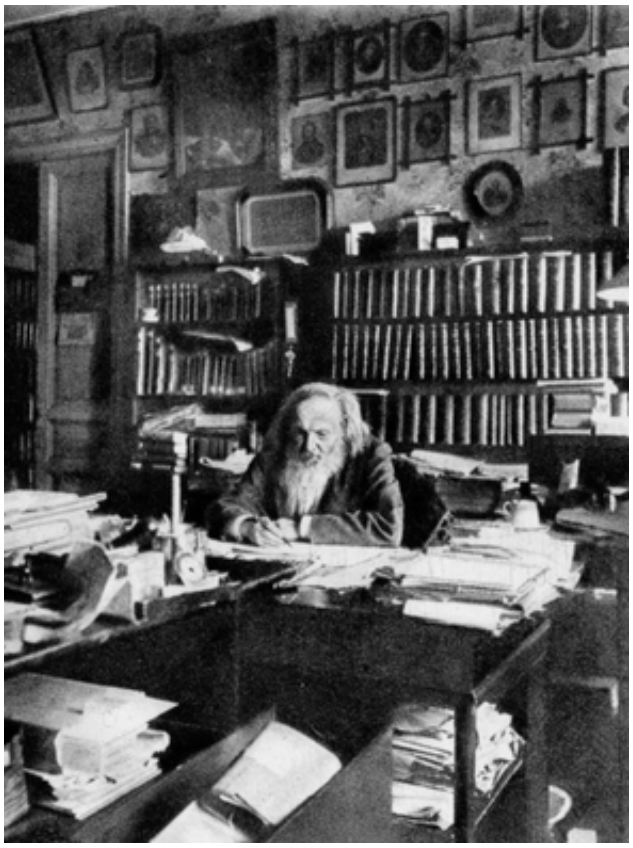


**Figure 8.** Mendeleev in his office: Russia's new Central Board of Weights and Measures (permission of Museum and archives of Dmitri Mendeleev in St, Petersburg).

copy of the periodic table of the elements as a tribute to his work.

## FINAL REMARKS

Mendeleev's priority in the discovery of the Periodic law and in the creation of Periodic Table of the chemical elements was definitively recognized by the International scientific community. In 1905 he was decorated with the Copley medal – the highest award from the Royal Society of London established in 1731 "For his contributions to chemical and physical science". Mendeleev was elected as a member of the London Royal Society, the United States National Academy of Science and the Royal Swedish Academy of Sciences (Fig. 10).

In 1876 he was also elected as a corresponding member of the St. Petersburg Academy of Science. The academician A.M. Butlerov, one of the principal creators of the theory of the chemical structure, nominated Mendeleev as a candidate for the full member vacancy

**Figure 9.** The tomb of Mendeleev in Volkov's cemetery in St Petersburg (permission of Museum and archives of Dmitri Mendeleev in St, Petersburg).

in March 1980. Two other well-known Russian chemists Friedrich Konrad Beilstein and Nikolai N. Beketov were also considered as challengers for the same vacancy. It is really touching that the relations between all of them were full of respect and estimation. However there was no doubt that Mendeleev should have been elected assuming his exceptional contribution to the science. Nevertheless the results of the voting in the Academy meeting on the 11th of November 1880 were really shocking: 10 votes – black, 9 votes – white [2, 28]. There were a lot of protests against this result but Mendeleev accepted it rather quietly and in his autobiographic notes he marked the events of 1880 with the single phrase: '… travelled with Volodia (his son) along Volga". Perhaps it is worthy to add that Anna Popova (later on she became his second wife) accompanied them…

Three times in 1905, in 1906 and in 1907 Mendeleev was nominated to Nobel Prize, however all the times it was done by 1 or 2 his foreign colleagues, whereas his opponents were supported by 20-30 scientists [6]. It is known that the Nobel Prize is conferred for the recently obtained outstanding results and therefore every time there was a controversy whether the creation of Periodic table could be considered as a state-of-the-art work. The discovery of the noble gas group and their very logic placement within the Periodic table were among the most convincing arguments to its urgency.

In 1905 apart from Mendeleev the Nobel Committee considered the works by two other chemists: A. von Bayer (organic chemistry, Germany) and H. Moissan (inorganic chemistry, France). As a result the voting was in favor of von Bayer. Next year the Nobel Committee in chemistry recommended D.Mendeleev to the Gener-

**Figure 10.** The participants of the 52nd meeting of British Association for the Advancement of Science, Manchester, 1887. 1st rank (from left to the right): Menshutkin N.A., Mendeleev D.I., Roscoe H.E.; 2nd rank: outside left - Joule J.P., president of Association, Shorlemmer C. (second from the right side), Thompson W., outside right (permission of Museum and archives of Dmitri Mendeleev in St, Petersburg).

al Assembly of the Royal Swedish Academy. The voting results for Mendeleev at the Committee meeting was 4:1. The only vote was for H. Moissan, who was again Mendeleev's competitor. The Swedish chemist Peter Klason, who was the member of Nobel Committee, supported him very actively. He positively estimated Mendeleev's contribution but emphasized that the creation of Periodic table could be impossible without the accurate values of atomic weights which were obtained by Cannizzaro. That is him who suggested considering both Mendeleev and Cannizzaro as the candidates for the Nobel Prize. At a first glance this suggestion seemed reasonable. However the inclusion of Cannizzaro into the list of candidates for the prize in 1906 was already impossible because the dead line for nomination was terminated on the 31st of January. Thus the H. Moissan received the prize in 1906. In 1907 both Mendeleev and Cannizzaro were nominated for the Noble prize. However that year Mendeleev passed away and according the statute of Nobel Prize it cannot be conferred posthumously.

Obviously the lack of Mendeleev's name in the list of Nobel Prize Laureates is a great historical mistake. His name is well-known all over the world and the Periodic table is in each classroom and auditorium where people study chemistry. On the 10th of June 1905 Mendeleev wrote in his diary: "Apparently the future does not threaten the Periodic Law by its destruction and on the contrary it promises the superstructure and its further development" [29, 30]. The last 150 years completely justified this prediction.

ACKNOWLEDGEMENT

The author is grateful to Mrs. J. Angelett for improving the English in the manuscript and to three anonymous referees for their valuable comments. A special gratitude is addressed to Professor Dmitriev I.S., the director of Museum and archives of Dmitri Mendeleev in St, Petersburg. This study was supported by the Russian Foundation for Basic Research (grant No. 18-05-00332).

REFERENCES

1. Pushcharovsky D.Yu. Dmitry Ivanovich Mendeleev and his discovery. *Nauka i Zhizn*', 2019, 2, 19-25 (in Russian).
2. Smirnov G.V. *Mendeleev.* Moscow, "Molodaya Gvardiya", 1974, 382 p. (in Russian).
3. Mendeleev D.I. The correlation between properties and atomic weights of the elements. *Journal of Russian Chemical Society*, 1869, 1, 60-77 (in Russian).
4. Mendeleev D.I. *Fundamentals of Chemistry. 8th edition, corrected and completed*. SPb: Frolova M.P. printing office and lithography, 1906, 816 p. (p. 612). (in Russian).
5. Mendeleev D.I. The Natural System of the Elements and Its Application for the Evidence to the Properties of Undiscovered Elements. *Journal of Russian Chemical Society*, 1871, 3, 25-56 (in Russian).
6. Hargittai B. and Hargittai I. Year of the periodic table: Mendeleev and the others. *Structural Chemistry*, 2019, 30(1), 1-7.
7. Wigner E.P. *City Hall Speech – Stockholm, 1963*. Reproduced in Wigner EP (1967) Symmetries and Reflections: Scientific Essays. Indiana University Press, Bloomington and London, 1963, pp. 262-263.
8. Avogadro A. Essai d'une manière de déterminer les masses relatives des molécules élémentaires des corps, et les proportions selon lesquelles elles entrent dans ces combinaisons. *Journal de Physique de Chimie et d'Histoire Naturelle* 1811, 73, 58-76.
9. Berzelius J. Table des Poids atomistiques des corps simples et de leurs oxides,d'après les analyses les plus exactes et les plus récentes. Paris 1828. *Annales de chimie et de physique*, 1828, 38. S.426-432.
10. Cannizzaro S. Lettera del prof. Stanislao Cannizzaro al prof.S. De Luca; Sunto di un corso di filosofia chimica, fatto nella R. Universita' di Genova. (1858). *Il Nuovo Cimento*, 7(1), 321-368.
11. Cannizzaro S. *Sunto di un corso di filosofia chimica. Nota sulle condensazioni di vapore dell autore stesso*. Pisa, Tipografia Pieraccini, 1858, 62 p.

12. Meyer L. Die Natur der chemischen Elemente als Function ihrer Atomgewichte. *Annalen der Chemie und Pharmacie, Supplementband* VII, 1870. S. 354-364.
13. Paneth F. Die Entwicklung und der heutige Stand unserer Kenntnisse iber das natiirliche System der Elemente (Zum 100-jährigen Jubilaum von Lothar Meyer's Geburtstag). *Die Naturwissenschaften*, 1930, Bd. 18. Heft 47, S. 964-976 (S. 968).
14. Newlands J. A. R. On Relations Among the Equivalents. *Chemical News*, 1864 (20 Aug.), Vol. 10, 94-95.
15. Newlands J. A. R. On the Law of Octaves. *Chemical News*, 1865 (Aug. 18, 1865). Vol. 12, 83.
16. Newlands. Extract from the report of the meeting of the Chemical Society, March 1st, 1866. *Chemical News* (March 9, 1866) 13, 113.
17. Newlands J.A.R. (1884) *On the Discovery of the Periodic Law, and on the Relations among the Atomic Weights*. E and FN Spon, London (it is a reprint collection).
18. Kedrov B.M. (ed.) (1958) *D. I. Mendeleev: The periodic law* (in Russian, D. I. Mendeleev: Periodicheskii zakon). Izd. Akad. Nauk SSSR, Moscow p. 314.
19. Mendeleev D.I. *Fundamentals of Chemistry. 8th edition, corrected and completed*. SPb: Frolova M.P. printing office and lithography, 1906, 816 p. (p. 613). (in Russian).
20. Belov N.V. *Essays on structural mineralogy.* Moscow, Nedra, 1976, 344 p. (in Russian).
21. S*tate and usage of mineral resources of Russian Federation in 2016 and 2017*. State report. The Ministry of Natural Resources and Environment of Russian Federation. Moscow 2018, pp. 370.
22. de Boisbaudran Lecoq. Caractères chimiques et spectroscopiques d'un nouveau métal, le gallium, découvert dans une blende de la mine de Pierrefitte, vallée d'Argelès (Pyrénées). *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 1875, 81, 493-495.
23. Winkler C. Germanium, Ge, ein neues, nichtmetallisches Element. *Berichte der deutschen chemischen Gesellschaft*, 1886, 19, S. 210-211
24. Moseley, H.G.J. The high-frequency spectra of the elements. *Philosophical Magazine*, 6th series. 1913, 26, 1024-1034.
25. Goldschmidt V. M. Die Gesetze der Krystallochemie. *Die Naturwissenschaften* 1926, 14, 21, 477-485. doi: 10.1007/bf01507527
26. Zambonini F. The isomorphism of albite and anorthite. *American Mineralogist*, 1923, 8, 81-85.
27. Zhang W., Oganov A.R., Goncharov A.F., Zhu Q., Boulfelfel S.E., Lyakhov A.O., Stavrou E., Somayazulu M., Prakapenka V.B., Konôpková Z. Unexpected Stable Stoichiometries of Sodium Chlorides. *Science* 20 Dec 2013, 342, 6165, 1502-1505. doi: 10.1126/science.1244989
28. Dmitriev I.S. *Boring Story*. In: Dmitriev I.S. *A Man of Alternation Epoch. (Essays of D.I.Mendeleev and His Time)*. St. Petersburg, Chimizdat, 2004, p. 397-458
29. Mendeleev D.I. – Archive, Vol. 1, Autobiographic materials, Collected Articles, Compilers Mendeleeva M.D., Kudryavtseva T.S. Editors: Shchukareva S.A. and Valka S.N. Leningrad, 1951, 34 p. (in Russian)
30. Evdokimov Yu. On the history of the periodic law. *Nauka i zhizn*', 2009, 5, 12-15 (in Russian).

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Historical Article

# Early contributions of crystallography to the atomic theory of matter

Giovanni Ferraris

*Dipartimento di Scienze della Terra, Università di Torino, Via Valperga Caluso 35, 10125 Torino, Italy*
E-mail: giovanni.ferraris@unito.it

**Abstract**. After briefly presenting early hypotheses on the submicroscopic origin of symmetry and polyhedral morphology in the crystals, the structural model proposed by Haüy in 1784, based on the periodic repetition of integrant molecules made up of simple molecules, is discussed. It is then highlighted how – through investigation of crystal hemihedry, isomorphism (mixed crystals) and optical activity – researches aiming at overtaking drawbacks of Haüy's model brought basic ideas to achieve the modern knowledge of the atomic structure of matter. The atomic-scale interpretation of properties of the crystalline state soundly contributed, among others, to properly define molecules and atoms, determine the atomic weights, hypothesize stereoisomerism, build the periodic table of elements and define ionic radii and bond.

**Keywords**. Isomorphism, stereoisomerism, integrant molecule, crystal morphology, atomic theory.



**Citation:** G. Ferraris (2019) Early contributions of crystallography to the atomic theory of matter. *Substantia* 3(1): 131-138. doi: 10.13128/Substantia-81



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Competing Interests:** The Author(s) declare(s) no conflict of interest.

## 1. INTRODUCTION

The introduction of the concept of atom as indivisible constituent of the matter dates back to the Greek philosopher Democritus (~ 460 b.C. - ~ 370 b.C.), but only at the end of the nineteenth century the modern science made the atom from a debated philosophical category definitively transit to a physical certainty. Towards the end of its long and troubled history, the concepts of atom and molecule intertwined and sometimes even clashed. Only the determination of the first crystalline structures – made possible after the discovery of the X-ray diffraction by Max von Laue (1879-1960) in 1912 – convinced the entire scientific community that atoms and molecules are different entities, both necessary to model the structure of matter at atomic scale.

The contribution of crystallography to the atomic theory of matter can certainly not be limited to the irrefutable evidence acquired through the aforementioned structural determinations. In fact, for centuries the geometric regularity (symmetry) of the crystal morphology has played a stimulating role to develop hypotheses on the submicroscopic structure of the matter suitable to explain the macroscopic observations. In this article, only contri-

butions of the pre-diffraction era, inspired by morphology-related investigations, are qualified as early ones.

Following ingenious but quite approximate earlier hypotheses on the internal structure of the crystals – mostly based on close packing of particles – and Nicolas Steno's (Niels Steensen, 1638-1686) statement on the constancy of the angles between corresponding faces in all crystals of the same mineral[1] – later assumed by Jean-Baptiste Romé de L'Isle (1736-1790) as a genuine law of nature[2] –, in the last quarter of the eighteenth century the French crystallographer Réné Just Haüy (1743-1822) proposed a revolutionary model based on the periodic repetition of a submicroscopic polyhedron named integrant molecule and comparable to the unit cell of the modern structural crystallography. This model preceded the atomistic theory of John Dalton (1766-1844) for over thirty years and represented a first modern and general attempt to reasonably represent the atomic structure of the matter. Although the integrant molecule was primarily intended as a tool to explain the crystal morphology, we shall see that Haüy's structural model contributed to inspire Amedeo Avogadro (1776-1856) and André Ampère (1775-1836) to draw fundamental theoretical consequences from the results published by Jean Louis Gay-Lussac (1778-1850) on the chemical combination of gases.

Subsequently, as illustrated in this paper, researches aiming to overcome drawbacks of Haüy's model brought sound contributions in issues such as: distinct roles of atoms and molecules; determination of the atomic weights; stereoisomerism of chemical groups; building of Mendeleev table; definition of ionic radii; nature of the chemical bond.

## 2. EARLY HYPOTHESES

Before the second half of the eighteenth century various conjectures on the internal structure of the crystalline materials had been proposed aiming to explain features of the crystals, such as their symmetry and polyhedral morphology. Among the scientists who investigated the property-structure relationships of crystals we find famous names, usually better known for their important contributions to frontier non-crystallographic problems.

The Italian polymath Gerolamo Cardano (1501-1576), inventor of several mechanical devices, including the Cardan shaft with universal joints, in 1550[3] noted the hexagonal symmetry common to the cells of the honeycombs and to the prismatic habit of quartz crystals and assumed for the latter an internal structure based on hexagonal particles.

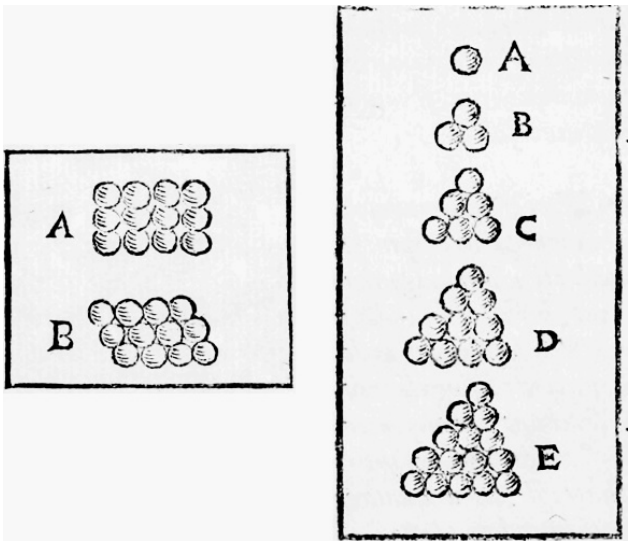



**Figure 1.** Close packing of spheres described by J. Kepler. (From reference 4, pp. 9-10).

Intriguing is the case of Johannes Kepler (1571-1630) – best known for his laws of planetary motion – who, in a booklet published in 1611[4], where he questions on the origin of snow crystals morphology, derives close packings of spheres (Figure 1) but, surprisingly, he does apply this finding to the investigated morphology.

Robert Hooke (1635-1703), discoverer of the law of elasticity that bears his name, in 1665[5] attributed the morphology of the rock alum ($KAl(SO_4)_2 \cdot 12H_2O$) crystals to a compact packing of submicroscopic spheres (Figure 2).

In 1690[6] Christiaan Huygens (1629-1695) proposed an anisotropic model of crystal structure based on a compact packing of ellipsoids (Figure 3) to explain – via his well known wave theory of light – the birefringence observed in calcite by Rasmus Bartholin (1625-1698) in 1669[7] (Figure 4). It might be worth to recall here that birefringence has been the first physical property to be explained via an anisotropic structure of the matter, i.e. an inherent characteristic of the crystalline state.[8]

Finally, in 1749 Michail Vasil'evič Lomonosov (1711-1765) – mineralogy was among his multifaceted interests – imagined that the morphology of the niter ($KNO_3$) crystals was related to a submicroscopic packing of hexagonal particles.[1]

---

[1] The original manuscript (M.V. Lomonosov, *Dissertatio de generatione et natura nitri, concinnata pro obtinendo praemio, quod illustris scientiarum Academia regia liberalitate Berolini florens proposuit ad 1-mum aprilis anni 1749*) is kept in the archives of the Berlin-Brandenburgische Akademie der Wissenschaften (Berlin, Germany). In 1934 it has been printed for the first time.[9]

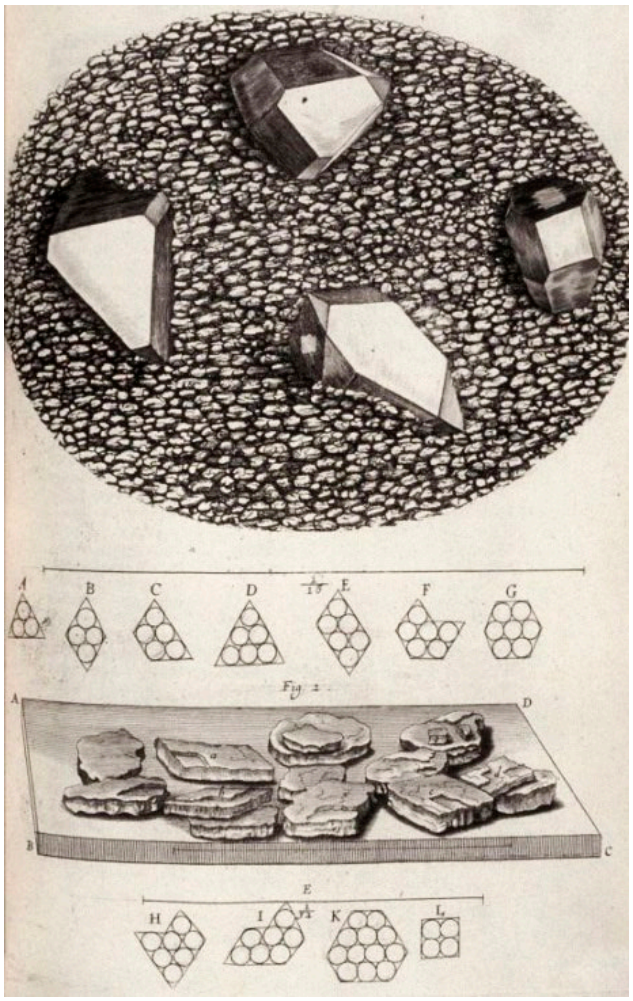


**Figure 2.** Morphology of the rock alum crystals related to close packing of spheres by R. Hooke. (From reference 5, opposite to page 82).

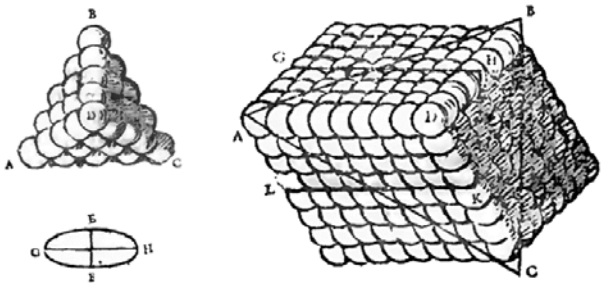


**Figure 3.** Anisotropic structure model based on close packing of revolution ellipsoids proposed by C. Huygens to explain the birefringence of calcite; a bidimensional section through the revolution axis of an ellipsoid is shown. (From reference 6, pp. 92-93).

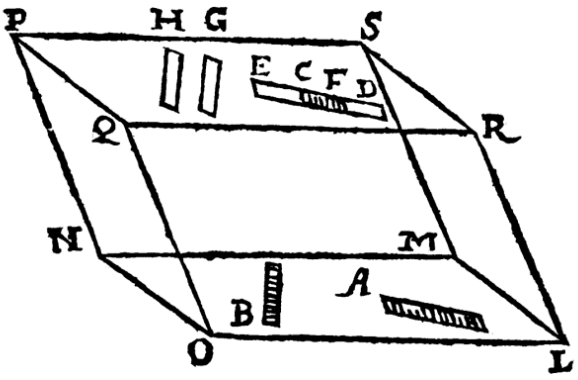


**Figure 4.** Birefringence in calcite observed by R. Bartholin. H and G, EC and FD are the double refracted images of B and A, in the order. (From reference 8, p. 12).

## 3. HAÜY AND THE INTEGRANT MOLECULE

Molecules and atoms were not yet clear and distinct concepts, when in 1766[10] Pierre Joseph Macquer (1718-1784) defined the *integrant molecule* as a submicroscopic particle consisting of *simple molecules* that, as a matter of fact, correspond to the modern atoms. In 1784 Haüy adopted the integrant molecule as polyhedral building block of his general and innovative model of crystal structure aiming to explain symmetry and morphology of the crystals.[11][2]

Haüy's model (Figure 5)[13] hypothesized a periodic arrangement of an integrant molecule, whose polyhedral shape and chemical nature of the constituent simple molecules are characteristic of each mineral species and, by extension, of any crystalline material.

The idea of an integrant polyhedral molecule as basic building block was suggested to Haüy by the cleavage polyhedron of crystals – originally observed in calcite – that can be reduced to microscopic dimensions by iterated cleavages. Haüy did not conceive empty spaces in the matter; therefore, both integrant and simple molecules had to be space-filling polyhedra. That clearly appears in a drawing (Figure 6), published (1822) in his treaty of crystallography[14], where, also considering critical comments received in the meantime, he improved his structural model of the 1784 *Essai* by graphically showing packing of simple molecules filling an integrant molecule.

The model was widely accepted, thus contributing to the advancement of a theory of matter based on molecules and atoms. In particular, the terminology of Haüy was adopted by Avogadro in his theoretical interpretation[15] of the experimental results obtained by Gay-Lus-

[2] Actually, in the reference 11 Haüy named constituent molecule (*molécule constituante*) the building block of his model and adopted the term integrant molecule (*molécule intégrante*) few years later.[12]

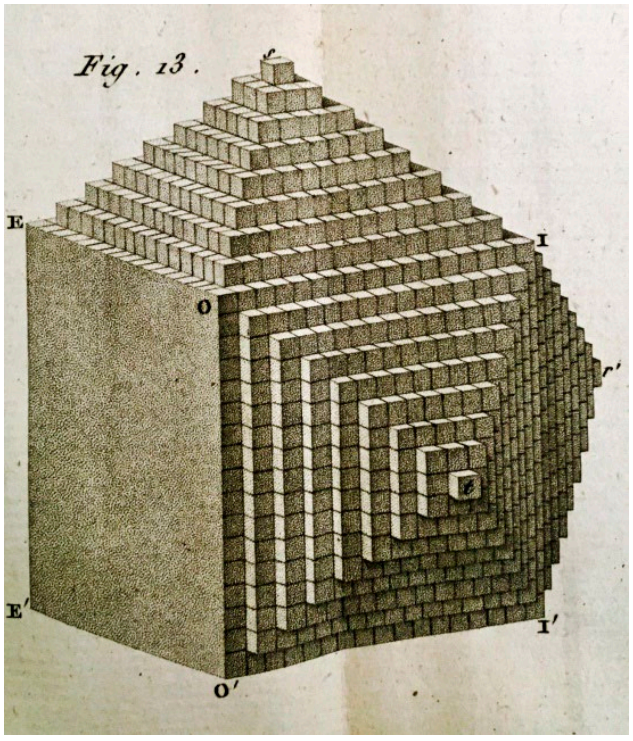



**Figure 5.** Haüy's model of a cubic crystal showing the periodic translation of a basic cubic building block (integrant molecule) and how different crystallographic forms can be obtained subtracting integrant molecules from an original cube. (From reference 13, vol. 5, fig. 13).

sac[16], which led him to hypothesize that, under the same conditions of pressure and temperature, equal volumes of different gases contain the same number of (integrant) molecules. Precisely, the keystone to reconcile the ideas of Gay-Lussac (simple ratio between the volumes of reagent gases) with those of Dalton[17] (fixed relationship between the reactant masses) was the distinction between the concept of integrant molecule (today molecule) and that of simple molecule (today atom); a step this, not made by Dalton who, instead, conceived atoms only. Independently, in 1814 André Ampère (1775-1836) proposed[18] the same hypothesis of Avogadro. Whereas, the latter did not quote Haüy's model – whose influence on his work seems, however, hardly deniable[19] – Ampère made explicit reference to this model with the variant of locating polyhedral atoms not inside, but at the vertices of the polyhedron representing the integrant molecule.

For long debated, but never sufficiently clarified reasons (cf., e.g., 20), the distinction between molecules and atoms affirmed by Avogadro was practically neglected for half a century. In fact, it was only at the first international congress of chemistry (Karlsruhe 1860) that Stanislao Cannizzaro (1826-1910) (cf., e.g., 21) brought

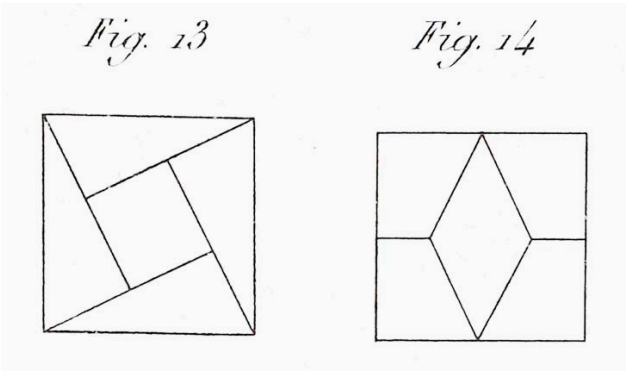



**Figure 6.** Two sections of a cubic integrant molecule filled with simple molecules. (From the *Atlas* of reference 14, pl. 69, figs. 13 and 14).

to general attention the paper that Avogadro had published since 1811 in a well known international journal (reference 15). Consequent to this revival, the majority of chemists and physicists accepted the idea of molecule as aggregation of chemically bound atoms.

## 4. ISOMORPHISM *VS.* ATOMS

Haüy's model was unable to explain hemihedral symmetries, polymorphism and isomorphism. Efforts to overtake these drawbacks, not only led to discover Bravais lattices, point and space groups, but also to reach agreement on the definition of molecule and atom.

Among early post-Haüy structure models able to explain hemihedral symmetries it is worth to quote the contribution by William Hyde Wollaston (1766-1828) who, in 1813, proposed mixed compact packages of spheres and ellipsoids, even of different color (nature), to explain relations between structure, morphology and chemical-physical properties of the crystals.[22] Hemihedral boracite, $Mg_3B_7O_{13}Cl$ (space group *mm*2, but pseudo-cubic -43*m*), was one of the minerals debated at that time for its puzzling morphology. It was investigated by Wollaston and again, a quarter of century later, by Gabriel Delafosse (1796-1878) who proposed a structure model based on a network of tetrahedra (Figure 7).[23] As pioneering models quoted in paragraph 2, at variance with Haüy's space-filling model, Wollaston's and Delafosse's models contain "empty" space between building blocks, thus prefiguring a situation shown by modern diffractometric and microscopic methods.

Research on the crystallization of compounds from water solutions with variable composition lead to discover mixed crystals (solid solutions), i.e. the co-crystallization of two (or more) compounds whose crystals bear

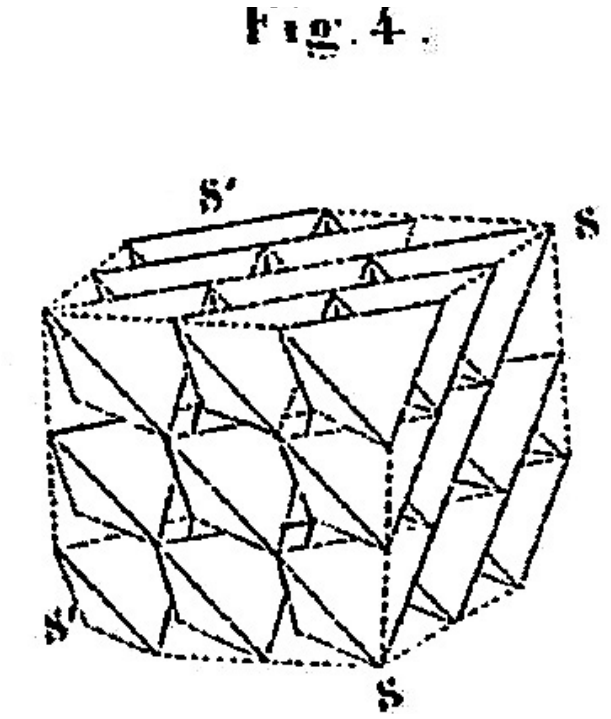



**Figure 7.** Model of boracite structure based on a framework of tetrahedra as proposed by G. Delafosse. (From reference 23).

the same morphology and chemical formulas differ only in the chemical nature of one element (isomorphism). In this field, pioneer results were published by Nicolas Leblanc (1742-1806)[24] and François Sulpice Beudant (1787-1850)[25] without reaching sound conclusions on the nature of the crystals showing mixed composition: either double salts or solid solutions? Beudant named crystalline mixtures (*mélanges cristallines*) these crystals that instead Wollaston considered to be solid solutions.[26]

Continuing the research of Beudant, Eilhard Mitscherlich (1794-1863) studied crystallization from water solutions of mixed salts with analogous chemical compositions. On the basis of chemical analyses and crystallographic measurements on the precipitated crystals, Mitscherlich definitively established what the above-mentioned authors had only glimpsed: compounds with similar chemical formula that display the same morphology can co-crystallize two by two and in some cases three by three (for example, ammonium and potassium salts with iron salts).[27] In a subsequent article[28] Mitscherlich, likely influenced by the morphology of the crystals, introduced the term isomorph (= same form) referring to the chemical elements that, substituting each other, give rise to a group of co-crystallizing compounds. He wrote: "The same number of atoms combined in the same way produces the same crystalline form. The latter is independent of the chemical nature of the atoms and is determined only by their number and arrangement"[3].

The atomistic interpretation of isomorphism was soon successfully tested by Mitscherlich and his master Jöns Jacob Berzelius (1779-1848), who determined atomic weights via a procedure suggested by the following reasoning. If the co-crystallizing AR and BR compounds differ in their chemical composition for the substitution of A for B atoms only, one can derive the ratio between the atomic weights of A and B from the following equality:

$$\frac{\text{(atomic wt of A)}}{\text{(atomic wt of B)}} = \frac{\text{(wt of A combined to R)}}{\text{(wt of B combined to R)}} .$$

The resulting first list of correct atomic weights was published by Berzelius in 1828.[30]

As further keystone of the atomistic theory of isomorphism, one can here recall that Dmitrij Ivanovič Mendeleev (1834-1907) fruitfully exploited it to build his periodic table of elements. He was particularly interested in relating macroscopic properties to microscopic proprieties and, in this context, he used suggestions from the crystal morphology to fill various boxes of the table with elements whose yet unknown chemical properties were hypothesized via isomorphism between their compounds and those of already well characterized elements.[31]

## 5. OPTICAL ACTIVITY VS. STEREOCHEMISTRY

The rotatory polarization (optical activity) discovered in quartz crystals by Jean Baptiste Biot (1874-1862)[32] was one of the properties not explainable by Haüy's structure model because it does not admit acentric crystals. The explanation, still essentially valid today, was given in 1824 by Augustin J. Fresnel (1788-1827) who, although adopting the term integrant molecule, went well beyond Haüy's model, which includes only periodic translations as repetition operations. Here are the words of Fresnel: "We conceive that this [optical activity] may result from a particular constitution of the refractive medium or of its integrant molecules, which establishes a difference between right-to-left and left-to-right. Such would be, for example, a helicoidal arrangement of the molecules of the medium, which offers inverse properties according to whether these helices are either dextrorsum or sinistrorsum".[33]

[3] A recent analysis of the background of Mitscherlich's work can be found in reference 29.

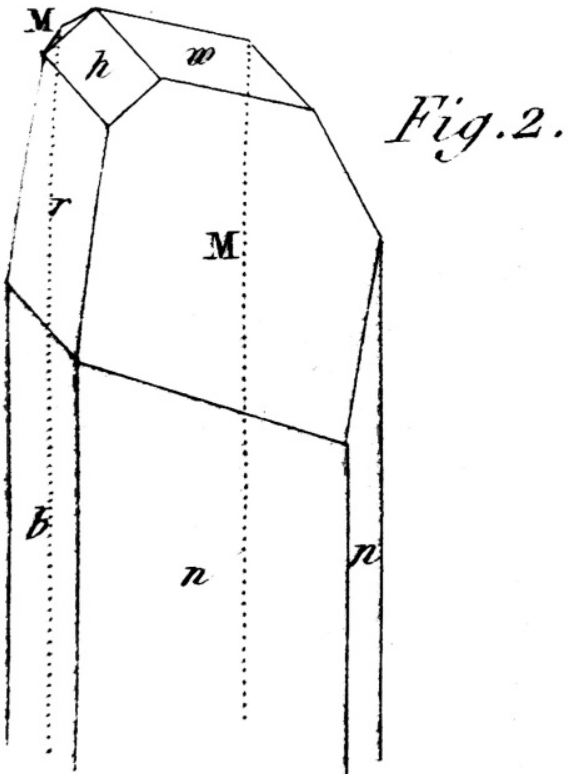



Fig. 2.

**Figure 8.** Hemihedral crystal of tartaric acid from the article of H. de la Provostaye (reference 36, pl. 1, fig. 2) that inspired the research of L. Pasteur on the optical activity (reference 34).

However, it has been necessary to wait for a further quarter of a century before, again by investigating optical activity, Louis Pasteur (1822-1895) proposed a mechanism able to explain the existence of substances showing rotatory polarization both in the crystalline state and in solution, while others display this property only in the solid state. For his results Pasteur is indebted to the theory of isomorphism too, and, as reported in his article[34], to observations by Mitscherlich on the optical activity of tartrates that Biot had published in 1844.[35] Precisely, Pasteur discovered that the racemes of tartrates are not solid solutions (mixed crystals), but fifty-fifty mechanical mixtures of left- and right-handed crystals. Having in mind the drawings of tartrate crystals published in 1841 by Hervé de la Provostaye (1812-1863)[364] (Figure 8), Pasteur identified in racemes the two types of opposite handed crystals via their specular morphologies and, patiently, under the microscope, separated them from one another.[5]

[4] This turns out to be the only article in which de la Provostaye describes tartrates, but for some unknown reason the reference made by Pasteur to the figures of this article does not match the numbering shown therein.

[5] For a recent analysis of Pasteur's work see references 37 and 38.

Pasteur's explanation at the atomic scale was that, as Fresnel had supposed in 1824, optical activity of a crystalline compound is determined by helical arrangement of groups of atoms / molecules in the structure. Considering that Biot had published since 1839 his observations on the absence of rotatory polarization in fused quartz and opal (i.e., in amorphous silica)[39], Pasteur concluded that a crystalline optically active compound preserves this property in solution only if it is due to the spatial arrangement of its atoms (stereoisomerism) in a group (molecule) which survives the structure collapsing. Thus, as a matter of fact, a distinction between molecular and non-molecular compounds was proposed; actually, this hypothesis will be fully accepted by the scientific community only after the evidence brought by experimental determination – via X-ray diffraction – of the crystal structure of NaCl and diamond in 1913.

## 6. FINAL REMARKS

The excursus through a selection of results achieved by crystallography between the second half of the eighteenth century and the first half of the following century, clearly highlights how – mainly thanks to the then leading French school[6] – the search for relationships between macroscopic (morphology) and submicroscopic features (structure) of crystals has contributed substantially to clarify the atomic structure of matter. In particular, the science, through the analysis of properties such as isomorphism (mixed crystals) and optical activity, moved from unspecified submicroscopic particles to a clear distinction between atoms and molecules.

Although this article is limited to the pre-diffraction period, it is worth to remember the influence that – via diffraction results – isomorphism had in establishing concepts such as ionic radius and its consequences in terms of definition of chemical bond and of relationships between structure and properties. In fact, the resolution of crystal structures, especially of minerals, followed to the discovery of X-ray diffraction in 1912, made available experimental data to define the radius of the sphere of influence of the elements linked by chemical bonding. Among several pioneers on this matter, one has to remember at least William Lawrence Bragg (1890-1971)[44] and Victor Moritz Goldschmidt (1888-1847)[45] who in 1920 and 1926, respectively, published detailed tables of ionic radii obtained via analysis of the interatomic distances.

The concept of radius connected with the length of predominantly ionic chemical bonds was success-

[6] Cf. references 40, 41, 42 and 43; in particular, reference 43 is fully dedicated to the nineteenth century French crystallography

ful, such that it was extended to other types of chemical bonds, defining atomic, metallic, covalent and van der Waals radii. Besides, Goldschmidt correlated the size of the cation ionic radii with the geometry of their coordination polyhedra, laying the groundwork for the five rules later established by Linus Pauling (1901-1994);[46] rules that are still useful tools for the validation and description of non-molecular crystalline structures. Finally, in 1921 Lars Vegard (1880-1963)[47] discovered that the cell parameters and the volume of the members of an isomorphous series are normally a linear function of the chemical composition.

## REFERENCES


1. N. Steensen, *De solido intra solidum naturaliter contento dissertationis prodromus*, Typographia sub signo Stellae, Florentiae, **1669**.
2. J.B. Romé de L'Isle, *Essai de cristallographie, ou description des figures géométriques, propres aux différents corps du règnes minéral, connus vulgairement sous le nom de cristaux*, Didot, Paris, **1772**.
3. G. Cardano, *De subtilitate rerum*, Iohan Petreium, Norimbergae, **1550**.
4. J. Kepler, *Strena, seu de nive sexangula*, Godefredum Tampach, Francofurti ad Moenum, **1611**.
5. R. Hooke, *Micrographia*, Jo. Martyn and Ja. Allestry, London, **1665**.
6. C. Huygens, *Traité de la lumière*, Pierre Vander Aa, Leide, **1690**.
7. R. Bartholin, *Experimenta crystalli islandici disdiaclastici, quibus mira et insolita refractio detegitur*, Danielis Paulli, Hafniae, **1669**.
8. G. Ferraris, *Dove la luce si fa in due*, in *La luce fra scienza e cultura. 2015, anno internazionale della luce*, Accademia delle Scienze, Torino, **2017**, pp. 3-14.
9. M.V. Lomonosov, *Sobranie sochineniy [Selected works]*, Vol. VI, pp. 111–152, Izd-vo AN SSSR, Leningrad, **1934**.
10. P.J. Macquer, *Dictionnaire de chimie*, Lacombe, Paris, **1766**.
11. R.J. Haüy, *Essai d'une théorie sur la structure des cristaux*, Gogué & Nèe de la Rochelle, Paris, **1784**.
12. R.J. Haüy, *Annales de Chimie* **1789**, *III*, 1.
13. R.J. Haüy, *Traité de Minéralogie*, 5 vols., Louis, Paris, **1801**.
14. R.J. Haüy, *Traité de cristallographie*, Bachelier et Huzard, Paris, **1822**.
15. A. Avogadro, *Journal de Physique de Chimie et d'Histoire Naturelle* **1811**, *73*, 58.
16. J.L. Gay-Lussac, *Mémoires de la Société de Physique et de Chimie de la Société d'Arcueil* **1809**, *2*, 207 and 252.
17. J. Dalton, *A new system of chemical philosophy, part 1*, S. Russel, Manchester, **1808**
18. A. Ampère, *Annales de Chimie* **1814**, *90*, 43.
19. G. Ferraris, *Amedeo Avogadro e la cristallografia*, in *A duecento anni dall'ipotesi di Avogadro*, Accademia delle Scienze, Torino, **2013**, pp. 41-52.
20. M. Morselli, *Amedeo Avogadro*, D. Reidel, Dordrecht, **1984**.
21. I. Guareschi, *Amedeo Avogadro e la sua opera scientifica*, in *Opere scelte di Amedeo Avogadro*, UTET, Torino, **1911**, pp. I-CXL.
22. W.H. Wollaston, *Philosophical Transactions of the Royal Society* **1813**, *103*, 51.
23. G. Delafosse, *Mémoires des Savants Étrangères* **1843**, *8*, 641.
24. N. Leblanc, *De la cristallotechnie, ou essai sur les phénomènes de la cristallisation*, no printer, Paris, **1802**.
25. F.S. Beudant, *Annales des Mines* **1817**, *2*, 1; *Annales des Mines* **1818**, *3*, 239 and 289.
26. W.H. Wollaston, *Annales de Chimie et de Physique* **1817**, *7*, 393.
27. E. Mitscherlich, *Abhandlungen der Königlichen Akademie der Wissenschaften Berlin* **1818–1819**, 427; *Annales de Chimie et de Physique* **1820**, *XIV*, 172.
28. E. Mitscherlich, *Annales de Chimie et de Physique* **1822**, *XIX*, 350.
29. S. Salvia, *Ambix* **2013**, 60 (3), 255.
30. J. Berzelius, *Annales de Chimie et de Physique* **1828**, *XXXVIII*, 426.
31. D.I. Mendeleev, *Zeitschrift für Chemie* **1869**, *12*, 405.
32. J.B. Biot, *Mémoires de la Classe des Sciences Mathématiques et Physiques de l'Institut Impérial de France* **1812**, *part 1*, 1.
33. A. Fresnel, *Bulletin des Sciences par la Société Philomathique de Paris* **1824**, 147.
34. L. Pasteur, *Annales de Chimie et de Physique* **1848**, *XXIV*, 442.
35. J.B. Biot, *Compte Rendu* **1844**, *19*, 720.
36. H. de la Provostaye, *Annales de Chimie et de Physique* **1841**, *3^e* série III, 129.
37. H. Flack, *Acta Crystallogr. Sect. A*, **2009,** *65*, 371.
38. J. Fournier, *Chimie Nouvelle* **2014**, *116*, 42.
39. J.B. Biot, *Comptes Rendus* **1839**, *8*, 683.
40. A. Authier, *Early days of X-ray crystallography*, Oxford University Press, Oxford, **2013**.
41. J.G. Burke, *Origins of the science of crystals*, University of California Press, Berkeley, **1966**.
42. J. Lima-de-Faria (Ed.), *Historical atlas of crystallography*, Springer, Dordrecht, **1990**.

43. S.H. Mauskopf, *Transactions of the American Philosophical Society* **1976**, *66*, 1.
44. W.L. Bragg, *Philosophical Magazine* **1920**, *40*, 169.
45. V.M. Goldschmidt, *Det Kongelige Norske Videnskabers Selskab Skrifter* **1926**, *no. 2*, 1; *Det Kongelige Norske Videnskabers Selskab Skrifte* **1926**, *no. 8*, 1.
46. L. Pauling, *The nature of the chemical bond and the structure of molecules and crystals: an introduction to modern structural chemistry*, Cornell University Press, Ithaca, **1939**.
47. L. Vegard, *Zeitschrift für Physik* **1921**, *5*, 17.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Historical Article

# Bringing Together Academic and Industrial Chemistry: Edmund Ronalds' Contribution

Beverley F. Ronalds

*University of Western Australia, Australia*
E-mail: beverley.ronalds@gmail.com

**Abstract**. Born 200 years ago, Edmund Ronalds (1819–1889) obtained his doctorate in Germany under Liebig, became a professor at Queen's College Galway and ran the little-studied but significant Bonnington Chemical Works in Edinburgh. His few mentions in the modern literature relate generally to the legacies of his actual and assumed academic supervisors of renown, yet his hitherto unknown mentors included family members and the important chemists Graham, Magnus, Tennant and Tennent. The novelty of his shift from university to manufacture has also been noted. With the aid of little-known primary sources, this biography details the evolution of Ronalds' career, exploring the context and influences for his diverse accomplishments and in particular the new and successful ways he bridged academia and industry through technological education and industrial research.

**Keywords**. Chemical technology, coal-tar processing.



**Citation:** B.F. Ronalds (2019) Bringing Together Academic and Industrial Chemistry: Edmund Ronalds' Contribution. *Substantia* 3(1): 139-152. doi: 10.13128/Substantia-211



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Competing Interests:** The Author(s) declare(s) no conflict of interest.

## UPBRINGING AND EDUCATION (1819-1842)

Edmund Ronalds, the eldest of at least twelve children, was born on 18 June 1819 at "No 1 Canonbury Square Islington", which then denoted the house on the west end of the partially-completed square (Figure 1).[1] His father Edmund Sr had lived his early years just down the road in Canonbury Place and now ran the family's large wholesale cheesemonger business in Upper Thames Street, London.[2] Edmund's mother Eliza Jemima was the only daughter of James Anderson,[3] a Scot who graduated from the University of Edinburgh and was awarded a Doctor of Laws there in 1794.[4] He ran a respected academy at Mansion House in Hammersmith offering a broad-based and vocationally-oriented curriculum.[5]

[1] The address of the house is given in Ronalds' birth registration at Dr Williams's Library (now in the National Archives) and its location can be discerned from the extended series of rate books held at the Islington Local History Centre.
[2] B. F. Ronalds, *Sir Francis Ronalds: Father of the Electric Telegraph*, Imperial College Press, London, **2016**.
[3] *Gentleman's Mag.* **1818**, *88:2*, 178.
[4] *Register of Laureations in the University of Edinburgh M.DLXXXVII–M.DCCC.*
[5] N. Hans, *New Trends in Education in the Eighteenth Century*, Routledge, London, **2001**, p. 111.

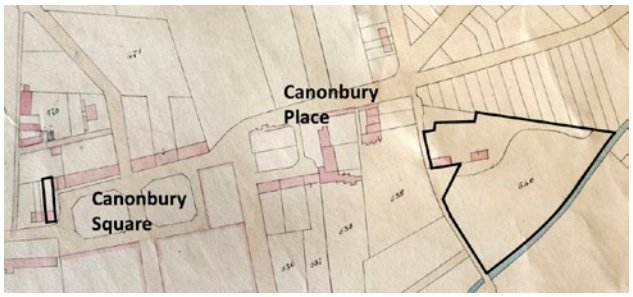



**Figure 1.** Locations of Ronalds' two homes in Canonbury, Islington. Source: Titheable Lands in the Parish of Saint Mary Islington, 1849, London Metropolitan Archives DL/TI/A/029/A. By permission of the Bishop of London and the London Diocesan Fund.

The family soon after moved to Brixton Hill, "nearly opposite the Telegraph",[6] where Edmund fell seriously ill[7] and a number of his siblings died. As a result, his surviving brothers were more than thirteen years his junior. Despite the spread of ages, it was a close and happy family, with later letters reminiscing of their "merry and boisterous" evenings.[8] They sang and played music together and conversation was informed by well-rounded education and their parents' friendships. Christmas Day was often spent with the Martineau family:[9] Edmund's aunt had married Peter Martineau, through whom they met his cousin, the sociologist Harriet Martineau. Edmund Sr and Eliza's associates included the early socialists and educational reformers Robert Owen and Fanny Wright. Edmund's brothers attended from about age five an "admirably-kept" preparatory boarding school[10] and his own education would have commenced in a similar manner, while his sisters were described by associates as "well educated" and read several languages.[11]

The Ronalds family being Dissenters – of the Unitarian faith – could not graduate from the English universities Cambridge and Oxford. Students of the first secular institution, University College London, were not awarded degrees until 1839. Any continuation of Edmund's studies of this kind would necessarily be undertaken elsewhere. His obituaries noted that he spent time in "Giessen, Jena, Berlin, Heidelberg, Zurich, and Paris",[12] a list that would have been provided by someone who knew him well. His entry in the *Dictionary of National Biography* and all but one of these obituaries (that written by his friend John Young Buchanan who lived near his widow and children) prefixed the descriptor "successively" to the names and, as a result, inaccurate assumptions have been made as to the identity and timing of his professors. The list is actually in a decreasing order of importance, based on such factors as stage in his education and length of attendance, and thus largely in reverse chronological order.

Edmund probably commenced his university studies in Paris as, like others in the family, he was most comfortable in French. Years later he edited a booklet for his uncle Sir Francis Ronalds in that language; Sir Francis – who was knighted for developing the first working electric telegraph – was a key influence for him and the two were always close and mutually supportive.[13] Once Edmund was sufficiently confident living abroad, and had shown his potential, he headed to the German regions and their renowned academics.

His teacher at Heidelberg could not have been Robert Bunsen as has on occasion been presumed,[14] as Bunsen was then elsewhere and Ronalds would still have been taking general courses. In late 1838 a family associate, the Unitarian diarist Henry Crabb Robinson, organised a letter of introduction to his botanist friend Professor Friedrich Siegmund Voigt at the University of Jena. Ronalds matriculated at this university on 29 April 1839 and remained three semesters, his major subject being philosophy with Jakob Friedrich Fries.[15] He had a break at home in April 1840, during which he was invited to breakfast with Robinson. His host, although admitting he did not understand science, noted in his diary that he "was pleased with him".[16]

Ronalds moved to the University of Berlin later in 1840 for the next three semesters.[17] He told his uncle Sir Francis that there it was Gustav "Magnus the professor of physicks & technology in whose laboratory I worked or rather idled a good deal of time", although he did

[6] E. Ronalds to R. Owen, 7 September 1829(?), Robert Owen Collection, National Co-operative Archive, Manchester, ROC/17/31/1.
[7] J. Lawe to E. Ronalds, 24 October 1834, Ronalds Family Papers, Harris Family Fonds, Western Archives, Western University, London, Ontario, Canada (hereafter WU), B1450.
[8] H. Ronalds to E. Ronalds, 28 March 1854, Alexander Turnbull Library, Wellington, New Zealand, qMS-1719 (hereafter ATL).
[9] S. Flower, *Great Aunt Sarah's Diary 1846–1892*, Printed privately, **1964**, p. 45.
[10] England Census, 1841; *Edmund Yates: his Recollections and Experiences*, Vol. 1, Richard Bentley, London, **1884**, p. 35.
[11] G. H. Scholefield, Ed., *Richmond-Atkinson Papers,* Vol. 1, NZ Government Printers, Wellington, **1961**, p. 473.
[12] *Proc. R. Soc. Edinburgh* **1889–1890**, *17*, xxviii; *J. Chem. Soc. Trans.* **1890**, *57*, 456; *Proc. Inst. Chem.* **1890**, *14*, 53.
[13] Ronalds, *Sir Francis Ronalds*.
[14] George Ronalds (unrelated to Edmund) studied with Bunsen at Heidelberg in the 1850s. See J. T. Krumpelmann, *Jahrbuch für Amerikastudien* **1969**, *14*, 167.
[15] University Archives Jena, Bestand BA, No. 815/9; Bestand G, Abt. 1, No. 67–72.
[16] H. C. Robinson, Diaries, 29 April 1840, Dr Williams's Library, London, with permission from the Trustees.
[17] *Amtliches Verzeichnis des Personals und der Studierenden der Königlichen Friedrich-Wilhelms-Universität zu Berlin*, Berlin, **1840–1841**, **1841**, **1841–1842**.

not neglect Magnus' colleague Heinrich Rose, whom he called "the great analytical chemist of the age".[18] It is of note that he was now orienting towards "technology"; this was already an academic field in Germany, associated with cameralism – administrative sciences promoting efficient stewardship of economic activity for the benefit of the state.[19] A short stay with Magnus' friend Justus Liebig at the University of Giessen formed the capstone of his formal education: he enrolled on 7 May 1842 and was awarded the degree of Doctor of Philosophy less than three months later on 2 August 1842.[20] He mentioned just these last two professors – Liebig and Magnus – and their laboratories in a brief statement of experience on his later professorial appointment.[21]

Ronalds' thesis, which contributed to Liebig's agricultural and physiological chemistry studies, addressed the analysis of wax by oxidation. He found that a crystalline material was produced after an extended reaction time with nitric acid; this proved to be succinic acid, which has biological functions. The work was published immediately in Liebig's journal under Ronalds' name, abstracted in *Pharmaceutisches Central-Blatt*, and quickly referenced by Charles Gerhardt, Bernhardt Lewy and Liebig himself in subsequent papers.[22]

The extent of his education and its subject matter indicate the family's affluence. When he embarked on his university training, there were few academic positions in chemistry in Britain (and even fewer for Dissenters) and these were not always salaried. It was largely his share of the family's accumulated wealth that would enable him to pursue his scientific interests while supporting a sizable future family and maintaining his accustomed lifestyle. Sir Francis had chosen this life of "gentleman scientist", determining his own research priorities and only taking on roles in an honorary capacity. Sir Francis' "chief amusement" in his youth had been chemistry.[23]

The family's religious and moral values in addition emphasised the application of knowledge acquired to bring benefit for society;[24] this ethos is apparent throughout Ronalds' career and is a central theme of this paper. The last two supervisors he chose were known for their laboratory-based teaching and gave him a strong grounding in practical science. Aided by his doctorate, a path in analytical consulting was thus also open to him. By way of example, Edmund Sr's cousin Silvanus Ronalds was Chemical Operator and a consultant with the Society of Apothecaries.[25] Another possible avenue was the growing manufacturing sector. Various members of his extended family were largescale industrialists – his uncle Peter Martineau owned and ran a sugar refinery.[26] Ronalds was to pursue all these options in the course of his career.

## ACADEMIA (1842-1856)

### *In London*

Immediately after completing his thesis, Ronalds returned home to his family, who were now living at a property of three acres called the Grove at the east end of Canonbury Place; its location is shown in Figure 1. Liebig visited him there right away – in mid-August 1842 – at the commencement of a trip around England, and kept his luggage there.[27] Liebig then met up with Thomas Graham, chemistry professor at University College, before heading to the regions.

A cousin reported the next step very soon afterwards. Ronalds had "most fortunately met with a situation exactly suited to him as assistant to a M$^{r}$ Graham the first Chemist in London which will occupy him from 11 Oclock to 5 every day and be the means of introducing him to become a popular man himself if he makes good use of the advantages he now enjoys".[28] Liebig must have been complimentary about Ronalds' abilities. Sir Francis could also have provided a recommendation to Graham: they knew each other quite well,[29] in part through their shared interest in the Kew Observatory that Sir Francis was beginning to set up for the British Association for the Advancement of Science (BAAS).

Just as his cousin recommended, Ronalds used every opportunity to meet other chemists and be helpful. Graham having begun his career in Glasgow, there

[18] E. Ronalds to F. Ronalds, 19 June 1858, Institution of Engineering and Technology Archives (hereafter IET), 1.9.1. See: A. W. Hofmann, *Allgemeine Deutsche Biographie* **1884**, *20*, 77.
[19] E. Schatzberg, *Technology: Critical History of a Concept*, UCP, Chicago, **2018**, p. 77–81.
[20] F. Kössler, *Register zu den Matrikeln und Inscriptionsbüchern der Universität Giessen 1807/08–1850*, Universitätsbibliothek, Giessen, **1976**, p. 155; Kössler, *Verzeichnis der Doktorpromotionen an der Universität Giessen von 1801–1884*, Universitätsbibliothek, Giessen, **1970**, p. 84.
[21] *Galway Vindicator*, 11 August 1849, 2.
[22] E. Ronalds, *Ann. Chem.* **1842**, *43*, 356. Summarised in *Pharmaceutisches Central-Blatt* **1842**, *2*, 926.
[23] F. Ronalds to S. Carter, 21 February 1860, University College London (UCL) Special Collections, GB 0103 MS ADD 206.
[24] Ronalds, *Sir Francis Ronalds*, pp. 53–54, 93–94.
[25] A. E. Simmons, *The Chemical and Pharmaceutical Trading Activities of the Society of Apothecaries, 1822 to 1922*, Ph.D. Thesis, The Open University, UK, **2004**.
[26] B. F. Ronalds, *Martineau Society Newsletter* **2018**, No. 41, 10.
[27] J. Volhard, *Justus von Liebig,* Vol. 1, Verlag, Leipzig, **1909**, p. 160.
[28] M. Ronalds to H. Ronalds, 12 October 1842, WU, B2284.
[29] Ronalds, *Sir Francis Ronalds*, p. 546.

was a steady stream of Scots to his laboratory. He was the founding president of the Chemical Society of London and the Cavendish Society,[30] and Ronalds joined both immediately, becoming a council member of the latter. Another original member of these organisations was John Tennent, denoted erroneously at times as "Tennant".[31] Both Johns – Tennent and Tennant – had grown up in the Glasgow area, studied chemistry under Thomas Thomson (as had Graham)[32] and became chemical manufacturers, and both would be prominent in Ronalds' future. The two men have been confounded over the years. For example, the Chemical Society's Jubilee Album featuring its founding members contains Tennant's rather than Tennent's portrait.[33]

Tennant (1796-1878) was the managing director of the "gigantic" Charles Tennant & Company established by his father, with its St Rollox chemical works that made bleaching powder.[34] Tennent (1813–1862) was the son of Barbara née Graham and Hugh Tennent, who helped run the famous Tennent Brewery. It was apparently the Tennent family who sold the land for St Rollox to the Tennants.[35] John Tennent and John Tennant partnered in the Bonnington Chemical Company in 1847, with the former being the manager of the facility.[36]

There was in addition a strong network of alumni from the universities Ronalds had attended. Former Giessen students Edward Frankland and Robert Angus Smith both asked him to be a referee when they applied for the professorship at Owens College, Manchester.[37] Ronalds also hosted numerous visitors that he had met abroad. Within weeks of arriving home, he had as guests "2 young Hungarians who could not speak one word of English but they were very animated & agreeable, both professors".[38] Fortunately several family members could contribute to the conversation in German.

Liebig visited again in 1844. It was Graham who took him to visit Sir Francis at the Kew Observatory on 4 September[39] and both also went to the BAAS annual meeting at York. This was the first BAAS conference that Ronalds and his uncle attended,[40] and he would have been proud to be associated with these mentors while meeting more of their associates. In 1851 Liebig visited him in Galway.[41]

Ronalds became a member of the BAAS in 1846[42] and, slowly gaining confidence, contributed increasingly to the technical discussions there.[43] He served as secretary of the chemical science section at the 1852 meeting held in Belfast and later as section vice-president at Edinburgh in 1871 and Sheffield in 1879. This was perhaps one of the ways he kept in touch with Magnus, who also visited him, his uncle and the Kew Observatory on a trip to England.[44] Ronalds in addition translated and summarised papers by his colleagues (as well as Liebig's) for publication in English journals.[45]

Already he had mix of experiences relevant for his later career path across academia and industry. He had started with a sojourn in Germany, where he received the best practical chemistry training in a culture of science utilisation, along with numerous contacts and associated kudos. He was now active in the overall chemical profession at its hub in his London hometown, with its links to commerce and government. He had friends and family from Glasgow and Edinburgh, important industrial centres that had close connection with their universities, and he was interacting with other chemists and industrialists at the BAAS. These built on the foundation of his Unitarian circle with its accent on societal benefit through education. Although the groupings overlapped significantly, as Bud and Roberts have illustrated through Lyon Playfair and others, Ronalds was unusual in having the influence of all of these education-practice networks early in his academic career.[46]

He now determined to develop his teaching skills and was soon giving lectures in London and further afield. On 19 February 1845, for example, he lectured on "Chemical principles of Gas Manufacture" at the Derby Mechanics' Institution and he taught at a school in Worksop, near Sheffield, that had a chemical laboratory.[47] Beginning in October 1845 he gave lectures at the

[30] W. H. Brock, *Ann. Sci.* **1978**, 35, 599.
[31] See for example: *Proc. Chem. Soc.* **1842**, *1*, 1.
[32] R. D. Thomson, *Edinburgh New Philosophical J.* **1853**, *54*, 86.
[33] *Jubilee of the Chemical Society of London*, Chem. Soc., London, **1896**, p. 24.
[34] *Glasgow Herald*, 18 April 1878, 4.
[35] Tennent Family Trees, University of Glasgow Archive Services, GB 248 T 13/1; G. Stewart, *Curiosities of Glasgow Citizenship*, James Maclehose, Glasgow, **1881**, p. 239.
[36] J. A. Anderson, Bonnington Chemical Works, 1851, National Records of Scotland (hereafter NRS), CS313/946; *Proc. Chem. Soc.* **1868**, *21*, xxix.
[37] E. Ronalds to E. Frankland, 10 May 1850, Papers of Sir Edward Frankland, Special Collections, University of Manchester, RFA OU mf 01.03.0900.
[38] E. Ronalds to H. Ronalds, 2 October 1842, WU, B558.
[39] Kew Observatory Diary and Accounts, 1844, National Meteorological Library and Archive, Exeter.
[40] Ronalds, *Sir Francis Ronalds*, p. 336.
[41] E. K. Muspratt, *My Life and Work*, John Lane, London, **1917**, p. 36.
[42] *Report of the 59th Meeting of the British Association for the Advancement of Science*, John Murray, London, **1890**.
[43] For example: *Annual of Scientific Discovery: or, Year-book of Facts in Science and Art*, Gould and Lincoln, Boston, **1850**, pp. 207–08; *Daily News*, 9 September 1852, 3.
[44] E. Ronalds to F. Ronalds, 19 June 1858.
[45] For example: *Philos. Mag.* **1846**, *28*, 161, and *29*, 25, 31.
[46] R. Bud, G. K. Roberts, *Science versus Practice: Chemistry in Victorian Britain*, MUP, Manchester, **1984**.
[47] *Derby Mercury*, 15 January 1845, 2; Muspratt, *My Life and Work*, p.

Aldersgate School of Medicine through the winter session and offered practical classes three days per week – this increased to four days the following year.[48] He was additionally lecturing regularly at the Middlesex Hospital School of Medicine and offering "Private Instruction in CHEMICAL MANIPULATION and ANALYSIS… at the Laboratory of the Hospital School" there.[49] The latter was affiliated with the nearby University College.

His role as "Lecturer on Chemistry at the Middlesex Hospital" was a continuing appointment and he began to use it as his affiliation for publications and in societies. The chemical laboratory was available to him to conduct consulting activities and research. He quantified the copper content of ores provided by the Australian Mining Company from their proposed Tungkillo mine near Adelaide, and published the results in the literature.[50] Mining continued there for some years.

He also devised and performed tests to assist medical questions. He discovered taurine in human bile, which was announced in the *Chemical Gazette* by his Giessen friend William Francis (who was later a partner in Taylor and Francis publishers).[51] Links between the impurities in water and its utility were beginning to be considered in this period and he undertook water quality analyses in several locations. These included the water supply for the new railway town of Wolverton, to help determine the best treatment for ailments experienced by residents, and spring water from the Colne Valley near Watford that was proposed to be pumped to Hampstead.[52] He also studied how the amount of organic matter taken up by water from peat increased with its temperature.[53]

On 18 June 1846 Golding Bird, a physician at Guy's Hospital, read a paper by Ronalds to the Royal Society. He had shown in what was viewed as "a series of well-devised experiments"[54] that urine contained sulphur and phosphorus in both unoxidised and oxidised states and quantified the amounts in 24-hour urine tests. The higher unoxidised sulphur in a diabetic patient illustrated the potential use of the results in diagnosis. The article was included in the *Philosophical Transactions* and published in the *Philosophical Magazine* and German journals.[55] The results were quickly picked up in summaries of medical advances and in pathology lectures and continued to be referenced into the twentieth century.[56]


**Figure 2.** Edmund Ronalds, photographed in May 1878 by George Shaw in Edinburgh. Source: Sir George Grey Special Collections, Auckland Library, New Zealand, NZMS 1235.

With his reputation growing, Ronalds (Figure 2) was given the opportunity to undertake two significant projects. Becoming secretary of the Chemical Society, he was the inaugural editor of its first journal. He was responsible for the first two volumes of the *Quarterly*

36. The school was founded on Johann Pestalozzi's educational philosophy, with which Ronalds' aunts had strong links, and the principal Dr Benjamin Heldenmaier was active in the Derby Mechanics' Institution.

[48] *Morning Chronicle*, 22 September 1845, 5; *Lancet* **1845**, *46*, 339; *Lancet* **1846**, *48*, 345.

[49] *Exeter Gazette*, 19 September 1846, 2; *Athenæum* **1846**, 1009; *Lancet* **1847**, *50*, 361.

[50] E. Ronalds, *Chemical Gazette* **1846**, *4*, 463.

[51] *Chemical Gazette*, **1846**, *4*, 281, 295; *Lancet*, **1848**, *52*, 335.

[52] G. Corfe, *Pharmaceutical Journal and Transactions* **1849**, *8*, 30, 71; *Morning Post*, 11 January 1850, 5.

[53] *Q. Rev.* **1850**, *87*, 479.

[54] G. Day, *Half-yearly Abstract of the Medical Sciences* **1847**, *5*, 285.

[55] E. Ronalds, *Philos. Trans. R. Soc. London* **1846**, *136*, 461. Also in: *Philos. Mag.* S3 **1847**, *30*, 253; *Journal Prakt. Chem.* **1847**, *41*, 185; *Notizen aus dem Gebiete der Natur- und Heilkunde* **1847**, *3*, 214.

[56] For example: A. B. Garrod, *Lancet* **1848**, *52*, 441, 469, 599; *Sci. Am.* **1869**, *21*, 249; J. J. Rae, *Biochem. J.* **1937**, *31*, 1622.

*Journal* published in 1849 and 1850, and received an honorarium of £50 each year. He incorporated a list of all papers published in chemistry locally and overseas and prepared abstracts of interesting papers appearing in foreign language journals.[57] When he retired to move to Galway, Henry Watts was employed as a paid editor but, unlike Ronalds, his name did not appear on the title page.

### *Chemical Technology*

The other project was a large book. Friedrich Ludwig Knapp, professor of technology at the University of Giessen, was preparing a text called *Lehrbuch der chemischen Technologie* and it would have been his brother-in-law Liebig who invited two of his past students, Ronalds and Thomas Richardson, to translate it into English. The preface to the first volume of their edition bore the same date of 1847 as Knapp's work and so they must all have been working in concert. In the English publication, entitled *Chemical Technology; or, Chemistry Applied to the Arts and to Manufactures*, Knapp was denoted as the author and it was "edited with numerous notes and additions" by Ronalds and Richardson; it was of credit to Ronalds to be the first-named of these two authors so early in his career. Their additions to the book included "excellent" figures[58] to give a total of over 550 illustrations. Knapp's first volume covering fuel, alkalies and earths was split into two, both appearing in 1848, and their third volume on food was completed in 1851.[59] They formed part of a new Library of Illustrated Standard Scientific Works published by Hippolyte Bailliere in London.

Although Ronalds downplayed the academic rigour of the book, calling it before it appeared "a merely popular treatise",[60] its research would have deepened his technical knowhow across the breadth of British chemical manufacture. Colleagues and family members like Tennent, Tennant and Martineau who owned processing plants offered assistance and in return benefited from the resulting amalgamation of current scientific thinking with industry best practice and trends.

Reviews in the press were very positive. The opinion of the *Athenæum* was that "To the manufacturer this publication must prove eminently useful" and it is also "most valuable as one of general reference".[61] The *Economist* highlighted "the good judgment of the translators, who have… done a great service to the public". "Scientific knowledge… is explained in a simple manner" while "Scientific men will hail with delight the quantity of practical information". "It is a book for everybody".[62] Even the *Lancet* gave a page-long review. Overseas, *Scientific American* called it a "great work" while the *Journal of the Franklin Institute* wrote that "The English editors have also performed their task with talent and faithfulness, as is evidenced by the large and judicious additions which they have made, describing British inventions and improvements, and giving us the latest results of British science and ingenuity".[63] An American edition of the first two volumes was quickly published in which Walter Rogers Johnson made further additions emphasising US industry.[64]

There was initially little mention of the potential value of the book in formal education. The authors had lamented in their preface the lack of higher education establishments with a technical emphasis. Chemist George Wilson was the first professor of technology in Britain and he explained in his inaugural lecture in Edinburgh in 1855 that "the word Technology has been introduced into our language" through the book.[65] Subsequent assessments suggest comparable conclusions on the text's novelty and significance.[66] "Technology" has a Greek etymology and, because it was then in few dictionaries, was described by the authors as "the systematic definition (λογος) of the rational principles upon which all processes employed in the arts (τεχνης) are based"; (after coming into use its meaning altered in the twentieth century as described by Schatzberg).[67] Their focus was thus a framework to aid understanding, use and development of plant processes, equipment etc. *Chemical Technology* can be considered to be a key early emphasis outside Western Europe on a distinct educational discipline of chemistry application for industry.[68]

Ronalds and Richardson soon began work on an updated edition of *Chemical Technology*. This became

---

[57] R. S. Cahn, *Proc. Chem. Soc.* **1958**, 157.
[58] *Sci. Am.* **1855**, *11*, 112.
[59] F. Knapp, E. Ronalds, T. Richardson, *Chemical Technology; or, Chemistry Applied to the Arts and to Manufactures*, Bailliere, London, **1848–1851**.
[60] E. Ronalds to L. L. Dillwyn, 20 September 1848, Swansea University Archives, GB 217 LAC/26/D/55.
[61] *Athenæum* **1849**, 321.
[62] *Economist*, 2 December 1848, 1364.
[63] *Sci. Am.* **1852**, *7*, 221; *J. Franklin Inst.* S3 **1848**, *15*, 449.
[64] Johnson's career is described in: G. E. Pettengil, *J. Franklin Inst.* **1950**, *250*, 93.
[65] G. Wilson, *What is Technology?* Sutherland and Knox, Edinburgh, **1855**. See also: R. G. W. Anderson, *Br. J. Hist. Sci.* **1992**, *25*, 169.
[66] Schatzberg, *Technology*, pp. 81–82, 91–94; J. M. van der Laan, *Narratives of Technology*, Springer, New York, **2016**, pp. 25–27; R. P. Multhauf, *The History of Chemical Technology: An Annotated Bibliography*, Garland, New York, **1984**; Bud, Roberts, *Science versus Practice*, p. 108.
[67] E. Schatzberg, *Technology and Culture* **2006**, *47*, 486.
[68] W. Schneider, *Neue Deutsche Biographie* **1979**, *12*, 151. Schatzberg, *Technology*, p. 81.

essentially a new work, much rewritten and enlarged. They were now the named authors, but noted that it "incorporated a revision of Dr Knapp's "Technology"". The text needed to be further divided, and the first two volumes covering Fuel and its Applications were published in 1855. They also received strong reviews, the *American Journal of Science* calling it "by far the most full, scientific and satisfactory exposition of the subjects of Fuel and Illumination to be found".[69]

Ronalds' priorities changed abruptly in this period, as explained below, and he stepped aside after these two volumes. Richardson and his new co-author Henry Watts completed the material on Acids, Alkalies, and Salts in 1867, which is the year Richardson died.[70] That it took twelve years to issue these later books hints at the scale of Ronalds' contribution to the earlier ones. The volume on food was not updated.

The overall book "became a standard work" internationally;[71] it was still advertised for sale in the *Chemical News* in the 1870s. Material was commonly quoted in other texts[72] and is referenced today in histories of the chemical industry to explain nineteenth-century processes.[73] It stood the test of time for over thirty years.

Watts had begun preparing an update before his death in 1884 and Charles Edward Groves, who replaced him as editor of the Chemical Society's journal, then took on the role of general editor for a new edition with oversight of numerous authors.[74] The first volume emerged in 1889 – the year Ronalds died – followed by three more in the period to 1903. The preface erroneously described them as being founded on Richardson and Watts' work but in fact they covered only fuel and lighting and thus used Ronalds and Richardson's volumes as their basis. This edition also received good reviews and maintained the strong reputation of the title. It is of interest that editors of the Chemical Society journals played a leading role in all the versions.

*Chemical Technology* featured increasingly in university education over time. It was included in the recommended library list published by the Canadian *Journal of Education* as early as March 1854.[75] Ronalds presented the 1848–1851 edition to the Queen's College Galway library and subsequent versions were acquired by the college as well. The 1855–1867 and 1899–1903 editions are held by innumerable universities around the world and Kikuchi has outlined how they would have been used in teaching.[76] Putting this progression into context, university chairs in chemical engineering were only established in the early twentieth century.[77]

### *In Galway*

Non-denominational higher education had commenced in Ireland in 1849 with the creation of the Queen's University of Ireland, which awarded degrees for the new Queen's Colleges of Belfast, Cork and Galway. These offered academic positions for which a Dissenter like Ronalds was eligible and he was appointed as the inaugural chemistry professor at Galway at age thirty. His salary would be £200 plus additional student fees.[78]

He asked Sir Francis to dine with him in Canonbury on 14 October 1849 to say farewell, along with Graham, and also Thomas Andrews, who was the first vice-president of Queen's College Belfast. He suggested his uncle's "advice about the purchases of physical apparatus would be of service to the irish colleges".[79] Ronalds and his sister left London immediately afterwards and were in Galway in a week.[80]

He gave his introductory chemistry lecture on 11 December.[81] Impatient to begin in earnest, he complained to Sir Francis the next February that "the intolerably dawdling habits of all workmen in this place has prevented me from yet getting to work in the laboratory. I do not think I shall be able to begin my course for some weeks".[82] Once up and running, he delivered up to 140 lectures each year at the college, around 40 being in practical chemistry in the laboratory,[83] and "he was

---

[69] *Am. J. Sci. Arts* S2 **1856**, *22*, 149.
[70] E. Ronalds, T. Richardson, H. Watts, *Chemical Technology; or, Chemistry in its Applications to the Arts and Manufactures*, Bailliere, London, **1855–1867**.
[71] "Richardson, Thomas (1816–1867)", *Oxford Dictionary of National Biography*.
[72] Muspratt, for example, referred to "the valuable treatise" numerous times in his *Chemistry, Theoretical, Practical & Analytical*, William Mackenzie, Glasgow, **1860**.
[73] For example: C. A. Russell, *Chemistry, Society and Environment: A New History of the British Chemical Industry*, Royal Society of Chemistry, Cambridge, **2000**.
[74] W. H. Brock, *The Case of the Poisonous Socks: Tales from Chemistry*, Royal Society of Chemistry, London, **2011**, p. 247.
[75] *Journal of Education for Upper Canada*, **1854**, *7*, 33.
[76] Y. Kikuchi, *History of Science* **2012**, *50*, 289. See also: *Anglo-American Connections in Japanese Chemistry: The Lab as Contact Zone*, Palgrave Macmillan, New York, **2013**, p. 44.
[77] C. Divall, S. F. Johnston, *Scaling Up: The Institution of Chemical Engineers and the Rise of a New Profession*, Kluwer, Dordrecht, **2000**.
[78] A. J. Ryder, *An Irishman of Note: George Johnstone Stoney*, Printed privately, **2012**, pp. 89–92.
[79] E. Ronalds to F. Ronalds, 12 October 1849, IET, 1.3.332.
[80] *Freeman's Journal*, 23 October 1849, 2.
[81] *Galway Vindicator*, 28 November 1849, 3.
[82] E. Ronalds to F. Ronalds, 9 February 1850, IET, 1.3.362.
[83] See for example: *Report of the President of Queen's College, Galway,* for the academic year 1852–53, HMSO, Dublin, **1854**, p. 7; and, for the year 1856, **1857**, p. 4.

remembered as a successful and inspiring teacher".[84] His first course outline and examination questions survive in the college calendar.[85] In 1854 he was able to take on Edward Divers as an assistant to help with the demonstrations.

Giving his new affiliation on the title page of *Chemical Technology* would have been a welcome boost to the reputation of the embryonic university. It was formally listed as a course textbook by Ronalds' successor.[86] Teaching of "chemistry applied to the arts and to manufactures" began to receive attention at various colleges from around mid-century, and Galway is an early example that has gone unnoticed in previous analyses of this curricular development. With his authorship and German education, Ronalds' approach was presumably more rational and balanced than efforts elsewhere in Britain, which matured only very slowly as alluded to above. Donnelly and others have discussed how this was in part because academics argued that their preferred "pure" chemistry was what industry needed, hinting at an academic elitism that appears again below. The technology chair at Edinburgh lapsed with Wilson's death in 1859 for similar reasons.[87] Ronalds suffered the disadvantage however of Galway having limited manufacturing industry and thus needing to rely on the book to illustrate how different chemical processes could be deployed at scale.[88]

He pursued other teaching opportunities as well. He gave a course of nine public lectures illustrated by "a series of beautiful and highly-successful experiments" under the auspices of the Board of Trade and the Royal Galway Institution. The press was most complementary about "the able and talented lecturer" – "we have never attended any Lectures with more pleasure". One commentator did regret however that he "does not avail himself of the opportunities… of directing the attention of the hearers to that Great and Almighty Being".[89] This was a reflection of widespread antipathy towards the new "godless colleges".[90]

He quickly adopted a priority of investigating local natural resources with a view to possible new and enhanced industries for the area, which had suffered terribly during the recent potato famine; the results would also have informed his lectures. He analysed peat found in different situations in Galway, including the quantity and composition of its ash and how the water content varied with drying method, both of which affected its value. The results were summarised in *Chemical Technology* (1855), repeated almost verbatim in the 1889 edition and continued to be quoted into the next century.[91] He had earlier studied the ash of several coals and these data were included in both editions of the book as well. He also analysed a peat fertiliser and fungicide for a new company.[92] He later donated "Specimens illustrative of the products of the destructive distillation of wood, bones, and coal, &c" to the Museum of Irish History in Dublin.[93]

The Irish press was delighted to announce in September 1852 that "The eminent authoress" Harriet Martineau was "on a visit with Dr. Ronalds".[94] She described in the national *Daily News* and in her subsequent book that the "professor of chemistry" attempted to demonstrate how the local red seaweed could be burnt to produce iodine and potash salts to supplement its traditional use as a fertiliser.[95] The locals, after accepting his advance payment to conduct a trial, apparently declined to participate. The new industry did develop however and continued into the twentieth century.[96]

She also highlighted work he presented to the 1852 BAAS meeting on the oil of the basking shark, which was found off the Bay of Galway. The fish contained large quantities of a very light oil and Ronalds emphasised its unusual and valuable properties, including its bright flame and possible medicinal uses, in the hope that the fishermen might obtain a higher price for it in new applications. The results were summarised in the *Athenæum*, published in the *Chemical Gazette* and included in *Chemical Technology* and other texts.[97] He also advised Sir Francis in this period on oil lighting for the continuously-recording cameras he had developed. In return he later teased his uncle that he "may possi-

[84] *Dictionary of Irish Biography,* Vol. 8, CUP, Cambridge, **2009**, pp. 597–98.
[85] *Calendar of Queen's College, Galway*, Hodges and Smith, Dublin, **1851**.
[86] See for example: *Report of the President of Queen's College, Galway,* for the academic year 1863–64, HMSO, Dublin, **1865**, p. 22; and, for the year ending 31st March, 1867, **1867**, p. 24.
[87] J. F. Donnelly, *Social Studies of Science* **1986**, *16:2*, 195; J. F. Donnelly, *History of Education* **1997**, *26:2*, 125; Bud, Roberts, *Science versus Practice*; Schatzberg, *Technology*, pp. 64–65; J. F. Donnelly, *Chemical Education and the Chemical Industry in England from the Mid-Nineteenth to the Early Twentieth Century*, Ph.D. Thesis, University of Leeds, UK, **1987**; Anderson, *Br. J. Hist. Sci.*
[88] Kikuchi, *History of Science*.
[89] *Galway Vindicator*, 10 February 1855, 2; 5 May 1855, 2.
[90] J. O. Ranelagh, *A Short History of Ireland*, 3rd Ed. CUP, Cambridge, **2012**, p. 141.
[91] W. A. Kerr, *Peat and its Products*, Begg, Kennedy & Elder, Glasgow, **1905**, p. 27.
[92] *Galway Vindicator*, 28 August 1852, 3.
[93] *Fourth Report of the Department of Science and Art*, HMSO, London, **1857**, p. 94.
[94] *Freeman's Journal*, 3 September 1852, 2.
[95] *Daily News*, 3 September 1852, 4; H. Martineau, *Letters from Ireland*, John Chapman, London, **1852**, pp. 82–91.
[96] G. H. Kinahan, *Q. J. Sci.* **1869**, *6*, 331.
[97] E. Ronalds, *Chemical Gazette* **1852**, *10*, 420. Also in: *Athenæum* **1852**, 1042. Summarised in: H. Watts, *Dictionary of Chemistry and the Allied Branches of other Sciences*, Vol. 5, Longmans, London, **1868**, p. 404.

bly… find time to make me that glass float w$^{h}$ has been five & twenty years in process".[98] Ronalds was presumably wanting a better hydrometer.

## FROM UNIVERSITY TOWARDS INDUSTRY

The 1850 BAAS meeting had been held in Edinburgh. On 23 December that year Ronalds married his friend Tennent's sister Barbara Christian at her mother's home: 128 Wellington Street, Glasgow.[99] The couple went on to have three daughters followed by three sons.

Not long afterwards, the Ronalds family suffered a major change of fortune. With Edmund Sr's younger sons now completing their schooling, he wished to fund their establishment in life. He had borrowed £12,000 from his elderly mother during the economic recession of the late 1840s and, on her death in 1852, the family cheesemonger business was sold and he invested his inheritance in a large silk mill in Derby that was in debt. The idea was that his son Hugh would learn the business and then start running it. Instead the current managers apparently absconded with the money.[100] A cousin summed up the outcome for Edmund Sr: "he must be much reduced in circumstances as two of his daughters have been obliged to go out as Governesses".[101] One went on to establish a respected school and another became a nursing sister. Their brother Hugh later reminisced about "the careless way I thought of money and time… no care or anxiety for the future" in the years before "the smash".[102]

Another of the sons had attended Queen's College Galway for a year, but did not continue his studies.[103] The three young men, aged eighteen, nineteen and twenty, set sail for New Zealand in February 1853 with their fares and early subsistence funded by Uncle Martineau. It was intended that the rest of the family would follow once they were settled as it was "mother's wish… to fly from all society" and escape her embarrassment. After arriving, however, Hugh quickly warned her not "to induce Edmund to come out, the settlement is too young and poor to attempt any experiments… I suppose there is no chance of his thinking of giving up his chymistry".[104] In the meantime Edmund Sr and Eliza joined Ronalds in Galway. Eliza's death there altered plans – two of Ronalds' sisters joined their brothers but the rest of the family remained in Britain.

The brothers took labouring work to support themselves while clearing a farm in the bush outside New Plymouth. Ronalds tried to help as he could, sending money and practical agriculture books. He was elected Examiner across the three Queen's Colleges, which supplemented his income by £100, and became Dean of Science and a member of the Galway College Council.[105]

This same year, 1853, his brother-in-law Tennent became a partner in Charles Tennant & Company and manager of the St Rollox works.[106] Ronalds had the opportunity to move into a much more remunerative role running the Bonnington chemical works. With him having other commitments however, Tennent's brother Hugh Brown Tennent, the assistant manager, cared for the facility until his death two years later.

## CHEMICAL MANUFACTURE (1856-1878)

In March 1856 Ronalds and Barbara were able to leave their home at Nun's Island in Galway and relocate to Bonnington:[107] he had extricated himself from his academic duties, the two *Chemical Technology* volumes were printed, and their new baby was three months old. Tennant, Tennent and Ronalds had all been on the chemical science committee for the BAAS meeting in Glasgow the previous September (with Liebig also being an attendee),[108] which is perhaps where the handover was organised. Ronalds became a partner in the Bonnington Chemical Company, with his contribution being the management of the facility. Tennant and Tennent remained non-active partners, the company being under the Tennant corporate umbrella.[109]

That Ronalds' career change was atypical has been noted by Fox and Guagnini in their discussion of applied science, but without comment on the context.[110] There were many interactions between universities and industry in his education-practice networks outlined earlier, and elsewhere, but it was very rare to swap sec-

---

[98] E. Ronalds to F. Ronalds, 30 March 1858, IET, 1.9.1.
[99] *Glasgow Herald*, 27 December 1850, 2.
[100] *Derby Mercury*, 27 April 1853, 4.
[101] H. Ronalds, Diary, 1851–1854, WU, B1462.
[102] H. Ronalds to M. Ronalds, 14 November 1854, ANL; E. Ronalds to J. Greg, 7 October 1928, Ronalds Family Papers, Sydney, Australia.
[103] *Queen's Colleges (Ireland), Return to an Order of The House of Commons dated 25 May 1857*, p. 22.
[104] H. Ronalds to M. Ronalds, 14 November 1854, H. Ronalds to E. Ronalds, 19 September 1853, ATL.
[105] *Cork Examiner*, 26 June 1854, 2; *Nenagh Guardian*, 29 October 1853, 1.
[106] *One Hundred and Forty Years of the Tennant Companies 1797-1937*, Tennant Companies, London, **1937**, p. 2.
[107] *Galway Mercury*, 15 March 1856, 3.
[108] *Athenaeum* **1855**, 1092.
[109] Bonnington Chemical Company v. Gibson and Walker, 1868, and 1874, NRS, CS242/203, CS242/208.
[110] R. Fox, A. Guagnini, *Hist. Stud. Phys. Biol. Sci.* **1998**, *29*, 55, esp. 75–76.

tors and integrate an academic experience base into the running of an established manufacturing business. Generally in such interactions the academic passed across scientific knowledge while ensuring their distinctive position: "they presented themselves above all as the theorists of industry… without becoming wholly assimilated in the industrial world"; they were the "elite".[111] Indeed, it has been presumed on occasion that Ronalds must have been "a chemist" or "consultant" at Bonnington rather than the managing partner.[112]

His closest university associates adopted comparable approaches, even in Germany with its cameralist links between state, commerce and science. Magnus supported "technology" through university teaching and research in experimental science, by visiting factories and advising government. In enthusiastically promoting industrial application of his research ideas, Liebig provided scientific guidance (often through his assistants), while also seeking commercial returns to supplement his academic income. Knapp aided Liebig in several of these endeavours and held the position of technical director at a government porcelain manufactory for a time – together with his professorship. Richardson's career was the other way round: he specialised in industrial chemistry at several different plants, and after a few years also took an appointment as a lecturer. Another Giessen associate, August Wilhelm Hofmann, Director of the Royal College of Chemistry, proudly associated himself with a further and oft-quoted model of technology transfer – his student William Perkin discovered the coal-tar dye mauveine in 1856; Perkin became what Homburg has called an "inventor-entrepreneur" when he established a factory and entered into production.[113] As a final example, Kranakis has identified academics who melded theory and practice in noteworthy "hybrid careers", but they did so while remaining attached to the university.[114]

Ronalds contrasts with these and other cases in that he moved at top level and permanently from academia to an operating manufacturing firm where he had little first-hand experience, and took responsibility overall rather than for technical aspects. Sharing scientific knowledge was part of his role but the imperative was to quickly acquire quite different skills while building credibility as the manager. Universities and manufacturing facilities were highly disparate entities in this era, which made the transfer demanding and risky. It was only later when industrial companies had research laboratories, universities became businesses, and the class structure changed that advantages could be seen in senior staff cross-fertilisation.[115]

An early ramification of Ronalds' move was an altered standing in the community in comparison with being a professor: he quipped to Sir Francis that he was now "completely ignored, as a tradesman, by the entire society".[116] Fortunately, as outlined below, status was of little concern to him. In the same light-hearted vein, he explained: "I have entirely changed my mode of life & have (with a view to the future of the bairns) taken seriously to money grubbing, an occupation sufficiently disgusting & only tolerable in consideration of the results which I hope may be successful". Like his uncle, he was unaccustomed to the marketing, sales and negotiation side of business and also ill-suited to it with his retiring nature. More importantly, there was a lot to learn about the plant and he admitted (with some self-deprecation) that he had "been kept & am still very hard at work, having hardly had time to master the details of manufacture & trade".

Despite these challenges, he welcomed his new opportunity. Not only could he now better support the Ronalds family, but he was responsible himself for the type of largescale manufacture he had before only written about and could trial ideas suggested by his studies. Barbara would also have enjoyed returning to family and friends in Scotland. It can be surmised however that without the trigger of financial distress he would not have taken on the job and also that its risks would have been too great if he not researched *Chemical Technology* and had the support of his relationship with Tennent. His partners, having studied at university, would also have appreciated that his alternative skillset could bring plant innovations. A career change from the academic to the manufacturing world at that time almost certainly required special circumstances, notwithstanding the potential benefits it brought.

The Bonnington chemical works was located close

---

[111] Fox, Guagnini, *Hist. Stud. Phys. Biol. Sci.* 79; See also: Bud, Roberts, *Science versus Practice*; E. Homburg, *Isis* **2018**, *109*, 565; E. Schatzberg, *Isis* **2012**, *103*, 555; *Technological Development and Science in the Industrial Age: New Perspectives on the Science-Technology Relationship*, (Eds.: P. Kroes, M. Bakker), Kluwer, Dordrecht, **1992**, pp. 1–15.

[112] W. H. Brock, *Ambix* **2013**, *60*, 203; W. H. Brock, *Justus Von Liebig: The Chemical Gatekeeper*, CUP, Cambridge, **1997**, p. 349.

[113] Hofmann, *Allgemeine Deutsche Biographie* ; Brock, *Justus Von Liebig*; Schneider, *Neue Deutsche Biographie* ; "Richardson, Thomas", *Oxford Dictionary of National Biography*; L. F. Haber, *The Chemical Industry during the Nineteenth Century*, OUP, Oxford, **1958**, pp. 80–87; Donnelly, *Social Studies of Science*; E. Homburg, *Br. J. Hist. Sci.* **1992**, *25*, 91.

[114] E. Kranakis in *Technological Development and Science in the Industrial Age*, pp. 177–204.

[115] On when and how manufacturing firms developed research arms, and their links with academia, see for example: Homburg, *Br. J. Hist. Sci.*, and D. A. Hounshell and J. K. Smith, Jr., *Science and Corporate Strategy: Du Pont R&D, 1902–1980*, CUP, Cambridge, **1995**.

[116] E. Ronalds to F. Ronalds, 30 March 1858.

to the Water of Leith on Newhaven Road, Edinburgh. It was a pioneer coal-tar processing facility established around 1822 to distil naphtha from the residues of the Edinburgh gasworks for Charles Macintosh's eponymous waterproof fabrics; Macintosh's firm was a special customer for two decades and probably longer.[117] The plenteous residues were transported from the gasworks to Bonnington by a dedicated pipeline over Calton Hill. In the words of Ronalds' Giessen friend Professor Frederick Penny, the processing works were "so extensive and so important" and were now run by "a distinguished scientific and practical chemist".[118]

Within months of arriving, Ronalds donated a large series of specimens to the Industrial Museum of Scotland showing the numerous intermediate, final and by-products created from gasworks waste.[119] The collection formed a valuable companion to the descriptions and illustrations of coal-tar processing in *Chemical Technology* and was used by the museum director (technology professor Wilson) as a teaching aid. From Ronalds' perspective, by looking outward to support technological education he was seemingly already in command of his role, which indicates both his prior understanding of industry practices and his adaptability.

His detailed summary of plant operations was published in the *Cyclopædia of Useful Arts*.[120] Bonnington's most important products were rectified naphtha, creosote, sal ammoniac (ammonium chloride), ammonium sulphate, and anticlor (sodium thiosulphate). He noted that "we have a good deal of business with the owners of the steamers"[121] exporting these commodities around the world and indeed George Seater, the director of the Leith, Hull & Hamburg Steam Packet Company, christened his son "Edmund Ronalds". He also made all his sulphuric, hydrochloric and sulphurous acid requirements and a new acid plant was the first facility he commissioned. Figure 3 shows the plan of the facility from the 1876 ordnance survey map. Comparing this with the first survey in 1852 indicates the extent of his alterations, with the facility's footprint increasing from two to approaching three acres. One of the motivations for the enhancements he made (including waste-gas capture equipment and a large new chimney) was to reduce emissions, which was an emphasis in *Chemical Technology*. The gamble of his appointment had paid off.

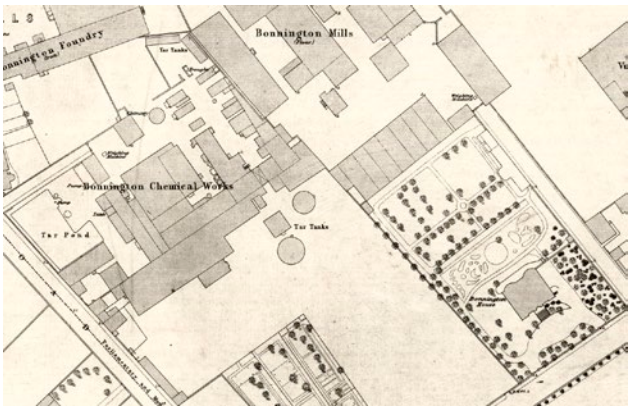



**Figure 3.** Bonnington chemical works near Edinburgh. Bonnington House is at the southeast corner of the overall site. Source: Ordnance Survey, Edinburgh, Sheet 16, 1876, National Library of Scotland.

Ronalds and Richardson had noted in the preface to the second edition of their book that "the valuable constituents of coal-tar [have not] yet been fully worked up into a merchantable form" and the chance to be part of a rapidly developing sector was another inducement to come to Bonnington. His longer-term aim would have been to build on the current efforts of Hofmann and others in fossil fuel chemistry and its applications by conducting in-house research. In the early years he had little time "for prosecuting my chemical enquiries connected with the manufacture which, however, exist in sufficient abundance & would well repay the time expended upon them, could it only be afforded by the more pressing demands of everyday business".[122] Unfortunately details are relatively scant on the science he was able to oversee when circumstances allowed, and how it was utilised in plant operations.

He was however elected a Fellow of the Royal Society of Edinburgh in 1862, proposed by Professor Peter Guthrie Tait,[123] and quickly served on the council. Interested to explore both the composition and handling risks of the light petroleum recently discovered in Pennsylvania in comparison with coal tar, he read a non-proprietary research paper to the society on its volatile components in February 1864. He discovered several lower members of the methane series dissolved in the crude: ethane, propane and butane. He described the proper-

[117] B. F. Ronalds, "Bonnington Chemical Works (1822–1878): Pioneer Coal Tar Company", Submitted. The Bonnington works is not listed in P. J. T. Morris, C A. Russell, *Archives of the British Chemical Industry 1750–1914*, BSHS, Faringdon, **1988**, but considerable archival material has now been identified.

[118] F. Penny, Report to the Provost, Magistrates, & Council of Leith on the Bonnington Chemical Works, 1865, Edinburgh City Archives, E32, MYBN U140G Box 00 01 20.

[119] *Fourth Report of the Department of Science and Art*, pp. 162–63.

[120] C. Tomlinson, *Cyclopædia of Useful Arts*, Vol. 1, James Virtue, London, **1862**, pp. 751–52.

[121] E. Ronalds to F. Ronalds, 19 June 1858.

[122] E. Ronalds to F. Ronalds, 30 March 1858.

[123] Royal Society of Edinburgh, *Biographical Index of Former Fellows of the Royal Society of Edinburgh 1783–2002*, **2006**.

ties of the last, also for the first time[124] – with a specific gravity of 0.600 at zero degrees Celsius, it was the lightest liquid known and it began boiling at that temperature. The paper was included in the society's transactions, reprinted in the Chemical Society's journal and the German literature, and was referenced numerous times as petroleum research progressed.[125] He presented product samples to the industrial museum in Edinburgh.[126]

Another aim was to investigate the properties of the pyridine series, which were very minor constituents of coal tar. He prepared a significant quantity of these bases by repeated fractionation but, perhaps due to time constraints, he then gave the various fractions to James Dewar. Dewar's analyses of this "liberal supply" enabled him publish the proposal that pyridine had a ring formula.[127]

Ronalds was able to determine that the tar he received from the gasworks contained almost no anthracene, and its relatively little benzene was often uneconomic to separate from the methane series of compounds also in the naphtha. This precluded him from contributing to the new synthetic dyestuff industry that was commencing to manufacture the dyes alizarin and mauveine from these components following Perkin's discovery. He also ascertained how the detailed properties of his coal tar varied with the coal mix and retort temperature being used at the gasworks. When Bonnington was closed he provided the results for Lunge's respected treatise on coal-tar processing.[128] These examples suggest that Ronalds had succeeded in building up advanced research capability, with experimental apparatus that was unusually sophisticated for a manufacturing environment.

In the meantime, his brothers and sisters in New Zealand had become embroiled in the Maori Wars in 1860 and their timber cottage and farm were destroyed. He encouraged Hugh, the most despondent and unsettled of the siblings, to return to Britain[129] and he became a partner in the firm in 1867 after Ronalds had trained him in the business.[130] Ronalds' eldest son Edmund Hugh later became an assistant chemist.[131] His other sons, christened Tennent and Frank, became respectively a fellow of the Edinburgh Obstetrical Society[132] and a merchant. His daughters attended the respected Rowdon House school for ladies in London until their late teens, continuing the family's emphasis on education.[133]

## LAST YEARS (1878-1889)

The Bonnington chemical works closed in 1878. Tennent and Tennant were dead, Ronalds had been "afflicted with very bad health" for some years that a spell on the North Berwick coast did not alleviate,[134] and his family members did not wish to take on the management responsibility. Since 1868 he had lived in the "beautiful" Bonnington House (Figure 4) with large ornamental gardens close to the works.[135] It and several smaller houses had been purchased by the chemical company before he joined and now became his personal property.[136] Hugh lived nearby at another "good house" called Hillhousefield.

This part of the family had become very wealthy – Ronalds had assets to the value of £136,000, exclusive of his recent real estate acquisition.[137] In addition to his portion of Bonnington's worth over two decades, Barbara and the children had been the major beneficiary of her brother Tennent's estate, which included his £54,000 share of St Rollox, an £8,300 contribution from Bonnington, plus real estate.[138] Hugh had married into Samuel Greg's family, renowned for their large cotton spinning mills. Ronalds repaid his good fortune by continuing to support other siblings through trust funds.

He occupied his last years in an "admirably appointed laboratory" he established,[139] denoting himself as a "scientific chemist".[140] It was a lifelong goal to pursue science of interest in a private facility in the mould of Magnus' teaching and research laboratory in Ber-

---

[124] Ronalds' discoveries are noted, for example, in: Watts, *Dictionary of Chemistry*, Vol. 4, p. 385; H. E. Roscoe, C. Schorlemmer, *Treatise on Chemistry*, Vol. 3, Macmillan, London, **1881**, pp. 144–45; W. T. Brannt, *Petroleum*, Henry Carey Baird, Philadelphia, **1895**, pp. 56–80; C. F. Maybery, *Proc. Am. Acad. Arts Sci.* **1896**, *31*, 1.

[125] E. Ronalds, *Trans. R. Soc. Edinburgh* **1864**, *23*, 491. Also in: *J. Chem. Soc.* **1865**, *18*, 54; *J. Prakt. Chem* **1865**, *94*, 420.

[126] Edinburgh Museum of Science and Art, *Catalogue of Industrial Department*, Neill, Edinburgh, **1869**, p. 94.

[127] J. Dewar, *Trans. R. Soc. Edinburgh* **1872**, *26*, 189.

[128] G. Lunge, *Treatise on the Distillation of Coal-tar and Ammoniacal Liquor*, John van Voorst, London, **1882**, pp. 12–13.

[129] H. Ronalds to E. Ronalds, 25 September 1860, ATL.

[130] *Inquirer* **1911**, 821.

[131] Scotland Census, 1881.

[132] *Trans. Edinburgh Obstetrical Society* **1888–1889**, *14*, xiii.

[133] Barbara, Eliza and Emily Ronalds, England Census, 1871.

[134] *Proc. Inst. Chem*. **1890**, *14*, 53.

[135] Property descriptions are in the Midlothian Ordnance Survey Name Books 1852–1853, ScotlandsPlaces, OS1/11/87.

[136] Ground Belonging to the Trustees of the Late Dr. Ronalds, Bonnington, Historic Environment Scotland, EDD 804/1–3; Valuation Rolls, 1885–1886, ScotlandsPeople, VR005500031-/386–387.

[137] Edmund Ronalds, Inventory, 1889, ScotlandsPeople, SC70/1/278.

[138] John Tennent, Will and Testament, 1867, ScotlandsPeople, SC36/48/58, SC36/51/52.

[139] *Proc. R. Soc. Edinburgh* **1889–1890**, *17*, xxviii.

[140] Scotland Census, 1881. In the 1861 and 1871 Censuses Ronalds called himself a "manufacturing chemist" and a "chemist and manufacturer", respectively.

**Figure 4.** Bonnington House, Ronalds' home in the period 1868-1889. Source: J. Grant, *Cassell's Old and new Edinburgh: Its History, its People, and its Places*, Vol. 3, Cassell, Petter, Galpin, London, **1887**, p. 93.

lin and the well-equipped workshop that Sir Francis set up at each of his homes.[141] He had come full circle and was now enjoying the life he had imagined he would lead when he was studying. "[H]e made any chemist welcome"[142] in the laboratory and, according to his obituary, was well known and remembered "with affection" by all chemists who had resided in Edinburgh. Little is known of the work conducted there, although his son's analyses aided George Beilby in the production of ammonia from shale and coal.[143]

In 1875 he was appointed a foundation trustee of the Ronalds Library at the Institution of Electrical Engineers bequeathed by Sir Francis – he had always cherished his copy of Sir Francis' 1823 booklet describing his telegraph.[144] He joined the new Society of Chemical Industry, became a Fellow of the Institute of Chemistry of Great Britain and Ireland when it was formed,[145] and along with other former professors was awarded the honorary D.Sc. degree by the Queen's University of Ireland in 1882.[146]

"He was a constant attendant at the meetings" of the Royal Society of Edinburgh and "always took a lively interest in everything", "although he rarely took an active part in its proceedings".[147] Similarly, when Tait invited him to help found a new learned club, he replied that he would be "be delighted to join if Smoking & good listening without much talk will qualify".[148] Like Sir Francis and other members of his Unitarian family, he was an introvert, with no interest in status or recognition and avoiding public roles. He was motivated in his work simply by the personal knowledge of achieving scientific and technical goals. There are therefore few institutional records of his contributions and this, together with his small portfolio of academic papers, helps to explain his comparative absence in the history of science literature.

Never recovering his health, he died on 9 September 1889 and was buried in Rosebank Cemetery diagonally opposite Bonnington House.[149]

## CONCLUSION

Ronalds had a highly advantageous entry to his lifelong field of chemistry through his international education and initial work experience and he brought significant talent and energy to his subsequent career. From his relative obscurity today it could be construed that he did not fulfil this early promise. He has been categorised in studies of the students of Liebig (and Bunsen) as an academic and his traditional metrics of science output are not strong.[150] He himself found that his focus on industrial interests equated in Britain to a lowered status, and even today dual academic and industrial achievement is not commonly embraced and quantified, despite the culture of science utilisation these students were exposed to in Germany.

Ronalds in fact had an unconventional two-stage career, spending fourteen years in academia and then twenty-two years in the quite different setting of large-scale manufacture. His change was abrupt but cogent because he always linked scientific insight and industry practice. As an academic, his research and teaching addressed local problems and facilitated the study of chemical technology through a seminal book that synthesised theory and application. He then put his advanced knowledge of technology into practice while also bringing research into a manufacturing firm, and

[141] C. Jungnickel, R. McCormmach, *Mastery of Nature: The Torch of Mathematics 1800–1870*, UCP, Chicago, **1986**, pp. 107–10; Ronalds, *Sir Francis Ronalds*, p. 95.
[142] *J. Chem. Soc. Trans.* **1890**, *57*, 456.
[143] G. Beilby, *J. Soc. Arts 33* (**1885**): 313; also *J. Soc. Chem. Ind.* **1884**, *3*, 216.
[144] Trust Deed of the Ronalds Library, 1875, IET; E. Ronalds to J. Fahie, 26 April 1882, IET, 1.9.2.119.
[145] *Proc. Inst. Chem.* **1878**, *2*, 13.
[146] *Belfast Newsletter*, 2 February 1882, 8.
[147] *Proc. R. Soc. Edinburgh* **1889–1890**, *17*, xxviii.
[148] E. Ronalds to P. Tait, 25 October 1869, National Library of Scotland, Archives & Manuscript Collections, MS.1704 f.74 v1.
[149] C. Napier, *Scottish Genealogist* **2012**, *59*, 176.
[150] For example: J. S. Fruton, *Proc. Am. Philos. Soc.* **1988**, *132*, 1; Brock, *Ambix*.

this resulted in new discoveries, plant improvements, business expansion and significant profits. These two cross-sector unions – technological education and industrial research – were then very novel but presaged what became key trends into the twentieth century, yet his accomplishments have been largely overlooked by historians.

Finito di stampare da
Logo s.r.l. – Borgoricco (PD) – Italia

March 2019 No walls. Just bridges

# Substantia

An International Journal of the History of Chemistry

Vol. 3 – n. 1

## Table of contents