March 2018
Vol. 2 - 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. 2, n. 1 - March 2018

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: **Charles Beck "Winter Tree" Woodcut, 1985.**
Charles Beck was born in Fergus Falls, Minnesota, USA, in 1923 and died there in 2017 at the age of 94. Fergus Falls is a small town in western Minnesota near the Red River Valley. The landscape consists of rolling hills, lakes, and country farms with diverse flora and fauna. The region was settled to a large extent by Scandinavian immigrants in the 19th century. Besides stints in the U.S. Navy at the end of WWII and the University of Iowa, Charles lived his whole life in Fergus Falls. To his relatives, it seemed completely normal to be a working artist in a small rural town, but of course this was not the case. Through his life's work, Charles became a highly recognized regional artist whose work at the boundary of realism and abstraction captured the natural world in this part of the country. His paintings and woodcuts are housed in homes and galleries around the world. One early painting was exhibited in the Metropolitan Museum in New York. In spite of his rising fame later in life, he maintained low prices for his works so that average people could enjoy his art in their homes. In this way he became part of the cultural fabric of the region. Remarkably, he continued his work during the last two years of his life, producing around a hundred new paintings while disabled in a local nursing home. He attended an opening of his new works in a local art gallery two months before he died. His art utilizes natural materials (wood, paints, paper) and metallic hand tools, and the paintings display complexity emerging from simplicity as seen throughout the chemical and broader natural worlds.

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.
We proudly welcome our new Associate Editors: Carin Berkowitz, Neil R. Cameron, Stephen Hyde and Ernst Kenndler.

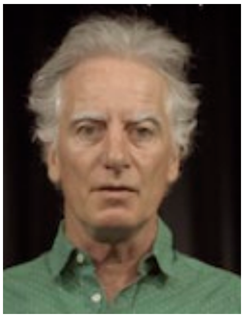
EDITOR-IN-CHIEF

Pierandrea Lo Nostro
Department of Chemistry "Ugo Schiff"
University of Florence, Italy
phone: (+39) 055 457-3010
email: substantia@unifi.it - pierandrea.lonostro@unifi.it

ASSOCIATE EDITORS

Carin Berkowitz
Chemical Heritage Foundation, USA

Neil R. Cameron
Monash University, Australia
University of Warwick, UK



Stephen Hyde
Australian National University, Australia

Ernst Kenndler
University of Vienna, Austria

SCIENTIFIC BOARD

as of 18 November 2016

EDITORIAL BOARD

MANAGING EDITOR

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Editorial

Why Chemists Need Philosophy, History, and Ethics

Since many years national and international science organizations have recommended the inclusion of philosophy, history, and ethics courses in science curricula at universities. Chemists may rightly ask: What is that good for? Don't primary and secondary school provide enough general education such that universities can focus on chemistry alone? Is that only a conservative call back to an antiquated form of higher education? Or do they want us to learn some "soft skills" that can at best improve our eloquence at the dinner table but is entirely useless in our chemical work?

The answers depend on what you understand by chemistry, philosophy, history, and ethics. Let's begin with chemistry.

If the prototypical chemist were somebody who secludes himself in his laboratory, ponders on some self-imposed questions, and once in a while comes up with an idea to impress his colleagues, there would perhaps be little need. However, modern chemical research is a highly connected activity, conducted in teamwork that is typically interdisciplinary. It is project-based, that is, it seeks a solution to a problem that the scientific community or society at large, or both, consider important, and which mostly aims at the improvement of material conditions of life. The results are likely to have an impact on future research and the technological world we live in.

Similarly, if chemical research consisted in following simple routines, in doing some minor modification here and there to produce easily predictable results, there would be little need either. However, scientific research results are expected to be novel in the proper sense, i.e. they cannot be predicted, derived, automatically produced, or bought with grant money, contrary to the expectations of many science policy makers and managers, and unlike the usual rhetoric of grant proposals. Such creative, and even more so groundbreaking, work requires questioning the received wisdom, what is taken for granted in science at the moment. Thus, if you want to be a successful chemist, you cannot just apply what you have learned in your chemistry class. On the contrary, you must be able to challenge exactly what has been taught to you to be the edifice of science, and take it only as a provisional state in the course of the ongoing research process of which your work is meant to become a part.

Next let's see what kind of philosophy, history, and ethics is needed for chemical research, and what not.

If philosophy of science were the marveling at theories from physics, as the popularization of physics has long articulated it, it would be of little use for chemistry. Indeed, most chemists have a much better understanding of the benefits and limits of quantum mechanics in their own field. However, philosophy is a way of asking questions about what is taken for granted but badly understood, and it usually aims at a better understanding. Such as science (which historically emerged out of philosophy) asks question about nature that don't bother ordinary people, so does philosophy of science asks question about science that scientists once stopped asking. What are the goals of science? How do scientists develop and establish knowledge? How is their knowledge organized, on which presuppositions does it depend? Which fundamental concepts do they use and how can those be defined? And so on. While the professional philosopher has learned to take any intellectual edifice apart within minutes, a little training in philosophy helps scientists to raise the right questions at their research frontiers where the received taking-for-granted view is just deadlocked.

Moreover, many chemists are inclined to say that everything is chemistry, as do many physicists, biologist, engineers, mathematicians, etc., each for their own discipline, claiming a privileged access to the world. However all such disciplinary chauvinism is not only built on ignorance of the diversity of modern science, it is also poisonous to any interdisciplinary teamwork. Only if you take your own chemical (or physical, biological, engineering, etc.) way of asking questions and solving problems no longer for granted but understand its disciplinary peculiarities that might essentially differ from that of other, equally acceptable, disciplines, you will be prepared for interdisciplinary teamwork. Abstract as it

Substantia. An International Journal of the History of Chemistry 2(1): 5-6, 2018
ISSN 1827-9635 (print) | ISSN 1827-9643 (online) | DOI: 10.13128/substantia-36

is, the philosophical understanding of different disciplinary approaches helps break the interdisciplinary barriers of conceptual misunderstandings and lack of mutual appreciation, which is more needed than ever. As a side-effect, you will understand your own field, chemistry, much better if you are able to look upon it from the outside, such as you understand your own culture much better once you have spent some time abroad.

If history consisted in setting up a fact sheet of who-did-first-what-and-when, that would not help either, other than to commemorate the ancestors and give them due credit. However, professional historians of science try to understand the scientists of the past from their own perspective, how they saw the world, what goals, beliefs, and methods they had, and in which social and cultural context they worked. Because all that frequently differs considerably from our present perspective, history trains our capacities of thinking science differently, exactly what the creative mind needs as a starter. Furthermore, history turns the static textbook view of the scientific edifice into a processual view of scientific evolvement, with all its entwined paths, dead ends, and prematurely given-up alternatives. History thus teaches you to understand science as a complex process, which you need in order to make creative and convincing contributions to it in the presence. Moreover, only by looking at chemistry in its social and cultural context and their interactions over time, you understand what chemistry means in a broader sense, what role it plays in society, what societal expectations, hopes, and fears it raises.

If ethics were a form of moral indoctrination, of making people comply with fixed rules, we would better do without. However, ethics, one of the oldest philosophical disciplines, is a technique of abstract reasoning about norms and values, of balancing different values, and of building moral arguments that try to justify why this is better than that. Chemists who are engaged in research projects that aim to improve the material conditions of life, must be able to understand all ethically relevant aspects of their work and to develop moral justifications for what they do – if they really aim at improvement in the full sense. They must so in three different senses. First, they are morally obliged to do so because they will be held accountable for all possible adverse effects of their research work. Second, because all technological innovations transcend traditional life forms, their moral assessment cannot simply follow traditional norms tailored to ordinary life contexts. Instead for each possible innovation we have to develop moral deliberations anew, which of course requires being acquainted with the tools of moral reasoning. Finally, if the goal of chemical research is material improvement according to general values, chemists can only be successful if they know all these values and are able to connect them in a balanced way to their research projects. Chemical success thus depends as much on ethical competence as on chemical knowledge.

In sum, education in philosophy, history, and ethics, each rightly understood, helps improve chemistry by making it more creative, more open to teamwork, and more aware of the social and ethical contexts that partly define it. It is therefore no additional luxury but in the self-interest of chemistry as a science to open itself to these fields. That has already been done in many countries, albeit mostly upon the request of accreditation agencies or governments, because society needs a stronger chemistry for the solution of many of its current problems. For the same reason, a journal like *Substantia* that aims to broaden the chemical horizon is particularly important and welcome.

Joachim Schummer

Editor-in-chief of *HYLE: International Journal for Philosophy of Chemistry*
(www.hyle.org), js@hyle.org

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

OPEN ACCESS

Citation: S.E. Friberg (2018) Emulsion Thermodynamics – In from the Cold. *Substantia* 2(1): 7-16. doi: 10.13128/substantia-37

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.

Feature Article

Emulsion Thermodynamics – In from the Cold

Stig E. Friberg

Ugelstad Laboratory, NTNU, Trondheim, Norway
Email: stic30kan@gmail.com

Abstract. Thermodynamics has played virtually no role in traditional emulsion research, because emulsions are inherently thermodynamically unstable. The problem with commercial emulsions needing to exist with none or only small changes during use and the industrial stability problem was resolved by formulating *colloidally stable* emulsions, i.e. the *rate* of destabilization was reduced. This approach was successful for single-oil emulsions, but encountered problems for double emulsions, for which the simultaneous stabilization of several interfaces within one drop encountered difficulties. Naturally, even for such an emulsion, colloid stability is the only option to stabilize the outer surface towards the continuous phase. In fact, the destabilization by flocculation/coalescence proceeds similarly to a single-oil emulsion. But experiments have demonstrated that complex emulsions with a thermodynamically stabilized inner interface *retain* the individual drop topology during the process. This result opens an avenue to significantly facilitate the formulation of a group of commercially important emulsions, because the cumbersome multiple emulsion stabilization is reduced to the more trivial single-oil emulsion case.

Keywords. Emulsion stability, colloidal stability, interfacial thermodynamics, janus emulsions, emulsion coalescence.

INTRODUCTION

The first meal humans and many animals enjoy is an oral *emulsion* that replaces the nutrients through the umbilical cord. The emulsion in question is a convenient and efficient means to provide the needed mixture of fats, sugars and proteins, which otherwise would be solid and not useful to a toothless newborn. Since these emulsions are truly essential, baby milk emulsions have been commercially available for 150 years for the cases, where the mother does not have milk, or does not want to breast feed. The initial modest introduction by Henry Nestlé has grown to a company with almost $100 billion in yearly sales, reflecting the commercial expansion to other food emulsions, such as cream, butter and ice cream, to mention the most obvious (Among the miniscule, but common, applications of milk is to remove the bitter taste from coffee - and in some societies from tea - an insult to the refined Oriental cuisine culture). In addition, emulsions also have a long history in personal care and cosmetics with Cleopatra taking baths in donkey milk about 2000 years ago.

ISSN 1827-9635 (print) | ISSN 1827-9643 (online) | DOI: 10.13128/substantia-37

This pioneering use of emulsions has developed into a Schueller/Bettencourt company, L'Oreal (actually partly owned by Nestlé), which has grown from a modest start in 1909 to an internationally large company with sales at the level of tens of billions of dollars. These are examples of large applications of emulsions for humans, like emulsions in the paint and coatings industry, which add another huge sum.

All these tremendous volumes of products, which are daily handled all over the world, are, in fact, thermodynamically unstable compositions, doomed to final separation of the components. Nonetheless, the compositions are required to remain virtually unchanged for specific times, varying from seconds to years, depending on the actual purpose. Hence, an extensive volume of research[1-3] is found on their properties and especially their "stability". This "stability", is defined as an unchanged appearance reflecting the commercial needs. A more quantitative measure of emulsion stability would be its half-life, e.g. the time for the number of drops to be reduced to one half by coalescence, Figure 1.

As a contrast, the so called microemulsions are not emulsions; they are equilibrium colloidal *solutions*[4,5] and have been investigated for different aspects of their interfacial properties.[6-8] Because they are thermodynamically stable, their preparation is completely different from that of emulsions. These latter are prepared, in most cases, by mechanically dividing a liquid into drops and dispersing these in a continuous liquid, while mircroemulsions are spontaneously formed,[5] requiring only the most minute mixing for large volumes. The microemulsions are of considerable commercial interest, (reflected in about half a million Google hits) but are not the subject of the present contribution, which is focused on emulsions.

Their comprehensive and advanced treatments[1-3], in turn, examine emulsion stabilization from the viewpoint of *colloidal stability*, applying the concepts from the DLVO theory,[9-11] in which a *repulsive force* is introduced between drops to reduce *the rate* of flocculation and coalescence. This action naturally does not make the emulsions thermodynamically stable, but tender serviceable information to retain the properties of traditional emulsions for commercial purposes. This colloidal stability approach was also employed for double emulsions,[12] whose more complex topology caused challenging stability problems.[13,14]

In summary, except for in the area of microemulsions, there was no significant need, nor even a role, for thermodynamics in emulsion research; a mindset that has recently experienced an initial conversion, when Janus emulsions were introduced.[15,16] These emulsions consist of combined drops of two mutually insoluble oils, Figure 2, in which total free interfacial energy of the single drop is significantly less than that of the oils in separate drops. Hence, their "inner" topology, Figure 2, is thermodynamically stabilized,[17-19] and actually has been shown to survive most of the flocculation/coalescence process.[20]

A brief comment at this stage is necessary to avoid misinterpretation. Janus emulsions are *not* thermodynamically stable, like microemulsions. The Janus emulsions, as a contrast, undergo a flocculation/coalescence process and, *if left for sufficiently long time, will separate into three, or four liquid layers, depending on the relative density,* because of gravitational forces. Nevertheless, the feature mentioned in the last part of the former paragraph is significant, because it brings to light a new approach to formulating emulsions with two-oil drops.

Traditionally, commercial double emulsions are prepared either by first forming an oil/oil emulsion followed

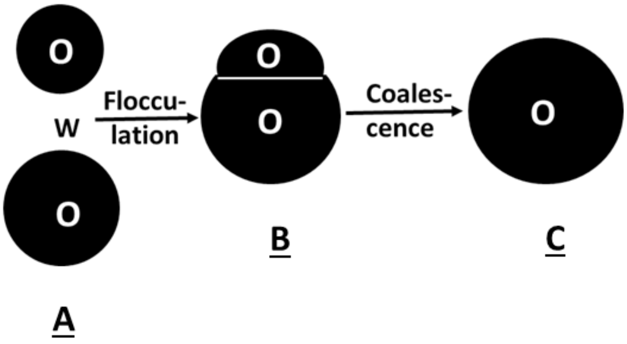


Figure 1. Two emulsion drops of one oil, A, are colloidally stabilized and, when they aggregate (Flocculation, A to B), they form an aggregated drop of transitory stability, B. The transient interface dividing them (white line, B) is destabilized and coalescence (B to C) occurs to a single drop, C.

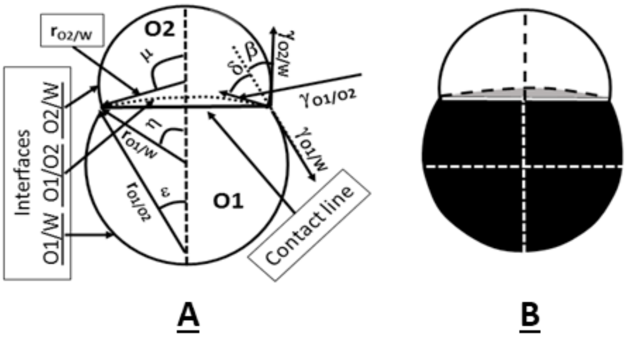


Figure 2. A. The geometry and interfacial tensions in a Janus drop. B. Black, O1 volume limited by the contact line, White, O2 volume limited by the contact line, Grey, O1 volume penetrating the O2 space from the contact line.

by its emulsification in an aqueous phase (O1/O2)/W or by emulsifying one oil in water, followed by emulsification of this emulsion into the second oil plus a final emulsification into an aqueous phase (O1/W/O2/W). All the preparations necessitate *independently* stabilizing the different interfaces by colloid stabilization; a methodology, beset with a main problem. The transfer of surfactant stabilizers between the liquid phases, which would markedly reduce their stabilizing action, was successfully overcome for individual emulsions after a substantial amount of outstanding research.[13,14] Conversely, in the alternative methodology using thermodynamic stabilization, the stabilizer is allowed freely to equilibrate between phases, as long as *the level of surface tensions are in the order* defined in later sections. With this approach, colloidal stabilization is limited to the outer interface towards the aqueous phase, a simpler problem, which has been extensively treated.[1-3]

In summary, the focus of the attention is shifted from colloidal stabilization, a kinetic phenomenon, to interfacial tensions, a quantifiable equilibrium entity. The transition from one preparation method to the other requires a change in the mind set and the present contribution is an attempt partially to outline the relevant thermodynamic framework.

THERMODYNAMIC STABILIZATION AND INTERFACIAL TENSIONS

The basis for the thermodynamic framework is the equilibrium of three tensions in one plane, Figure 3. These tensions act at the contact line, Figure 2, and, combined with the relative volume of the two oil lobes, determine the equilibrium topology of the combination drops. As a result, they are central to this contribution and a summary of their ramifications is useful for the continued analysis with appropriate limitation. The features of Figure 3 are relevant only for a case in which $\gamma_{O1/W} < \gamma_{O2/W} + \gamma_{O2/O1}$ and $\gamma_{O1/W} > \gamma_{O2/W}$ and $\gamma_{O1/W} > \gamma_{O2/O1}$.

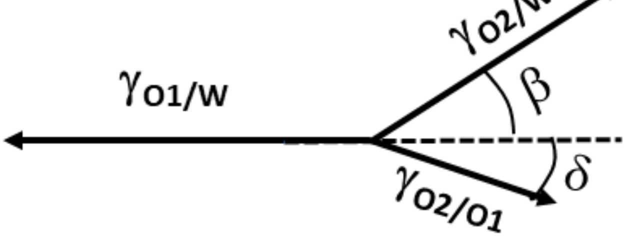


Figure 3. Angles β and δ for three equilibrium tensions in one plane.

An analysis of the consequences from inequalities, outside the ones mentioned, is of algebraic interest, but offer no contribution of value to the problem at hand.

Instead, the main theme of this examination is to survey the thermodynamic effect on the topology of Janus and double emulsion drops and it is convenient to divide the range of tensions into two parts. The first part of the range covers $\gamma_{O1/W} > \gamma_{O2/W} + \gamma_{O2/O1}$ (relevant for double emulsion drops), followed by $\gamma_{O1/W} < \gamma_{O2/W} + \gamma_{O2/O1}$, defining the equilibrium in a Janus drop, Figure 2. In addition, the intermediate case of $\gamma_{O1/W} = \gamma_{O2/W} + \gamma_{O2/O1}$ has some features of interest, which will be briefly mentioned. The stipulation $\gamma_{O1/W} > \gamma_{O2/W} + \gamma_{O2/O1}$ portends a non-equilibrium *spreading* of O2 or O1 on W, while $\gamma_{O1/W} < \gamma_{O2/W} + g_{O2/O1}$ means a stable *equilibrium* with defined angles β and δ, Figure 2 and 3. For these, one has, with the ratios $\gamma_{O2/W}/\gamma_{O1/W}$ = a and $\gamma_{O2/O1}/\gamma_{O1/W}$ = b, equations [1] and [2].

$$\beta = \operatorname{acos}((1 + a^2 - b^2)/2a) \qquad [1]$$

$$\delta = \operatorname{acos}((1 - a^2 + b^2)/2b) \qquad [2]$$

It should be observed that these equations are applicable only to the equilibria in Figure 3, and may be used to calculate the interfacial free energy as such for a selected drop topology, but the calculations fail to identify the thermodynamically *preferred* topology. This topology is obtained first, when the free energy quantity is contrasted with that of a counterpart. The free energy number per se actually implies a counterpart with *no interfacial free energy* and the approach would show *any* selected topology to be thermodynamically disfavored and is of no use to judge thermodynamic stability. In addition, the free energy of two-oil emulsion drops also depends on the volume ratios of the two oils and an equitable, but injudicious, choice of a reference topology will result in erroneous conclusions. This fact will later be brought to light.

Equations [1] and [2] relate the angles β and δ to tensions at equilibrium, but not in an explicitly illustrative manner and a few numbers from a model system are informative as a graphic. Figure 4 shows the limitations of the angles β and δ versus the $\gamma_{O2/O1}/\gamma_{O1/W}$ (b) with $\gamma_{O2/W}/\gamma_{O1/W}$ (a) as parameter. The range of the two variables in the figure is limited in order to reflect the conditions in Figure 3. So, are numbers for b > 1 excluded, because they would represent a reorganization of the angles in the figure. The δ limit for a =1 varies as δ = 60 + 30(1 - b) (degrees) for the same reason.

The numbers for the d angle shows the development of the equilibrium/spreading border, Figure 4 C,

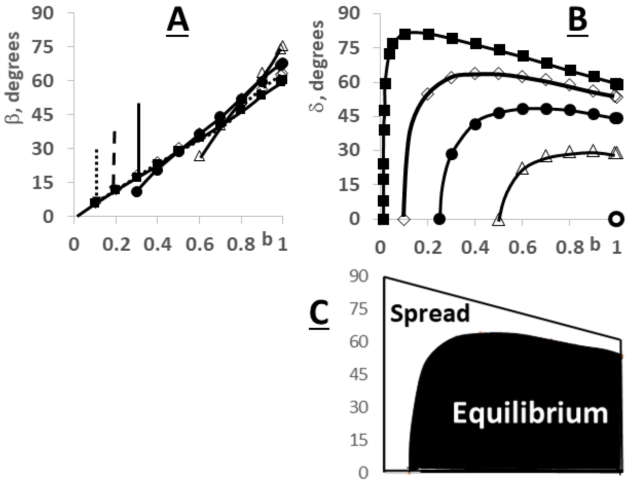


Figure 4. Angles β (A) and δ (B) and areas of equilibrium and spreading (C) versus the ratio $\gamma_{O2/O1}/\gamma_{O1/W}$ (b). The ratio $\gamma_{O2/W}/\gamma_{O1/W}$ (a) is the parameter with the following numbers. ■, 0.99; ◇, 0.90; ●, 0.75; △, 0.5. The minimum b number for each a is marked with an arrow; ◇, 0.90, dotted; ●, 0.75, dashed; △, 0.5, full line. For ○, $g_{O1/W}$ implicitly equals unity.

with varied $\gamma_{O2/W}/\gamma_{O1/W}$ (a). The artificial δ value for a = 0 (The oils are mutually completely soluble) is a single point with δ = 0, because, since $\gamma_{O2/W} = 0$ and $\gamma_{O2/O1} = \gamma_{O1/W}$, the number for angle b becomes irrelevant. In fact, the entire area (except b = 1) denotes spreading. Increasing the $\gamma_{O2/W}/\gamma_{O1/W}$ (a) values from zero, results in an expansion of the equilibrium part of the area, reaching the entire area for a = 1. Any number for a < 1, however close, e.g. 1 - ε (epsilon small) means a 2ε wide area of spreading along the b = 0 axis and ε broad along the axis for maximum δ.

After this brief review of the ramifications for the conditions in Figure 3, the following sections analyze the interfacial free energy of a single *drop*, omitting the emulsion inter-drop dynamics. The analysis of the double emulsion drop is built on the inequality $\gamma_{O1/W} > \gamma_{O2/W} + \gamma_{O2/O1}$, and the spreading means that the O2/W interface does not exist. The drop instead consists of a larger drop O1/W with an O2/O1 drop inside, in accordance with the thermodynamic condition. Conversely, the O/W interface is thermodynamically unstable for virtually all positive $\gamma_{O1/W}$ and an assembly of such drops will coalesce like the simpler single-oil emulsions.

Hence, the potential thermodynamic stabilization is limited to the *inner* interface of the drop, while the initial coalescence is exclusively concerned with its colloidally stabilized outer surface, as has been shown for Janus emulsions.[20] Hence, at a first glance, the thermodynamic stabilization of the inner interface may be considered only of minor importance, but the extensive research on double emulsions[12-14] indicates otherwise. In fact, with the "inner" interface of Janus or double emulsion drops thermodynamically stabilized, the only stabilization needed for a commercial double emulsion would be for the interface towards the continuous phase; i.e. a problem, that has been solved using the colloidal stability approach.[1-3] The only requirement for the O1/W stabilizers is that the interfacial tensions obey the stated inequality; a non-specific condition.

Hence, considering the future potential for a new line of formulations and the fact that virtually no information exists, a review of the thermodynamics of both a double emulsion and a Janus drop has merits. The examination is initiated at a double emulsion drop, because of the extensive technical and commercial relevance for such emulsions.[12-14]

DOUBLE EMULSION DROP

Double emulsions have wide use within a number of industries and technologies, as described in a recent and comprehensive review.[12] The extensive and high-quality research within the area[13,14] has applied the colloidal stability approach, illustrating the enhanced difficulties, compared to those of simple emulsions. The problem arises, because of the fact that two interfaces have to be *independently* stabilized.[13,14] A surfactant, acting effectively on the O/W interface, has to be prevented from diffusing towards the O/O interface and vice versa, reducing its stabilizing action.

However, the results of recent studies of the destabilization of Janus emulsions[20] have demonstrated a substantial effect of interfacial thermodynamic factors to modify the coalescence process of these emulsions. In fact, the Janus topology was retained during coalescence until the very last stages.[20] No such experimental results have been reported for double emulsions, but the potential for such an outcome is estimated sufficiently positive to justify a section on their interfacial thermodynamics in the present publication. Hence, the interfacial thermodynamics of a double emulsion drop is examined to outline its prospective stabilizing effect, with a view towards the effect found for Janus emulsions.

The interfacial free energy basis for a double emulsion drop is the inequality $\gamma_{O1/W} > \gamma_{O2/W} + \gamma_{O2/O1}$, which, as mentioned, shows non-equilibrium *spreading* with an appealing application of this condition to estimate the thermodynamic stability of a double emulsion drop. However, this kind of interpretation has to be made with caution, because the term thermodynamic stability is defined only against a specific counterpart. A seemingly attractive such system is two separate drops of the single oils, but leads to thermodynamic contradictions, showing this stabilization outside the mentioned inequality. Nonetheless, such a choice per se has distinct algebraic interest, combining a well delineated interfacial free energy and an easily comprehended connection to the destabilization of physical emulsions. The following evaluation focuses on the overall free energy difference between a double emulsion drop and two single-oil drops. The volume *fraction* of oil 2 in the drop is v_{O2} and the interfacial free energy of the drops are offered in Table 1.

Table 1. Interfacial free energy of double emulsion drops.

Configuration, drops	IFE, O1	IFE, O2
Double drop	$4\pi(0.75/\pi)^{(2/3)}\gamma_{O1/W}$	$4\pi(0.75v_{O2}/\pi)^{(2/3)}\gamma_{O2/O1}$
Separate drops	$4\pi(0.75(1 -v_{O2})/\pi)^{(2/3)}\gamma_{O1/W}$	$4\pi(0.75v_{O2}/\pi)^{(2/3)}\gamma_{O2/W}$

Regrettably, there is no direct algebraic expression for the relationship between free energy and volumes for the calculation.[18,21] Instead a realistic example was selected to illustrate the variation in free energy during a coalescence process. The particular emulsion consists of $1.09{\cdot}10^{9}$ internally thermodynamically stabilized double emulsion drops, each with a volume of $4.188{\cdot}10^{-15}$ m^3. The two oils O1 and O2, with interfacial tensions 0.004N/m ($\gamma_{O1/W}$), 0.00262 N/m ($\gamma_{O2/W}$) and 0.00116 ($\gamma_{O1/O2}$) each occupy one half. The drops are coalesced, two and two, 30 times, leaving only one final drop with a volume of $4.05{\cdot}10^{-6}$ m^3 and retained topology. The coalescence covers the free energy change due to interface size increase of both the outer sphere and the inner one, of which the former contributes a majority of the free energy reduction. As is obvious and expected, the emulsion interfacial free energy is exponentially reduced, Figure 5, during coalescence; a reflection of the thermodynamic overall instability of emulsions; even when the drops contain more than one interface.

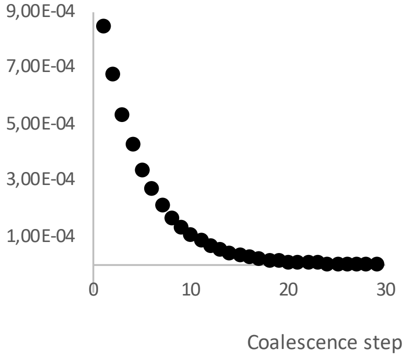


Figure 5. The difference in interfacial free energy (See text).

The overall reduction in interfacial free energy is certainly expected, but a more essential issue is a comparison of the free energy change, when two double emulsion drops coalesce into one double emulsion drop or to two single oil drops, O1 and O2, during the coalescence. Figure 6 depicts this difference between interfacial free energy change from two Janus drops, when a double emulsion drop, ●, or two separate drops, △, form.

The results in Figures 5 and 6 are unequivocal; the coalescence to a single double emulsion drop implies an expected *reduction* of free energy, while the alternative means an *increase.* These results are remarkable and conclusions would unquestionably be tempting; both about the general validity of the result and vis-à-vis the technical and commercial effects. Nevertheless, such inferences would be premature at this stage, because a more complex geometry after coalescence will, in some cases, lead to a modified result. Even so, the results encourage future experimental and numerical evaluations. Leaving that aspect temporarily aside, even the more fundamental aspects offer some unexpected results, illustrated by the ratio, $R_{C/D}$ between the interfacial free energy of the combination drop and that of the two separate drops, equations [3] and [4]. v_{O2} is fraction of O2 volume.

$$R_{C/D} = IFE_{comb\ dr}/IFE_{sep\ dr} \quad [3]$$

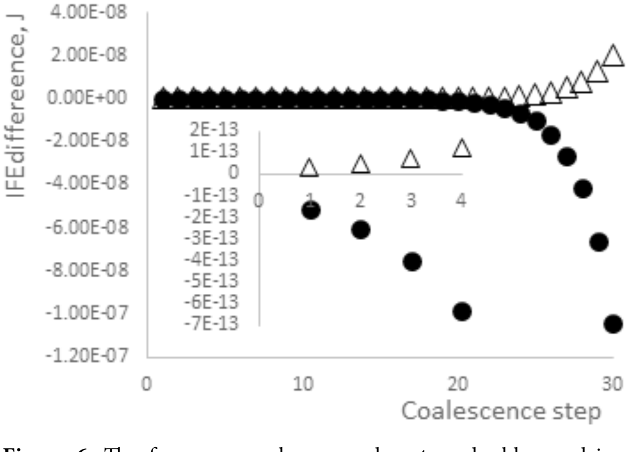


Figure 6. The free energy changes, when two double emulsion drops (Example in text) coalesce to form one double emulsion drop, ●, or two single oil drops, △, of O1 and O2, respectively.

$$R_{C/D} = [(1 - v_{O2})^{2/3}\gamma_{O1/O2} + \gamma_{O2/W}] /$$
$$[(1 - V_{O2})^{2/3}\gamma_{O1/W} + v_{O2}{}^{2/3}\gamma_{O2/W}] \qquad [4]$$

According to these conditions, there are examples, Figure 7, which indicate ranges outside the condition $\gamma_{O1/W} > \gamma_{O2/W} + \gamma_{O2/O1}$, at which the combined drop is thermodynamically favorable to two separate drops.

The interfacial free energy of the double emulsion drop is unquestionably less than that of the two separate drops for *all* volume ratios, when $\gamma_{O1/W} > \gamma_{O2/W} + \gamma_{O2/O1}$. Then again, for $\gamma_{O1/W}/(\gamma_{O2/W} + \gamma_{O2/O1}) = 0.85$ (curve ◇ in Figure 7) the ratio in question is still less than one; an obvious thermodynamic contradiction. If there is no spreading, it is difficult to accept that two separate drops should spontaneously unite to a double emulsion drop. Nonetheless, neglecting the fundamentals, focusing on the algebra per se, the trend as such of the curves in Figure 7 is anticipated from equations [3] and [4]. The initial increase of the first term of the equation denominator is less than that of the numerator of the equation, giving a downward slope, while the final change is the opposite, giving the minimum of the curve. As a result of the shape of the curves, the $R_{C/D} < 1$ within a limited range of relative volumes, even for $\gamma_{O1/W} < \gamma_{O2/W} + \gamma_{O2/O1}$, a purely algebraic result. However, the indisputable conclusion is that the use of two separate spheres is not the correct counterpart to gauge the thermodynamic stability of a double emulsion drop. In fact, the choice of a correct counterpart to evaluate the thermodynamic stability of the double emulsion drop needs a more comprehensive evaluation of the entire tension range.

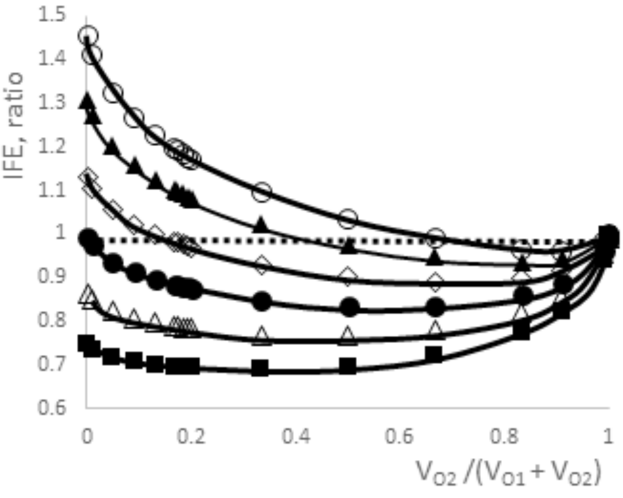


Figure 7. The IFE ratios between the interfacial free energy of a double emulsion drop and two separate single-oil drops with $\gamma_{O1/W} = 1$.

Symbol	$\gamma_{O2/W}$	$\gamma_{O2/O1}$
■	0.5	0.25
△	0.58	0.29
●	2/3	1/3
◇	0.78	0.39
▲	0.88	0.44
○	0.98	0.49

Another small detail in Figure 7 might be mentioned, for which the inequality $\gamma_{O1/W} > \gamma_{O2/W} + \gamma_{O2/O1}$ actually is directly applicable to a physical emulsion. The final outcome of the coalescence of a double emulsion, whose interfacial tensions obey the inequality in question, is in the form of three layers of the liquids. At that point, gravity decides the order of the layers in the container. If the liquid densities vary as $\rho_W > \rho_{O2/W} > \rho_{O1/W}$, the order of the three layers will be O1, O2 and W from the top and three layers are found. Conversely, if the densities are ranged $\rho_W > \rho_{O1/W} > \rho_{O2/W}$, four layers are found, since the O1 layer is not in direct contact with the water layer, but separated from it by an (infinitely thin) O2 layer, because O2 spreads on W. Needless to say, these results are based on single-oil drops as the alternative to the double emulsion drop. In reality, a double emulsion drop will not change to two individual drops, if interfacial tension ratio is moved outside the condition $\gamma_{O1/W} > \gamma_{O2/W} + \gamma_{O2/O1.}$ This will become evident in the analysis of *the entire range* of interfacial tensions, which shows that, for $\gamma_{O1/W} < \gamma_{O2/W} + \gamma_{O2/O1,}$ the preferred topology becomes that of a Janus drop.

JANUS DROP

The introductory publication on Janus emulsions[15] was based on microfluidics[16] emulsification. This preparation guarantees emulsions at virtual internal equilibrium and led to a number of investigations into the different aspects of Janus drops,[17-22] the results of which confirmed the agreement between equilibrium predictions and experimental results. These studies formed the basis for an extensive foray into several important aspects of emulsions in biology and medicine, led by Weitz.[23] The method, as such, enabled the preparation of emulsions, to all intents and purposes, of *any* complex topology, but was inherently limited to diminutive volumes, preventing applications into commodities. This condition was changed in 2011, when Hasinovic et al prepared Janus emulsions by traditional vibrational emulsification,[24] opening an avenue to large scale production. This pioneering contribution showed microscopy photos of well-defined Janus emulsions, Figure 8.

The image shows well defined drops of an O/W Janus emulsion of a vegetable oil, weight fraction 0.18

and a light silicone oil, weight fraction 0.72, while the continuous aqueous phase comprises only 0.1. It is noteworthy that an O/W emulsion with such limited volume of continuous phase is formed in a standard vibrational emulsification; an early indication of the unexpected effect of interfacial thermodynamics on vibrational emulsification process. Furthermore, but equally important, the regular Janus topology was first achieved by shear *after* the initial emulsification. Both processes were necessary in most cases and deserve separate comments.

The emulsification as such is a process, in which a large number of transitory small drops of irregular shape are formed and the freshly prepared emulsion is a result of these drops rapidly coalescing to larger entities.[25] This process will favor irregular Janus drops for kinetic reasons, because there is virtually no colloidal stability effect involved. Assume an equal number, n, of equally sized drops of two mutually insoluble oils, which are allowed to coalesce at a rate, which is independent of specific drop topology. Subsequent coalescence of these drops leads to an overwhelming fraction of irregular shape Janus drops, in addition to their larger sizes. During the ensuing shear the small attached O2 drops coalesce to a regular Janus lobe. The effect of shear was cursory illustrated [26,27] by optical microscopic images, before and after a cover glass was applied on the microscope slide, Figure 9. The minute shear from the cover glass resulted in fewer drops, as expected, but also in an extensive topology change to better defined Janus drops.

In addition, the results of shear also – albeit indirectly – serve to confirm the internal thermodynamic stability of the Janus drops. Contrary to the case for single-oil emulsions, for which the effect of shear *at low rates* is to form larger spherical drops, low rate shear of the initial Janus drops, left micro-photo Figure 9, leads to coalescence of the attached drops and a more regular Janus drop. As such, the information in Figure 9 also complements and supports later experimental proofs of the thermodynamic stability of the structure.[20]

These results are concerned with the kinetic factors of the process, leaving the thermodynamics unexamined. The equilibrium angles and tensions of the Janus drop are given in Figure 2A and the algebra for equilibrium has been reported[16-19,27,28] with the following summary.

Balancing the forces in Figure 2A along and perpendicularly to the $\gamma_{O1/W}$ direction gives the angles β and δ, Figure 2A, which, in turn, define the angles μ and ε.

$$\mu = \eta + \beta \qquad [5]$$

$$\varepsilon = \eta - \delta \qquad [6]$$

Furthermore, assuming $r_{O1/W} = 1$, the radii $r_{O2/W}$ and $r_{O2/O1}$ are

$$r_{O2/W} = \sin\eta/\sin\mu \qquad [7]$$

$$r_{O2/O1} = \sin\eta/\sin\varepsilon \qquad [8]$$

These equations control the equilibrium at the contact line, while the entire drop configuration, Figure 2B, also depends on the relative volumes of the two dispersed liquids. Unfortunately, the latter feature is not easily calculated from given volume fractions. The expressions become prohibitively complex and Guzowski et al[18] opted to use a computer program to correlate volumes and topology. As an alternative, the volumes are calculated in the present contribution from the geometrical features in Figure 2 and the correlations between

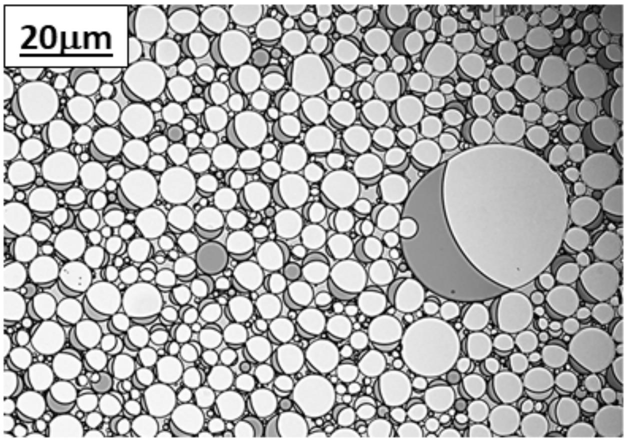


Figure 8. An optical microscopy image of a Janus emulsion, prepared by vibrational emulsification.

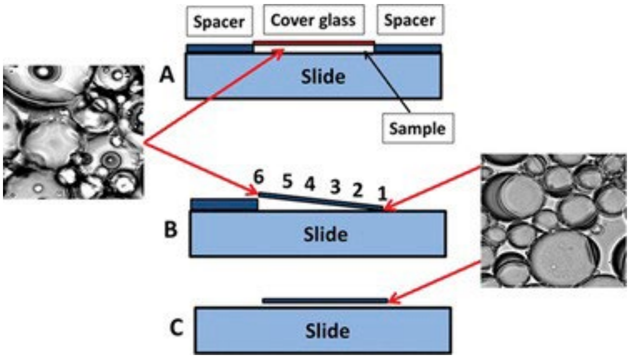


Figure 9. A simple experiment illustrating the effect by shearing on a Janus emulsion, prepared by vibrational emulsification[27]. (From reference 27 with permission).

volume ratios and topology are evaluated ex post facto.[26,28]

The volumes of O1 and O2 are calculated (equations [12] and [13]) via pre-volumes, φ_{O1}, black, Figure 2B, and φ_{O2} white + grey, Figure 2B, separated by the plane through the visible contact line.

$$\varphi_{O1} = \pi(1 + \cos\eta)^2(2 + \cos\eta)/3 \quad [9]$$

$$\varphi_{O2} = \pi(r_{O2/W} - \cos\mu)^2(3 - r_{O2/W} + \cos\mu)/3 \quad [10]$$

The volumes V_{O1} and V_{O2} are attained from $\varphi_{O1,}$ φ_{O2} and $\varphi_{O1/O2}$ (Grey, Figure 2B),

$$\varphi_{O2/O1} = \pi r_{O2/O1}{}^3(1 - \cos\varepsilon)^2(2 + \cos\varepsilon)/3 \quad [11]$$

$$V_{O1} = \varphi_{O1} + \varphi_{O2/O1} \quad [12]$$

and

$$V_{O2} = \varphi_{O1} - \varphi_{O2/O1} \quad [13]$$

In the comparison of interfacial free energies, the single-oil drops as counterparts are now replaced by a direct comparison between the free energies of double emulsion and Janus drops. As will be demonstrated, this new comparison is more relevant, removing the "anomalous" results in Figure 7. This figure showed the double emulsion drop to have lower interfacial free energy than two separate single-oil drops in a limited range of volume fractions, even for $\gamma_{O1/W} < \gamma_{O2/W} + \gamma_{O2/O1}$. The reason for this result is that the free interfacial free energy of the Janus drop was neglected, as has commonly been the case. Figure 10 shows the ratio of the free interfacial energies of a Janus drop to those of a double emulsion drop of identical volume. The ratio was limited to the tensions $\gamma_{O1/W} < \gamma_{O2/W} + \gamma_{O2/O1}$; since the drop equilibrium free energies for Janus drops with $\gamma_{O1/W} > \gamma_{O2/W} + \gamma_{O2/O1}$ is outside the equilibrium conditions and cannot be exactly calculated.

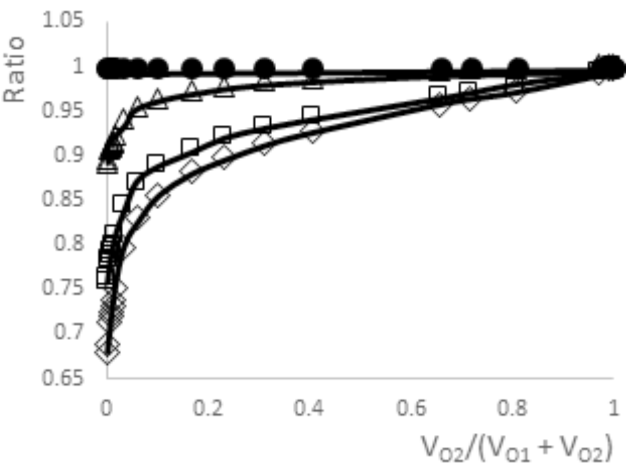


Figure 10. The ratio of Interfacial free energy of a Janus drop and a double emulsion drop of identical oil volumes.

The figure demonstrates the Janus drop to have a lower free energy than the double emulsion drop in the inequality range $\gamma_{O1/W} < \gamma_{O2/W} + \gamma_{O2/O1.}$ The faulty conclusion from using two single-oil drops as counterpart is now corrected.

In addition, there are two details that are of interest. The topology change, when the inequalities change from $\gamma_{O1/W} < \gamma_{O2/W} + \gamma_{O2/O1}$ to $\gamma_{O1/W} > \gamma_{O2/W} + \gamma_{O2/O1}$ is of special interest, because it is highly prominent. A complete analysis involves a large number of variables and in the present contribution a simplified example is used to graphically illustrate the phenomenon. The basis for the example is the specific and well defined case $\gamma_{O1/W} = \gamma_{O2/W} + \gamma_{O2/O1}$. In the calculation the expression is divided by $\gamma_{O1/W}$, giving $\gamma_{O2/W} = a$ and $\gamma_{O2/O1} = 1 - a$. The change of inequalities mentioned is represented by $\gamma_{O1/W}$ altered from 1 - ε to 1 + ε, in which ε is a small positive quantity. The angles b and d are calculated with the $\gamma_{O2/W}$ and $\gamma_{O2/O1}$ equal as are the O2 and O1 volumes. The specific selection of these is not essential for the central theme, and the angles β and δ are calculated versus ε.

$$\cos\beta = \cos\delta = 1 - (2 - \varepsilon)\varepsilon \quad [14]$$

The radical topology change, caused by the minute alteration of $\gamma_{O1/W}$ from 0.995 to 1.005 is illustrated in Figure 11.

The minute increase (1%) of the interfacial tension $\gamma_{O1/W}$ causes a drastic topology change with O1 spreading on the large sphere of O2. The activity is the same for smaller ε, but the graphics becomes less instructive, because the visible contact line is transferred to greater η angles with reduced ε.

Another, perhaps even more drastic consequence, is found of the interfacial tension variation for model sys-

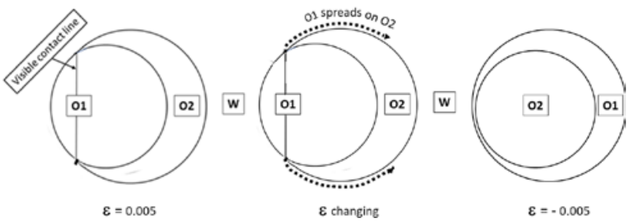


Figure 11. With e = 0.005, $\gamma_{O1/W} < \gamma_{O2/W} + g_{O2/O1,}$ a Janus drop is thermodynamically preferred. When ε changes to -0.005, O1 spreads on O2 and a double emulsion drop is favored.

tems with realistic interfacial tensions.[28] As an example, the variation of $r_{O2/W}$ and $r_{O2/O1}$ ($r_{O1/W}$ =1) with η is truly remarkable, Figure 12, for $\beta = 27.8°$, $\delta = 35.4°$, $\gamma_{O1/W}= 1$ and $r_{O1/W} = 1$. Each curve has a discontinuity, at which an infinite radius switches to the opposite sign with changing η. The $r_{O2/O1}$ versus η switches from $-\infty$ to $+\infty$ at $\eta = \delta = 35.4°$, while r_{O2} versus η discloses a discontinuity at $\eta = \pi - \beta$, approximately $\eta = 152.4°$, Figure 12.

The discontinuity of $r_{O2/O1\ 2}$ at $\eta = 35.4°$, has only a small effect on the volumes of the two lobes, Figure 2. Instead, the influence is felt on the Laplace pressure correlation.

$$\Sigma 2\Delta P_{XX}/r_{XX} = 0 \qquad [15]$$

The radius $r_{O2/O1}$ is negative for $\eta < 35.4°$, Figure 2. In this η range the terms $\gamma_{O2/W}/r_{O1/W} < 1$ and hence necessitates a negative $r_{O2/O1}$ to satisfy the LaPlace requirements, equation [15]. For $\eta = 35.4°$, the $\gamma_{O2/W}$ and $r_{O2/W}$ both are 0.65 and, again the Laplace pressure condition complies with those in equation [15]. For $\eta > 35.4°$, the sign of the radius is opposite to that at $\eta < 35.4°$. Nonetheless, as mentioned, the effect is not decisive for the size of the volumes, since $\varphi_{O2/O1}$ is usually small compared to volumes φ_{O1} and φ_{O2}, equations [9] and [10].

Conversely, the change in r_{O2} is accompanied by an extreme change in the O2 volume. As shown by Friberg[28] as well as by Ge et al[26] in different examples, at $\eta \approx 154.2°$ for the Janus drop in question, the change represents a partial inversion of the Janus drop from (O1 + O2)/W, η = 140°, Figure 13, to (O1 + W)/O2 η = 164°.

At η = 140°, the $r_{O2/W}$ has reached a value of 3.0 with $r_{O1/W} = 1$ and $r_{O2/O1} = 0.66$. The O1 drop is formed by two lobes, one reaching 0.23 (fraction of $r_{O1/W}$) into W, O1/W, and a second lobe entering 0.83 fraction into O2, O2/O1. There is, needless to say, no interface between the lobes. Together they form a *non-spherical* drop, O1, with an abrupt sign change of the radius as well as its dimension at the contact line. As expected, the added Laplace pressure over the interfaces O2/O1 and O2/W equals the pressure over the O1/W interface.

When η is increased to 152.4°, the O1 drop is located at the interface between two infinite phases W and O2. Nor in this case is the O1 drop symmetrical, since the $\gamma_{O1/W}$ is different from $\gamma_{O2/O1}$ and the LaPlace pressure is equal across the two interfaces of O1. Increasing γ to 164°, Figure 13, shows an (O1 + W)/O2 Janus emulsion and a further reduction of the O1 drop size.

These results and those in the preceding paragraphs are correct illustrations of the drop topology, as directed by the thermodynamic requirements. However, the applications to a physical emulsion are fraught with uncertainties and a few comments on the prerequisites for the model system are useful. The $\gamma_{O1/W} = 1$ is not a cause of concern; it only implies that instead of numerical values for $\gamma_{O2/W}$ and $\gamma_{O2/O1}$, their ratios $\gamma_{O2/W}/\gamma_{O1/W}$ and $\gamma_{O2/12}/\gamma_{O1/W}$ are used to simplify the algebra. Conversely, the second condition $r_{O1/W} = 1$ causes artificial restrictions on the physical image. It indicates that for each η change, a slice of O1 between the η:s in question is removed and a modified section of O2 is added at the contact line with exactly correct b angle. As is obvious, the conditions, in spite of being thermodynamically correct, are difficult to reconcile with any physical system, especially to the close packing of drops. The second alternative, retaining the O1 volume constant, gives a result similar to the one for constant $r_{O1/W}$, while the more artificial choice of keeping the entire drop vol-

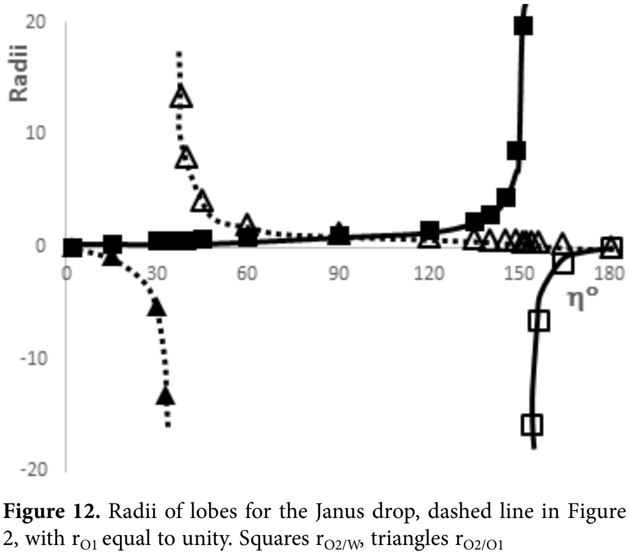


Figure 12. Radii of lobes for the Janus drop, dashed line in Figure 2, with r_{O1} equal to unity. Squares $r_{O2/W}$, triangles $r_{O2/O1}$

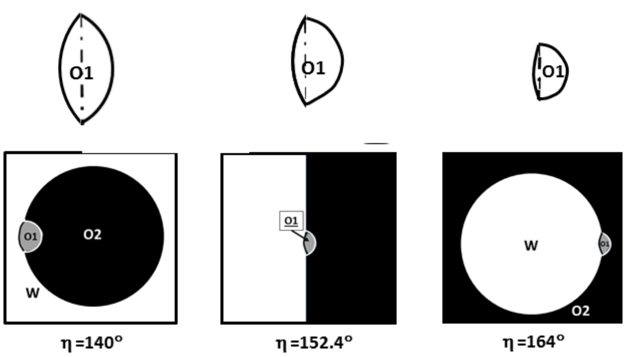


Figure 13. Schematic representation of the inversion of the Janus drop in the range η = 140° - 164°, Black areas are O2, grey areas O1 and white W. Expanded O1 areas with correct radii ratios are shown on top with the extension of the contact line as dotted/dashed.

ume constant leads to some modification. Nevertheless, from a physical point of view, the second alternative is the most realistic with O2 added to an already formed O1/W emulsion. For this case, adding O2 brings about a greater and greater O2 lobe of the drop, but causes no emulsion inversion, until the volume of O2 is greater than that of the initial continuous phase O1.

As a summary, adding the oils to an initial aqueous liquid gives an O1/W emulsion. Adding O2 to this emulsion results in a Janus emulsion, (O1+ O2)/W, an emulsion with increasingly larger O2 lobes. When the O2 volume exceeds that of the W, an inversion takes place to an (O1 + W)/O2 emulsion. Continued addition of O2 gives rise to a diminution of the relative size of the (O1 + W) drop.

CONCLUSIONS

The conclusions to include Janus emulsions as a counterpart, when considering the thermodynamic factors for double emulsion drops have been proven correct for selected examples.

The extension of these conclusions to Janus and double emulsion drops in general would be premature, but, so far, the indications are that the inference has more general validity.

ACKNOWLEDGMENT

The author is deeply grateful to his wife, Susan for her unfailing loyalty and support during the research and to the Ugelstad Laboratory, Trondheim, Norway for support.

REFERENCES

1. B. P. Binks, (Ed), *Modern Aspects of Emulsion Science*, The Royal Society of Chemistry, Cambridge, **1998**
2. J. Sjöblom, (Ed), *Emulsions and Emulsion Stability. 2 ed*. Taylor and Francis; **2006**
3. Y. Liu, in Th. F. Tadros (Ed), *Phase Inversion, Encyclopedia of Surface and Colloid Science, 2nd Ed.*, Springer, Amsterdam, **2013**
4. D. J. Mitchell, B. W. Ninham, *J. Chem. Soc. Faraday Trans.* **1981**, *77*, 601.
5. S.I. Ahmad, S.E. Friberg, K. Shinoda, *J. Colloid Interface Sci.* **1974**, *47*, 32.
6. G. Horvath-Szabo, j. H. Masliyah, J. A. W. Elliott, H. W. Yarranton, J. Czarnecki, *J. Colloid Interface Sci.* **2005**, *283*, 174.
7. F. Eslami, J. W. A. Elliott, *J. Phys. Chem. B* **2014**, *118*, 14675.
8. J. Korozs, G. Kaptay, *Colloids Surf. A* **2017**, *533*, 296.
9. B. Derjaguin, L. Landau, *Acta Physico Chemica URSS* **1941**, *14,* 633.
10. E. J. W.Verwey, J. Th. G. Overbeek, *Theory of the stability of lyophobic colloids, Elsevier, Amsterdam,* **1948**
11. M.Boström , V. Deniz, *G.V.* Franks, *B.W.* Ninham. *Adv. Colloid Interface Sci.* **2006,** *123–126*, 5.
12. A. Aserin, *Multiple Emulsions: Technology and Applications*, John Wiley & Sons, New Jersey, **2008**
13. S. Magdassi, M. Frenkel, N. Garti, *J. Dispersion Sci. Technol.* **1984**, *5*, 49.
14. J. Jiao, D. J. Burgess, *J. Colloid Interface Sci.* **2002**, *250*, 444.
15. N. Nisisako, S. Okushima, T. Torii, *Soft Matter* **2005**, *1*, 23.
16. G. M. Whitesides, *Nature* **2006**, *442*, 368.
17. N. Pannacci, H. Bruus, D. Bartolo, I. Etchart, T. Lockhart, Y. Hennequin, H. Williame, P. Tabeling, *Phys. Rev. Lett.* **2008**, *101*, 164502.
18. J. Guzowski, P. M. Korczyk, S. Jakiela, P. Garstecki, *Soft Matter* **2012**, *8*, 7269.
19. M. J. Neeson, D. Y. C. Chan, R. F. Tabor, F. Grieser, R. R. Dagastine, *Soft Matter* **2012**, *8*, 11042.
20. H. Hasinovic, S. E. Friberg, I. Kovac, J. Koetz, *Colloid Polym. Sci.* **2014**, *292,* 2319.
21. S-B. Zhang, X-H. Ge, Y-H. Geng, G-S. Luo, J. Chen, J-H. Xu, *Chem. Eng.* ***Sci.*** **2017**, *172*, 100.
22. Z. Chen, W. T. Wang, F.N. Sang, J. H. Xu, G. S. Luo, Y. D. Wang, *AIChE J.* **2016**, *62*,3685.
23. C.-H. Chen, R. K. Shah, D. A. Weitz, *Langmuir* **2009**, *25*, 4320.
24. H. Hasinovic, S. E. Friberg, G. Rong, *J. Colloid Interface Sci.* **2011**, *354*, 424.

25 H. Hasinovic, S. E. Friberg, *Langmuir* **2011**, *27*, 6584.

26 L. Ge, S. Shao, G. Lu, G. Rong, *Soft Matter* **2016**, *10*, 4498.

27 H. Hasinovic, C. Boggs, S. E. Friberg, I. Kovach, J. Koetz, *J. Dispersion Sci. Technol.* **2014**, *33*, 613.

28 S. E. Friberg, *J. Colloid Interface Sci.* **2014**, *416*, 167.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

OPEN ACCESS

Citation: H.-J. Apell (2018) Finding Na,K-ATPase. I - From Cell to Molecule. *Substantia* 2(1): 17-28. doi: 10.13128/substantia-38

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.

Feature Article

Finding Na,K-ATPase

I - From Cell to Molecule

Hans-Jürgen Apell

Dept. of Biology, University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Germany
Email: h-j.apell@uni-konstanz.de; Telephone: +49 7531 882253

Abstract. The oppositely oriented concentration gradients of Na^+ and K^+ ions across the cell membrane as found in animal cells led to the requirement of an active ion-transport mechanism that maintains this steady-state condition. As solution of this problem the Na,K-ATPase was identified, a member of the P-type ATPase family. Its stoichiometry has been defined as 3 Na^+/2 K^+/1 ATP, and a class of Na,K-ATPase-specific inhibitors, cardiac steroids, was established, which allow the identification of this ion pump. In an effort lasting for several decades structural details were uncovered down to almost atomic resolution. The quaternary structure of the functional unit, either αβ heterodimer or $(\alpha\beta)_n$ complexes with $n \geq 2$, is still under discussion.

Keywords. Sodium pump, active transport, discovery, physiological role, structure.

I. HISTORY OF THE NEED FOR A SODIUM PUMP

In the 1930s it was already well known that inside living cells the composition of the ionic contents was significantly different from that of the extracellular space. At that time the physiological investigations were focused mainly on muscle and red blood cells, and the evident asymmetry of high K^+ and low Na^+ concentrations inside and vice versa outside was well documented. During this period hardly any functional properties of the cell (or 'cytoplasmic') membrane were known, not to mention understood. It is not surprising that various concepts were developed to explain the asymmetry and how it was sustained. The initially preferred and widely supported idea was that the cell membrane is an almost perfectly impermeable barrier for ions. Some scientists were even willing to sacrifice the validity of the second law of thermodynamics in the field of biology in order to explain the experimental observations. A monograph on the historical development of our understanding of membrane transport, which is comprehensive and worth reading, was published about twenty years ago by Joseph D. Robinson.[1]

Major breakthroughs were achieved when new experimental techniques became available, such as the use of radioactive isotopes of K^+ and Na^+ that allowed the detection of unidirectional fluxes. Probably medical requirements of the Second World War also contributed, since physiologists were forced to study every detail on the optimization of blood preservation.

When $^{24}Na^+$ was introduced into physiological experiments it was shown that this radioactive isotope readily and rapidly exchanged with the stable isotope on the other side of the cell membrane. The membranes of the muscle cells were evidently permeable for Na^+.[2,3] It was also demonstrated that the cytoplasmic K^+ concentration could be very well modulated by the extracellular K^+ concentration, a clear indication of a K^+ permeability of the cell membrane.[4] Nevertheless, the asymmetric distribution of both cation species remained preserved despite the fact that both ion species were able to permeate through the membrane. As an inevitable consequence, a counter-movement of Na^+ and K^+ had to be assumed that keeps up the concentration gradients. In 1940 H. Burr Steinbach mentioned in a contribution to a Cold Spring Harbor Symposium for the first time the request for a "pumping out the sodium" from the cytoplasm.[5] It became clear that ion translocation driven by their electrochemical potential gradients, the so-called passive ion transport, had to be counteracted by energy-consuming transport processes (or "active transport") that ensured the indispensable condition of stationary high K^+ and low Na^+ concentrations inside the cells and even the electric membrane potential. Figure 1 shows a schematic representation of transport pathways and the transporters that were identified as leading actors during decades of investigations of cell membrane properties.

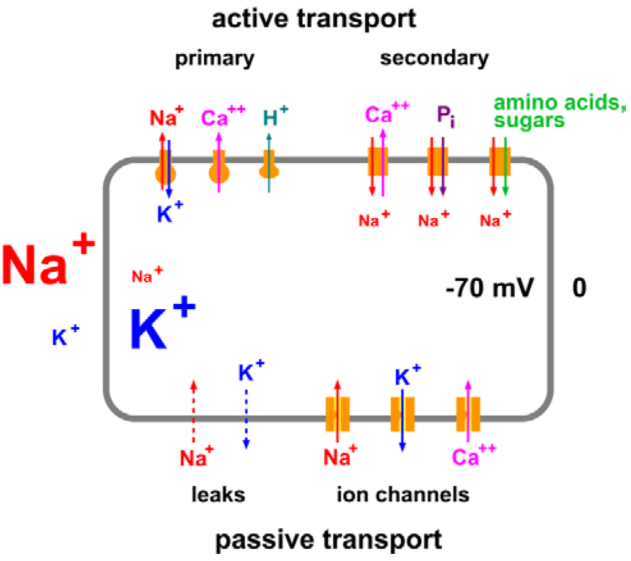


Figure 1. Ion transport pathways in animal cells. Common to all cells is the inside negative electric potential and the ion-concentration gradients oppositely oriented for Na^+ and K^+, with high Na^+ concentrations outside and low in cytoplasm and vice versa in the case of K^+. In principle, two different categories of transport mechanisms have to be discriminated, active and passive transport. The active transport is split into primary and secondary active transport. In primary active transport so-called ion pumps utilize ATP as energy source to translocate ions "uphill", i.e. against their electrochemical potential gradient. Secondary active transport is performed by antiporters (e.g. the Na,Ca-exchanger) or cotransporter (e.g. the Na,P_i- or Na,glucose-cotransporter) which translocate one of their substrates "uphill" while the other substrate, usually Na^+, provides the necessary free energy by its transport "downhill". A selection of examples is shown in this figure. Passive ion transport occurs either by leak conductance which is minimized by the nature and structure of the cell membrane (but unavoidable) or it is facilitated by channels or carriers. These transporters are regulated by different mechanisms to meet the metabolic needs of the cells. Key players for the passive cation transport are Na^+, K^+ and Ca^{2+} channels.

A further fruitful approach to advance the understanding of transport processes across the cell membrane was contributed by investigations of red blood cells. Already in the late 1930s the blood banks tried to find optimized preservation conditions of red blood cell batches. It was found that the stored cells lost their internal K^+ in the course of time, and the concentration of free K^+ in the blood plasma reached toxic levels when blood was preserved in the cold. In addition, it was shown that the cytoplasmic Na^+ concentration increased and that these changes were not primarily caused by deteriorated blood cells.[6] The net outflow of K^+ in the cold could be reversed at a temperature of 37 °C and in the presence of glucose. This observation indicated that K^+ and Na^+ ions were transported across the membrane against their concentration gradient. In the end, the physiological asymmetry was restored and this action was dependent on glycolysis.[7-9] In addition, inhibition of glycolysis by incubation with fluoride caused a delayed loss of K^+ from the red blood cells which was no longer balanced by K^+ uptake, even at physiological temperature.[9] This observation pointed out that glycolysis may not be immediately responsible for K^+ inward transport. At that time the underlying metabolic functions were still subject to speculation.

Later in the 1940s sufficient experimental evidence had been collected to conclude convincingly that active transport of Na^+ maintains the Na^+/K^+ asymmetry across the cell membrane and that this condition is a steady state and not an equilibrium.[10] Another step forward was the realization that the Na^+ flux out of the cell is coupled to the presence of external K^+.[11]

In the early 1950s a focus was set on the energy sources that fuel the concentration asymmetry for Na^+ and K^+ across the cell membrane of the red blood cell. The fact that glucose was metabolized but was not the direct source of energy had been made evident already ten years earlier.[9] It was also discussed that glycolysis plays a role as ATP-generating process.[12] In 1954, first experimental studies demonstrated that the K^+ accumulation in red blood cells was an ATP-requiring process which was activated by Mg^{2+}.[13] A few years later, the positive proof was provided that active K^+ transport occurred only when ATP was present,[14] and a corresponding finding was made also in squid giant axons.[15] In studies of frog skin significant experimental evidence was collected that active Na^+ transport was a forced exchange of Na^+ against K^+,[16] like in the case of red blood cells.[11]

Constructive findings concerning the sodium pump were made possible by another crucial discovery in the early 1950s which eventually turned out to be of eminent importance for all following investigations of the Na,K-ATPase. When studying the cation transport in red blood cells, Hans J. Schatzmann found that cardiac steroids blocked the K^+ uptake and the Na^+ outflux. He named the causative process "Na-K-Pumpe".[17] He was interested in identifying the mode of the blockers' action and with detailed experiments he proved in his study that these compounds did not affect glycolysis and oxygen consumption. He also concluded that there was a direct blockade of the transport mechanism.[17] This finding was supported and fortified by others.[18,19] Transport was blocked from the outside of the cell,[20] and external K^+ had an antagonistic effect on cardiac steroids.[19] Eventually, it became clear that Schatzmann had discovered a class of compounds, of which ouabain was the most well-known, that provides highly selective inhibitors of the Na,K-ATPase. Cardiac steroids have become the "custom-made" tool to discriminate the sodium pump from all other ATP-hydrolyzing enzymes in whatever biological tissues.[21]

Since at that time enough experimental evidence was collected that Na^+ and K^+ were coupled to maintain steady-state concentrations for both ion species inside the cells, the question was raised whether there exists a constant coupling ratio. To find the answer to that question was not so simple. It was necessary to discover a reliable experimental approach since the active ion transport balanced passive leak fluxes that in turn were dependent on the prevailing ion concentrations on the outside of the cells. From experiments with squid axons it was concluded that because of the passive ion permeabilities an active "secretory mechanism driven by metabolism" had to be present that moved Na^+ and K^+ against their electrochemical gradients.[22] The experiments also supported strongly the idea of a coupled system in which Na^+ was moved out of the cell "on one limb of the cycle" and K^+ taken up on the other.[22] In 1956 Ian M. Glynn was able to present experimental data from red blood cells using radioactive tracers, in which active and passive transport could be separated distinctly and Na^+ efflux and K^+ influx were tightly coupled in the active transport. He suggested a one-to-one exchange of Na^+ and K^+.[23] A year later Robert L. Post was able to measure net fluxes of Na^+ and K^+ with an unprecedented accuracy and determined a ratio of 3 Na^+ for 2 K^+, which was constant over the whole experimental concentration range of the transported ions.[24] This coupling ratio withstood all challenges and was found to be generally valid (except for a few extremely unphysiological electrolyte compositions).

II. TRACING THE PROTEIN

In the 1950s it was quite a challenge to propose the concept that a single protein molecule comprised both enzymatic function, in this case ATPase activity only known so far to soluble proteins, and transport function, i.e. vectorial ion movements across the cell membrane. Although during these years evidence was accumulated that there exists a tight coupling between both functions, the final proof, the identification of a protein (or protomer) that unites ATPase activity and ion transport was still pending. In 1957, the state of the art – with respect to red blood cells – was presented in a meticulously elaborated review by Glynn.[25] However, even on the basis of this amassed knowledge, no convincing experimental approach was developed to solve the puzzle of the Na,K-pump.

The breakthrough was eventually provided by a scientist from a completely different field. Jens Christian Skou studied the effect of local anesthetics on nerve conduction with the view of finding a membrane preparation that could be used in monolayer experiments, in which a well-defined enzymatic activity should be detected as function of applied local anesthetics. He finally attained his goal in 1956 by a membrane-fragment preparation from crab nerves and showed Mg^{2+}, Na^+ and K^+ dependent ATPase activity.[26] The history of this development is elaborately described in a comment on his original paper, published in 1989[27], in Skou's Nobel Lecture in 1997[28] and in his autobiographical book, "Lucky Choices. The Story of my Life in Science", published recently.[29] Since active transport of ions was not his field of interest, initially he was not aware of the contribution he had made. Only after Post triggered the crucial test when both met at a conference in 1958, namely to confirm inhibition of the enzyme activity by ouabain, it was decided that he had identified the Na,K-ATPase.[30] In retrospect Skou described his position in those early years: "I felt like an intruder in a field that was not mine."[28] The break-through was possible because he serendipitously worked with a membrane preparation from crab nerves that consisted of open membrane fragments in which both sides of the membranes were accessible simultaneously. In contrast, cell membranes of most other cells formed closed vesicular structures upon homogenization. Those preparations needed to be treated with detergents before the desired simultaneous access to both sides access of the membrane was obtained. With this information Post was able to identify briefly afterwards the Na,K-ATPase also in red blood cells.[31] In 1965 Skou had already published a review with reference to numerous tissues in which the presence of Na,K-ATPase had been verified too.[32] Today we know that the Na,K-ATPase is present in virtually all animal cells. A schematic biochemical characterization of the Na,K-ATPase is shown in Figure 2.

Because cell membranes contain scores of different proteins, at that time the question still remained open, whether the protein which performs ATPase activity was also responsible for ion pumping or whether more than one protein had to be coupled to a functional complex. It became a prominent task to isolate and purify the (minimal) enzyme complex that performed as Na,K-ATPase and analyze its components. In the early 1960s this project was, however, a major challenge because on one hand no standard methods were available to isolate and purify membrane proteins and keep them concurrently functional. On the other hand in isolated complexes no sidedness was given, and therefore, ion transport could not be proven. To overcome this problem, the solubilized complexes had to be reconstituted back into membranes that formed the interface of two compartments which were separate from each other and provided the required sidedness.

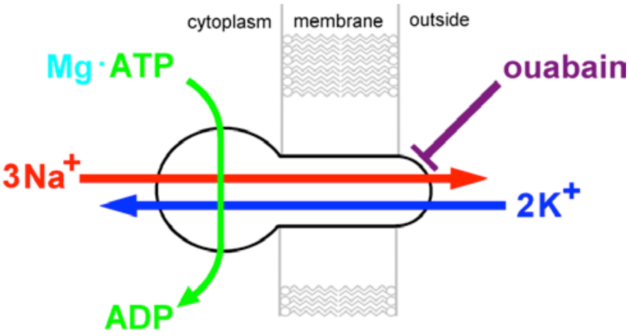


Figure 2. Biochemical characterization of the Na,K-ATPase. The ion pump is an integral membrane protein of animal cell membranes with its enzymatic machinery located on the cytoplasmic side. It hydrolyzes one MgATP complex into ADP and inorganic phosphate, P_i, and utilizes the released free energy to expel 3 Na^+ ions from the cytoplasm and to translocate 2 K^+ ions into the cytoplasm. Ouabain, a cardiac glycoside, is a specific inhibitor that completely blocks the Na,K-ATPase from the extracellular side of the membrane.

Tissues from which the Na,K-ATPase can be isolated in reasonable amounts are found either in mammalian brain and electroplax from fish that both contain excitable cells or in tissues specialized to transport sodium, such as the outer medulla of kidney, rectal glands of shark or salt glands of ducks.[33,34] Early findings showed that the outer medulla of mammalian kidneys is a fairly easily accessible and convenient source. The basolateral membranes of the cells forming the thick ascending limbs of the loops of Henle are specifically abundant in Na,K-ATPase.[35,36]

When cells rich in Na,K-ATPase are broken up by homogenization, a membrane preparation can be separated by centrifugation, the so-called crude microsomal fraction of which ATPase activities were measured in the order of 200 – 500 μmol inorganic phosphate (P_i) released per mg protein and hour. These membranes form vesicular structures with the cytoplasmic surface facing the outside.[37] To obtain a purified preparation from such a vesicle suspension that still contains all proteins of the plasma membrane, Peter L. Jørgensen elaborated in 1969 a specific treatment using a low concentration of the detergent sodium dodecyl sulfate (SDS). This method provides, after separation by differential centrifugation, open membrane fragments containing Na,K-ATPase in high density, the so-called purified microsomal preparation.[38,39] By this treatment most of the other proteins and a considerable fraction of the membrane lipids are removed. This approach became subsequently – with minor improvements – the standard routine to isolate and purify the Na,K-ATPase.[40] The protein density is so high that these fragments become stiff enough that they are no longer able to form vesicles. The hydrophobic membrane interior at the edge of the fragments is apparently covered by a layer of SDS molecules which prevent contact of the hydrophobic core to the aqueous phase. The resulting fragments contain ion pumps with densities of up to 10^4 per μm^2 as determined from electron micrographs.[41] When Jack Kyte applied in 1971 the shortly before introduced SDS polyacrylamide gel electrophoresis to a purified microsomal preparation, he proved a purity of better than 95% and that the Na,K-ATPase is a protomer of two distinct polypeptides with proposed molar masses of 84 kDa and 57 kDa.[42] The purified microsomal preparations from pig kidney attained ATPase activities of up to 2400 μmol P_i per mg protein and hour.[43] When incubated in solutions with Mg^{2+} and sodium vanadate, an inhibitor of the Na,K-ATPase,[44] the proteins spontaneously form in the microsomal membranes large two-dimensional crystal lattices in the membrane fragments.[43]

III. DEFINING PROPERTIES

In the 1970s advancing biochemical techniques promoted the study of membrane proteins. These were applied successfully to characterize the Na,K-ATPase with increasing precision in the following decades. A detailed review of the progress yielded during this period has been compiled in a 1979 review by Robinson and Flashner.[45]

At first, the most fundamental open question was probably about the subunit composition of the Na,K-ATPase: Does the functional ion pump consist of a single polypeptide chain or of a complex of two or more subunits? Although purified preparations displayed two different proteins in SDS gel electrophoresis,[42] it was not clear whether the lighter glycoprotein "is a true component of the NaK ATPase" or a tenaciously bound or coincidentally co-purified unrelated component.[46] From the gel-analysis method, applied to numerous purified protein preparations from different sources, the determined ratios of both subunits varied between 2:1 and 1:2 when the heavier subunit was compared with the lighter.[42,47,48] These diverse stoichiometries were obtained from experiments in which Na,K-ATPase preparations from different tissues were used, different amounts of proteins were applied on the gels, and at that time the behavior of glycoproteins on gels was not well understood. It lasted another few years until common agreement was reached that the stoichiometry is 1:1 and that only both subunits together form the active Na,K-ATPase.[33,49,50] In 1980 the final notation was introduced, in which the large polypeptide was named α subunit and the smaller glycoprotein β subunit.[49] It was established that the α subunit is phosphorylated by ATP[51] and that cardiac glycosides bind to it.[52] The location of the ion-binding sites was assumed to be also in the α subunit, but this problem was still under discussion in 1988.[53] The role of the β subunit was largely unknown at that time. Although it does not carry out enzymatic functions, it is crucial for the activity of the Na,K-ATPase.[54] Reduction of a single disulfide bridge that the β subunit possesses in its extracellular C-terminal part leads to a complete loss of the pump's activities.[55] How far various functions of the Na,K-ATPase are modulated by this subunit was under scrutiny for a long time. Convincing evidence, however, was compiled throughout the years that the β subunit is crucial for structural and functional maturation, trypsin resistance in ER preparations, appropriate trafficking in the cells and cell-cell adhesiveness.[56,57] Only after the Na,K-ATPase could be expressed from cloned cDNA it has been shown directly that the α subunit alone is incapable to perform Na,K-ATPase specific functions.[58]

The question whether an αβ complex is sufficient to perform Na^+ and K^+ transport fueled by ATP hydrolysis could be answered only after the Na,K-ATPase has been purified functionally and reconstituted as single protein species in a lipid membrane. This experimental approach was successfully introduced in 1974, when the purified Na,K-ATPase was incorporated in lipid vesicles and inside-out oriented reconstituted pumps transported ^{22}Na ions into the vesicles.[59] The transport was inhibited when ouabain was present inside the vesicles. As will be shown later, the application of Na,K-ATPase containing vesicles (or 'liposomes') turned out to become an extremely useful tool to investigate functional properties of the ion pump.[60]

In 1978 for the first time reliable evidence was presented that Na,K-ATPase isolated from pig kidneys contained a third subunit, a small polypeptide with a molar mass in the order of 12 kDa.[61] It took until the 1990s before more systematic investigations of this third subunit started, and during the following years a tissue specific distribution was found. Some tissues lacked of a third subunit and in others, where it was present, it was a member of the so-called FXYD protein family.[62-64] This family consists of seven members. These proteins possess less than 165 amino acids, have a single membrane-spanning segment, and they share the (eponymous) extracellular motive FXYD (with X as place holder for either T, E, Y, F). They interact with the α subunit and their function is a modulation of the ion-transport kinetics that allows short-term adaptation to specific metabolic needs of the cells.[64]

Skou reported already rather early experimental results that the Na,K-ATPases of rabbit kidney and brain exhibited different sensitivity to g-strophantin (ouabain). It was higher by a factor of 5 in the enzyme from the brain than from the kidney.[37] In 1976 experimental results were published from ouabain-binding studies using enzyme isolated from ox brain that are explained by the existence of two (or more) enzyme populations.[65] In 1979 Kathleen Sweadner advanced the field considerably when she proved by SDS-gel electrophoresis that two forms of the Na,K-ATPase existed in the brain that differ by 2 kDa in molar mass[66] which differed in their sensitivity to strophanthidin by almost a factor of 1000.

By 1989 three isoforms of the α subunit were identified and a common nomenclature fixed (α1 – α3).[67,68] In 1994 a fourth isoform, α4, was identified that is specific to testis.[69,70] α1 is the dominant isoform in kidney and heart, the tissue-specific distribution is catalogued extensively.[68] For the β subunit three isoforms (β1 – β3) were found.[71,72]

More precise access to the molar masses became available when amino-acid sequences were obtained from exploiting the analysis of complementary DNA that was introduced in the late 1970s[73] and the molar mass of proteins could be calculated precisely. In 1985, accurate numbers were published: 1016 amino acids (AAs) were determined for the (mature) α subunit of sheep kidney,[74] and 1022 AAs for the α subunit of *Torpedo californica*.[75] That led to a calculated molar mass of 112,177 Da for the sheep subunit. For the β subunit the first sequences were published in 1986: 302 AAs (sheep kidney)[76] with a calculated molar mass of 34.937 Da, and 305 AAs (*Torpedo californica*).[77] Due to the fact that glycosylation varies between animals and tissues, a detectable variation in total molar mass has to be expected.[78] With this technique and its success, further sequences of α and β subunits from several other tissues were published in 1986: α,β pig kidney,[79] β rat brain,[80] β human tumor cells.[81] Thereafter, numerous additional sequences followed in quick succession.[82,83] Whole families of Na,K-ATPase genes were identified and their transcriptional competence confirmed.[84,85] In 1987 the cloned cDNAs of both subunits from *Torpedo californica* were used to produce mRNAs by transcription in vitro. These were transferred by microinjection into *Xenopus* oocytes, expressed and trafficked functionally into the cell membrane.[58] It was shown that both subunits were necessary for correct folding and transfer to the cell membrane. With an increasing variety of molecular-biological tools that became available, the vast field of amino-acid mutations and protein expression in “foster cells”, such as oocytes,[86] yeast,[87] or various cell lines,[88,89] became accessible. This leap in development opened a completely new dimension of experimental investigations, and they enabled us to gain a major part of our contemporary understanding of function and structure function relationship (see subsequent part II).

With the knowledge of the gene sequence not only of the Na,K-ATPase but also of other ion motive ATPases a comparison of their sequences revealed that there is a whole family of ATPases which were named P-type ATPases because of their covalently phosphorylated intermediate.[90] Throughout the years increasingly complex phylogenetic trees of the P-type (super) family have been compiled[91,92], in which the Na,K-ATPase belongs to the type-II ATPases and its closest family members are the H,K-ATPase and the SR Ca-ATPase.

IV. INSIGHTS INTO THE STRUCTURE

Before the amino-acid sequence (or primary structure) became available, information on the structure of the Na,K-ATPase was mostly restricted to gross spatial features from electron-microscopical images[41,93] or from spectroscopic studies that allowed an estimation of the percentage of α helices and β sheets present in the protein at various substrate compositions.[94] More detailed concepts on the secondary structure were proposed after the tool of hydropathy analysis of the amino-acid sequence of proteins was introduced by Jack Kyte in 1982[95] and applied to the α subunit of the Na,K-ATPase. Between 1985 and 1994 many groups tried to derive from the primary structure a spatial organization of the protein in the membrane and especially the number of transmembrane segments. The count varied between 6 and 10 membrane-embedded α helices.[53,74,75,79,82,83,96] The proposal of an odd number of transmembrane segments, 7 or 9,[79,97] was quickly ruled out, because experimental evidence was presented that both N and C terminus of the α subunit were located on the cytoplasmic side of the membrane.[96,98] Eventually, consensus was obtained at a count of 10 helices,[99] which was confirmed in the end, when detailed tertiary structures became available by X-ray structure analysis from crystals of the complete Na,K-ATPase.[100,101] As shown in Figure 3, two major cytoplasmic loops were identified between the second and third transmembrane segment with about 140 amino acids, and between the fourth and fifth transmembrane segment with about 440 amino acids. In the latter loop the phosphorylation site as well as the FITC and IAF binding sites were located, which play important roles for function and analysis of the enzymatic activity of the pump.[102-105]

In the case of the β subunit it was accepted from the beginning that this small protein has no more than one transmembrane segment. The larger extracellular part with the C terminus[76] carries the essential disulfide bridges and is glycosylated at three asparagines.[53,78]

For a long time the insight to gain understanding of the tertiary structure was confined to low-resolution data provided by electron microscopy, starting with the identification of knob-like structures with a diameter of about 45 Å.[107,108] A few years later a ‘stalked knob’ was resolved on the cytoplasmic side of the catalytic subunit of the Na,K-ATPase.[93] The next step was the investigation of vanadate-induced two-dimensional Na,K-ATPase crystals[109] and a three-dimensional model reconstructed thereof,[110,111] as shown in Figure 4.

No further real improvement in the revelation of structural details was achieved until almost a decade later, the first structure of the SR Ca-ATPase was solved with a resolution of 2.6 Å.[112] The structure of this closely related enzyme was used for homology modeling of the α subunit of the Na,K-ATPase. The resulting structure proposals gained a lot of popularity and were used quite successfully to identify crucial amino acids as targets for mutation studies. Yet, another seven years had to pass until in 2007 the first original crystal structure of the Na,K-ATPase became available at a resolution of 3.5 Å in an E_2P-analogous conformation with two K^+ ions bound.[100] Two years later, another structure of the Na,K-ATPase in the same conformation became available at a resolution of 2.4 Å,[113] as well as another four years later a complex with a Mg^{2+} and a ouabain bound.[114] Since then, structures in the E_1 conformation with 3 Na^+ ions in their binding sites were resolved and published.[115,116] As already introduced in the analysis of the first Ca-ATPase structure, the cytoplasmic portion of the α subunit is subdivided into three domains named *N*, *P*, and *A*.[112] This organization was found similarly for all P-type ATPases studied so far. The *N* domain is the largest of the three domains and contains the nucleotide-binding site to which the Mg-ATP complex binds in a specific orientation that subsequently enables phosphorylation of the enzyme. The *P* domain includes the conserved aspartate that is phosphorylated by ATP. This domain is formed from two segments of the large cytoplasmic loop (between M4 and M5). The *A* domain is formed by the loop between transmembrane helices M2 and M3 and part of the sequence before M1. It is assumed to be an actuator that moves the phosphate hydrolysis machinery in and out of the active site by large rotational motions. A comparison of the crystal structures in the E2P and E1 conformation is shown in Figure 5. These represen-

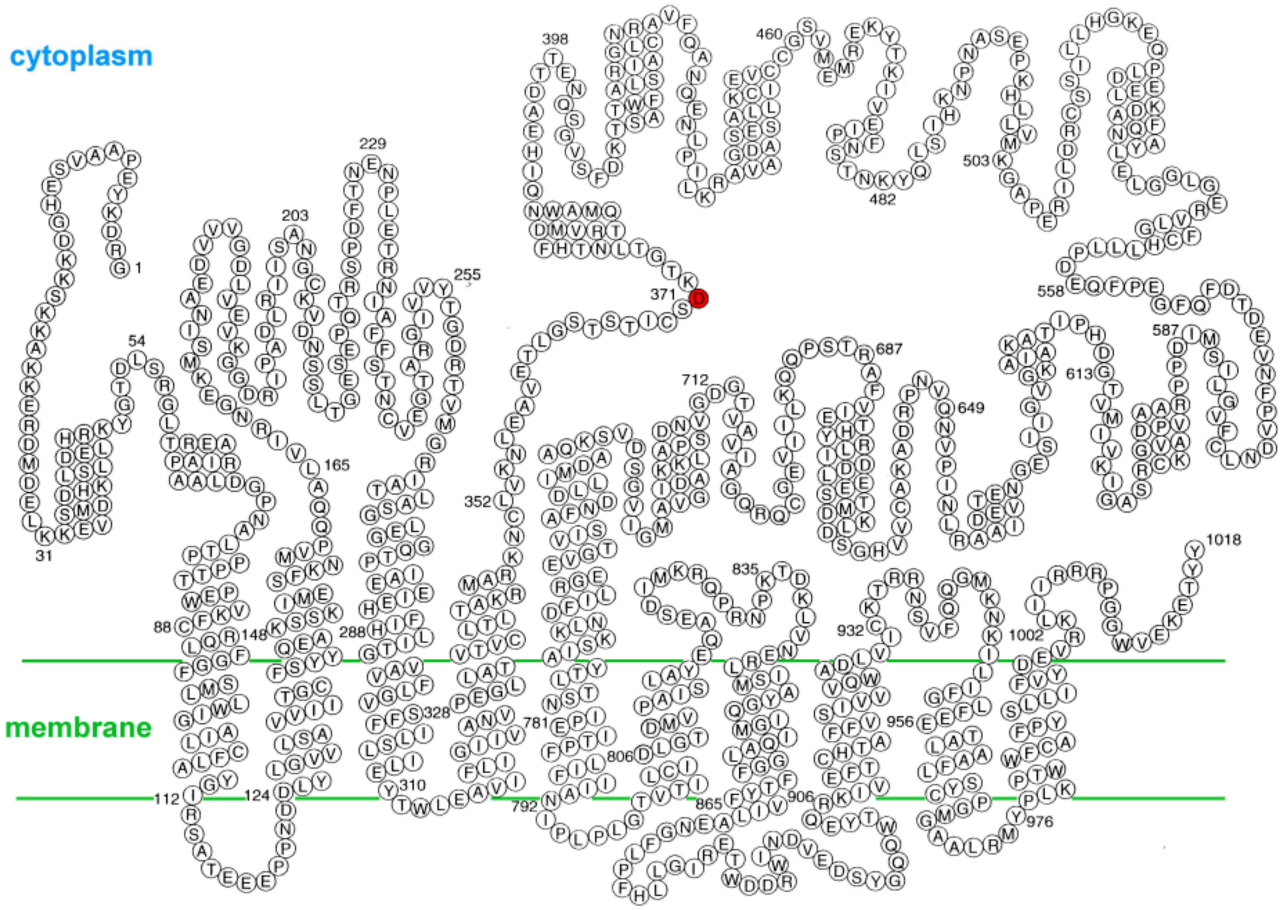


Figure 3. Secondary structure of the rat Na,K-ATPase α1 subunit with ten transmembrane helices. The high-lighted aspartate 371 is the amino acid phosphorylated by ATP. The drawing is adapted from Vilsen et al.[106] with permission.

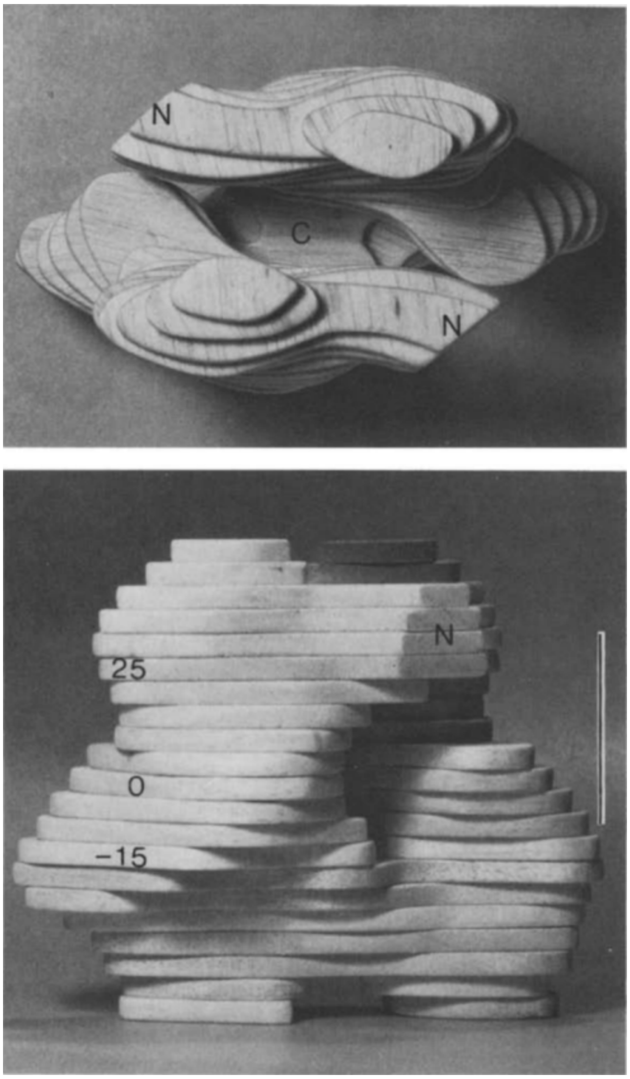


Figure 4. Reconstruction of a three-dimensional model of a Na,K-ATPase dimer from a tilt series of electron-microscopical images taken from a two-dimensional crystal of Na,K-ATPase in membrane fragments. Upper panel: top view, lower panel side view. The vertical bar indicates the assumed position of the lipid membrane, the cytoplasmic protrusion of the protein is on the bottom. (Figure taken from Ref. 109, with permission)

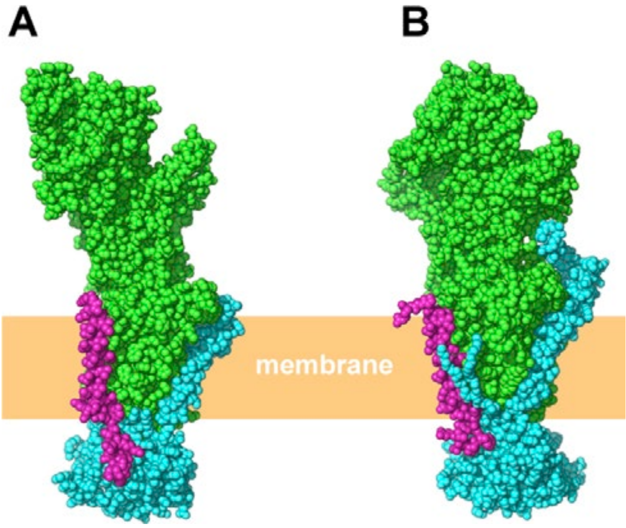


Figure 5. Crystal structure of the Na,K-ATPase in two conformations. The ion pump consists of the α (green), β (cyan) and an regulatory FXYD subunit (magenta). **A:** E_2 conformational state with 2 K^+ bound in a E_2P-like state in which phosphate is replaced by MgF_4^{2-} (pdb ID 2ZXE).[101] The analyzed crystal had a resolution of 2.4 Å. The enzyme was isolated and purified from shark rectal glands. It contains FXYD10 as regulatory subunit. **B:** Transition state of the Na,K-ATPase preceding the E_1P conformation with 3 Na^+ ions occluded after binding from the cytoplasmic side (pdb ID 3WGU).[115] The analyzed crystal had a resolution of 2.8 Å. The enzyme was isolated and purified from pig kidney. It contains FXYD2 as regulatory subunit.

tations of the Na,K-ATPase confirmed also that the structures derived by homology studies of the α subunit based on the SR Ca-ATPase structure have been rather well-suited. Furthermore, the original structures of the Na,K-ATPase revealed how the β and the additional regulatory FXYD subunit are connected to the α subunit. In combination with the biochemical and biophysical studies on the kinetics of the sodium pump, these structures with almost atomic resolution (and those in further different conformations that hopefully will come) are extremely useful to advance the comprehension of the molecular mechanism of enzyme and transport activity of the sodium pump.

For a long time a passionate discussion was carried out on the composition of the functional Na,K-ATPase, the protein's quaternary structure. The main opposing proposals were that under physiological conditions the ion pump consists either of a single αβ heterodimer or of an oligomer $(\alpha\beta)_n$ with n = 2 or even larger. This controversial issue was triggered by the rather early experimental finding that the Na,K-ATPase exhibits during the course of its catalytic activities two distinguishable affinities for ATP binding which were assigned to a high and low affinity site, accordingly.[33] This observation may be explained by three different concepts: First, a single αβ heterodimer has one ATP-binding site that changes its properties when the enzyme switches its conformation during the pump cycle.[34,117] The second proposal was that a single αβ heterodimer has two spatially different ATP-binding sites, one to perform the energizing enzyme phosphorylation while the other acts as a regulatory site, used to modulate the Na,K-ATPase activity,[118] a function found in numerous other ATP-controlled enzymes. The third concept was that of an Na,K-ATPase oligomer, $(\alpha\beta)_2$, in which both heterodimers are (tightly) coupled to provide synergistic effects.[119] The experimental evidence collected over many years was multifarious and seemed often to be in favor of one proposal but rarely refuted the other(s).[120]

On one hand, it has been shown that an isolated, monomeric αβ was able to run the enzymatic reaction cycle and to occlude Na and Rb ions.[121-123] In this condition, however, there was no membrane present, hence no sidedness was given and ion transport through the Na,K-ATPase could not be proven. On the other hand, in membranes Na,K-ATPase molecules were frequently clustered, and $(\alpha\beta)_n$ complexes were found when isolated with mild detergents. It was easily possible to crosslink the subunits of two different pumps,[124] and in the presence of vanadate, and when the pumps were in a E_2 conformation, the emergence of long rows of dimeric two-dimensional crystals was detected.[125] Such a crystal formation was observed also in purified membrane preparations of Na,K-ATPase after a treatment with phospholipase A_2 by which the lipid content of the membrane fragments was reduced.[126] Extensive crosslinking studies were performed in the group of Amir Askari and, modulated by the chosen substrate conditions, they substantiated various links between α-α, α-β, and β-β subunits. Their interpretation of the results from their crosslinking studies led to the conclusion that the "minimum association state within the membrane must indeed be $(\alpha,\beta)_4$".[127] But no direct evidence was presented that furnished proof of a functional cooperation between αβ heterodimers. Robinson wrote in his book (p.160): "$(\alpha\beta)_n$ formulations were popular in the 1970s, but αβ was favored in the 1980s when *almost* all the observations were revised and reinterpreted and when new data favoring the αβ interpretation was reported".[1]

Conceptual discussions about the requirements of monomeric or oligomeric structures of the Na,K-ATPase are multifaceted. The finding that all reaction steps needed for enzyme and transport functions have been documented in the case of monomers confronts the frequently observed spatial arrangement of the Na,K-ATPase as patches of two or more αβ protomers in close neighborhood or even in a tight contact that would allow functional interaction. Why should such contacts be stabilized or maintained if they are not advantageous? And indeed, the list of published experimental findings is notably that resorts to functional interaction between two αβ protomers. For example, it was used to describe complex kinetic behavior detected in measured substrate dependencies. A first option would be that tight coupling of two protomers could promote substrate binding to one αβ and then modulate the interaction of the other αβ with a second substrate. Since in both αβ protomers the pump cycles can (or will) be out of phase, one substrate may affect the interaction with a different one, even on the other side of the membrane. With such additional degrees of freedom and additional selectable parameters that are provided by modeling functional coupling, a complex kinetic behavior may be represented much easier by mathematical reaction schemes.[128] Furthermore, instead of a permanent tight coupling between two αβ protomers, it may be proposed that an ATP-dependent aggregation and separation of the $(\alpha\beta)_2$ oligomer exists which leads to different turnover rates when acting as $(\alpha\beta)_2$ or (temporarily) separated αβ under control of the ATP concentration in the cell.[129] It is tempting to see how more complex reaction schemes result in equation systems that allow almost perfect fits of experimental results by mathematical modeling. However, such a success does not justify the reverse conclusion that the underlying model is the right one.

Another reason to ask for dimers of Na,K-ATPases is to consider synergistic effects. When the energetics of the ion transport in the sodium pump was analyzed in terms of basic free energies,[130] i.e. the amount of free energy needed or being released at each single (experimentally accessible) reaction step of the pump cycle, no 'power stroke' was found.[131] This fact implies that at no single reaction step of the pump cycle, a driving thrust of energy was released to the "out-side world". The steps of the pump cycle consuming most energy were found to be the dislocation steps for both K^+ ions from their binding sites to the cytoplasm in the E_1 conformation. The absence of a power stroke could have its origin, on the one hand, in the mismatch of the application of our comprehension of macroscopic motors to molecular machines. But on the other hand, it could be also the consequence of an energetic coupling between two ion pumps running around their cycles out of phase in a way that energy production and consumption in coupled protomers compensate each other largely like in coupled chemical reactions.

For a comprehensive view of quaternary-structure formation it is, however, necessary to consider also the fact that the ion pumps are embedded in a lipid bilayer which is in a liquid-crystalline phase. This two-dimensional liquid has a life of its own that is under control of entropy. The various species of lipid molecules forming the cell membrane differ in the nature of their polar head groups as well as in the lengths of their hydrophobic fatty-acid chains and their number of double bonds. Especially the latter properties affect the degree of membrane fluidity. The presence of large rigid particles such as integral membrane proteins is able to promote separation into liquid-disordered and liquid-ordered phases. Studies of the molecular interaction mechanisms between proteins and lipids have shown that the match between the thickness of the hydrophobic domain of the integral protein and the bilayer core leads to an accumulation of specific lipids around the protein molecules.[132] This observation led to the conclusion that "the proteins end up in the membrane that provides for the best hydrophobic matching".[132] In addition, Na,K-ATPase has been found clustered in stable and rigid lipid rafts 'floating' in the membrane. Those rafts form spontaneously in membranes of (at least) ternary lipid mixtures.[133] Based on those findings an alternative line of arguments to explain clustering the Na,K-ATPase in cell membranes can be based on purely entropic effects that lead to an aggregation of Na,K-ATPase molecules without any functional coupling. But again, counter-arguments may be provided by the observation that in gastric acid-secreting cells an association between the K^+-Cl^- cotransporter-3a and the α1 subunit of the Na,K-ATPase was formed spontaneously in lipid rafts when cholesterol was present, and upregulated ATPase activity could be detected in a strictly cholesterol-dependent manner.[134] So far nothing is known about the molecular mechanism of the reported effects and about the role of cholesterol in the interactions between both ion transporters. Obviously, we still do not have all the pieces of a puzzle to recognize the complete raison d'être of Na,K-ATPase aggregation. "The paper is (still) open for discussion."

So far we have followed the trace of the discovery of the Na,K-ATPase, what it is good for and what it looks like, but the crucial question, how it works remains still open. Countless scientists contributed during decades to find answers to this question since this ion pump was identified and since protein preparations became available to work hands-on with physiological, biochemical and biophysical methods. An overview on this research will be presented in a subsequent article, "Finding Na,K-ATPase – From fluxes to ion movements."

ACKNOWLEDGEMENTS

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

REFERENCES

1. J. D. Robinson, *Moving Questions - A History of Membrane Transport and Bioenergetics*, Oxford University Press, New York, **1997**, p. 373.
2. L. A. Heppel; *Am. J. Physiol.,* **1940**, *128*, 449.
3. J. F. Manery, W. F. Bale; *Am. J. Physiol.,* **1941**, 215.
4. H. B. Steinbach; *J. Biol. Chem.,* **1940**, *133*, 695.
5. H. B. Steinbach; *Cold Spring Harb. Symp. Quant. Biol.,* **1940**, *8*, 242.
6. E. L. DeGovin, J. E. Harris, E. D. Plass; *J. Am. Med. Assoc.,* **1940**, *114*, 855.
7. J. E. Harris; *Biol. Bull.,* **1940**, *79*, 373.
8. M. Maizels, J. H. Paterson; *Lancet,* **1940**, *2*, 417.
9. J. E. Harris; *J. Biol. Chem.,* **1941**, *141*, 579.
10. A. Krogh; *Proc. R. Soc. Med.,* **1946**, *133*, 140.
11. F. Flynn, M. Maizels; *J. Physiol,* **1949**, *110*, 301.
12. M. E. Greig, W. C. Holland; *Arch. Biochem.,* **1949**, *23*, 370.
13. G. Gardos; *Acta Physiol Acad. Sci. Hung.,* **1954**, *6*, 191.
14. G. Gardos, F. B. Straub; *Acta Physiol Acad. Sci. Hung.,* **1957**, *12*, 1.
15. P. C. Caldwell, R. D. Keynes; *J. Physiol,* **1957**, *137*, 12.
16. V. Koefoed-Johnsen, H. H. Ussing; *Acta Physiol Scand.,* **1958**, *42*, 298.
17. H. J. Schatzmann; *Helv. Physiol. Pharmacol. Acta,* **1953**, *11*, 346.
18. A. K. Solomon, T. J. Gill, G. L. Gold; *J. Gen. Physiol,* **1956**, *40*, 327.
19. I. M. Glynn; *J. Physiol,* **1957**, *136*, 148.
20. P. C. Caldwell, R. D. Keynes; *J. Physiol.,* **1959**, *148*, 8P.
21. H. J. Schatzmann; *Protoplasma,* **1967**, *63*, 136.
22. A. L. Hodgkin, R. D. Keynes; *J. Physiol,* **1955**, *128*, 28.
23. I. M. Glynn; *J. Physiol,* **1956**, *134*, 278.
24. R. L. Post, P. C. Jolly; *Biochim. Biophys. Acta,* **1957**, *25*, 118.
25. I. M. Glynn; *Prog. Biophys. Biophys. Chem.,* **1957**, *8*, 242.
26. J. C. Skou; *Biochim. Biophys. Acta,* **1957**, *23*, 394.
27. J. C. Skou; *Biochim. Biophys. Acta,* **1989**, *1000*, 435.
28. J. C. Skou; *Biosci. Rep.,* **1998**, *18*, 155.
29. J. C. Skou, *Lucky Choices. The Story of my Life in Science*, U Press, Aarhus, **2017**, p. 213.
30. J. C. Skou; *Biochim. Biophys. Acta,* **1960**, *42*, 6.
31. R. L. Post, C. R. Merritt, C. R. Kinsolving, C. D. Albright; *J. Biol. Chem.,* **1960**, *235*, 1796.
32. J. C. Skou; *Physiol Rev.,* **1965**, *45*, 596.
33. P. L. Jørgensen; *Biochim. Biophys. Acta,* **1982**, *694*, 27.
34. J. C. Skou; *Meth. Enzymol.,* **1988**, *156*, 1.
35. J. Kyte; *J. Cell Biol.,* **1976**, *68*, 287.
36. J. Kyte; *J. Cell Biol.,* **1976**, *68*, 304.
37. J. C. Skou; *Biochim. Biophys. Acta,* **1962**, *58*, 314.
38. P. L. Jørgensen, J. C. Skou; *Biochem. Biophys. Res. Commun.,* **1969**, *37*, 39.
39. P. L. Jørgensen; *Biochim. Biophys. Acta,* **1974**, *356*, 36.

40. P. L. Jørgensen; *Quart. Rev. Biophys.,* **1975**, *7*, 239.
41. N. Deguchi, P. L. Jørgensen, A. B. Maunsbach; *J. Cell Biol.,* **1977**, *75*, 619.
42. J. Kyte; *J. Biol. Chem.,* **1971**, *246*, 4157.
43. E. Skriver, A. B. Maunsbach, P. L. Jørgensen; *FEBS Lett.,* **1981**, *131*, 219.
44. L. C. Cantley, L. Josephson, R. Warner, M. Yanagisawa, C. Lechene, G. Guidotti; *J. Biol. Chem.,* **1977**, *252*, 7421.
45. J. D. Robinson, M. S. Flashner; *Biochim. Biophys. Acta,* **1979**, *549*, 145.
46. J. L. Dahl, L. E. Hokin; *Annu. Rev. Biochem.,* **1974**, *43*, 327.
47. L. E. Hokin; *Ann. N. Y. Acad. Sci.,* **1974**, *242*, 12.
48. L. K. Lane, J. H. Copenhaver, Jr., G. E. Lindenmayer, A. Schwartz; *J. Biol. Chem.,* **1973**, *248*, 7197.
49. W. S. Craig, J. Kyte; *J. Biol. Chem.,* **1980**, *255*, 6262.
50. W. H. Peters, J. J. de Pont, A. Koppers, S. L. Bonting; *Biochim. Biophys. Acta,* **1981**, *641*, 55.
51. J. Kyte; *Biochem. Biophys. Res. Commun.,* **1971**, *43*, 1259.
52. A. Ruoho, J. Kyte; *Proc. Natl. Acad. Sci. U. S. A,* **1974**, *71*, 2352.
53. P. L. Jørgensen, J. P. Andersen; *J. Membr. Biol.,* **1988**, *103*, 95.
54. A. A. McDonough, K. Geering, R. A. Farley; *FASEB J.,* **1990**, *4*, 1598.
55. M. Kawamura, K. Nagano; *Biochim. Biophys. Acta,* **1984**, *774*, 188.
56. K. Geering; *J. Membr. Biol.,* **1990**, *115*, 109.
57. K. Geering; *Curr. Opin. Nephrol. Hypertens.,* **2008**, *17*, 526.
58. S. Noguchi, M. Mishina, M. Kawamura, S. Numa; *FEBS Lett.,* **1987**, *225*, 27.
59. S. Hilden, H. M. Rhee, L. E. Hokin; *J. Biol. Chem.,* **1974**, *249*, 7432.
60. H.-J. Apell, B. Damnjanovic; *Methods Mol. Biol.,* **2016**, *1377*, 127.
61. B. Forbush, III, J. H. Kaplan, J. F. Hoffman; *Biochemistry,* **1978**, *17*, 3667.
62. K. J. Sweadner, E. Rael; *Genomics,* **2000**, *68*, 41.
63. K. Geering, P. Beguin, H. Garty, S. Karlish, M. Fuzesi, J. D. Horisberger, G. Crambert; *Ann. N. Y. Acad. Sci.,* **2003**, *986*, 388.
64. H. Garty, S. J. Karlish; *Annu. Rev. Physiol,* **2006**, *68*, 431.
65. O. Hansen; *Biochim. Biophys. Acta,* **1976**, *433*, 383.
66. K. J. Sweadner; *J. Biol. Chem.,* **1979**, *254*, 6060.
67. G. E. Shull, J. Greeb, J. B. Lingrel; *Biochemistry,* **1986**, *25*, 8125.
68. K. J. Sweadner; *Biochim. Biophys. Acta,* **1989**, *988*, 185.
69. O. I. Shamraj, J. B. Lingrel; *Proc. Natl. Acad. Sci. USA,* **1994**, *91*, 12952.
70. G. Blanco, G. Sanchez, R. J. Melton, W. G. Tourtellotte, R. W. Mercer; *J. Histochem. Cytochem.,* **2000**, *48*, 1023.
71. J.-D. Horisberger, P. Jaunin, P. J. Good, B. C. Rossier, K. Geering; *Proc. Natl. Acad. Sci. USA,* **1991**, *88*, 8397.
72. J. B. Lingrel; *J. Bioenerg. Biomembr.,* **1992**, *24*, 263.
73. F. Sanger, S. Nicklen, A. R. Coulson; *Proc. Natl. Acad. Sci. U. S. A,* **1977**, *74*, 5463.
74. G. E. Shull, A. Schwartz, J. B. Lingrel; *Nature,* **1985**, *316*, 691.
75. K. Kawakami, S. Noguchi, M. Noda, H. Takahashi, T. Ohta, M. Kawamura, H. Nojima, K. Nagano, T. Hirose, S. Inayama; *Nature,* **1985**, *316*, 733.
76. G. E. Shull, L. K. Lane, J. B. Lingrel; *Nature,* **1986**, *321*, 429.
77. S. Noguchi, M. Noda, H. Takahashi, K. Kawakami, T. Ohta, K. Nagano, T. Hirose, S. Inayama, M. Kawamura, S. Numa; *FEBS Letters,* **1986**, *196*, 315.
78. M. J. Treuheit, C. E. Costello, T. L. Kirley; *J. Biol. Chem.,* **1993**, *268*, 13914.
79. Yu. A. Ovchinnikov, N. N. Modyanov, N. E. Broude, K. E. Petrukhin, A. V. Grishin, N. M. Arzamazova, N. A. Aldanova, G. S. Monastyrskaya, E. D. Sverdlov; *FEBS Letters,* **1986**, *201*, 237.
80. R. W. Mercer, J. W. Schneider, A. Savitz, J. Emanuel, E. J. Benz, Jr., R. Levenson; *Mol. Cell Biol.,* **1986**, *6*, 3884.
81. K. Kawakami, H. Nojima, T. Ohta, K. Nagano; *Nucl. Acids Res.,* **1986**, *14*, 2833.
82. J. B. Lingrel, J. Orlowski, M. M. Shull, E. M. Price; *Prog. Nucleic Acid Res. Mol. Biol.,* **1990**, *38*, 37.
83. L. A. Vasilets, W. Schwarz; *Biochim. Biophys. Acta,* **1993**, *1154*, 201.
84. N. N. Modyanov, K. E. Petrukhin, V. E. Sverdlov, A. V. Grishin, M. Y. Orlova, M. B. Kostina, O. I. Makarevich, N. E. Broude, G. S. Monastyrskaya, E. D. Sverdlov; *FEBS Lett.,* **1991**, *278*, 91.
85. M. M. Shull, J. B. Lingrel; *Proc. Natl. Acad. Sci. U. S. A,* **1987**, *84*, 4039.
86. Y. Hara, M. Ohtsubo, T. Kojima, S. Noguchi, M. Nakao, M. Kawamura; *Biochem. Biophys. Res. Commun.,* **1989**, *163*, 102.
87. B. Horowitz, R. A. Farley; *Prog. Clin. Biol. Res.,* **1988**, *268B*, 85.
88. J. R. Emanuel, J. Schulz, X. M. Zhou, R. B. Kent, D. Housman, L. Cantley, R. Levenson; *J. Biol. Chem.,* **1988**, *263*, 7726.
89. Y. Hara, A. Nikamoto, T. Kojima, A. Matsumoto, M. Nakao; *FEBS Letters,* **1988**, *238*, 27.

90. P. L. Pedersen, E. Carafoli; *TIBS,* **1987**, *12*, 146.
91. W. Kühlbrandt; *Nat. Rev. Mol. Cell Biol.,* **2004**, *5*, 282.
92. H. Chan, V. Babayan, E. Blyumin, C. Gandhi, K. Hak, D. Harake, K. Kumar, P. Lee, T. T. Li, H. Y. Liu, T. C. Lo, C. J. Meyer, S. Stanford, K. S. Zamora, M. H. Saier, Jr.; *J. Mol. Microbiol. Biotechnol.,* **2010**, *19*, 5.
93. F. Vogel, H. W. Meyer, R. Grosse, K. R. Repke; *Biochim. Biophys. Acta,* **1977**, *470*, 497.
94. T. J. Gresalfi, B. A. Wallace; *J. Biol. Chem.,* **1984**, *259*, 2622.
95. J. Kyte, R. F. Doolittle; *J. Mol. Biol.,* **1982**, *157*, 105.
96. S. J. Karlish, R. Goldshleger, P. L. Jørgensen; *J. Biol. Chem.,* **1993**, *268*, 3471.
97. Yu. A. Ovchinnikov, N. M. Luneva, E. A. Arystarkhova, N. M. Gevondyan, N. M. Arzamazova, A. T. Kozhich, V. A. Nesmeyanov, N. N. Modyanov; *FEBS Letters,* **1988**, *227*, 230.
98. N. N. Modyanov, N. M. Vladimirova, D. I. Gulyaev, R. G. Efremov; *Ann. NY Acad. Sci.,* **1992**, 134.
99. P. L. Jørgensen, P. A. Pedersen; *Biochim. Biophys. Acta,* **2001**, *1505*, 57.
100. J. P. Morth, B. P. Pedersen, M. S. Toustrup-Jensen, T. L. Sorensen, J. Petersen, J. P. Andersen, B. Vilsen, P. Nissen; *Nature,* **2007**, *450*, 1043.
101. T. Shinoda, H. Ogawa, F. Cornelius, C. Toyoshima; *Nature,* **2009**, *459*, 446.
102. F. Bastide, G. Meissner, S. Fleischer, R. L. Post; *J. Biol. Chem.,* **1973**, *248*, 8385.
103. M. O. Walderhaug, R. L. Post, G. Saccomani, R. T. Leonard, D. P. Briskin; *J. Biol. Chem.,* **1985**, *260*, 3852.
104. R. A. Farley, C. M. Tran, C. T. Carilli, D. Hawke, J. E. Shively; *J. Biol. Chem.,* **1984**, *259*, 9532.
105. P. A. Tyson, M. Steinberg, E. T. Wallick, T. L. Kirley; *J. Biol. Chem.,* **1989**, *264*, 726.
106. B. Vilsen, D. Ramlov, J. P. Andersen; *Ann. N. Y. Acad. Sci.,* **1997**, *834*, 297.
107. L. E. Hokin, J. L. Dahl, J. D. Deupree, J. F. Dioxon, J. F. Hackney, J. F. Perdue; *J. Biol. Chem.,* **1973**, *248*, 2593.
108. S. Uesugi, N. C. Dulak, J. F. Dixon, T. D. Hexum, J. L. Dahl, J. F. Perdue, L. E. Hokin; *J. Biol. Chem.,* **1971**, *246*, 531.
109. H. Hebert, P. L. Jørgensen, E. Skriver, A. B. Maunsbach; *Biochim. Biophys. Acta,* **1982**, *689*, 571.
110. H. Herbert, E. Skriver, A. B. Maunsbach; *FEBS Lett.,* **1985**, *187*, 182.
111. Maunsbach, A. B., Skriver, E., and Hebert, H. (1991) in *The Sodium Pump: Structure, Mechanism, and Regulation* (Kaplan, J. H. and de Weer, P., Eds.) pp 159-172, The Rockefeller University Press, New York.
112. C. Toyoshima, M. Nakasako, H. Nomura, H. Ogawa; *Nature,* **2000**, *405*, 647.
113. T. Shinoda, H. Ogawa, F. Cornelius, C. Toyoshima; *Nature,* **2009**, *459*, 446.
114. M. Laursen, L. Yatime, P. Nissen, N. U. Fedosova; *Proc. Natl. Acad. Sci. U. S. A,* **2013**, *110*, 10958.
115. R. Kanai, H. Ogawa, B. Vilsen, F. Cornelius, C. Toyoshima; *Nature,* **2013**, *502*, 201.
116. M. Nyblom, H. Poulsen, P. Gourdon, L. Reinhard, M. Andersson, E. Lindahl, N. Fedosova, P. Nissen; *Science,* **2013**, *342*, 123.
117. J. G. Nørby; *Chem. Scripta,* **1987**, *27B*, 119.
118. D. G. Ward, J. D. Cavieres; *J. Biol. Chem.,* **1996**, *271*, 12317.
119. Askari, A. and Huang, W.-H. (1985) in *The Sodium Pump* (Glynn, I. and Ellroy, O., Eds.) pp 569-573, The Company of Biologists Limited, Cambridge.
120. K. Taniguchi, S. Kaya, K. Abe, S. Mardh; *J. Biochem. (Tokyo),* **2001**, *129*, 335.
121. P. L. Jørgensen, J. P. Andersen; *Biochemistry,* **1986**, *25*, 2889.
122. B. Vilsen, J. P. Andersen, J. Petersen, P. L. Jørgensen; *J. Biol. Chem.,* **1987**, *262*, 10511.
123. Y. Hayashi, K. Mimura, H. Matsui, T. Takagi; *Biochim. Biophys. Acta,* **1989**, *983*, 217.
124. S. M. Periyasamy, W. H. Huang, A. Askari; *J. Biol. Chem.,* **1983**, *258*, 9878.
125. Maunsbach, A. B., Skriver, E., Söderholm, M., and Hebert, H. (1988) in *The Na^+,K^+-Pump, Part A* (Skou, J. C., Nørby, J. G., Maunsbach, A. B., and Esmann, M., Eds.) pp 39-56, Alan R. Liss, Inc., New York.
126. M. Mohraz, M. Yee, P. R. Smith; *J. Ultrastruc. Res.,* **1985**, *93*, 17.
127. A. V. Ivanov, N. N. Modyanov, A. Askari; *Biochem. J.,* **2002**, *364*, 293.
128. I. W. Plesner; *Biophys. J.,* **1987**, *51*, 69.
129. R. J. Clarke, X. Fan; *Clin. Exp. Pharmacol. Physiol,* **2011**, *38*, 726.
130. P. Läuger, *Electrogenic Ion Pumps*, Sinauer Associates, Inc., Sunderland, MA, **1991**, p. 313.
131. H.-J. Apell; *Ann. N. Y. Acad. Sci.,* **1997**, *834*, 221.
132. M. Ø. Jensen, O. G. Mouritsen; *Biochim. Biophys. Acta,* **2004**, *1666*, 205.
133. T. Bhatia, F. Cornelius, J. H. Ipsen; *Biochim. Biophys. Acta,* **2016**, *1858*, 3041.
134. K. Fujita, T. Fujii, T. Shimizu, N. Takeguchi, H. Sakai; *Biochem. Biophys. Res. Commun.,* **2012**, *424*, 136.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

OPEN ACCESS

Citation: L. Caruana SJ (2018) Mechanistic Trends in Chemistry. *Substantia* 2(1): 29-40. doi: 10.13128/substantia-39

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.

Feature Article

Mechanistic Trends in Chemistry

Louis Caruana SJ

Faculty of Philosophy, Pontificia Università Gregoriana, Rome, Italy
E-mail: caruana@unigre.it

Abstract. During the twentieth century, the mechanistic worldview came under attack mainly because of the rise of quantum mechanics but some of its basic characteristics survived and are still evident within current science in some form or other. Many scholars have produced interesting studies of such significant mechanistic trends within current physics and biology but very few have bothered to explore the effects of this worldview on current chemistry. This paper makes a contribution to fill this gap. It presents first a brief historical overview of the mechanistic worldview and then examines the present situation within chemistry by referring to current studies in the philosophy of chemistry and determining which trends are still mechanistic in spirit and which are not.

Keywords. Mechanism, Descartes, atomism, substance, teleology.

Chemistry can be described as the study of how matter adopts different forms and of how it changes from one form to another. Within this discipline, various conceptual issues arise. They are of interest to philosophers, to historians and sometimes to chemists themselves, especially to those chemists who seek greater clarity about the deeper assumptions of their work. One of the most interesting conceptual issues has to do with mechanism. To account for the way matter changes from one form to another, chemists often use mechanistic explanations. For instance, they may explain a chemical reaction in terms of a small number of steps from the initial reactants to the final products, the intermediates being conceived of as somewhat stable molecular combinations. The mechanistic explanation in these cases is like a set of snapshots taken at different stages of the transformation. The basic assumption is that the world functions like a complex machine and that every process can be analysed into definite steps that involve simple reconfiguration of parts and transfer of energy. Chemists tend to see a chemical reaction as a kind of sub-device within the larger machine of nature and tend to see their task as a kind of reverse engineering.[1]

[1] The clearest example of this method is probably E. J. Corey's retrosynthetic analysis. "*Retrosynthetic* (or *antithetic*) analysis is a problem-solving technique for transforming the structure of a synthetic target (TGT) molecule to a sequence of progressively simpler structures along a pathway which ultimately leads to simple or commercially available starting materials for chemical synthesis." E. J. Corey and Xue-Min Cheng, *The Logic of Chemical Analysis* (New York: John Wiley & Sons, 1989), p. 6.

ISSN 1827-9635 (print) | ISSN 1827-9643 (online) | DOI: 10.13128/substantia-39

The fascination with mechanisms has an interesting history. The origin of the so-called mechanistic worldview is often associated with Galileo Galilei and with the beginning of the scientific revolution. Even before his time, however, the proliferation of machines had been a distinctive feature of the intellectual milieu of the Renaissance. About a hundred years before Galileo, Leonardo da Vinci had filled notebook after notebook with designs of various contraptions intended to satisfy all kinds of human needs. Leonardo's contraptions qualify as machines because he proposed them as artefacts made up of components that function together for an overall positive effect. The most rudimentary machines, like levers and pulleys, have been with us since the dawn of civilization but, in the course of European history, from the Renaissance onwards, we see the importance of machines increasing at an accelerating rate, with human society becoming progressively dependent upon them. The rise of mechanistic thinking left a significant mark on the wider cultural, philosophical, and religious contexts. It affected the way people understood the world and their place within it, and it gave rise to a distinctive worldview, a cosmology in which God became increasing seen as the chief engineer responsible for the greatest and most intricate machine of them all, the entire universe. The specific philosophical features and assumptions of this worldview were not completely clear from the start. It took philosophers and scientists of the early modern period many generations to explore and articulate such assumptions often after lengthy disputes with theologians.

During the twentieth century, the major features of this worldview came under attack mainly through the rise of quantum mechanics but some of the basic characteristics still survive today in some form or other. For example, it is arguable that the research programme of reductive physicalism within the brain sciences is a direct descendant of the mechanistic materialism of the late seventeenth century. Many scholars have produced interesting studies of such significant mechanistic trends within current physics and biology but very few have bothered to explore the effects of this worldview on current chemistry. In this paper, I intend to make a contribution precisely in this neglected area, primarily by seeking an answer to the question, "How mechanistic is current chemistry?" In the first section, I will present a brief historical overview of the mechanistic worldview with the aim of extracting its main philosophical characteristics. In the second section then, I will examine the present situation by referring to current studies in the philosophy of chemistry and determining which trends are still mechanistic in spirit and which are not.

1. THE MECHANISTIC VIEW IN HISTORY

The rise of a distinctively mechanistic worldview would not have been possible without the position often called corpuscularianism, according to which macroscopic bodies should be described, and their behaviour explained, in terms of microscopic corpuscles — a view not very different from traditional atomism.[2] The novelty of corpuscularianism arose from the conviction of fifteenth and sixteenth century thinkers that the Aristotelianism of the Middle Ages needed urgent revision. These thinkers were convinced moreover that a revision could only come about by using mathematics to quantify in some way the various attributes of the ultimate constituents of the world. Such insights pointed towards something that would be definitely new. In spite of this novelty, however, some affinity between the emerging, new mechanistic paradigm and the old scholastic Aristotelianism remained. Elements of continuity were especially evident in the way major proponents of the new paradigm justified their project. They adopted an attitude that corresponds exactly to what one would expect from an Aristotelian Scholastic thinker: they referred to an underlying essence. Aristotle had constructed his entire edifice of natural philosophy on the idea of substance. Descartes, one of the paradigmatic mechanistic philosophers, adopted a similar approach: he constructed the entire edifice of his mechanistic view upon the idea that the essence of matter was extension. It is clear therefore that, since the change from the old to the new paradigm included elements of both discontinuity and continuity, a responsible historiography needs to be sensitive to both.[3] To determine those features of the new worldview that affected chemistry, we need to investigate such complex conceptual transformations and legacies with special attention. Let us start by considering the contribution of three prominent protagonists of the mechanistic worldview.

Pierre Gassendi (1592-1655) left his mark on the history of philosophy because he not only endorsed and developed the atomistic philosophy of Epicurus but also attempted to produce a Christianized version of it. With his competence in both philosophy and theology, he managed to produce a sophisticated natural philosophy that was on a par with the Cartesian proposal.

[2] Ancient Greek philosophers used to assume that atoms were indivisible; corpuscles however were assumed microscopic building blocks of everyday objects, just like atoms, but without the condition of indivisibility.

[3] See R. Ariew, "Descartes and Scholasticism: the intellectual background to Descartes' thought," in *The Cambridge Companion to Descartes*, edited by J. Cottingham, Cambridge University Press, 1992, pp. 58–90.

Now, many features of Epicureanism seem, at first sight, completely irreconcilable with religious belief, especially Christianity. For instance, the kind of atomism defended by Epicurus denies creation and divine providence, assumes the infinity and eternity of atoms, rejects final causes, and gives a central role to chance. Gassendi knew this well. Nevertheless, he engaged in reconciliatory work by launching a critical evaluation of Aristotle's objections to this kind of natural philosophy. Gassendi was convinced that what Aristotle had attacked was not the genuine version of atomism but a caricature of it. The genuine Epicurean philosophy was indeed reconcilable with religious belief primarily because it included an element of wisdom. Like many other Ancient Greek philosophers, Epicurus had produced his theoretical proposal ultimately as an ethical way of life, his main aim being that of grasping the correct structure of the world so as to do away with imaginary fears arising from animistic cosmology. This element was certainly reconcilable with Christianity. Gassendi was aware of this point but did not limit his defence of Epicurus to these considerations. He turned his attention to other aspects as well. According to Gassendi, Epicurus and Christianity were opposed primarily as regards materialism and divine providence. Unlike Democritus, Epicurus had accepted as real not only the atoms themselves, but also the complex compounds that these atoms constitute when combined in various configurations. Epicurus had not insisted however that, by perceiving the macroscopic object, in other words, the combination of atoms, we have access to that object's essence, as Aristotle was to do after him. For Epicurus, in cognition we do not grasp the hidden essence of things but merely their appearances. Gassendi's first move was therefore to try to retrieve and defend this pre-Aristotelian Epicurean position.

Epicurus had also insisted that all physical processes, including those of perceivable macroscopic objects, are nothing more than interaction between atoms. He had complicated his picture however by assuming that all atomic motion was downwards, if not disturbed by swerving. In this somewhat strange assumption, Gassendi found his opportunity to introduce the element of divine providence. He argued that God provides atoms with different initial attributes: different motions and different sizes. Thus God gives the atoms the ability "to disentangle themselves, to leap away, to knock against other atoms, to turn them away, to move away from them, and similarly the capacity to take hold of each other, to attach themselves to each other, to join together, to bind each other fast, and the like, all this to the degree that he [God] foresaw would be necessary for every purpose and effect that he destined them for".[4] With such a proposal, Gassendi seemed to distort Epicurean thinking considerably. He was effectively replacing the fundamental idea of chance with goal-directed atomic behaviour. In this move, we see why Gassendi represents a radical departure from both Epicureanism and Aristotelianism. He brought God's action into Epicurean atomism and he also removed Aristotelian final causes from within nature. He ceased seeing goal-directed behaviours or final causes as intrinsic to the nature of things. Instead, he started to treat goal-directedness as an extrinsic factor, deriving not from nature but from God. With such an idea of providence, Gassendi had to respond to the problem of personal freedom. How could an externally goal-directed universe leave space for freewill? The answer for him involved the idea of flexibility: he argued that the intellect is precisely that kind of complex combination of atoms that produces a flexible nature, one that can judge various aspects of the same object, and can evaluate different future possibilities. This proposal does account for what we observe but seems *ad hoc*. Overall, we can say that Gassendi's attempt to arrive at a synthesis of Epicureanism and religious belief was brave but not without its own problems.

We move on now to another defender of the mechanistic worldview, Thomas Hobbes (1588-1689) who proceeded in Gassendi's steps but was not concerned as much as Gassendi with retaining consistency with religious belief. Hobbes embraced materialism and determinism, and consequently expressed an overall view that nowadays we would call a materialist theory of the mind. He developed a comprehensive version of mechanistic philosophy that aspires to explain the entire universe in terms of matter and motion only, without reference to other features or forces, not even space and time. Within this picture, there is place neither for spiritual substance nor for religious belief in the traditional sense. In his book *De Corpore*, which contains most of his ideas on the workings of nature, he adopts an overall reductive approach. For him, any object's capacity to produce motion is nothing more than the motion of the constituent corpuscles. Space for him is neither substantial, enjoying a separate existence, nor a container, as Plato had suggested. It is merely a subjective frame of refer-

[4] P. Gassendi, *Opera Omnia*, Lugduni: 1658 (sumpt. Laurentii Anisson, & Ioan B. Bapt. Devenet), volume I, page 280: "congruam sese movendi, ciendi, evolvendi; et consequenter sese extricandi, emergendi, prosiliendi, impingendi, retundendi, regrediendi; itemque ses invicem apprehendendi, complectendi, continendi, revinciendi, et cetera quasenus ad omneis fineis effectusque quos tum destinabat necessarium providit." Such anthropomorphic language seems inevitable and is still present in chemistry today. In current literature, molecules attack each other, nuclear spins "flip," and electrons push other electrons.

ence, a mental abstraction. To explain an object's tendency to move, Hobbes developed the notion of *conatus*, which roughly refers to the object's inherent directionality or vectorial aspect. He developed also the correlative notion of *impetus*, which roughly refers to the measured conatus of any given object.[5] He used these two notions to describe not only any given object's motion but also its capacity to produce sensation in rational creatures. Overall, his worldview was clearly materialistic, but most commentators agree that the arguments he put forward to defend his materialism were never very strong. It seems likely that what convinced him of materialism was not sustained reflection or a knockdown argument but confidence in the new method of empirical inquiry, which was making fast progress during his time. There is no doubt that religious belief played a significant role in his philosophy, especially in his political philosophy, but this does not mean that he was an orthodox believer. Some surprising ideas that he expressed, for instance that God could be material, suggest that he was an outright atheist. The issue however remains unclear. The best way of seeing him is probably as a heavy-handed, revisionist, religious believer, a sceptic about much that organized religion proposed; in other words, a very critical theist.[6]

Compared to Gassendi and Hobbes, René Descartes (1596-1650) stands out as the one who produced the most characteristic expression of the mechanistic worldview of the modern period, influencing nearly all areas of culture. He presented his views for the first time in the book *Principia Philosophiae* of 1644, where he focused on the nature of human cognition. For him, the very nature of natural philosophy obliges us to see the characteristics of mind, and of God, as essentially distinct from the physical world. This affirmation for him was a typical "clear and distinct idea", something that can offer guidance to inquiry because we perceive it with the mind rather than with the senses. A non-deceiving God who is responsible for all existence will ensure that what we perceive by the mind clearly and distinctly is in general true rather than systematically misleading. This principle throws light also on the essential features of substances that make up the world. The two basic attributes of substances are extension and thought. Extension can show variations, for instance when an object is now in one position and later in another. Similarly, thought can show variations, for instance when the mind remembers one thing and then another. Such variations constitute what he calls modes. For him therefore, motion is a mode; and so is the lack of it, the state of rest.

We notice here some fundamental novelties with respect to Aristotelian thinking. For Aristotle, there was an asymmetry between motion and rest, at least when the motion is not heavenly. For non-celestial objects, all motion showed a natural tendency to come to rest. The motion of such non-celestial objects therefore needed an explanation, while their state of rest did not. As opposed to this, Descartes sees a symmetry between uniform motion and rest. Both are modes. For him therefore, uniform motion does not need an explanation in terms of a force. He follows Aristotle and says that God is the primary cause of motion but adds that God maintains a constant quantity of motion within the entire universe. What may change is the distribution of motion and of rest within the universe. The overall amount of motion, however, remains the same. This is a law of conservation, justified just like all genuine laws of nature, by God's immutability. Descartes concludes that "in general, we evidently cannot see this otherwise than as follows: that God himself, who set the parts of matter in motion or at rest when he first created them, now, through his sole ordinary attention, preserves in all of it the same quantity of both motion and rest".[7] Another important novelty with respect to Aristotelian physics concerns the laws of nature. For Descartes, laws are causes: "From God's immutability, we can also know certain rules or laws of nature, which are the secondary and particular causes of the various motions we observe in individual bodies."[8] Consequently, the natural regularities we discover and formulate in mathematical form are not, as Aristotelians had assumed, descriptions of the intrinsic activity of the various substances. They are rather extrinsic causes affecting extended substance that, on its own, is inert. A third innovation worth mentioning here deals with the idea of a vacuum. For Descartes, the idea of a vacuum is mistaken. Motion is not movement across empty space but displacement of one part of the universe by another.

[5] For further details, see H. Bernstein, "Conatus, Hobbes, and the Young Leibniz", *Studies in History and Philosophy of Science*, 11 (1980): 25–37.

[6] For a specific study of Hobbes's mechanistic philosophy, see F. Brandt, *Thomas Hobbes' Mechanical Conception of Nature* (Copenhagen; London, 1928); C. Leijenhorst, *The Mechanisation of Aristotelianism: The Late Aristotelian Setting of Thomas Hobbes' Natural Philosophy* (Leiden: Brill, 2002).

[7] R. Descartes, *Principia Philosophiae*, Part II, sec. 36: "Et generalem quod attinet, manifestum mihi videtur illam non aliam esse, quam Deum ipsum, qui materiam simul cum motu & quiete in principio creavit, jamque, per solum suum concursum ordinarium, tantundem motus & quietis in ea tota quantum tunc posuit conservat." See *Oeuvres de Descartes*, ed. C. Adam and P. Tannery, Paris: Vrin, 1996, volume VIII, p. 61, my translation.

[8] *Ibid*., sec. 37: "Atque ex hac eadem immutabilitate Dei, regulae quaedam sive leges naturae cognosci possunt, quae sunt causae secundarae ac particulares diversorum motuum, quos in singulis corporibus advertimus." *Oeuvres de Descartes*, ed. C. Adam and P. Tannery, Paris: Vrin, 1996, volume VIII, p. 62, my translation.

For any part of the universe to move, other parts need to squeeze out of the way accordingly. This is the direct consequence of Descartes's idea that the entire cosmos is a *plenum*. This means that, at any point, there is either a body or the fluid medium that fills up the space between bodies. This fluid medium causes bodies to move or come to rest. It reconfigures the overall distribution of motion and rest within the universe, which, on the large scale, is therefore a system of adjacent whirlpools carrying planets around their respective centres that are occupied by a central star. The Sun is just one of these central stars. This Cartesian cosmology is often referred to as a vortex theory, because it is modelled on what we see when a liquid moves round in a whirlpool. He argues that "the matter of the heavens, in which the planets are situated, revolves unceasingly, like a vortex having the sun as its centre, and that those of its parts that are close to the sun move more quickly than those further away."[9] Descartes thus offers a serious contender to Aristotle's cosmological system, which had assumed that the sub-lunar region is essentially different from the supra-lunar regions. Some historians highlight the fact that Descartes was not conceptually innovative on all counts. As I mentioned briefly before, he remained committed to giving explanatory priority to deductive arguments and essentialist thinking. Nevertheless, his original cosmological system had an enormous impact and remained the major point of reference for many generations of thinkers, even after the publication of Newton's *Principia Mathematica*.

The three philosophers mentioned up to now are by no means the only defenders of the mechanistic worldview. For a full list of philosophers who contributed to the detailed articulation of this worldview, we need to include people who were more directly associated with the new methods of empirical inquiry, figures like Galileo Galilei, Isaac Newton, and Pierre-Simon de Laplace.[10] It may be interesting to note that, even if we add all these, the list will not contain the name of anyone who was definitely against religious belief. In some form or other, religion was never completely absent in the work and life of these mechanistic thinkers.[11]

With the hindsight we enjoy today, after about four centuries since the emergence of the mechanistic worldview, what can we say about its basic conceptual ingredients? To answer this question, some scholars adopt the method of first identifying an ideal type of mechanistic philosophy, and then seeing the major sixteenth and seventeenth century thinkers as expressing some specific aspect or aspects of this ideal type. For instance, according to Stephen Gaukroger, the ideal mechanistic philosophy is one that reduces all physical processes to the motion of inert particles, fully describable in mechanistic and geometric terms.[12] The ideal mechanistic worldview assumes that we can fully explain any macroscopic object and its behaviour in terms of such particle motion only. The solid corpuscles are all of the same shape and size, while causation occurs between them only on contact. What we call matter is space that is full to capacity with such solid corpuscles. All macroscopic features of matter, such as observable variations in density, arise because of variations in the distribution of the constitutive corpuscles in space. The universe is causally closed, with no possibility of processes beginning or ending spontaneously. Given this basic picture, the main research programme of a mechanistic natural philosopher is to determine the laws of nature that allow a mathematical explanation of all observable changes. Of course, within such an explanation, corpuscles have passive attributes only. They are driven around according to the laws of nature. This modest list of assumptions is all we need to produce an exhaustive account of all kinds of motion and change, whether organic or inorganic. An important consequence here is that such a worldview has no place for Aristotelian final causes. It involves no intrinsic goals or purposes: neither within the corpuscles themselves nor within the complex composites.

This last point may give the impression that the mechanistic worldview represents a clear breach from ancient and medieval cosmology, but this is not completely true. Some important features of the old style of explanation did remain, as manifested by the example already mentioned, namely the recourse to first principles within the explanation. Descartes resorted to precisely this kind of explanatory strategy when *deriving*

[9] *Ibid*., Part III, sec. 30. The fuller explanation is as follows. "Sic itaque sublato omni serupulo de Terrae motu, putemus totam materiam coeli in qua Planetae versantur, in modum cuiusdam vortices, in cuius centro est Sol, assidue gyrare, ac eius partes Soli viciniores celerius moveri quam remotiores, Planetasque omnes (e quorum numero est Terra) inter easdem istius coelestis materiae partes semper versari. Ex quo solo, fine ullis machinamentis, omnia ipsorum phaenomena facillime intelligentur."

[10] For a fuller historical treatment of the mechanistic worldview, see E.J. Dijksterhuis, *The mechanization of the world picture,* Oxford 1961.

[11] Admittedly, some historians today present Laplace as a champion of religious unbelief. He showed mathematically that the solar system is stable on its own, without the need of the occasional Divine readjustment as Newton had proposed, and he famously affirmed that he had no use of the "divine hypothesis". He thus produced the complete mechanistic worldview and allegedly pushed God out of a causally closed universe. This interpretation however neglects the fact that even Laplace retained a form of Deism and endorsed the idea that we should consider God the Supreme Being responsible for the laws of nature.

[12] S. Gaukroger, *The Emergence of a Scientific Culture: Science and the Shaping of Modernity 1210-1685*, Oxford University Press, 2006. I am drawing especially from chapters 8 and 9.

observations from his basic principle of matter as *res extensa*. Is this strong element of deduction an essential ingredient of the mechanistic worldview? Some historians distinguish between a mechanistic philosophy that is highly dependent on deduction from another kind that is less dependent. This second kind of mechanistic philosophy highlights observation and experiment, and minimizes the role of speculation about what might lie hidden. Gaukroger argues that these two kinds of methods of approaching nature depend on whether one gives explanatory priority to the formal element of the explanation or to the observations themselves. The mechanistic style of natural philosophy described so far puts the emphasis on first principles, which it then considers the building blocks of the new worldview. It then interprets the phenomena to fit that logical structure. As opposed to this, the experimental style of natural philosophy gives the priority to the observations and experiments, highlighting the importance of empirical evidence and reliability. In this latter style, first principles are not the engine of inquiry. They do not play the role they had within the Cartesian mechanistic philosophy, the role of ensuring the organization and unity of knowledge. In the experimental style of mechanistic explanation, what drives the inquiry is rather the effort to arrive at piecemeal, local explanations of the phenomena at hand.

An interesting example is the explanation of colour. Descartes, as a typical mechanistic philosopher of the deductive style, produced a theory rationally grounded on his geometrical optics and microscopic corpuscles, whereby he explained white light as a homogenous collection of corpuscles whose spin could be differentially affected by passing through a prism. This explains our sensation of seeing different colours. Isaac Newton, on the contrary, did not feel constrained to start his explanation from an alleged underlying hidden principle, from which the observations could be derived. He concentrated exclusively on the relations between observable aspects at the phenomenal level. He thereby arrived at the idea that white light is indeed heterogeneous, composed of different colours that can be separated by passing through a prism. Descartes therefore had looked for underlying causal links while Newton looked for manifest causal links at the phenomenal level without the need for foundational assumptions regarding the hidden dimension of reality. Newton realized that, if he focused on the phenomenal relations only, he had to suspend judgment as regards the correctness of the theory of corpuscles. This introduced a new attitude within the mechanistic philosophy, an attitude that Newton excellently summarized in his famous comment regarding the origin of gravity:

> *I have not as yet been able to discover the reason for these properties of gravity from phenomena, and I do not feign hypotheses. For whatever is not deduced from the phenomena must be called a hypothesis; and hypotheses, whether metaphysical or physical, or based on occult qualities, or mechanical, have no place in experimental philosophy. In this philosophy particular propositions are inferred from the phenomena, and afterwards rendered general by induction.*[13]

We note here how Newton distinguishes his views from mainstream mechanistic ideas by calling his own philosophy experimental.[14] In spite of this new terminology, however, there is much that keeps all mechanistic natural philosophers together. The Cartesian style and the Newtonian style are therefore better seen as two versions of the same worldview rather than as two different worldviews. It may be interesting to add here that, as regards consistency with religious belief, there was no significant difference between Descartes' deductive style and Newton's experimental style. Both versions were open to belief in God more or less in line with the standard Western religious tradition.[15]

[13] Newton wrote this in the General Scholium, which was an appendix to his book *Philosophae Naturalis Principia Mathematica*. The final version of this Scholium appeared in the 1726 edition of the *Principia*. "Rationem vero harum gravitatis proprietatum ex phaenomenis nondum potui deducere, et hypotheses non fingo. Quicquid enim ex phaenomenis non deducitur, hypothesis vocanda est; et hypotheses seu metaphysicae, seu physicae, seu qualitatum occultarum, seu mechanicae, in philosophia experimentali locum non habent. In hac philosophia propositiones deducuntur ex phaenomenis, et redduntur generales per inductionem." The translation used here is from I. Newton, *The Principia: mathematical principles of natural philosophy*, trans. I. B. Cohen and A. Whitman, Berkeley: University of California Press, 1999, p. 943.

[14] M. Ben-Chaim, "The Discovery of Natural Goods: Newton's Vocation as an 'Experimental Philosopher'" *The British Journal for the History of Science* 34 (2001): 395-416.

[15] Descartes presented and justified his most famous work, the *Meditations*, as a way of defending the Catholic Faith: "I have always considered the two questions, the one regarding God and the other the Soul, to be the main ones that ought to be answered by the help of Philosophy rather than of Theology. For, although to us, the faithful, faith is enough to believe that the human soul does not cease to exist with the body, and that God exists, it surely seems impossible ever to convince infidels of the reality of any religion, or almost even any moral virtue, unless, first of all, those two things be proved to them by natural reason." (Semper existimavi duas quaestiones, de Deo et de Anima, praecipuas esse ex iis quae Philosophiae potius quam Theologiae ope sunt demonstrandae: Nam quamvis nobis fidelibus animam humanam cum corpore non interire, Deumque existere, fide credere sufficiat; certe infidelibus nulla religio, nec fere etiam ulla moralis virtus, videtur posse persuaderi, nisi prius illis ista duo ratione naturali probentur.) R. Descartes, "Sapientissimis Clarissimisque Viris Sacrae Facultatis Theologiae Parisiensis Decano & Doctoribus" (Letter of Dedication to the very sage and illustrious, the Dean and Doctors of the sacred faculty of theology of Paris), in *Oeuvres de Descartes*, ed. C. Adam and P. Tannery, Paris: Vrin, 1996, vol. VII, pp. 1-2, my translation. Newton justified his major work just as Descartes had done before him, by referring to its

So far, I have tried to show how the mechanistic worldview emerged slowly, how it remained in step with the new empirical methods of the natural sciences, and how it then eventually took definite shape towards the end of the seventeenth and early eighteenth centuries. Most of its fundamental tenets enjoyed considerable popularity during the nineteenth century but started to experience serious setbacks during the twentieth century. Scientific advances, especially in the area of quantum mechanics, obliged physicists to abandon the idea of elementary particles as tiny blobs of matter. The new paradigm in physics became incompatible with most of the features of classical mechanistic thinking. One of the most surprising novelties was the way in which the new physics undermined the materialistic basis of the mechanistic worldview. It brought about what N. R. Hanson called the "dematerialization of matter".[16] This expression does not mean that we should now reject the word "matter" as useless. It means rather that we need to retrieve its original philosophical sense: matter as a principle of individuation. The new paradigm obliges us to refrain from assuming that "matter" refers to some elemental stuff situated in space and time. This and other relevant shifts of meaning regarding fundamental terms show that, in the course of the twentieth century, the support that the mechanistic worldview used to receive from physics decreased considerably. It seems fair to say that, compared to what this support used to be in the seventeenth and eighteenth centuries, it is at present of minor importance.[17]

2. THE MECHANISTIC VIEW WITHIN CHEMISTRY

How has this mechanistic worldview affected chemistry? Does the present state of this discipline still show traces of mechanistic thinking? To answer these questions, we can first prepare the ground by considering the two fundamental concepts at work here, namely the concept of nature and the concept of mechanism, as they appear in chemistry.

For many centuries before the rise of natural science, the concept of nature was determined by Aristotelian philosophy and included a strong element of teleology or finality. Moreover, the distinction between natural and artificial, between *physis* and *technē*, was clear and important for the understanding of the world and of our place within it. With the Christian assimilation of these Aristotelian ideas, "natural" started to mean "in line with God's will", but, when the mechanistic view took over, final causes lost much of their importance, God was sidelined, and nature itself started being seen as the ultimate basis of explanation. These shifts caused some prominent thinkers to examine carefully how God should be referred to by the new science. The first book-length study of the concept of nature, written by Robert Boyle in 1682 and entitled *A free inquiry into the received notion of nature*, argued against the replacement of God by nature. For Boyle, it was a mistake to see nature as an agent. What scientists call the laws of nature should really be called the laws of God.[18] In line with this understanding, going against the laws of nature, or doing the unnatural, becomes sinful. When chemists create substances that are not found in nature, they therefore seem to transgress God's will, because, if God had wanted such substances to exist, He would have included them in creation. This kind of argument is obviously simplistic. Chemists could indeed be held responsible for going against God's will but their transgression would not lie in their having added something new, which God had not created before. It would lie rather in their intention to cause harm via the use of that new substance. Since humans are themselves part of nature, created by God like the rest of creation, their chemical ingenuity is not in itself something that goes against God's will. The chemists' endeavor to bring out, to actualize, the hidden potentialities of creation is perfectly natural. These considerations show how the dis-

value as an apology for religion: "When I wrote my treatise about our [solar] system, I had an eye upon such principles as might work with considering men, for the belief of a Deity, and nothing can rejoice me more than to find it useful for that purpose. [...] To make this system therefore, with all its motions, required a Cause which understood and compared together the quantities of matter in the several bodies of the sun and planets, and the gravitating powers resulting from thence; the several distances of the primary planets from the sun, and of the secondary ones from Saturn, Jupiter, and the earth; and the velocities with which these planets could revolve about those quantities of matter in the central bodies; and to compare and adjust all these things together in so great a variety of bodies, argues that Cause to be not blind and fortuitous but very well skilled in mechanics and geometry." I. Newton, *Four letters from Sir Isaac Newton to doctor Bentley, containing some arguments in proof of a Deity*, London: R. & J. Dodsley, 1756, digitized 2007, Letter I, p. 1; p. 7-8.

[16] N. R. Hanson, "The Dematerialization of Matter," *Philosophy of Science* 29 (1962): 27-38.

[17] As regards the fundamental constituents of nature, the present majority-view seems to involve a version of structural realism according to which what scientific theories ultimately refer to are not objects in space and time but patterns of relations expressed in the form of mathematical equations. See for instance Anjan Chakravartty, *A Metaphysics for Scientific Realism: knowing the unobservable* (Cambridge University Press, 2007).

[18] See Joachim Schummer, "The notion of nature in chemistry," *Studies in the History and Philosophy of Science* 34 (2003): 705–736. This paper presents a good overview of how chemistry resisted some of the fundamental assumptions of the mechanistic worldview. It is important to add however that the way the author identifies the Christian worldview with an odd narrative that he extracts from the non-canonical *Book of Enoch* shows considerable ignorance in this area.

tinction between natural and artificial can be misleading. In fact, we can observe that, from Newton onwards, the distinction starts losing its significance in philosophical and theological works about science and technology.

If we can say that the Aristotelian heritage regarding the pair *physis-technē* loses its importance, we cannot say the same thing as regards final causes. The literature about modern chemistry, especially during the interesting period of the artificial production of organic substances (roughly between the 1840s and the 1870s) shows that chemists increasingly assumed a teleological notion of nature and thereby distanced themselves more and more from physicists. Chemists readily made use of expressions like "imitating nature" and "learning from nature". Of course, the meaning of such expressions can oscillate between two extremes. The meaning may be that chemists see themselves as apprentices of nature or as its rivals. In spite of this possible semantic ambiguity however, we can safely conclude that, especially with the discovery of how to produce organic compounds artificially, chemists started seeing nature as active, as *doing* something. They thus reinstated some elements of teleology within the notion of nature, liberating themselves from the strictures of the classical mechanistic worldview. In this respect therefore, the chemists' idea of nature lies apparently midway between the finality-free mechanistic worldview and the teleologically rich, biological worldview. Current chemical literature confirms this point. A recent study affirms that, "the fact that we can so easily attribute the old metaphors to each of the branches [of current drug research] – learning from Nature, imitating Nature, improving Nature, competing with Nature, and controlling Nature – is hardly pure chance. It is more likely that these metaphors have actually been effective in shaping research traditions until today."[19]

Like the concept of nature, that of mechanism has had significant recent developments, some of which are relevant for chemistry. For lack of space, I will highlight two main features only. The first one deals with the idea of mechanism as corresponding to the form of acceptable explanations. Basically, a mechanism is an explanation that has the form of "nested hierarchies".[20] In line with the classic mechanistic worldview, the explanatory style I am referring to here assumes that objects are complex arrangements of smaller units, which are themselves made up of even smaller units, and so on. For current chemists, this should sound familiar. The explanation of a given phenomenon consists in supplying a description of a lower-level set of objects together with the push-pull relations between them and then supplying another even-lower-level set of smaller objects and their relations, and so on until we bottom out at the level of fundamental non-reducible elements. We support the entire explanatory ladder by assuming that the fundamental elements have some dispositions that do not need any further explanation. We affirm, for instance, that the electron has a negative charge, period. In such an explanatory process, the challenge is to reduce the number of inexplicable dispositions to a minimum. Of course, to arrive at a satisfactory set of nested hierarchies in this sense, we are entitled to use all the knowledge at our disposal. We can use previous knowledge of other systems and subsystems. We can use also knowledge that we may have acquired from situations that have nothing to do with the phenomenon that we are trying to explain. Moreover, the particularity of the phenomenon we are studying could be a stepping-stone for broader understanding. If we manage to extract the abstract form of the mechanism, if we manage to extract it out of the particularity of the one phenomenon we are studying, we could then use it for understanding other similar phenomena.[21]

This is one feature of mechanism within current chemistry. Another important feature is the mechanism's inherent directionality. The Hempel-inspired discussions on the structure of explanation of the late 1960s and 70s supported the idea that to explain a phenomenon is to provide some information about general laws and about its causal history. Given this background, we can conceive of a mechanism simply as a particular subset of causal relations that contribute to the appearance of the phenomenon. For any given phenomenon, the entire causal history is a vast network of mutually interacting cause-effect relations. The subset of this network that deserves to be called a mechanism is, according to this view, that subset that researchers consider relevant for their discipline. This understanding of mechanism, however, remains unsatisfactory. It seems too subjective. Different researchers would carve up the causal history in different ways. Some philosophers therefore have defended the claim that a mechanism is not just any sub-

[19] Schummer, "The notion of nature in chemistry", p. 726.

[20] Peter Machamer, Lindley Darden and Carl F. Craver, "Thinking about Mechanisms," *Philosophy of Science* 67/1 (2000): 1-25; the quote is from p. 13.

[21] The abstract version of a mechanism, usually in a diagrammatic form, is sometimes called a mechanism schema. Such schemas help in the effort to unify the knowledge that we derive from different situations, regarding both macroscopic properties and microstructure. "Higher level entities and activities are thus essential to the intelligibility of those at lower levels, just as much as those at lower levels are essential for understanding those at higher levels. It is the integration of different levels into productive relations that renders the phenomenon intelligible and thereby explains its." Machamer et al., p. 23. See also James A. Overton, "Mechanisms, Types, and Abstractions," *Philosophy of Science*, 78/5 (2011): 941-954.

set of the causal history. Something more is needed. A causal subset deserves to be called a mechanism when it is clearly directional, when it is clearly productive of the specific effect that we are investigating. The causal subset needs to be a complex system consisting of mutually interacting sub-systems that *function together* to produce the specific effect. The specific effect, in this case, would be what the mechanism is for.[22] On this view therefore, we are not entitled to call a set of nested hierarchies of systems a mechanism if we do not know what it is for.

We notice immediately here the affinity with the biological concept of function. These developments therefore are suggesting that the concept of mechanism in chemistry should depend on that of function, just as in biology.[23] When explaining living organisms, we can talk about the mechanism involved in a specific organ only when we know the function of that organ within that living thing. The simple causal-role view of mechanism therefore is not enough. We do not pick any set of events that have a causal role within the production of an effect. To refer to a biological mechanism, we first determine the specific task that the effect represents, in other words, we determine its function, and then spell out, step by step, how that function is realized. Not every change in a living organism is associated with a function. Changes can be accidental or even pathological. We do not take a pathology to be a mechanism. We take it to be a mechanism that has broken down. A malfunction is, as the word implies, a mechanism that went wrong. This shows how intimately related is the idea of mechanism to that of function. Now, the functional view of mechanism is typical of biology. In physics, the situation is different. Here, final causes have no significant role and the causal-role view of mechanism is therefore the only one available. What about chemistry? As one would expect, chemistry lies somewhere between these two positions. In the course of the seventeenth and eighteenth centuries, the mechanistic worldview gave priority to physics over the other sciences, it emphasized the causal-role view of mechanism, and it convinced many scientists to apply the causal-role view without alterations to chemistry and even to biology. The indispensable role of functional explanations within biology however, together with the onset of organic chemistry, has persuaded recent philosophers of science that the functional view of mechanism is indispensable not only for biology but for chemistry as well. The present situation therefore is interesting because, within the one discipline of chemistry, we find features that are definitely mechanistic and others that are not.[24]

Let us consider one current feature that is definitely mechanistic in character, namely the way chemists espouse atomism in some form or other. Just as the early mechanistic philosophers had their version of atomism, according to which all things were made up of small corpuscles, so nowadays chemists have their own version. They think of substances as combinations of smaller units and these units as combinations of even smaller units, and so on. The basic idea of postulating building blocks or elements to explain the great variety of things in the world has a long history going back to the Ancient Greek philosophers for whom there were only four elements: earth, water, fire and air.[25] In the course of history, alchemists adopted this assumption of the four elements and used it extensively in their somewhat confused talk about the transformation of substances. Subsequent studies became more systematic and started to involve the categorization of substances and the study of controlled changes, especially through the invention and betterment of the distillation apparatus. No doubt, technological advances continued to increase our knowledge of how substances react but the deeper mechanisms behind the observed changes remained indefinite. Sometimes, alchemists referred to animistic powers or occult forces to explain the hidden mechanisms, but such explanations were never a substitute for the basic idea of the four elements. As innovation progressed, interest in uncovering what lay hidden waned. Metallurgical manuals of the mid 1500s adopted an instrumentalist view, concentrating on how-to-do rather than on the underlying mechanisms that might explain the production of useful materials like glass, acids and gunpowder.[26] When natural philosophers started formulating the mechanistic worldview in terms of corpuscles, early chemists tended to combine the doctrine of the four elements with the new atomism, postulating the existence of four kinds of atoms, one for each element. On this view, the transformation of substances became a reconfigura-

[22] Stuart Glennan, “Rethinking Mechanistic Explanation,” *Philosophy of Science* 69/S3 (2002): S342-S353.

[23] Justin Garson, “The Functional Sense of Mechanism,” *Philosophy of Science* 80/3 (2013): 317-333.

[24] The functional view of mechanism and the idea of “nested hierarchies” are not the only important features of current philosophical research concerning mechanism. For other features, see for instance Stuart Glennan, *The New Mechanical Philosophy* (Oxford University Press, 2017); Carl Craver and James Tabery, “Mechanisms in Science”, *The Stanford Encyclopedia of Philosophy* (Spring 2017 online edition), Edward N. Zalta (ed.). These studies deal with the broad picture including all the sciences. In my paper, I focus on chemistry.

[25] For more on how Ancient Greek philosophy paved the way for modern theories about atoms, see Andrew G. van Melsen, *From atomos to atom: the history of the concept Atom*, trans. H. J. Koren (Pittsburg: Duquesne University Press, 1952).

[26] Aaron J. Ihde, *The Development of Modern Chemistry* (New York; Evanston; London: Harper and Row, 1964), p. 24.

tion of these four types of atoms.[27] The fact that the atomic theory continued to develop, to receive empirical confirmation and to arrive finally at the remarkable achievement of the Periodic Table did not change the basic explanatory strategy. The idea that chemists should reduce every process to a mechanistic picture that involves nothing more than motion of electrons, atoms or molecules, together with energy-transfer between states, continued well into the twentieth century and is still dominant today, even as regards organic chemistry. This explanatory strategy gained considerable support within organic chemistry through the discovery of the DNA structure in 1953, a discovery that refreshed hopes that chemists would soon be able to explain the reproduction of cells via a mechanism in the classical sense.[28] The twentieth century formulation of quantum mechanics did affected research strategies in chemistry but it did not eliminate all traces of the mechanistic and reductive style that this discipline had inherited from atomism.

At this point, we need to consider an important conceptual issue that lies behind the entire approach, an issue that arises whenever we seek to explain a phenomenon by referring to some lower-level microstructure. The problem arises because of the relation between parts and wholes. In general, we can say that there are different ways for parts to be together to form a whole. We can have loosely associated agglomerations, likes heaps. We can have mixtures. We can have compounds. And we can have strongly associated agglomerations like living cells. We can even have very intricately associated agglomerations that reiterate themselves, forming a whole of wholes, as in the case of the human body made up of organs, each of which is constituted of living tissue. What is the difference between these degrees of unity? Philosophers have discussed this question since ancient times. It is certainly not a new question resulting from the mechanistic worldview or from modern chemistry.[29] In nature, we find examples of all these kinds of combinations. As regards chemical thinking, a strictly mechanistic attitude would imply that the behavior of a chemical compound is exhaustively explainable in terms of our knowledge of the constituent atoms. Current knowledge however does not seem to support this view. Obviously, a compound like H_2O is not just a mixture of hydrogen and oxygen. It represents a specific state of togetherness that is different from that of mixtures. It is also different from that of organisms. Its state lies somewhere between these two grades of constitution. The individual elements, hydrogen and oxygen, can indeed exist separately, and, when combined, they are not destroyed. In philosophical terms, we can say that their ontological identity is not annihilated by the identity of the whole of which they are now part. We need to add however that they do not have any longer all of the attributes that they had before the formation of the compound. We do not seem capable of explaining all of the properties of the compound in terms of those of the constituents. Some philosophers argue that, with the formation of the compound, the individual elements undergo a kind of ontological promotion. The hydrogen atom in its combined state is not *primarily* a hydrogen atom any longer; it is now *primarily* a part of the water molecule. These last decades, philosophers have been exploring these issues in terms of emergent properties but we need not stray too far away from our main argument in this paper.[30] Suffice it to say that current explanatory strategies within chemistry include some persistent conceptual issues but, in spite of this, still show some traces of the classic mechanistic philosophy especially as regards the trend to explain marco-properties in terms of micro-properties.[31]

[27] This was defended especially by the seventeenth century Wittenberg professor of medicine, Daniel Sennert (1572-1637).

[28] For a useful historical study of how the mechanistic worldview had a role in the emergence of molecular biology, especially in the contrasting explanatory strategies of microbiologist Oswald T. Avery and theoretical physicist Max Delbrück, see Ute Deichmann, "Different Methods and Metaphysics in Early Molecular Genetics — A Case of Disparity of Research?" *History and Philosophy of the Life Sciences* 30/1 (2008): 53-78.

[29] See, for instance, Aristotle, *Metaphysics* 1040 b, 5-10; *On generation and corruption*, Book I, chapter 10. A more recent study worth mentioning is Pierre Duhem, *Le mixte et la combinaison chimique. Essai sur l'évolution d'une idée* (Paris, 1902). See also Paul A. Bogaard, "After Substance: How Aristotle's question still bears on the philosophy of chemistry," *Philosophy of Science* 73/5 (2006): 853-863.

[30] The philosophical literature on emergent properties is considerable. I offer a short overview with special attention to the concept of nature in chapter 7 of *Nature: its conceptual architecture* (Peter Lang, 2014). What I am calling ontological promotion is more evident in the case of a biological whole. When I eat a loaf of bread, I am not adding bread as such to myself. Nevertheless, some parts of the bread do indeed became part of me. What used to be part of an inanimate thing becomes part of a living thing. If we accept the idea of higher and lower forms of unity, higher and lower kinds of wholes, then we should take the chemical example of H and O combining into H_2O as analogous to the organic example. For the specific question of how major biologists Ernst Mayr, Theodosius Dobzhansky, and George Gaylord Simpson defended biology from the encroachment of the physics-inspired mechanistic approach, see Erika Lorraine Milam, "The Equally Wonderful Field: Ernst Mayr and Organismic Biology," *Historical Studies in the Natural Sciences*, 40/3 (2010): 279-317.

[31] My insistence on persistent atomistic assumptions within chemical thinking might suggest that the way chemists resort to explanations in terms of micro-attributes is diametrically opposed to the way physicists do so. When chemistry resorts to the microscopic, it reveals itself as mechanistic while, when physics resorts to the microscopic, it reveals itself as non-mechanistic, especially because of its indeterminism, non-locality and wave-particle duality. It is good to recall however that this is correct only to the extent that chemistry focuses mainly on what happens from the level of electrons, protons and neutrons upwards, and

What about other features that seem diametrically opposed to the mechanistic worldview? I will focus on three points only. Consider first the way chemistry as a discipline is related to physics. There is, of course, the trend to see chemistry as part of the far-reaching physicalist research program that seeks to reduce all objects and all motion to the fundamental interactions now acknowledged in physics, namely the electromagnetic, strong, weak and gravitational interactions.[32] In current chemistry, some reduction of this kind is always present, as is most evident in the sub-discipline of computational quantum chemistry. The results of using computers instead of chemicals have been important but we cannot take these methods to be a substitute for practical, experimental work. Computational chemistry is the theoretical counterpart of concrete practice, accounting for what is already known and exploring new possibilities, but always in need of calibration with reference to experimentally observed data. For the theory to be useful, approximations are inevitable. As the complexity of the system increases, so also the need to make approximations. For heuristic reasons therefore, it seems better not to limit chemistry to strictly reductionist explanatory methods but to assume that chemical explanation enjoys a certain degree of autonomy with respect to physics. The forms of explanation in both camps show similarities but remain distinct. For instance, a theory in physics may include theoretical entities whose existence is justified because of the theory's explanatory success. This occurs also in chemistry. Chemical theories have their own theoretical entities, entities like atomic and molecular orbitals, but these entities are different from anything that physics deals with.[33] We have here, therefore, a feature of current chemistry that is opposed to the classical mechanistic worldview. The point can be summarized as follows. If we take the classical mechanistic worldview as equivalent to today's physicalism and if we take physicalism as the idea that all scientific disciplines are reducible to physics, then chemistry today, even in its computational form, is not straightforwardly mechanistic. It is no wonder that some current philosophers working in this area are convinced that we "must abandon the *a priori* assumptions and ontological commitments of traditional mechanistic epistemology and go beyond the physicalistic reference frame [...]. Mechanistic doctrine is even a barrier for understanding the epistemology of chemistry."[34]

A second novel feature worth mentioning here is the shift of interest from the internal microstructure of substances to relations. The classical mechanistic worldview suggests that we should see chemistry as the study of substances and their constitution. The interest of current chemists however is not primary in substances as such but in relations between them. The emphasis is on the rules that govern the combinatorial possibilities of substances. These rules are comparable to the rules of grammar that determine how language can function properly. Some philosophers of chemistry call them "semiotic rules" and equate them to reaction mechanisms.[35] On this view, chemistry is "the science of the rules of possible chemical substances".[36] The term "mechanism" therefore is changing its meaning. According to these philosophers, a mechanism for chemistry is not a physical system of particles in motion but the set of signs and their rules of combination. For instance, the valence of an element, as the measure of its combining power, serves as a rule within the writing of a chemical equation. Revising the meaning of mechanism in this way implies a major shift from the classical stance. Previously, we used to assume that the highest form of understanding of a given phenomenon was the determination of the primary qualities of the entities involved and, when possible, the determination of its accurate pictorial representation. This attitude apparently implies that a direct photograph of a molecule, as we can sometimes obtain via X-ray crystallography, would be the best that chemistry could achieve. Such a photograph however would be useless for modern chemistry because what constitutes the important focus of chemical mechanisms is the set of rules of combination. A significant transformation is happening here within the very concept

rarely considers elementary particles. Physics, on the contrary, had to abandon its mechanistic foundations precisely because of its tackling phenomena at the level of elementary particles. The relatively recent sub-discipline of quantum chemistry reduces this opposition to some extent because it uncovers quantum effects at the atomic and sometimes even at the molecular level. The overall point is that there are areas of chemistry that are not influenced by quantum mechanics. These retains a mechanistic character.

[32] The way physics dominates other disciplines, and the reasons behind this phenomenon, constitute an interesting area of study; for more about this effect on chemistry in the early 1900s, see Kostas Gavroglu, "Philosophical Issues in the History of Chemistry," *Synthese* 111/3 (1997): 283-304.

[33] In philosophy of science, the term "theoretical entity" refers to an unobservable thing that scientists assume to exist so that their theory predicts observations successfully. For further discussion on this point as regards chemistry, see Eric R. Scerri and Lee McIntyre, "The Case for the Philosophy of Chemistry," *Synthese* 111/3 (1997): 213-232. A typical theoretical entity in current physics is the electron. In chemistry, molecular orbitals were first stipulated as a mathematical construct to help solve a particular set of quantum mechanical problems. They were then co-opted by organic chemists as an explanatory framework, and are now said to have been "observed" via the visualization of electron density.

[34] Joachim Schummer, "Towards a Philosophy of Chemistry," *Journal for General Philosophy of Science / Zeitschrift für allgemeineWissenschaftstheorie* 28/2 (1997): 307-336; the quoted text is from pp.308-309.

[35] E.g. J. Schummer. See *Ibid*. p. 324.

[36] *Ibid*. p. 327.

of mechanism. From the idea of a faithful pictorial representation of material elemental objects and the push-pull relations between them, mechanism has become the abstract idea of a set of rules. Although we use the same term "mechanism", the way chemists today use this word would hardly be recognizable by the mechanistic natural scientists of the modern period.

The third point of departure of modern chemistry from the mechanistic worldview concerns the persistent importance of macro-properties with respect to microstructural explanation. The classical mechanistic view, as has been shown in the first part of this paper, emphasized the importance of microstructure. It emphasized the way the corpuscles were configured in a specific way. It deconstructed the idea of substance inherited from ancient and medieval philosophy, substances as persistent macro-objects, and substituted it with that of a combination of elemental units. This is what the classic mechanistic philosophers defended. Does modern chemistry still depend upon this kind of deconstruction? It seems not. Modern chemistry, of course, still considers atomic structure of capital importance. Nevertheless, we have some clear indications that it does not make the idea of substance redundant. It does not substitute the idea of substance by a discourse about atoms. Philosopher Jaap Van Brakel argues persuasively that at least two chemical definitions of pure substance remain fully operational within current chemistry. They are in fact independent of one's convictions regarding atoms or quantum mechanics. First, "a pure substance is a substance of which the macro-properties (of one of its phases), such as temperature, density and electric conductivity, do not change during a phase-conversion (as in boiling a liquid or melting a solid)". Second, "pure chemical substances are the relatively stable products of chemical analysis and synthesis: nodes in a network of chemical reactions".[37] The plausibility of such definitions shows that, for chemistry, the way we quantify and understand the macroscopic world remains indispensable. We need not resort always to the microscopic world. The macroscopic world, in fact, remains indispensable for calibrating the microscopic. The macroscopic world guides the explanation in terms of microstructure and not the other way round, as reductionists sometimes assume. This point recalls the crucial distinction between what philosophers call the manifest image of the world, which refers to what I am here calling the macroscopic world, and the scientific image of the world, which refers to microstructure. For chemistry, the manifest image remains indispensable. We are entitled to say this because, as Van Brakel puts it, "if quantum mechanics turns out to be wrong, it would not affect all chemical knowledge […] What there is, are chemical and physical descriptions of macroscopic entities, whose identity conditions are grounded in the end in the manifest image".[38]

CONCLUSION

My original aim was to determine the extent to which current chemistry is still mechanistic in spirit. Through my initial historical overview, I illustrated that the mechanistic worldview involved some basic ingredients, such as the assumption that we can fully explain any macroscopic object and its behaviour in terms of corpuscular motion only, that the scientist's task is to determine the laws of nature, that the universe is causally closed, and that there are no final causes. This set of assumptions experienced some setbacks during the twentieth century but some of its explanatory maxims remained. In the second part of my paper, I focused on chemistry, analysed the notion of nature and that of mechanism within this discipline and determined which trends in current chemistry are still mechanistic in spirit and which are not. The results show that, as regards the urge to explain phenomena by resorting to lower-level ontological units, chemistry is still in line with some of the major tenets of the mechanistic worldview. It is not mechanistic, however, as regards its acceptance of higher-level properties that are not fully reducible to lower-level properties, as regards its assumption of some form of finality within nature, as regards its heightened focus on rules of combination, and as regards its notion of substance that is primarily associated with macroscopic attributes. Of course, much more can be said about many of the points I discuss in this paper. Moreover, my evaluation of the current situation has probably not considered all the significant trends in current chemical thinking. I hope however that what I did present here is enough to support the conclusion that current chemistry still involves some traces of mechanistic thinking but it does so without adopting the entire philosophical baggage of the seventeenth and eighteenth centuries. Today, chemists seem to use mechanistic explanatory strategies just like any other instrument. In their overall project of studying substances and their properties, they use this instrument when it helps and reject it when it hinders.[39]

[37] J. Van Brakel, "Chemistry as the Science of the Transformation of Substances," *Synthese* 111/3, (1997): 253-282; the quote is from p. 253.

[38] *Ibid*. p. 273. The distinction between manifest and scientific images of the world is, and has been, the object of sustained philosophical study. The most prominent philosophers in this area are probably Edmund Husserl and Wilfred Sellars.

[39] Thanks to Prof. Michelle Francl-Donnay and to an anonymous reviewer for *Substantia* for helpful comments to a previous version of this paper.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

OPEN ACCESS

Citation: F.T. Arecchi (2018) Cognition and Reality. *Substantia* 2(1): 41-50. doi: 10.13128/substantia-40

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.

Research Article

Cognition and Reality

F. Tito Arecchi

Emeritus of Physics-Università di Firenze and INO-CNR, Firenze, Italy
E-mail: tito.arecchi@ino.it

Abstract. We discuss the two moments of human cognition, namely, *apprehension* (A),whereby a coherent perception emerges from the recruitment of neuronal groups, and *judgment*(B),that entails the comparison of two apprehensions acquired at different times, coded in a suitable language and retrieved by memory. (B) entails *self-consciousness,* in so far as the agent who expresses the judgment must be aware that the two apprehensions are submitted to his/her own scrutiny and that it is his/her task to extract a mutual relation. Since (B) lasts around 3 seconds, the semantic value of the pieces under comparison must be decided within that time. This implies a fast search of the memory contents. As a fact, exploring human subjects with sequences of simple words, we find evidence of a limited time window , corresponding to the memory retrieval of a linguistic item in order to match it with the next one in a text flow (be it literary, or musical, or figurative). While apprehension is globally explained as a Bayes inference, judgment results from an inverse Bayes inference. As a consequence, two hermeneutics emerge (called respectively circle and coil). The first one acts in a pre-assigned space of features. The second one provides the discovery of novel features, thus unveiling previously unknown aspects and hence representing the road to reality.

Keywords. Human language, homoclinic chaos, synchronization of neural spike sequences, Bayes inference, inverse Bayes inference, circle hermeneutics, coil hermeneutics.

1. PERCEPTION, JUDGMENT AND SELF-CONSCIOUSNESS

Figs 1 and 2 introduce the difference between A-apprehension or perception that rules the motor reactions of any brainy animal, and B-language ,only humans, and that provides judgments.

Figs 3 and 4 show why the scientific program is a linguistic one and what is the reason of its success.

With this in mind, we explore whether and how Cognition unveils Reality…

Following the philosophy of cognition of Bernard Lonergan [Lonergan], I discuss two distinct moments of human cognition, namely, *apprehension* (A) whereby a coherent perception emerges from the recruitment of neuronal groups, and *judgment* (B) whereby memory recalls previous (A) units coded in a suitable language; these units are compared and from comparison it follows the formulation of a judgment.

ISSN 1827-9635 (print) | ISSN 1827-9643 (online) | DOI: 10.13128/substantia-40

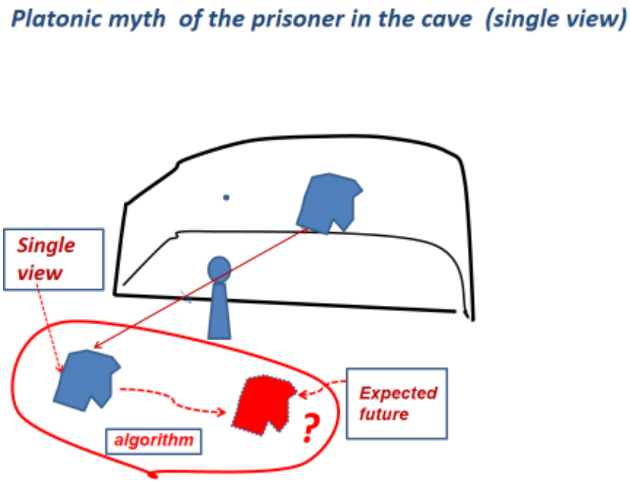


Figure 1. Plato said that we see the shadows of things, like a prisoner constrained to view the end of a cave and forbidden to turn and see the outside world. This occurs indeed in perceptual tasks, where the sensorial stimuli are interpreted by "algorithms" and generate (within1 sec) a motor reaction. The procedure is common to all brainy animals.

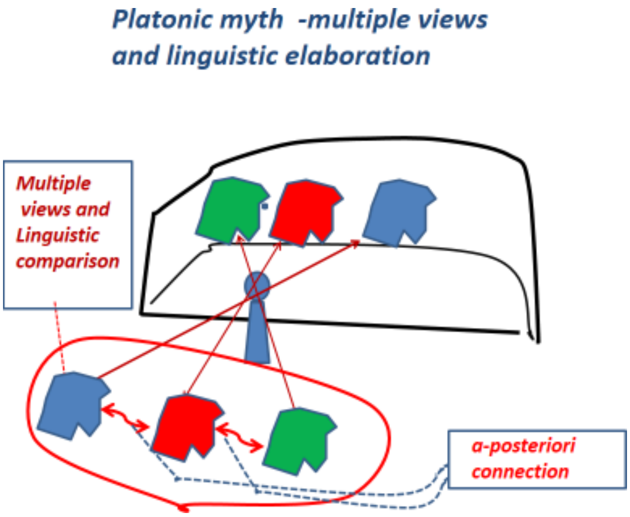

Platonic myth -multiple views
and linguistic elaboration
Multiple views and Linguistic comparison
a-posteriori connection

Figure 2. In linguistic operations, humans code a perception in a linguistic code andretrieve it by short term memory (around 3 sec)comparing it to a successive coded perception. From the comparisonit emerges a connection that increases the details of the observed thing.

The first moment, (A), has a duration around 1 sec; its associated neuronal correlate consists of the synchronization of the EEG (electro-encephalo-graphic) signals in the so-called gamma band (frequencies between 40 and 60 Hz) coming from distant cortical areas .It can be described as an interpretation of the sensorial stimuli on the basis of available algorithms, through a Bayes inference [Arecchi,2007; Doya et al.].

Precisely, calling ***h***(h= hypothesis) the interpretative hypotheses in presence of a sensorial stimulus ***d*** (d=datum), the Bayes inference selects the most plausible hypothesis ***h****,that determines the motor reaction, exploiting a memorized algorithm ***P(d|h),*** that represents the conditional probability that a datum ***d*** be the consequence of an hypothesis ***h.***

The ***P(d|h)*** have been learned during our past; they represent the equipment whereby a cognitive agent faces the world. By equipping a robot with a convenient set of ***P(d|h),*** we expect a sensible behavior.

The second moment, (B),entails a comparison between two apprehensions (A) acquired at different times, coded in a given language and recalled by the memory. If, in analogy with (A), we call ***d*** the code of the second apprehension and ***h**** the code of the first one, now – at variance with (A) ***h**** is already given; instead, the relation ***P(d|h)*** which connects them must be retrieved; it represents the ***conformity*** between ***d*** and ***h*,*** that is, the best interpretation of ***d*** in the light of ***h*.***

Thus, in linguistic operations, we compare two successive pieces of the text and extract the conformity of the second one on the basis of the first one. This is very different from (A), where there is no problem of conformity but of plausibility of ***h**** in view of a motor reaction.

Let us make two examples: a rabbit perceives a rustle behind a hedge and it runs away, without investigating whether it was a fox or just a blow of wind.

On the contrary, to catch the meaning of the 4-th verse of a poem, we must recover the 3-d verse of that same poem, since we do not have a-priori algorithms to provide a satisfactory answer.

Once the judgment, that is, the ***P(d|h)*** binding the codes of the two linguistic pieces in the best way, has been built, it becomes a memorized resource to which to recur whenever that text is presented again. It has acquired the status of the pre-learned algorithms that rule (A).

However-at variance with mechanized resources-whenever were-read the same poem, we can grasp new meanings that enrich the previous judgment ***P(d|h).*** As in any exposure to a text (literary, musical, figurative) a re-reading increases our understanding.

(B) requires about 3 seconds and entails *self-consciousness,* as the agent who expresses the judgment must be aware that the two successive apprehensions are both under his/her scrutiny and it is up to him/her to extract the mutual relation.

At variance with (A), (B) does not presuppose an algorithm, but rather it builds a new one through An ***inverse Bayes procedure*** [Arecchi, 2010]. This construction of a new algorithm is a sign of

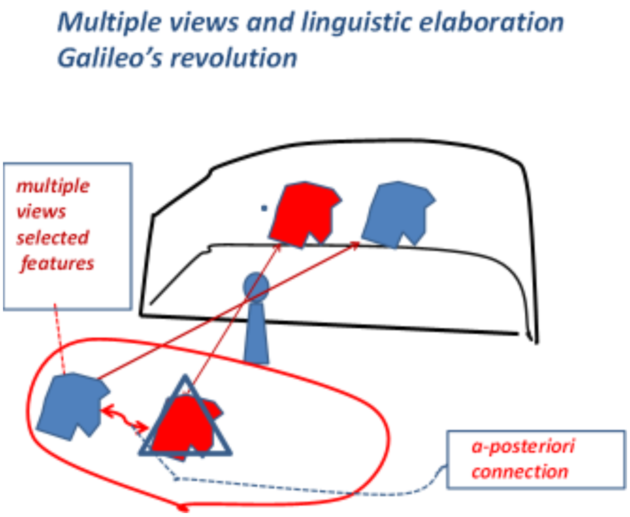

Multiple views and linguistic elaboration
Galileo's revolution
multiple views selected features
a-posteriori connection

Figure 3. The scientific program is a linguistic task. Galileo's approach consists of extracting mathematical features; it implies a linguistic operation, according to Fig.2.

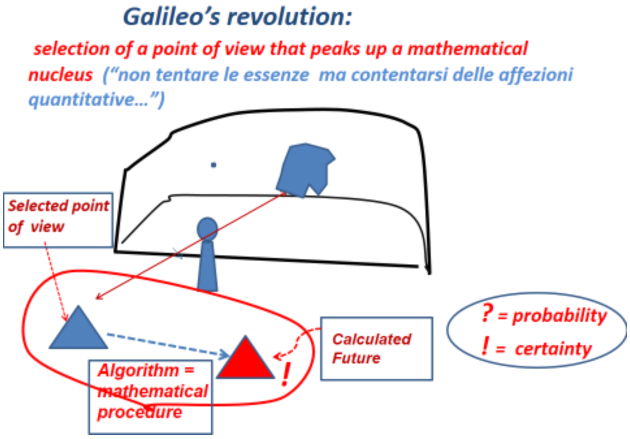

Galileo's revolution:
selection of a point of view that peaks up a mathematical nucleus ("non tentare le essenze ma contentarsi delle affezioni quantitative…")
Selected point of view
? = probability
! = certainty
Calculated Future
Algorithm = mathematical procedure
!

Figure 4. Once the object under investigation is reduced to a collection of mathematical features, one applies millennia of mathematical wisdom and predicts the future behavior.

Creativity and *decisional freedom*

Here the question emerges: can we provide a computing machine with the (B) capacity, so that it can emulate a human cognitive agent, as expected in theTuring test? The answer is NOT, because (B) entails non-algorithmic jumps, insofar as the inverse Bayes procedure generates an *ad hoc* algorithm, not available previously.

The scientific endeavor can not be carried on by an AI (artificial intelligence) device, since it entails a linguistic step, as shown in Fig. 3. Fig. 4 explains why Galileo's program provides certainties, rather than probabilistic expectations.

2. THE BRAIN OPERATIONS-ROLE OF HOMOCLINIC CHAOS

Let us introduce "deterministic chaos".[Arecchi,2004 b] Since Poincaré (1890) we know that a dynamical system is extremely sensitive to the initial conditions. That yields the so called "butterfly effect" whereby a tiny shift in the initial conditions yields a large difference in course of time . Precisely, a difference of initial conditions induces a divergence of the dynamical trajectories in course of time. In the case of a meteo dynamical model, accounting for most atmospheric features (wind, pressure, humidity, etc) but neglecting the motion of a butterfly wing could lead to a wrong prediction(from sunny to rainy).Loss of the initial information occurs over a time t whose inverse is called ***K*** (Kolmogorov entropy). In the meteo model such a ***t*** may be days, in a dynamical model of the solar system it takes millions of years.

We call *geometric chaos* the above trajectory divergence.

Another type of chaos,that we call *temporal chaos,* consists of regular closed orbits that however repeat at irregular times (Fig. 5).

A single neuron in the brain undergoes temporal chaos and its electrical output consists of a train of spikes (each one high 100mV and lasting 3 ms). The minimal inter-spike separation is 3 ms; the average separation is 25 ms in the so-.called g band of the EEG(electro-encephalo-gram).

A neuron communicates with other neurons in two ways [Arecchi,2004 a, Singer,Womelsdorf. & Fries]:

- either directly , by coupling its spike train to another neuron via an electric line called *axon,*
- or indirectly, by building with nearby neurons a local potential (detectable as an EEG signal) and providing a signal χ to a distant neuron, that consequently re-adjusts its firing rate.

Fig. 6 shows the direct synchronization of two trains over a time Dt. The neurons involved in the coupling are confined in a thin layer of the brain (thickness 2 mm) called the *cortex*. Groups of nearby neurons contribute to a common task forming a specialized area that builds global interactions with other areas (Fig. 7).

The areas are visualized via the amount of oxygenated blood required by a working region and visualized by *f-MRI* (functional magnetic resonance imaging).

Fig. 8 visualizes the competition between two neuron groups ***I*** and ***II*** fed by the same sensorial (***bottom-up***) stimulus, but perturbed ***(top-down)*** by different interpretational stimuli provided by the long term

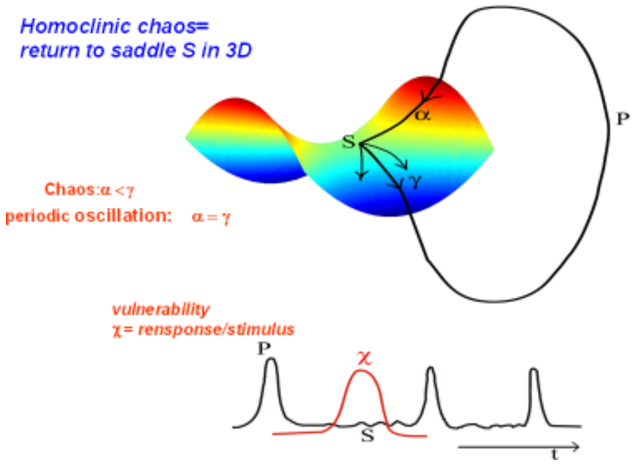


Figure 5. Homoclinic chaos; the dynamic trajectory is a closed orbit starting from S (saddle point) and returning to it. Projecting on a single direction, we observe spikes P repeating in time. The time separation between two spike occurrences depends on the relations between α and g, thus it can be controlled by a voltage applied to S, as the signal χ.

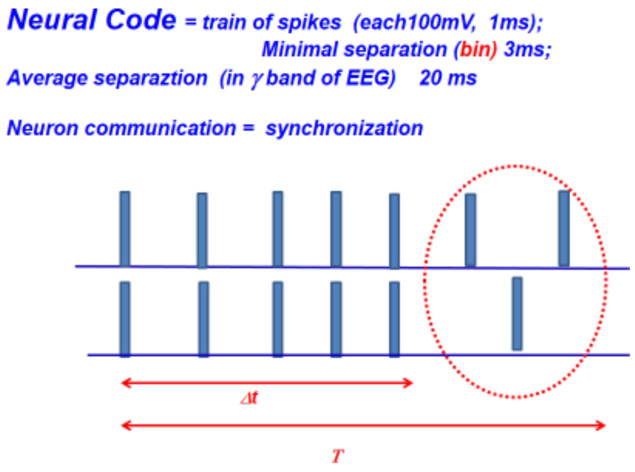
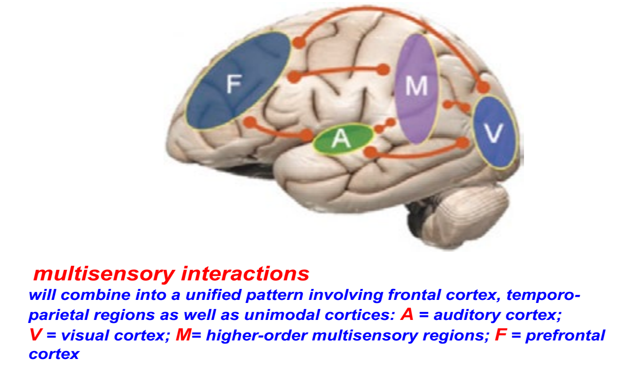

Neural Code = train of spikes (each100mV, 1ms);
Minimal separation (bin) 3ms;
Average separaztion (in γ band of EEG) 20 ms
Neuron communication = synchronization
Δt
T

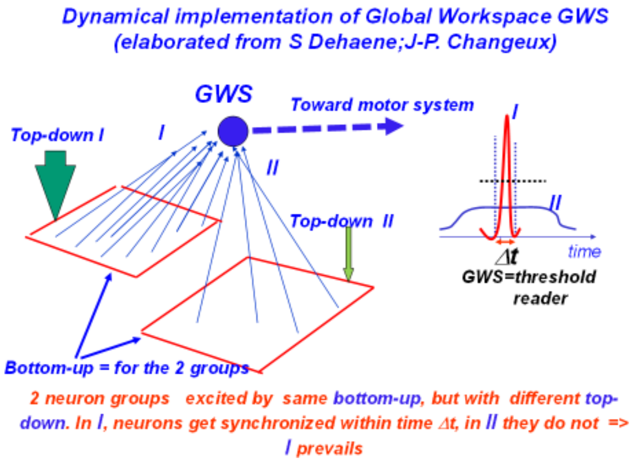
Figure 6. Direct coupling of two neurons by synchronized spike trains; synchronization missed after Dt for an extra-spike in the upper train.

memory. ***I*** wins, as the corresponding top-down stimulus succeeds in synchronizing the neuron pulses of this group better than in group ***II.*** This means that, over a time interval Δ***t,*** neurons of ***I*** sum up coherently their signals, whereas neurons of ***II*** are not co-ordinated, hence yielding a smaller sum. As a consequence a reader GWS (= global workspace, name given to the cortical area where signals from different areas converge; it is located in area F of Fig.7) reads within Δ***t***a sum signal overcoming a suitable threshold and hence eliciting a motor response [Dehaene].



multisensory interactions
will combine into a unified pattern involving frontal cortex, temporo-parietal regions as well as unimodal cortices: A = auditory cortex; V = visual cortex; M= higher-order multisensory regions; F = prefrontal cortex

Figure 7. Topology of specialized cortical areas ,each one being active as a large collection of synchronized neurons; mutual communication occurs via EEG signals.


Dynamical implementation of Global Workspace GWS
(elaborated from S Dehaene;J-P. Changeux)
GWS
Toward motor system
Top-down I
I
II
Top-down II
Bottom-up = for the 2 groups
I
II
Δt
time
GWS=threshold reader

2 neuron groups excited by same bottom-up, but with different top-down. In I, neurons get synchronized within time Δt, in II they do not => I prevails

Figure 8. Competition of two cortical areas with different degrees of synchronization.

Thus, the winning interpretation driving the motor system is that provided by ***I***.

What represented in Fig. 8 models the mechanism *(A)* common to any animal with a brain.

3. PERCEPTION AS A BAYES INFERENCE

Neurosciences hypothesize a collective agreement of crowds of cortical neurons through the mutual synchronization of trains of electrical pulses (spikes) emitted individually by each neuron [Singer et al., Dehaene et al.] .The neuro-scientific approach is summarized in Fig. 8.

However, a global description of the above process can be carried on in probabilistic terms, without recurring to the details of the process.

In 1763, Thomas Bayes, looking for a reliable strategy to win games, elaborated the following probabilistic argument[Bayes]. Let us formulate a manifold of hypotheses *h* about the initial situation of a system, attributing to each hypothesis a degree of confidence expressed by an a priori probability

$$P(h).$$

Any hypothesis, introduced as input into a *model* of evolution, generates data. Let us assume that we know the model and, hence, can evaluate the probability of the *data conditioned* by a specific hypothesis *h*; we write it as

$$P(data|h).$$

The model is like an instruction to a computer, thus we call it ***algorithm***; it generates different data for different *h*. If then we perform a measurement and evaluate the probability

$$P(data)$$

of the data, we must conclude that there is an *h* more plausible than the other ones, precisely the one that maximizes the probability conditioned by the data

$$P(h|data),$$

that we call the a posteriori probability of *h*and denote as *h**.

This procedure is encapsulated in the formula, or theorem, of Bayes, that is

$$P(h^*)=P(h|data) = P(h)\ [P(data|h)/P(data)]$$

To summarize, the a posteriori probability of *h*, conditioned by the observed data, is given by the product of the a priori probability of *h*, times the probability *P(data|h)* of the data conditioned by a given *h*, that we call the *model,* and divided by the probability *P(data),* based on a previous class of trials (Fig. 9).

Fig. 10 summarizes the whole perception procedure, that is initiated by an external stimulus and concluded by a motor reaction [Arecchi, 2007].

Successive applications of the theorem yield an increasing plausibility of *h**; it is like climbing a mountain of probabilities along its maximum slope, up to the peak. After each measurement of the data and consequent evaluation of the a posteriori *h**, we reformulate a large number of new a priori *h* relative to the new situation, and so on (Fig. 11). Notice that Darwinian evolution by *mutation* and successive *selection* of the best fit mutant is a sequential implementation of Bayes theorem.

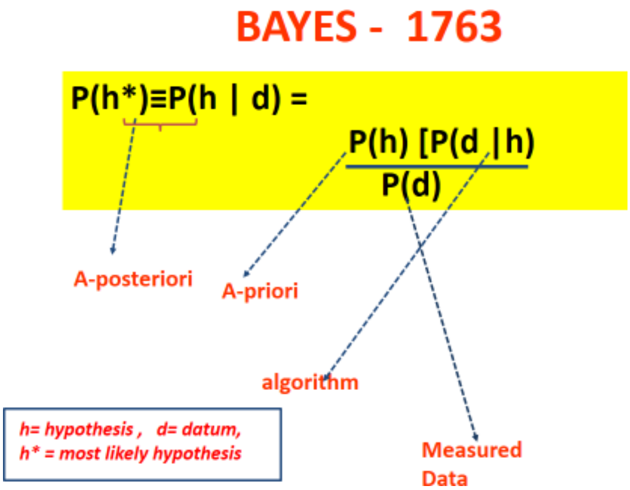


Figure 9. Bayes inference.

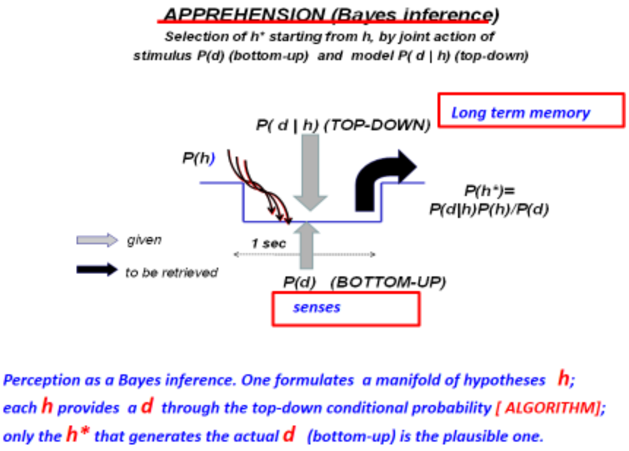
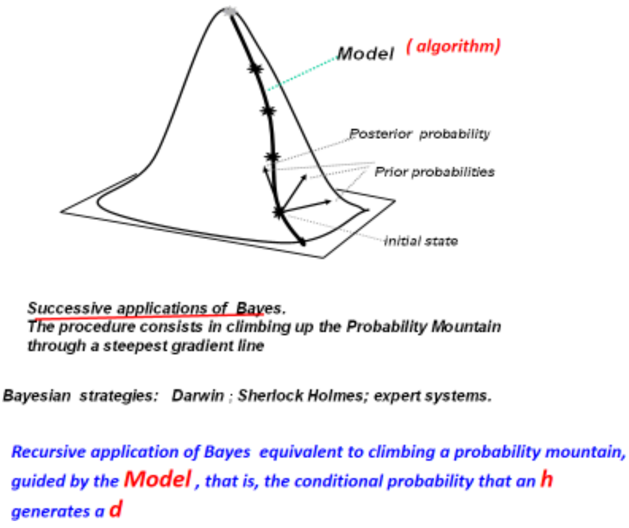

APPREHENSION (Bayes inference)
Selection of h* starting from h, by joint action of
stimulus P(d) (bottom-up) and model P(d | h) (top-down)
Long term memory
P(d | h) (TOP-DOWN)
P(h)
P(h*)=
P(d|h)P(h)/P(d)
given
1 sec
to be retrieved
P(d) (BOTTOM-UP)
senses
Perception as a Bayes inference. One formulates a manifold of hypotheses h;
each h provides a d through the top-down conditional probability [ALGORITHM];
only the h* that generates the actual d (bottom-up) is the plausible one.

Figure 10. Starting with a large number of presumed hypotheses h, the occurrence of the data selects the *h** that satisfies the above relation and drives a suitable reaction.

4. LINGUISTIC OPERATIONS AS INVERSE BAYES

In Fig. 11 the recursive application of Bayes using the same algorithm- or model- is visualized as climbing a probability mountain. The bit length of the algorithm is the Algorithmic Complexity of the cognitive task.

However, in everyday life we experience jumps toward different algorithms, that means going to climb different mountains (Fig. 12). The associated multiplicity of choices corresponds to attributing different meanings

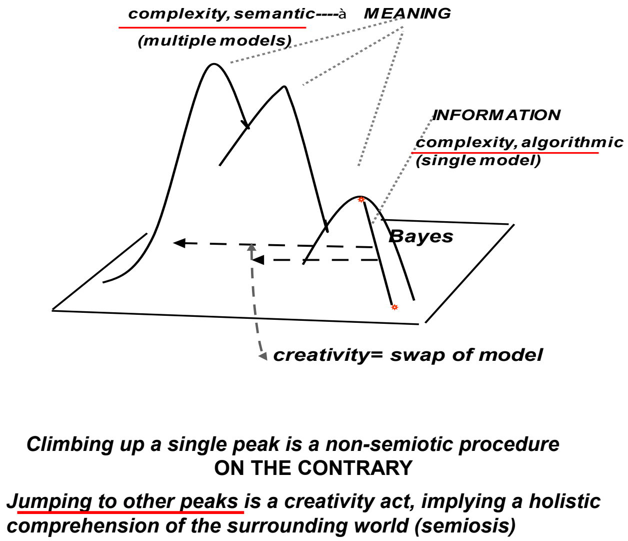


Successive applications of Bayes.
The procedure consists in climbing up the Probability Mountain through a steepest gradient line

Bayesian strategies: Darwin ; Sherlock Holmes; expert systems.

Recursive application of Bayes equivalent to climbing a probability mountain, guided by the Model , that is, the conditional probability that an h generates a d

Figure 11. Recursive application of Bayes is equivalent to climbing a probability mountain, guided by the Model ,that is, the conditional probability that an hypothesis generates a datum. This strategy is common e.g. to Darwin evolution and to Sherlock Holmes criminal investigation; since the algorithm is unique, it can be automatized in a computer program (expert system).



Climbing up a single peak is a non-semiotic procedure
ON THE CONTRARY
Jumping to other peaks is a creativity act, implying a holistic comprehension of the surrounding world (semiosis)

Figure 12. Comparison of two different complexities, namely, i)the algorithmic C. , corresponding to the bit length of the program that enables the expert system to a recursive Bayes; and ii)semantic C., corresponding to the occurrence of different models.

to the input data; the number of alternative choices will be called Semantic Complexity [Arecchi, 2007].

This swap of the model is a creative jump proper of language operations.

It is the root of Goedel-1931 incompleteness theorem and Turing-1936 halting problem for a computer, as discussed in a previous paper [Arecchi, 2012].

Altogether different from (A) is the situation for *(B),* that – implying the comparison between different apprehensions coded in the same language (literary, musical, figurative, etc.) – represents an activity exclusively human.

In fact, the second moment *(B)* entails the comparison of two apprehensions acquired at different times, coded in the same language and recalled by the memory.

(B) lasts around 3 sec; it requires ***self-consciousness,*** since the agent who performs the comparison must be aware that the two non simultaneous apprehensions are submitted to his/her scrutiny in order to extract a mutual relation.

At variance with*(A), (B)* does not presuppose an algorithm but it rather builds a new one through an ***Inverse Bayes procedure*** introduced by Arecchi [Arecchi,2010]. This construction of a new algorithm is the source of *creativity* and *decisional freedom.*

Language indeed permits an infinite use of finite resources [Humboldt].

It is the missing step in Turing's claim that human intelligence can be simulated by a machine [Turing].

The first scientist who explored the cognitive relevance of the 3sec interval has been Ernst Pöppel [Pöppel 2004, 2009].

This new temporal segment has been little explored so far. All the so-called *"neural correlates of consciousness" (NCC)* are in fact electrical (EEG) or functional magnetic resonance *(fMRI)* tests of a neuronal recruitment stimulating a motor response through a GWS (see Fig. 8); therefore they refer to (A). In such a case, rather than *consciousness*, one should call itperceptual *awareness,*that we have in common with brainy animals.

Fig. 13 shows how an inverse Bayes procedure provides the best comparisons of two successive pieces of a linguistic text, thus generating a judgment.

While in perception we compare sensorial stimuli with memories of past experiences, in judgment we compare a piece of a text coded in a specific language (literary, musical, figurative) with the preceding piece, recalled via the short term memory. Thus we do not refer to an event of our past life, but we compare two successive pieces of the same text.

Such an operation requires that:

i) The cognitive agent be aware that he/she is the same examiner of the two pieces under scrutiny;

ii) The interpretation of the second piece based upon the previous one implies to have selected the most appropriate meanings of the previous piece in order

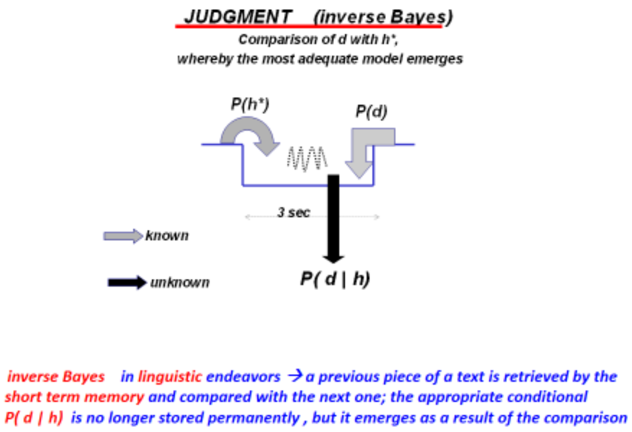

JUDGMENT (inverse Bayes)
Comparison of d with h*,
whereby the most adequate model emerges
P(h*)
P(d)
3 sec
known
unknown
P(d | h)
inverse Bayes in linguistic endeavors → a previous piece of a text is retrieved by the
short term memory and compared with the next one; the appropriate conditional
P(d | h) is no longer stored permanently , but it emerges as a result of the comparison

Figure 13. The inverse Bayes procedure that occurs in linguistic endeavors, whereby a previous piece ofa text is retrieved by the short term memory and compared with the next one: the appropriate conditional probability is no longer stored permanently but it emerges as a result of the comparison (judgmentand consequent decision).

to grant the best conformity (from a technical point of view, this conformity is what in the philosophy of cognition of Thomas Aquinas was defined as truth = ***adaequatio intellectus et rei** (loosely translated as : conformity between the intellectual expectation and the object under scrutiny)*

In Fig. 10 we have generically denoted as *top-down* the bunch of inner resources (emotions, attention) that, upon the arrival of a bottom-up stimulus, are responsible for selecting the model *P(d|h)* that infers the most plausible interpretation *h** driving the motor response. The *focal attention* mechanisms can be explored through the so-called NCC (Neural Correlates of Consciousness) [Koch] related to EEG measurements that point the cortical areas where there is intense electrical activity producing spikes, or to f-MRI (functional magnetic resonance imaging) that shows the cortical areas with large activity which need the influx of oxygenated blood.

Here one should avoid a current confusion. The fact that a stimulus elicits some emotion has NOTHING to do with the judgment that settles a linguistic comparison. As a fact, NCC does not reveal self-consciousness, but just the awareness of an external stimulus to which one must react.

Such awareness is common to animals, indeed many tests of NCC are done on laboratory animals.

It is then erroneous to state that a word isolated from its context has an aesthetical quality because of its musical or evocative power. In the same way, it is erroneous to attribute an autonomous value to a single spot of color in a painting independently from the comparison with the neighboring areas.All those "excitations" observed by fMRI refer to emotions related to apprehension and are inadequate to shed light on the judgment process.

The different semantic values that a word can take are associated with different emotions stored in the memory with different codes (that is, spike trains). Among all the different values, the cognitive operation "judgment" selects that one that provides the maximum synchronization with the successive piece.

Thus emotions are necessary but not sufficient to establish a judgment. On the other hand, emotions are necessary and sufficient to establish the apprehension as they represent the algorithms of the direct Bayes inference. This entails a competition in GWS (Fig. 8) ,where the winner is the most plausible one; whereas in the judgment-once evoked the panoply of meanings to be attributed to the previous piece- these meanings do not compete in a threshold process, but they must be compared with the code of the next word in order to select the best interpretation.

Recent new terms starting with *neuro*-(as e.g. neuro-ethics, neuro-aesthetics, neuro-economics, neuro-theology) smuggle as shear emotional reactions decisions that instead are based on judgments. The papers using those terms overlook the deep difference between apprehensions and judgments. The question is discussed in detail in the Conclusions.

A very successful neurological research line deals with *mirror neurons*, that is, neurons that activate in subjects (humans or higher animals) observing another subject performing a specific action, and hence stimulate mimetic reactions [Rizzolatti]. Here too, we are in presence of mechanisms(empathy)limited to the emotional sphere, that is, very useful for formulating an Apprehension, but not a Judgment.

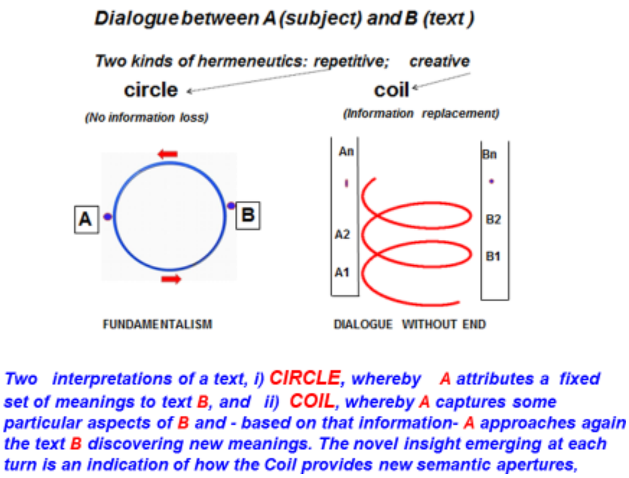
5. TWO DIFFERENT HERMENEUTICS, THAT IS, INTERPRETATIONS OF COGNITIVE DATA

Fig. 14 shows how a cognitive agent ***A*** reads an object ***B***. The *CIRCLE* refers to a Bayes cognition, whereby an algorithm is taken as necessary and sufficient to generate knowledge of B.

Whenever ***A*** reconsiders ***B***, he/she finds the same ***B*** already memorized.

On the contrary, expressing the knowledge in a language and comparing successive pieces by inverse Bayes, entails an increase of details of ***B (B1,B2***, etc.) that improve the cognition of the agent ***(A1,A2***, etc).

As for the *CIRCLE*, in information science, an **ontology** is a formal definition of the properties, and


Dialogue between A (subject) and B (text)
Two kinds of hermeneutics: repetitive; creative
circle
coil
(No information loss)
(Information replacement)
A
B
An
Bn
A2
B2
A1
B1
FUNDAMENTALISM
DIALOGUE WITHOUT END

Two interpretations of a text, i) CIRCLE, whereby A attributes a fixed set of meanings to text B, and ii) COIL, whereby A captures some particular aspects of B and - based on that information- A approaches again the text B discovering new meanings. The novel insight emerging at each turn is an indication of how the Coil provides new semantic apertures,

Figure 14. Two kinds of interpretation of a text, or hermeneutics, namely, the CIRCLE, whereby the interpreter A attributes a finite and fixed set of meanings to the text B, and the COIL, whereby A captures some particular aspects of B and-based on that information- A approaches again the text B discovering new meanings. The novel insight provided at each coil is an indication of how language provides new semantic apertures.

mutual relationships of the entities that exist for a particular domain of discourse. An ontology lists the variables needed for some set of computations and establishes the relationships between them. For instance, the booklet of the replacement parts of a brand of car is the ontology of that car. The fields of artificial intelligence create ontologies to limit complexity and to organize information. The ontology can then be applied to problem solving. Nothing is left out; we call this cognitive approach *"finitistic"* as no new insight is provided by repeated trials.

On the contrary, in any human linguistic endeavor (be it literary, or musical or figurative) ***A*** starts building a provisional interpretation ***A1*** of the text ; whenever ***A*** returns to ***B***, he/she has already some interpretational elements to start with, and from there ***A*** progresses beyond , grasping new aspects ***B2, B3***…and hence going to ***A2*** and so on *(COIL)*. To carry on a COIL program, we do not need a large amount of resources; language makes an infinite use of finite resources [Humboldt].

The *COIL* hermeneutics describes also the inter-personal dialogue. If the object ***B*** of cognition is a human person as ***A***, then the changes ***B1,B2,*** etc are not only due to an increased knowledge by ***A***, but also to an intrinsic change of ***B*** who re-adjusts his/her relation with ***A.***

Thus, if ***B*** is another human subject, then ***B*** undergoes similar hermeneutic updates as ***A***; this is a picture of the dialogical exchange between two human beings. (persons).

6. CONCLUSIONS- TWO ASPECTS OF LINGUISTIC CREATIVITY

We conclude by stressing two well known aspects of linguistic creativity. First, if we start a linguistic endeavor, a wealth of possible situations emerge , giving rise to ambiguous behaviors as it occurs in most products of human creativity, that – like the Etruscan Chimera – display apparently contradictory behaviors, from Ulysses to Dom Quixote (Fig. 15). The onset of Chimeras is explained in Fig. 16 as the lack of an external referent ***B***.

Altogether different is the what takes place when the language is interpreting scientific observations. In fact, the repeated comparison extracts elements of reality, as hinted in the COIL hermeneutics.

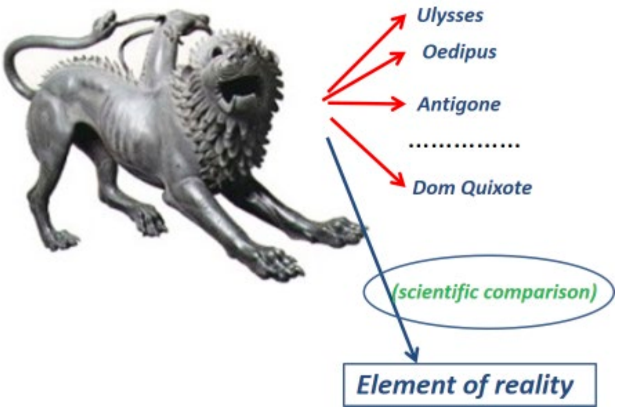
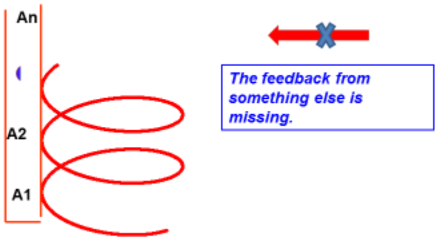


Figure 15. A linguistic action that proceeds from a known piece toward an unknown one is like the Etruscan Chimera: it can generate mutually conflicting behaviors , as it occurs in most characters ,from Ulysses to Dom Quixote.When instead the linguistic comparison regards two observed items (as it occurs in reading the verses of a Poem, but also in scientific observations), then we really increase our personal knowledge with an element of reality.

In the COIL hermeneutics, A compares with B, where B is the external reality that drives the successive choices of A. On the contrary, in the creation of new pieces as in literary creation, A evolves but without an external referent



The feedback from something else is missing.

Figure 16. How chimeras emerge in linguistic creations.

Applying our hermeneutics to the scientific program, we have two possible approaches. As the size of the observed world increases from a few particles to many, within a universal scientific description associated with a fixed-algorithm (as an AI tool would operate) we witness an exponential increase of the size ***C*** of the computational program (called the *algorithmic complexity*) as well as a reduction of the time interval ***t*** over which predictions are reliable, that is, an increase of the *Kolmogorov entropy* ***K=1/t*** (Fig. 17).

A more efficient scientific program consists of linguistic comparisons of different situations, with the help of inverse Bayes inference, applying non-algorithmic jumps as the horizontal lines of Fig. 12. We are somewhat manipulating the set of attributions of the item under study, emphasizing novel aspects and overlooking some previous ones. Such a change of paradigm [Kuhn] leads to novel theories with lower ***C*** and ***K***. A very familiar example is the formulation of Maxwell's electromagnetic equations, unifying electric, magnetic and optical phenomena. The novel values we attribute to some features leads to the so called *effective science* [Hartmann, 2001]. An outstanding example is offered by Landau theory of phase transitions.

In Fig. 18 we list the so-called ***neuro-by products*** *(neuro-ethics, neuro-aesthetics, neuro-economics,* etc.) meaning that the decisions in that specific matter result from brain processes signaled by f-MRI. One would attribute to emotions the role that is instead proper of a linguistic act, thus requiring an inverse Bayes.

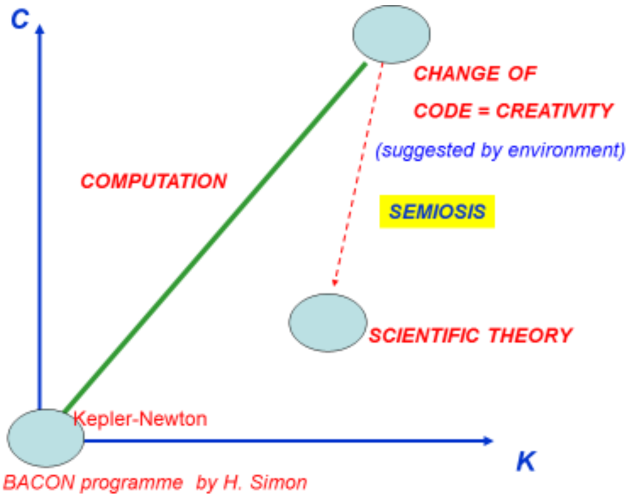
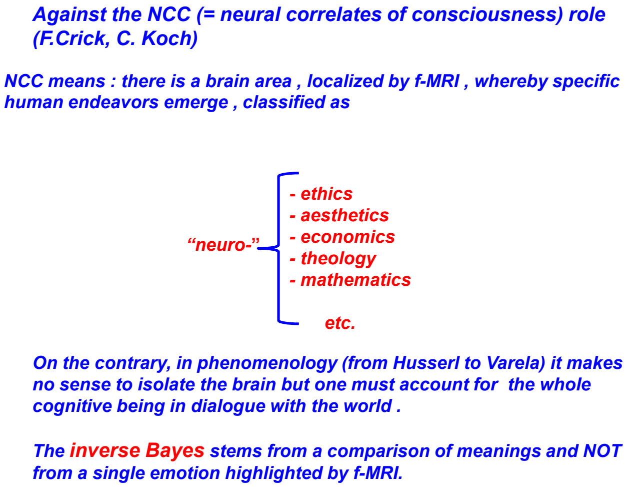


Figure 17. Normal science vs. paradigm shift → effective science [Hartmann, Kuhn]. Let **C** be the bit length of the algorithm and **K** the Kolmogorov entropy, i.e., the inverse of the time t beyond which the initial information is lost by dynamical chaos. A computer program that evaluates Kepler's orbits, as BACON, has small **C** and **K**. As the physical system gets richer, both **C** and **K** increase and a scientific search carried on by an AI system would be affected by higher and higher **C** and **K**. However, a linguistic actor as a human scientist can act by a "jump of paradigm", that is, change code and introduce a new scientific theory (effective description) with low **C** and **K**.



Against the NCC (= neural correlates of consciousness) role (F.Crick, C. Koch)

NCC means : there is a brain area , localized by f-MRI , whereby specific human endeavors emerge , classified as

"neuro-"
- ethics
- aesthetics
- economics
- theology
- mathematics
etc.

On the contrary, in phenomenology (from Husserl to Varela) it makes no sense to isolate the brain but one must account for the whole cognitive being in dialogue with the world .

The inverse Bayes stems from a comparison of meanings and NOT from a single emotion highlighted by f-MRI.

Figure 18. It is fashionable to speak of neuro-ethics, neuro-aesthetics, neuro-economics, etc., overlooking the essential role of inverse Bayes. Hence, the neuro-xxxx should be regarded as scientific misunderstandings.

Precisely, the emotions select a particular meaning to be assigned to the piece of text *h** in order to optimize its matching with the next one *d;* they are crucial for maximizing *P(h*|d)* but they are not all , just a piece of the whole tapestry (see Fig. 13).

Final consideration. Does AI operate by inverse Bayes in a linguistic elaboration? Answer: only in a very limited way. Indeed, AI refers to a built-in *"ontology"*, consisting of a large, yet finite, list of properties of each item. (See e.g. the informational use of the term *ontology* to list the component parts of a car). Thus AI can build *P(h*|d)* for each *h** and *d* , and it can do it in a much faster way than a human. However the human exploits emotions in selecting the meaning of *h**, thus he/she can go beyond the large, yet limited ontology available to AI and attribute to *h** novel aspects previously unknown. This is the creativity of human language, already addressed by Humboldt , and absent in AI.

BIBLIOGRAPHY

Arecchi F.T. (2004a). Chaotic neuron dynamics, synchronization and feature binding, *Physica A* 338: 218-237.

Arecchi F.T. (2004b). Caos e complessità nel vivente (lezioni tenute alla Scuola Universitaria Superiore, Pavia), IUSS Press-Pavia, pp. 248

Arecchi F.T. (2007). Complexity, Information Loss and Model Building: from neuro- to cognitive dynamics, *SPIE Noise and Fluctuation in Biological, Biophysical, and Biomedical Systems,* Paper 6602-36.

Arecchi F.T. (2010). Dynamics of consciousness: complexity and creativity, *The Journal of Psychophysiology*, 24(2), 141-148.

Arecchi F.T. (2012). Fenomenologia della coscienza: dall'apprensione al giudizio, in "*...e la coscienza? FENOMENOLOGIA, PSICO-PATOLOGIA, NEURO-SCIENZE*" a cura di A. Ales Bello.

Bayes T. (1763). An Essay toward solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society of London* 53, 370-418

Dehaene S., Sergent C. & Changeux J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception, *Proceedings of the National Academy of Science (USA)*, 100(14), 8520-8525.

Doya K., Ishii S. & Pouget A. (Eds.) (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding.* MIT Press.

Galilei G. (1612). Terza lettera a M. Welser sulle macchie solari, *Opere, vol. V* (pp. 187-188). Firenze: Edizione Nazionale, Barbera 1968.

Hartmann S. (2001). Effective field theories, reduction and scientific explanation, *Studies in History and Philosophy of Modern Physics* 32B, 267-304.

Humboldt W.(1836). *On Language: On the Diversity of Human Language Construction and its Influence on the Mental Development of the Human Species.* Trans. Peter Heath, Cambridge: CUP, 1988.

Koch C. (2004). *The quest for consciousness: a neurobiological approach*. Englewood, US-CO: Roberts & Co. Publishers.

Kuhn T. (1962 1st ed.; 1996, 3rd ed.). *The Structure of Scientific Revolutions,* University of Chicago Press.

Lonergan B.S.J (1970). *Insight, A study of human understanding,* Philosophical Library, New York.

Pöppel E. (2004). Lost in time: a historical frame, elementary processing units and the 3-second window, *Acta Neurobiologiae Experimentalis*, 64: 295-301.

Pöppel E.(2009). Pre-semantically defined temporal windows for cognitive processing, *Philosophical Transactions of the Royal Society B*, 364,1887-1896.

Rizzolatti G. et al. (1996). Premotor cortex and the recognition of motor actions, *Cognitive Brain Research*, 3(2), 131-141.

Singer W. & Gray C.M. (1995). Visual feature integration and the temporal correlation hypothesis, *Annual Reviews of Neuroscience* 18, 555-586.

Singer W. (2007). Binding by synchrony, *Scholarpedia,* 2(12), 1657.

Tommaso D'Aquino (1269). *Summa theologica, Pars Prima,* Quaestio 16.

Turing A. (1950). Computing Machinery and Intelligence. *Mind* 59, 433-460.

Womelsdorf T. & Fries P. (2007). The role of neuronal synchronization in selective attention, *Current Opinion in Neurobiology*, 17, 154-160.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Citation: B.D. Hughes, B.W. Ninham (2018) A Correspondence Principle. *Substantia* 2(1): 51-76. doi: 10.13128/substantia-41

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.

Permission to reproduce. Republished from *Physica A*, **2016**, *443*, 495-517. With permission from Elsevier. Copyright 2016.

Research Article

A Correspondence Principle

Barry D. Hughes[1,*] and Barry W. Ninham[2]

[1] *School of Mathematics and Statistics, University of Melbourne, Victoria 3010 Australia.*
**Corresponding author*
[2] *Department of Applied Mathematics, Research School of Physical Sciences and Engineering, Australian National University, ACT 0200, Australia*
E-mail: barrydh@unimelb.edu.au (Barry D. Hughes),
barry.ninham@anu.edu.au (Barry W. Ninham)

Abstract. A single mathematical theme underpins disparate physical phenomena in classical, quantum and statistical mechanical contexts. This mathematical "correspondence principle", a kind of wave–particle duality with glorious realizations in classical and modern mathematical analysis, embodies fundamental geometrical and physical order, and yet in some sense sits on the edge of chaos. Illustrative cases discussed are drawn from classical and anomalous diffusion, quantum mechanics of single particles and ideal gases, quasicrystals and Casimir forces.

Keywords. Classical analysis, quantum mechanics, statistical mechanics, random walks and Lévy flights, quasicrystals, Casimir forces.

Physics is not just Concerning the Natures of Things, but Concerning the Interconnectedness of all the Natures of Things [1]

1. INTRODUCTION

One of the more insightful critics of relatively recent mathematics–from inside the profession–is Morris Kline, who has made the following observation [2]. "It behooves us to learn why, despite its uncertain foundations and despite the conflicting theories of mathematicians, mathematics has proved so incredibly effective". The views of Wigner [3] and Hamming [4] on the "unreasonable effectiveness of mathematics" are perhaps better known, are warmer towards the mathematical profession, and have likely been better received. Philosophers of mathematics have perhaps placed undue emphasis on the apparent rightness of mathematics for the formulation of physical theories.

The essential point of this article is that there is a single theme–though one which can be recast in many superficially distinct ways–that reappears in a bewildering array of mathematical and physical contexts. Its appearance is seldom in the direct formulation of models, but rather arises in the working out of the implications of those formulations. We venture to suggest, though with some diffidence, that this mathematics internal to theories may itself contain some measure of physical insight, and perhaps even of physical reality.

Some of the ways in which the theme presents itself are collected in Table 1. It is particularly striking that the formulae in Table 1 vary from highly specific results about particular mathematical functions to results involving arbitrary functions, and include formulae that make sense in relatively elementary calculus, formulae that necessarily involve the theory of functions of a complex variable, and formulae that make no sense in classical real or complex analysis and need to be interpreted in the sense of generalized functions. The mathematical equivalence of the results in Table 1 has been addressed twenty years ago in a paper of Ninham, Hughes, Frankel and Glasser [5], and the reader may refer to that paper for a fuller account of the mathematical inter-relations and some relevant references that are not repeated here. What we offer here is a more compelling case for centrality of these relations to physics, rather than to mathematics.

In the context of physics, the "correspondence principle", first enunciated by Bohr [6], requires quantum mechanics to be consistent with classical mechanics in an appropriate limit, initially in Bohr's case in the limit of large quantum numbers, but now interpreted more broadly [7]. The "complementarity principle", also due to Bohr [8], was enunciated in the context of the problem of measurement in quantum mechanics, and its consequence of most interest in the present paper (loosely expressed as "wave–particle duality") is the requirement that quantum mechanical systems exhibit both wave and corpuscular characteristics, though never both at the same time. Echoes of these principles may be discerned in the discussion that follows.

In Section 2.1 we discuss various perspectives on the common theme underlying the entries in Table 1, which we regard, perhaps controversially, as the deepest "correspondence principle" in mathematical physics. There is an elegance and a tidiness in the formulae of Table 1, but these formulae are in some sense at the edge of chaos, as we discuss in Section 2.2. Moving towards specific physical contexts, we discuss time-evolving classical and quantum processes (Section 3), before turning our attention to questions of dilatational symmetry motivated by scattering data from quasicrystals (Section 4).

The examples in Sections 3 and 4 all involve intrinsically linear, non-cooperative phenomena and there is no explicit temperature dependence. In Section 5 we consider problems of quantum statistical mechanics, before concluding with perhaps the most elegant and intriguing appearance of our common theme in the context of Casimir forces (Section 6).

A collection of useful formulae for the theta functions is given in Appendix A. The variety of contexts from which our examples are drawn have their own popular notations and characteristic terminologies. For the most part we are able to avoid different uses of the same symbol, however force of habit and prevailing idiom oblige us to use τ in two different ways: as a complex number in the upper half-plane for the theory of theta functions and (in Section 4 and Appendix B) as the golden ratio $(1 + \sqrt{5})/2$. For brevity we use the usual notations $\mathbb{Z}$, $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{C}$ for the integers, natural numbers (i.e., the strictly positive integers), real numbers and complex numbers, respectively. All computations were performed with Mathematica.

2. THE MATHEMATICAL CONTEXT

2.1. Variations on a theme

An infinite sum of periodically spaced delta functions, $\sum_{n=-\infty}^{\infty} \delta(x-2nL)$, corresponding to equally spaced "points" or "atoms" on a line, is one of the simplest conceptualizations of the atomic-scale granularity of real matter. Finite segments of such a function, stacked in two and three dimensions form visualizations of elementary crystals and at large scales, where the granularity cannot be resolved, produce apparently smooth structures.

By purely formal Fourier analysis–though a proper derivation within the theory of generalized functions is available [9]–we shall represent $\sum_{n=-\infty}^{\infty} \delta(x-2nL)$ as a (classically divergent) series of classical functions. As $\sum_{n=-\infty}^{\infty} \delta(x-2nL)$ is periodic, computing its Fourier expansion in the usual manner using

$$f(x) = \sum_{n=-\infty}^{\infty} \Big[\frac{1}{2L}\int_{-L}^{L} e^{-in\pi y/L} f(y)dy\Big] e^{in\pi x/L} \tag{1}$$

yields

$$\begin{aligned}\sum_{n=-\infty}^{\infty} \delta(x-2nL) &= \frac{1}{2L}\sum_{n=-\infty}^{\infty} e^{in\pi x/L} \\ &= \frac{1}{2L} + \frac{1}{L}\sum_{n=1}^{\infty} \cos\Big(\frac{n\pi x}{L}\Big).\end{aligned} \tag{2}$$

Although Eq. (2) is valid only in the sense of generalized functions, it arises very cleanly as an extrapolation from a very classical result. Where

$$\theta_3(z|\tau) = \sum_{n=-\infty}^{\infty} \exp\{\pi i n^2 \tau + 2inz\} \tag{3}$$

Table 1. Five essentially equivalent results, identifiable as a single theme that is central to a broad range of problems in classical and quantum physics.

the correspondence principle or wave–particle duality	$\sum_{n=-\infty}^{\infty}\delta(x-2nL)=\frac{1}{2L}\sum_{n=-\infty}^{\infty}e^{in\pi x/L}=\frac{1}{2L}+\frac{1}{L}\sum_{n=1}^{\infty}\cos\left(\frac{n\pi x}{L}\right)$
theta function transformations (many equivalent or related forms)	$\frac{1}{2L}\sum_{n=-\infty}^{\infty}\exp(-\epsilon\pi n^2+in\pi x/L)=\sum_{n=-\infty}^{\infty}\frac{1}{\sqrt{4L^2\epsilon}}\exp\left[-\frac{\pi}{4L^2\epsilon}(x-2nL)^2\right]$
(classical) Poisson summation formula	$\sum_{n=-\infty}^{\infty}f(n)=\sum_{n=-\infty}^{\infty}\int_{-\infty}^{\infty}e^{2\pi inx}f(x)dx$
Riemann relation for the analytic continuation of $\zeta(s)=\sum_{n=1}^{\infty}n^{-s}$	$\zeta(s)=2^s\pi^{s-1}\sin(\frac{1}{2}\pi s)\Gamma(1-s)\zeta(1-s)$
transformation of Euler's product	$\prod_{k=1}^{\infty}(1-e^{-kx})=\left(\frac{2\pi}{x}\right)^{1/2}\exp\left(\frac{x}{24}-\frac{\pi^2}{6x}\right)\prod_{k=1}^{\infty}(1-e^{-4\pi^2k/x})$

denotes the third of the Jacobi theta functions [10] (see Appendix A) in what is now the traditional notation [11], the Jacobi theta function transformation

$$\theta_3(z|\tau)=\exp\left(\frac{i\pi}{4}-\frac{iz^2}{\pi\tau}\right)\tau^{-1/2}\theta_3\left(\frac{z}{\tau}\middle|-\frac{1}{\tau}\right) \tag{4}$$

is valid for all $z\in\mathbb{C}$ and for $\mathrm{Im}\{\tau\}>0$. If we divide Eq. (4) by $2L$ and set $\tau=i\varepsilon$ ($\varepsilon>0$) and $z=\pi x/(2L)$, we find that

$$\begin{aligned}&\frac{1}{2L}\sum_{n=-\infty}^{\infty}\exp(-\epsilon\pi n^2+in\pi x/L)\\&=\frac{1}{2L}+\frac{1}{L}\sum_{n=1}^{\infty}\exp(-\epsilon\pi n^2)\cos\left(\frac{n\pi x}{L}\right)\\&=\sum_{n=-\infty}^{\infty}\frac{1}{\sqrt{4L^2\epsilon}}\exp\left[-\frac{\pi}{4L^2\epsilon}(x-2nL)^2\right].\end{aligned} \tag{5}$$

For each fixed value of x, every term in the sum converges rapidly to zero as $\varepsilon\to0$, unless we have $x=nL$, in which case the nth term diverges, but we have

$$\begin{aligned}&\int_{-\infty}^{\infty}\frac{1}{\sqrt{4L^2\epsilon}}\exp\left[-\frac{\pi}{4L^2\epsilon}(x-2nL)^2\right]dx\\&=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}e^{-X^2}dX=1,\end{aligned}$$

showing that the right-hand side converges in an appropriate sense to $\sum_{n=-\infty}^{\infty}\delta(x-2nL)$. We show the series (5) for $\varepsilon=1/16$, $1/4$, 1 and 4 in Fig. 1. The elegant identity (5) equates two conceptually distinct viewpoints: a sum of smooth waves and a sum of pulses that may be identified as individual particles [12]. The particle interpretation becomes increasingly more attractive as ε is reduced. The three other Jacobi theta functions have analogous transformations that connect classically divergent trigonometric series to periodically spaced delta functions [5]. Indeed a more general result can be obtained by writing

$$\vartheta_{a,b}(z|\tau)=\sum_{n=-\infty}^{\infty}\exp\{\pi i[(n+a)^2\tau+2(n+a)(z+b)]\} \tag{6}$$

(so, for example, $\theta_3(z|\tau)=\theta_{0,0}(z/\pi\,|\tau)$) and noting that the generalization of Jacobi's transformation,

$$\vartheta_{a,b}(z|\tau)=\exp\left(\frac{i\pi}{4}-\frac{\pi iz^2}{\tau}+2\pi iab\right)\tau^{-1/2}\vartheta_{-b,a}\left(\frac{z}{\tau}\middle|-\frac{1}{\tau}\right), \tag{7}$$

leads to

$$\begin{aligned}&\lim_{\epsilon\to0}\frac{1}{2L}\sum_{n=-\infty}^{\infty}e^{-\pi(n+a)^2\epsilon}e^{\pi i(n+a)(x/L+2b)}\\&=\sum_{n=-\infty}^{\infty}e^{2\pi ina}\delta(x-2(n-b)L).\end{aligned} \tag{8}$$

The periodic delta function structures associated with the standard Jacobi theta functions θ_1 θ_2, θ_3 and

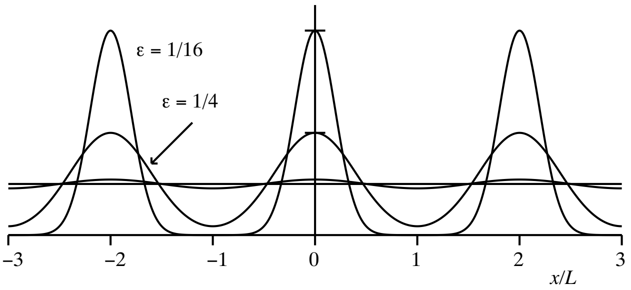


Figure 1. The series (5) interpolates between a uniformly flat profile (ε → ∞), a continuous wave (finite ε) and a train of particles (ε → 0). We illustrate this with the cases ε = 1/16 (highest peaks), 1/4, 1 and 4 (nearly flat).

θ_4 are recovered by replacing (a, b) with (1/2, 1/2), (1/2, 0), (0, 0) and (0, 1/2), respectively. Some of these choices yield coefficients with alternating plus or minus signs in the string of delta functions, and so can represent microscopically charged but macroscopically neutral matter. The generalized function $\sum_{n=-\infty}^{\infty} \delta(x - 2nL)$ is sometimes called a (or the) "Dirac comb" and its implications on the interpretation of diffraction data from solid crystals have received some attention [13, 14], especially in the context of its invariance (up to dilation and multiplication) under Fourier transformation:

$$\int_{-\infty}^{\infty} e^{2\pi i f x} \sum_{n=-\infty}^{\infty} \delta(x - 2nL) dx = \sum_{n=-\infty}^{\infty} e^{4\pi i n L f}$$

$$= \frac{1}{2L} \sum_{n=-\infty}^{\infty} \delta\Bigl(f - \frac{n}{2L}\Bigr). \qquad (9)$$

The generalized function identity (2) is sometimes called the Poisson summation formula, a forgivable appropriation of terminology [15] that we shall not adopt. For us the Poisson summation formula is [16]

$$\sum_{n=-\infty}^{\infty} f(n) = \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} e^{2\pi i n x} f(x) dx. \qquad (10)$$

This follows immediately from the observation that

$$\sum_{n=-\infty}^{\infty} f(n) = \int_{-\infty}^{\infty} f(x) \sum_{n=-\infty}^{\infty} \delta(x - n) dx$$

$$= \int_{-\infty}^{\infty} f(x) \sum_{n=-\infty}^{\infty} e^{2\pi i n x} dx$$

$$= \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} f(x) e^{2\pi i n x} dx.$$

Jacobi's transformations, which we have seen produce such things as generalized function identity (2), can also be regarded as consequences of Eq. (10). For example, by taking $f(x) = \exp(-\pi x^2 \varepsilon)$ in Eq. (10), we obtain $\theta_3(0|i\varepsilon) = \varepsilon^{-1/2}\theta_3(0|i\varepsilon^{-1})$. Riemann [17, 18] used this relationship to establish his famous functional relationship

$$\zeta(s) = 2^s \pi^{s-1} \sin(\tfrac{1}{2}\pi s)\Gamma(1 - s)\zeta(1 - s), \qquad (11)$$

where the Riemann zeta function $\zeta(s)$ and the gamma function $\Gamma(s)$ are defined initially by

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s} \quad \text{for } \mathrm{Re}\{s\} > 1, \qquad (12)$$

and

$$\Gamma(s) = \int_0^{\infty} e^{-t} t^{s-1} dt \quad \text{for } \mathrm{Re}\{s\} > 0, \qquad (13)$$

and extend by analytic continuation to functions holomorphic except for simple poles at $s = 1$ and at $s = 0, 1, 2, \ldots$, respectively [11, 17, 19].

The Riemann zeta function is profoundly important in number theory, but surprisingly frequently encountered also in physics [20]. Although a rigorous account of those of its properties that are rigorously established requires serious work [11, 19], some results fall out very simply [21]. Since

$$\zeta(s) = 1 + \frac{1}{2^s} + \sum_{m=2}^{\infty} \frac{1}{(m + 1)^s},$$

inserting the binomial expansion

$$(m + 1)^{-s} = \sum_{j=0}^{\infty} \frac{(-1)^j s(s + 1) \cdots (s + j - 1)}{j! m^{j+s}}$$

and interchanging orders of summation (the double series is absolutely convergent for Re{s} > 1) we find after a little algebra that

$$\sum_{j=1}^{\infty} \frac{(-1)^{j-1} s(s + 1) \cdots (s + j - 1)}{j!} [\zeta(s + j) - 1] = \frac{1}{2^s}. \qquad (14)$$

Analytic continuation of this result, which we obtained initially on the assumption that Re{s} > 1, shows immediately that

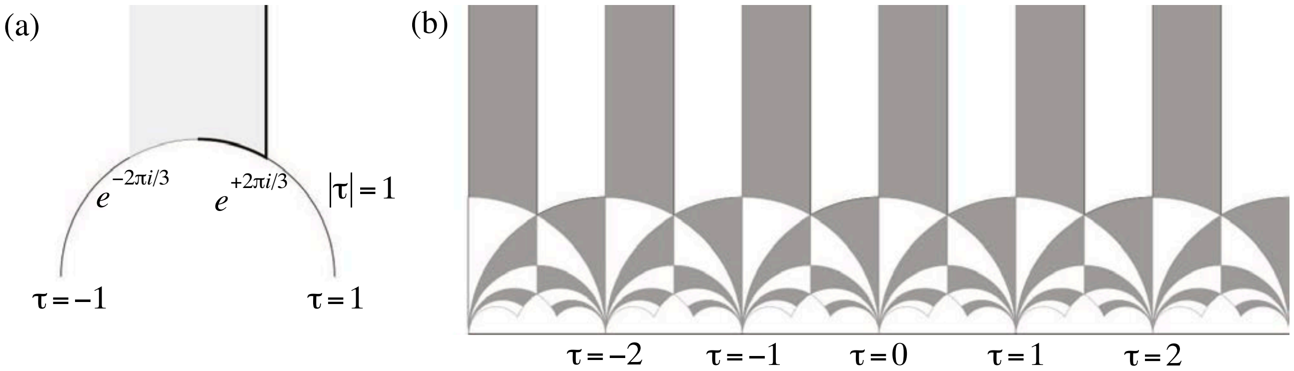


Figure 2. (a) The shaded region shows the fundamental set 𝕄 for the modular group. We include only the right half of the boundary, that is, boundary points with $0 \leq \text{Re} \leq \{\tau\} \leq ½$, shown as a dark curve. There are no modular transformations connecting any pair of distinct points in 𝕄. (b) The images of the fundamental set 𝕄 under the modular group tessellate the plane. We have shaded the right half of 𝕄 and all its images, while the left half of 𝕄 and all its images are left white, though their boundaries are drawn in gray. The images of 𝕄 shown here were obtained from those modular transformations with $a = 0$, $b = -1$, $c = 1$ and $-2 \leq d \leq 2$ (corresponding to $\tau' = \tau + d$, followed by $\tau'' = -1/\tau'$), or these transformations followed by a translation. The region close to the real axis is progressively filled as further transformations of 𝕄 are considered, but it become increasingly hard to portray the images without increasing the magnification of the figure.

$$\lim_{s\to 0} s\zeta(1 + s) = 1,\ \zeta(0) = -\frac{1}{2},\ \zeta(-1) = -\frac{1}{12},\ \zeta(-2) = 0,$$

and so on, and the Riemann relation (11) then yields $\zeta(2) = \pi^2/6$, $\zeta(4) = \pi^4/90$, …, although closed-form elementary evaluations of $\zeta(3)$, $\zeta(5)$, … have never been found. Some of the known simple exact values of the zeta function will be needed in Section 6, as will the equation

$$\Gamma(s)\zeta(s) = \int_0^\infty \frac{x^{s-1}dx}{e^x - 1} \tag{15}$$

that results from writing $t = nx$ in Eq. (13) and summing over n.

As noted in the Introduction, but worth emphasising again, the five results collected in Table 1–four of which we have already discussed, with the fifth (an infinite product transformation)–are essentially a single result [5]. It is possible to obtain all of the results from any one of them, and they establish a link between many substantial fields of mathematics, including complex analysis, number theory, harmonic analysis and numerical analysis [5, 22, 23]. It is the centrality of this common theme to physics that we begin to address in Section 3.

2.2. Analytical irregularity

There is surprising irregularity and complexity lurking behind the five equivalent identities in Table 1. We illustrate this first by considering the special case of the theta function $\theta_3(z|\tau)$ with $z = 0$. If we write for brevity $\theta(\tau) = \theta_3(0|\tau)$, then $\theta(\tau)$ is well-defined as a holomorphic (that is, complex-differentiable) function of the complex variable τ in the upper half plane $\text{Im}\{\tau\} > 0$. From Eqs (3) and (4) we find that

$$\theta(\tau + 1) = \theta(\tau) \quad \text{and} \quad \theta(\tau) = \left(\frac{i}{\tau}\right)^{1/2}\theta(-\tau^{-1}). \tag{16}$$

Both of the transformations $\tau \to \tau + 1$ and $\tau \to -\tau^{-1}$ are bijections of the upper half plane (that is, one-to-one correspondences between two copies of the upper half plane). These two fundamental transformations are the generators of a group of transformations of the upper half-plane known as the modular group [24]. Modular transformations have the form $\tau \mapsto (a\tau + b)/(c\tau + d)$, where $a, b, c, d \in \mathbb{Z}$ and $ad - bc = 1$.

Figure 2(a) shows a subset 𝕄 of the upper half-plane known as the fundamental region for the modular group. Every point in the upper half-plane is the image of a point in 𝕄 under a modular transformation, but there is no modular transformation connecting any two points of 𝕄. Figure 2(b) shows the remarkable way in which successive applications of simple modular transformations carry M into regions of progressively smaller total area, located closer and closer to the real τ axis. It follows that along the line segment defined by $\tau = \sigma + i\varepsilon$, with $-1 \leq \sigma \leq 1$ and $0 < \varepsilon \ll 1$, there is enormous variation in $\theta(\sigma + i\varepsilon)$, as shown in Fig. 3.

If we write $q = e^{i\pi\tau}$, the upper half-plane $\text{Re}\{\tau\} > 0$ corresponds to the disk $|q| < 1$ and we have

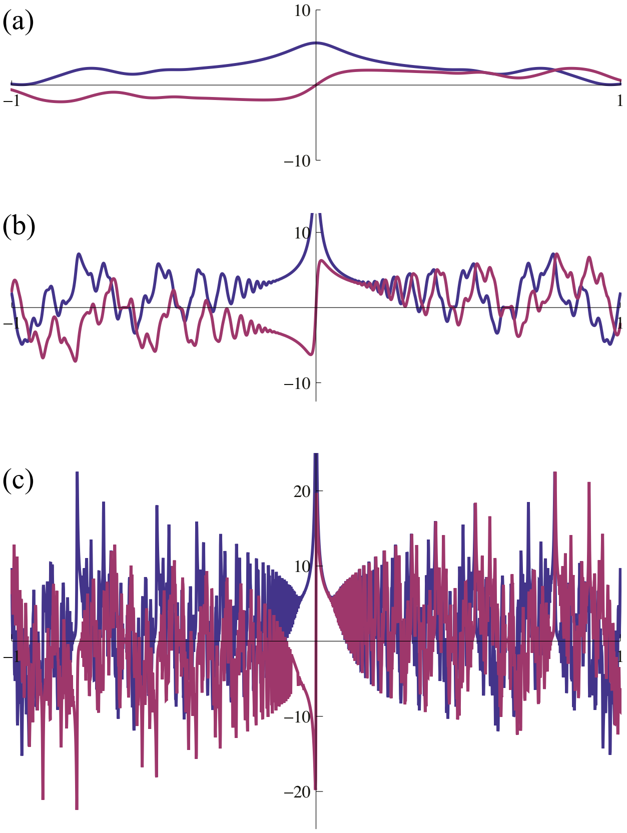


Figure 3. We show the real part (blue curves) and the imaginary part (red curves) of $\theta_3(\sigma + i\varepsilon)$ for $-1 \le \sigma \le 1$: (a) $\varepsilon = 0.1$; (b) $\varepsilon = 0.01$; (c) $\varepsilon = 0.001$.

$$\theta_3(\tau) = 1 + 2\sum_{n=1}^{\infty} q^{n^2} = 1 + 2q + 2q^4 + 2q^9 + \cdots . \tag{17}$$

This is a power series in q, with the unit circle as its circle of convergence, and gaps of rapidly increasing length between powers of q with nonzero coefficients. Indeed, for fixed z, all of the theta functions $\theta_k(z|\tau)$, $k \in \{1, 2, 3, 4\}$, have the form

$$\theta_k(z|\tau) = q^{\kappa} \sum_{n=0}^{\infty} a_{k,n}(z) q^{\lambda_n} \tag{18}$$

where $\kappa \in \{0, 1/2\}$ and either $\lambda_n = n^2$ or $\lambda_n = n(n+1)$, with the series always convergent for $|q| < 1$ and always divergent for $|q| > 1$.

More generally, if λ_n is a strictly increasing sequence of non-negative integers, then $\sum_n a_n q^{\lambda_n}$ is a power series in the complex variable q. If $\lambda_n/n \to \infty$ as $n \to \infty$ the power series is called a "lacunary series", the name referring to the gaps between powers of q that have nonzero coefficients. A beautiful theorem of Fabry [25, 26] states that if $\sum_n a_n q^{\lambda_n}$ is a lacunary power series with radius of convergence 1, then the function defined by $f(q) = \sum_n a_n q^{\lambda_n}$ for $|q| < 1$ cannot be continued analytically beyond $|q| = 1$. As functions of q, the theta functions meet the conditions of Fabry's Theorem. Analytic continuation across the unit circle $|q| = 1$ is prevented by the presence of a dense fence of singular points on this circle. Figure 3 manifests the existence of this fence.

It is interesting that the five equivalent identities in Table 1, which involve either smooth functions or periodic functions, are the gateway to revealing dense, nonsmooth behavior.

3. TIME-EVOLVING CLASSICAL AND QUANTUM PROCESSES

Our point of departure in Section 2.1 was already associated with physical concepts, namely periodically spaced point masses or point charges, but no physical models or processes have really been addressed.

3.1. Classical Diffusion

For $-\infty < z < \infty$ and $\mathrm{Im}\{\tau\} > 0$, all four Jacobi theta functions satisfy the partial differential equation

$$\frac{\pi i}{4}\frac{\partial^2 u}{\partial z^2} + \frac{\partial u}{\partial \tau} = 0, \tag{19}$$

as indeed does the more general function $\theta_{a,b}(z/\pi\,|\tau)$.

If we take $\tau = (4D/\pi)it$ with t real, replace z by x, and write $u(x, \tau) = v(x, t)$, Eq. (19) reduces to the one-dimensional diffusion equation

$$\frac{\partial v}{\partial t} = D\frac{\partial^2 v}{\partial x^2}. \tag{20}$$

The theta function transformations connect optimally structured short-time and long-time solutions of one-dimensional diffusion problems in finite domains, with one theta function expression corresponding to an expansion of the solution in spatial trigonometric functions with exponentially decaying time-dependent coefficients (a good solution from at long times) and the other corresponding to a "method of images" solution constructed from Gaussian propagators (a good solution at short times) [22]. For example, if we write $\varepsilon = \pi Dt/L^2$, then Eq. (5) equates these two solutions in the case of impenetrable reflecting boundaries (zero flux: $-D\partial v/\partial x = 0$) at $x = \pm L$, and initial condition $v(x, 0) = \delta(x)$:

$$\frac{1}{2L}+\frac{1}{L}\sum_{n=1}^{\infty}\exp\Bigl(-\frac{n^2\pi^2Dt}{L^2}\Bigr)\cos\Bigl(\frac{n\pi x}{L}\Bigr)$$

$$=\sum_{n=-\infty}^{\infty}\frac{1}{\sqrt{4\pi Dt}}\exp\Bigl[-\frac{1}{4Dt}(x-2nL)^2\Bigr]. \tag{21}$$

3.2. Anomalous diffusion

In one dimension and in the absence of boundaries, the mean-square displacement for the diffusion process (20) grows linearly with time [27]:

$$\int_{-\infty}^{\infty}x^2v(x,t)dx=\int_{-\infty}^{\infty}x^2v(x,0)dx+2Dt. \tag{22}$$

The study of diffusion processes based on Eq. (20) was initiated by Fick [28] in 1855. Much more recently there has been intense interest in transport processes that are not diffusive in character [29, 30, 31, 32]. In one-dimensional unbiased non-diffusive processes the mean-square displacement may grow more slowly that linearly with time (sub-diffusive processes), or faster than linearly (super-diffusive processes). An extreme case of one-dimensional super-diffusion has an infinite mean-square displacement and this can lead to a statistically self-similar or fractal [33] footprint structure (the set of points visited has a fractal dimension less than 1).

Hughes, Shlesinger and Montroll [34] considered a random walk model in which the random displacement made at any step has the probability density function

$$p(x)=\frac{a-1}{2a}\sum_{n=0}^{\infty}a^{-n}[\delta(x-\Delta b^n)+\delta(x+\Delta b^n)], \tag{23}$$

with $a > 1$ and $b > 1$. Since motions on the length scale Δb^n are a times more abundant than motions on the next shortest length scale Δb^{n+1}, the stepping law has fractal character built in (with fractal dimension $\mu = \ln(a)/\ln(b)$), and the only question is whether fractal footprints are left visible at long times, or the legacy of the walk is smeared. If $\mu < 2$ the mean-square displacement per step is infinite, the central limit theorem fails, and the continuum limit of the process does not have the standard Gaussian propagator familiar from classical diffusion [30, 35]. The walk is transient if $\mu < 1$ (any interval is visited only finitely many times with probability 1) and fractal footprints are left.

To analyze features of this model, it is necessary to understand the behavior near the origin of the Fourier transform of the probability density function (23), and this is equivalent to requiring the small-k behavior of

$$\lambda(k)=\frac{a-1}{a}\sum_{n=0}^{\infty}a^{-n}\cos(b^nk). \tag{24}$$

It is easy to see that $\lambda(k)$ satisfies the rather simple-looking functional equation

$$\lambda(k)=\frac{a-1}{a}\cos(k)+a^{-1}\lambda(bk), \tag{25}$$

which is reminiscent of equations obtained in real-space renormalization treatments of lattice spin systems [36, 37]. The apparent simplicity of the functional equation is illusory. Hughes et al. [34] were able to show using the Poisson summation formula (10) that for $k > 0$ and $\ln(a)/\ln(b) \notin \{2, 4, 6, \cdots\}$,

$$\lambda(k)=\frac{a-1}{a}\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n)!}\frac{k^{2n}}{1-b^{2n}/a}+k^{\mu}Q(k), \tag{26}$$

where [38]

$$Q(k)=\frac{a-1}{a\ln b}\sum_{n=-\infty}^{\infty}\Gamma(s_n)\cos\Bigl(\frac{\pi s_n}{2}\Bigr)\exp\Bigl(-\frac{2n\pi i\ln k}{\ln b}\Bigr) \tag{27}$$

and we have written for brevity $s_n = -\mu + 2n\pi i/\ln b$. The appearance in $Q(k)$ of "log–periodic oscillations" (periodic in $\ln k$ with period $\ln b$) is striking (see Fig. 4). Similar oscillations occur in real-space renormalisation group transformations for lattice spin systems [39], in a model for the distribution of family names in a society [40] and in a variety of other systems that exhibit a form of discrete scale invariance [41].

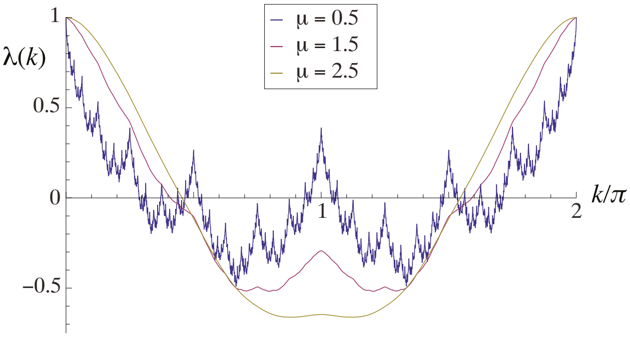


Figure 4. The structure function (24) of the Weierstrass random walk step probability density function (23). In each case, $b = 2$, and we choose a values so that so that $\mu = \ln(a)/\ln(b)$ takes the values 0.5 [blue (most irregular) curve], 1.5 (red curve) and 2.5 (gold curve).

If b is a positive integer and $b \geq 2$ then we can recognize $\lambda(k)$ as a constant multiple of the real part of the lacunary power series $\sum_{n=0}^{\infty} a^{-n} z^{b^n}$ evaluated on its circle of convergence ($|z| = 1$), so a dense set of singular points must be present and indeed for appropriate values of a and b the series for $\lambda(k)$ is the celebrated nowhere-differentiable function of Weierstrass [42].

There is a second perspective on Eq. (23) that is also worth considering [30, 34]. By considering the contour integral of $e^{-z}z^{s-1}$ around a simple closed contour in the z-plane consisting of the arc of the circle $|z| = R$ within the first quadrant, and straight lines along the real and imaginary axis linking the ends of the arc to the origin, it is easy to prove that for $0 < \text{Re}\{s\} < 1$,

$$\int_0^\infty e^{ix} x^{s-1} dx = \Gamma(s) \exp(\tfrac{1}{2} i\pi s). \tag{28}$$

Adding this equation and its complex conjugate we find that for $0 < \text{Re}\{s\} < 1$,

$$\int_0^\infty \cos(x) x^{s-1} dx = \Gamma(s) \cos(\tfrac{1}{2}\pi s). \tag{29}$$

The definition of the Mellin transform and the associated inversion formula [16],

$$\overline{f}(s) = \int_0^\infty x^{s-1} f(x) dx, \tag{30}$$

$$f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} x^{-s} \overline{f}(s) ds, \tag{31}$$

with the Bromwich contour Re $\{s\} = c$ placed inside a strip in which the Mellin transform integral converges, are another manifestation of the relations collected in Table 1, since they can be used to obtain both the Riemann relation and the theta function transformation in relatively straightforward ways.

Using Eqs (29) and (31) we can write

$$\cos(b^n k) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} b^{-ns} k^{-s} \Gamma(s) \cos(\tfrac{1}{2}\pi s) ds, \tag{32}$$

for $k > 0$ and $0 < c < 1$. Inserting this representation into Eq. (24), interchanging the order of integration and summation and recognizing a geometric series, we find [30, 34] that

$$\lambda(k) = \frac{a-1}{a} \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{k^{-s}\Gamma(s)\cos(\tfrac{1}{2}\pi s) ds}{1 - a^{-1}b^{-s}}. \tag{33}$$

Translating the contour of integration to the left and taking account of the residues at the poles crossed, we recover Eqs (26) and (27). The power series arises from the simple poles along the real axis at $s = 0, -2, -4, \ldots$, while $k^\mu Q(k)$ comes from the line of poles at $s_n = -\mu + 2n\pi i/ \ln b$. The small-$k$ behavior, which governs the limiting behavior of the random walk, is dominated by which pole or poles the contour next encounters after we have translated it past the origin. For $\mu > 2$ the next pole encountered is a simple pole at $s = -2$, so that $1 - \lambda(k) \propto k^2$ as $k \to 0$ (ensuring a diffusive limit). However for $0 < \mu < 2$ we meet the line of poles at $s = s_n$, and this is how the term $k^\mu Q(k)$ arises, precluding diffusion.

These kinds of calculations using Mellin transforms are closely connected to the powerful role of Mellin transforms in asymptotic analysis [43] and also give one link between several identities in Table 1. Whichever approach is used to reveal the small-k behavior of $\lambda(k)$, the simplest limiting behavior is obtained as $\Delta \to 0$ and $t_0 \to 0$ (where t_0 is the time between successive steps) if we also make $a \to 1$ and $b \to 1$, while holding both $\mu = \ln a/ \ln b$ and Δ^μ/t_0 constant. Then if $\mu < 2$, the evolution of the random position X_t of the moving agent satisfies

$$\frac{\partial}{\partial t} \mathsf{E}\{\exp(iqX_t)\} = -c|q|^\mu \mathsf{E}\{\exp(iqX_t)\}, \tag{34}$$

where c is a positive real constant, $q \in \mathbb{R}$, and E denotes mathematical expectation or averaging. The "characteristic function" $\mathsf{E}\{\exp(iqX_t)\}$ is just a spatial Fourier transform of the probability density function for the agent's location at time t. Solving the evolution equation (34) with the initial condition $X_0 = 0$ gives $\mathsf{E}\{\exp(iqX_t)\} = \exp(-c|q|^\mu t)$ and inverting the Fourier transform gives the celebrated symmetric stable densites [29, 30] of Lévy [44],

$$S_\mu(x,t) = \frac{1}{2\pi} \int_{-\infty}^\infty \exp(-iqx - c|q|^\mu t) dq. \tag{35}$$

The borderline case $\mu = 2$ corresponds to the Gaussian density, while for $\mu < 2$, the density decays algebraically rather than exponentially, with $\Pr\{X_t > x\} \propto x^{-\mu}$ as $x \to \infty$. The only other case where the symmetric stable density has a simple closed form expression [45] is $\mu = 1$, which is the Cauchy density $(c/\pi)(x^2 + c^2)^{-1}$.

Super-diffusive processes, such as the stable density, are naturally formulated in unbounded space, but it may be of interest to seek solutions in finite intervals. Appropriate boundary conditions for reflecting boundaries are debatable (for $\mu < 1$ the path is discontinuous), but we can use method of images arguments [cf. Eq. (21)]

to obtain a solution which conserves probability in the interval $(-L, L)$. The following analysis is very much in the sense of generalised functions, as we work with classically divergent series and use the identity (2):

$$\begin{aligned}&\sum_{n=-\infty}^{\infty} S_\mu(x-2nL)\\ &= \sum_{n=-\infty}^{\infty} \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-iq(x-2nL)-c|q|^\mu t)dq\\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\exp(2inLq)\exp(-iqx-c|q|^\mu t)dq\\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\pi}{L}\sum_{n=-\infty}^{\infty}\delta\Big(q-\frac{n\pi}{L}\Big)\exp(-iqx-c|q|^\mu t)dq\\ &= \frac{1}{2L}+\frac{1}{L}\sum_{n=1}^{\infty}\cos\Big(\frac{n\pi x}{L}\Big)\exp\Big[-\Big(\frac{n\pi}{L}\Big)^\mu ct\Big]. \qquad (36)\end{aligned}$$

It is perhaps curious that for this system, the relaxation to the uniform density $1/(2L)$ on the interval is a simple exponential, rather than some form of stretched exponential, despite the transport process being highly super-diffusive.

Clearly there are many subtleties that can arise when stochastic ideas intersect with self-similarity. For another manifestation of this, see Appendix B.

3.3. *One quantum particle*

Let h denote Planck's constant and $\hbar = h/(2\pi)$. If we write $\tau = -2\hbar t/(\pi m)$ with t complex (with a negative imaginary part) and $u(z, \tau) = \psi(z, t)$, we obtain from Eq. (19) the one-dimensional Schrödinger equation in free space,

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial z^2}. \qquad (37)$$

Thus linear combinations of theta functions with the complex time extrapolated to the real axis should be able to be used to construct non-trivial time-dependent solutions of Schrödinger equation. Although we are aware of no systematic study of this, for an investigation of some cognate issues the reader may consult the beautiful paper of Fulling and Güntürk [46] on the one-dimensional Schrödinger equation with periodic boundary conditions. Less direct applications of theta functions to solving Schrödinger's equation have been considered by Gaveau and Schulman [47].

The formal connection between theta functions and Schrödinger's equation (obtained by letting the artificial negative imaginary part of the time approach 0) corresponds to moving radially outwards towards the circle of convergence of a lacunary series, as discussed in Section 2.2. The highly irregular form of the propagator discussed by Fulling and Güntürk should therefore come as no surprise.

If we don't observe the connection to theta functions, and instead use an energy eigenfunction approach to solve the d-dimensional Schrödinger equation $i\hbar\,\partial\psi/\partial t = -[\hbar^2/(2m)]\nabla^2\psi$ in the box $(0, L)^d$ (with the wave function vanishing on the boundary) we obtain the general solution

$$\psi(\mathbf{x}, t) = \sum_{\mathbf{n}\in\mathbb{N}^d} c(\mathbf{n})\exp\Big[-\frac{iE(\mathbf{n})t}{\hbar}\Big]\prod_{j=1}^{d}\sin\Big(\frac{n_j\pi x_j}{L}\Big), \qquad (38)$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_d)$ and

$$E(\mathbf{n}) = \frac{\hbar^2\pi^2|\mathbf{n}|^2}{2mL^2} = \frac{h^2|\mathbf{n}|^2}{8mL^2}. \qquad (39)$$

If we take the initial condition $\psi(\mathbf{x}, t') = \delta(\mathbf{x} - \mathbf{x}')$ with $\mathbf{x}' \in (0, L)^d$, we obtain the (never classically convergent) generalized function propagator

$$\begin{aligned}\psi(\mathbf{x}, t\,|\,\mathbf{x}', t') = \Big(\frac{2}{L}\Big)^d \sum_{\mathbf{n}\in\mathbb{N}^d}\exp\Big[-\frac{iE(\mathbf{n})(t-t')}{\hbar}\Big]\\ \times\prod_{j=1}^{d}\sin\Big(\frac{n_j\pi x_j}{L}\Big)\sin\Big(\frac{n_j\pi x'_j}{L}\Big). \qquad (40)\end{aligned}$$

Perhaps the connection to theta functions makes the strangeness of this result easier to comprehend.

Despite the irregular propagator for finite intervals, the free-space Schrödinger equation does have some comparatively simple, well-behaved normalizable solutions on the real line, such as the spreading Gaussian wave packet found by Darwin [48] and Kennard [49] and the Airy function solution of Lekner [50].

4. QUASICRYSTALS

Diffraction experiments probe the structure of condensed matter using electrons, neutrons or X rays. In systems with long- range order, this order is revealed by observed intensity distributions exhibiting sharp peaks [51, 52]. Experimental realities and the finiteness of the atoms scattering the incident radiation broaden the peaks, but basically in an idealized but substantially correct way, the observed density in many cases is a set of delta functions, whose locations encode information about the atomic locations, which are also delta functions. This picture is clearest and most apt for crystals, where the diffraction data corresponds to the Fourier transform of the crystal [53]. For the one-dimensional case with equal spacing between atoms, Eq. (9) shows that the Fourier transform is simply the original lattice structure with a changed lattice spacing (the Fourier transform of the Dirac comb is a Dirac comb). Similar results hold in two and three dimensions [54].

The discovery in 1984 by Shechtman et al. [55] of a metallic phase with long-range orientational order but no translational symmetry challenged established paradigms in crystallography, which assume that crystals consist of unit cells of atoms of various species arranged periodically. Within six weeks, Levine and Steinhardt [56] had dubbed these structures “quasicrystals”–a name the structures have retained [57]–and suggested analogies to nonperiodic tilings of space with local pentagonal symmetry previously studied by Penrose [58]. Shechtman received the 2011 Nobel Prize in Chemistry for the discovery of quasicrystals.

The appearance in a physical context of long-range order without translational symmetry naturally motivated a number of fundamental studies of a purely mathematical character, including a careful definition of diffraction on aperiodic structures [59]. Many important observations concerning cognate mathematical issues can be found in Senechal [60, 61], Senechal and Taylor [62, 63] and Baake and Grimm [64].

When translational invariance breaks down in the observed crystallographic data, the unambiguous connection to the original lattice is lost. How is the structure of a quasicrystal to be inferred? To begin, the attempt to fit the diffraction data $\tilde{\rho}(\mathbf{k})$ to delta functions with a non-zero minimal spacing and recover well-spaced delta functions for $\rho(\mathbf{r})$ is doomed to failure [65].

Ninham and Lidin [66] suggested the possible relevance to quasicrystals of dilatational rather than translational symmetry, using the following example, which has interesting historical antecedents. The gnomon (γνώμων) is the shadow-casting blade on a sundial, but also refers to triangles or rectangles produced by internal subdivision of triangles or rectangles in a special way. In particular, if an isosceles triangle with two sides of length $\tau > 1$ and third side of length 1 is subdivided by drawing a straight line from one of the equal angles to the opposite face to create an isosceles triangle with two sides of length 1, the other triangle created in this subdivision is the gnomon. A simple argument based on similar triangles establishes that the gnomon is itself an isosceles triangle if and only if

$$\tau^2 - \tau - 1 = 0, \tag{41}$$

from which it follows that

$$\tau = (1 + \sqrt{5})/2 \approx 1.618 \tag{42}$$

(see Fig. 5(a), in which the gnomon is shaded in gray). For this special choice of τ, the internal angles of the triangles produced in the subdivision are all integer multiples of $\pi/5$, as shown. In Fig. 5(b) we take the scaled replica of the original triangle produced by the subdivision, subdivide it in a similar manner, and repeat this process several times, always producing isosceles triangles with the same angles, but with the lengths of sides diminishing by a factor of τ at each stage. The number τ is the famous golden mean, golden ratio or golden number, which figures prominently in aesthetics [67] and in nature [68]. The logarithmic spirals

$$\frac{\ln r}{\ln \tau} = \frac{\theta + a\pi/5}{\pi/5}, \quad a \in \{0, \pm 2, \pm 4, \pm 6, \pm 8\}, \tag{43}$$

are shown as grey curves in Fig. 5(c). Their intersections generate a distribution of points with five-fold rotational symmetry about the origin. The logarithmic spiral $\ln r/\ln \tau = \theta/(3\pi/5)$ passes through these points of intersection, with the distance from the origin increasing by a factor of τ between any two consecutive intersections. With suitable scaling and rotation, the inscribed triangles shown in Fig. 5(b) can be placed with their vertices located at the intersection points [66, 69].

We consider the Fourier-space signature of the mass distribution

$$\rho(\mathbf{r}) = \sum_{j=1}^{10} \sum_{m=-\infty}^{\infty} a^{-|m|} \delta\Big(x - \tau^m \cos(\pi j/5)\Big) \times \delta\Big(y - \tau^m \sin(\pi j/5)\Big), \tag{44}$$

which places all mass on rays through the origin, with angular separation $\pi/5$ between rays, and on each ray, we have dilational invariance in the locations of the masses, with a scaling factor τ. The convergence factor $a^{-|m|}$ (with $a > 1$) is present to keep finite total mass in the system. We find that where $\mathbf{k} = (k_1, k_2)$,

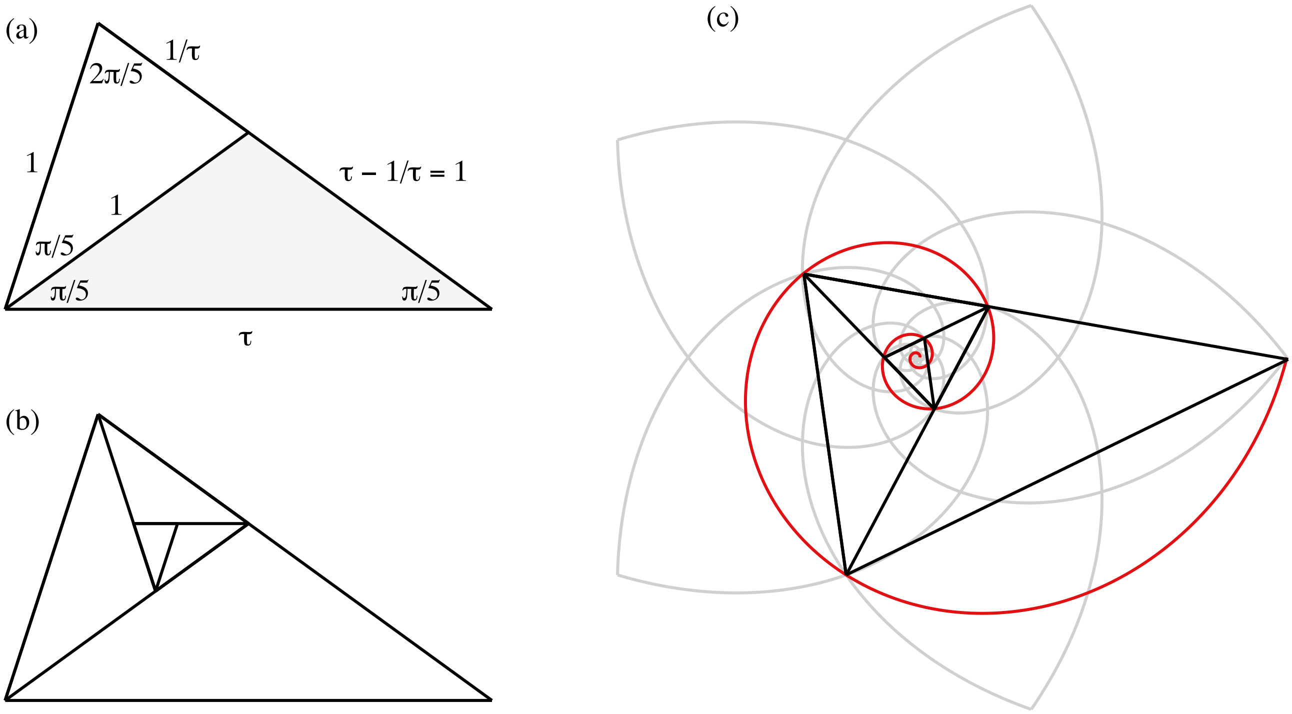


Figure 5. (a) With τ given by Eq. (42), an isosceles triangle with side lengths ratios $\tau : \tau : 1$, can be subdivided into two isosceles triangles, one of which (white interior) has side length ratios $\tau : \tau : 1$ but side lengths a factor τ smaller than those in the original triangle. (b) We can continue the process of subdivision to generate a nested set of isosceles triangles with side length ratios $\tau : \tau : 1$, but at each step of the process, the side length of the triangle just produced is reduced from that of its parent by a factor of τ. (c) The intersections of the logarithmic spirals (43) generate a distribution of points with five-fold rotational symmetry about the origin. The logarithmic spiral $\ln r / \ln \tau = \theta/(3\pi/5)$, shown in red, passes through these points of intersection, with the distance from the origin increasing by a factor of τ between any two consecutive intersections. With suitable scaling and rotation, the inscribed triangles shown in diagram (b) can be placed with their vertices located at the intersection points (figure adapted from Ninham and Lidin [66]).

$$\tilde{\rho}(\mathbf{k}) = \sum_{j=1}^{10} \sum_{m=-\infty}^{\infty} a^{-|m|} \exp\left\{ i\tau^{m}\left[k_1 \cos\left(\frac{\pi j}{5}\right) + k_2 \sin\left(\frac{\pi j}{5}\right)\right]\right\}. \quad (45)$$

We show $|\rho(\mathbf{k})|$ for $a = 1.1$ in Fig. 6. It is not surprising that the rotational symmetry in the mass distribution is reflected in the Fourier transform domain: this is clear from Eq. (45). What is more interesting, and more beautiful, is that in the Fourier transform domain, where the signal is continuous (rather than localized on lines, as in the original space domain), we see a rich structure with local intensity maxima occurring at many points in the sectors between the ten lines on which the brightest peaks are located. Also, it is by no means obvious from the formula (45) where the intensity maxima on the bright lines will occur. In Fig. 7, we show $|\rho(\mathbf{k})|$ on the vertical axis in the $\mathbf{k}$-plane. There are many local maxima, but a sequence of locally outstanding maxima can be identified at the k_2 values 4.775, 7.732, 12.51, 20.25, 32.77. The successive ratios of these k_2 values are all close to (but not exactly) 1.618, and we see the golden ratio from physical space recurring (to a decent approximation) in intensity maxima in Fourier space.

The convergence factor $a^{|m|}$ in Eq. (45) stops $|\tilde{\rho}(\mathbf{k})|$ from having exact dilatational symmetry. If we were able to set $a = 1$, then we would recover $|\tilde{\rho}(\mathbf{k})| = |\tilde{\rho}(\tau\mathbf{k})|$. Berry and Lewis [70] have considered what they call the Weierstrass–Mandelbrot fractal function

$$W(t) = \sum_{n=-\infty}^{\infty} \frac{\left[\left(1 - e^{i\gamma^{n} t}\right) e^{i\phi_n}\right]}{\gamma^{(2-D)n}}, \quad 1 < D < 2,\ \gamma > 1, \quad (46)$$

where ϕ_n represents a constant phase added onto each term. The series is convergent and, if ϕ_n is constant, has perfect self-similarity: $W(\gamma t) = \gamma^{2-D} W(t)$. Ninham and Lidin [66] have considered another way of overcoming the problem of infinite mass accumulating in the neighborhood of the origin by using the formal series

$$\sum_{m=-\infty}^{\infty} \left[\cos\left(2\pi r \tau^{-m}\right) - \cos\left(2\pi \tau^{-m}\right)\right]$$

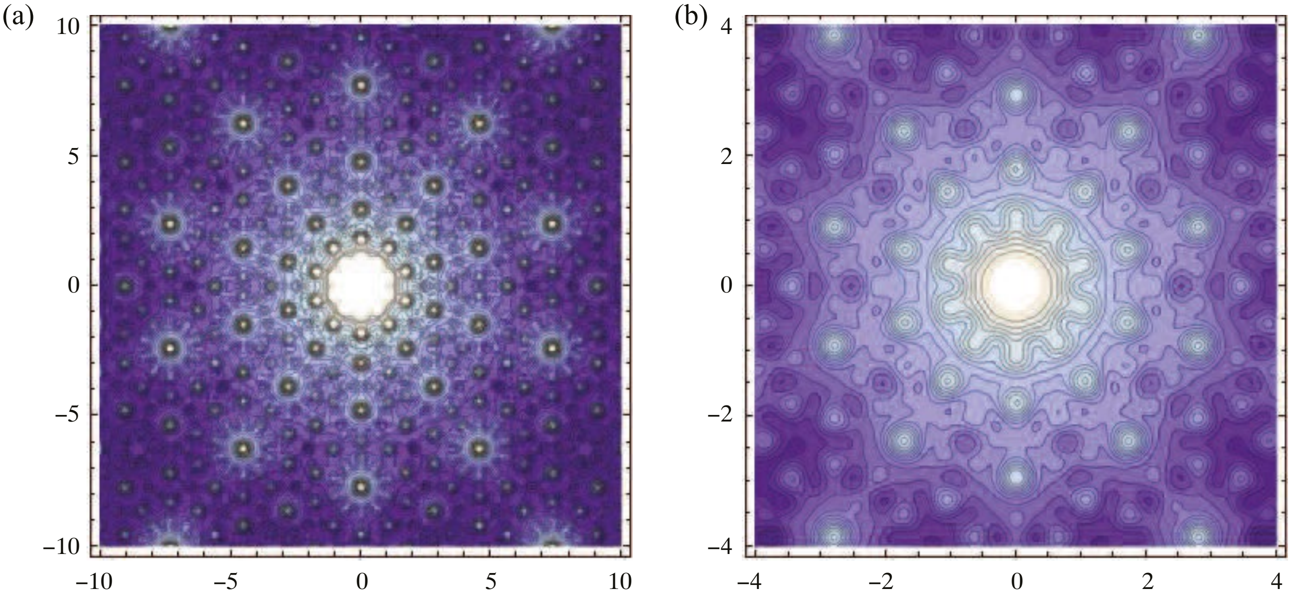


Figure 6. The **k**-plane is colored (with 20 levels) to show $|\tilde{\rho}(\mathbf{k})|$, where the Fourier transform $\tilde{\rho}(\mathbf{k})$, given by Eq. (45), arises from the mass distribution (44). Lighter shades represent larger values of $|\tilde{\rho}(\mathbf{k})|$. (a) $-10 \le k_1, k_2 \le 10$; (b) $-4 \le k_1, k_2 \le 4$. For the convergence factor $a^{|m|}$ we have taken $a = 1.1$.

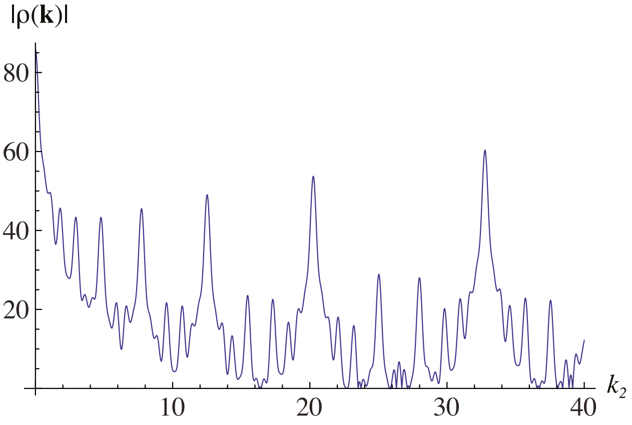


Figure 7. We show $|\tilde{\rho}(\mathbf{k})|$ on the line $\mathbf{k} = (0, k_2)$, where the Fourier transform $\tilde{\rho}(\mathbf{k})$, given by Eq. (45) arises from the mass distribution (44). For the convergence factor $a^{-|m|}$ we have taken $a = 1.1$.

for the mass distribution along a ray through the origin, where r is the distance from the origin.

Quasicrystals are not the only context in which wild oscillations and apparent self-similar structure arise in the amplitude of diffracted light. Berry [71] gives a beautiful example, in which theta functions play a key role.

5. QUANTUM STATISTICAL MECHANICS

We consider several models from quantum statistical mechanics, for which we use standard notation and terminology [72, 73], so that k is Boltzmann's constant and T is the absolute temperature.

5.1. The harmonic oscillator

The free energy $g(\omega)$ associated with a harmonic oscillator of frequency ω and energy levels $(n + 1/2)\hbar\omega$ ($n = 0, 1, 2, \ldots$) is given in terms of the canonical partition function $\mathcal{Z}(\omega)$ by

$$\exp\Big[-\frac{g(\omega)}{kT}\Big] = \mathcal{Z}(\omega) = \sum_{n=0}^{\infty} \exp\Big[-\frac{(n+1/2)\hbar\omega}{kT}\Big], \tag{47}$$

$$g(\omega) = kT\ln\Big[2\sinh\Big(\frac{\hbar\omega}{2kT}\Big)\Big] = \frac{\hbar\omega}{2} - \sum_{n=1}^{\infty}\frac{kT}{n}\exp\Big(-\frac{n\hbar\omega}{kT}\Big). \tag{48}$$

For brevity, we have suppressed in the notation the dependence of the free energy on the temperature. Hence

$$g'(\omega) = \frac{\hbar}{2} + \hbar\sum_{n=1}^{\infty}\exp\Big(-\frac{n\hbar\omega}{kT}\Big), \tag{49}$$

leading to the formal identification [74]

$$\mathrm{Re}\{g'(i\omega)\} = \frac{\hbar}{2} + \hbar \sum_{n=1}^{\infty} \cos\Big(\frac{n\hbar\omega}{kT}\Big)$$
$$= \pi kT \sum_{n=-\infty}^{\infty} \delta\Big(\omega - \frac{2\pi kTn}{\hbar}\Big). \tag{50}$$

This superficially bizarre result connecting the oscillator free energy to a string of delta functions, arising from the mathematical correspondence principle, proves surprisingly useful. If the modes of oscillation of a system are given by a secular equation of the form $D(\omega) = 0$, then the free energy can be computed as a sum over the contributions from the various modes by the contour integral

$$F = \frac{1}{2\pi i} \oint g(\omega) \frac{d}{d\omega} \ln[D(\omega)] d\omega, \tag{51}$$

the contour integral being taken over a simple closed contour that surrounds all zeros of $D(\omega)$ on the positive real axis. If there are infinitely many such zeros with the spacing bounded below as $\omega \to \infty$, an appropriate limiting construction is made. Integrating by parts, deforming the contour and making formal use of Eq. (50) enables the free energy to be computed conveniently [74, 75]. This is especially convenient in the calculation of dispersion (van der Waals) forces between dielectric media [74, 75].

5.2. Particle in a box

Using the energy eigenvalues (39), the free energy $\mathcal{G}$ associated with a single (non-elementary [76]) particle of mass m in the d-dimensional box $[0, L]^d$ is given by

$$\exp\Big[-\frac{\mathcal{G}}{kT}\Big] = \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} \cdots \sum_{n_d=1}^{\infty} \exp\Big[-\frac{h^2 \sum_{j=1}^{d} n_j^2}{8mL^2kT}\Big]$$
$$= \Big\{\sum_{n=1}^{\infty} \exp\Big[-\frac{h^2n^2}{8mL^2kT}\Big]\Big\}^d$$
$$= \Big\{\frac{1}{2}\theta_3\Big(0, \exp\Big[-\frac{h^2}{8mL^2kT}\Big]\Big) - \frac{1}{2}\Big\}^d. \tag{52}$$

Here we adopt the usual notational convenience of writing $\theta_k(z|\tau) = \theta_k(z, q)$, where $q = e^{i\pi\tau}$. If we consider N identical non-interacting particles in the same box, Eq. (52) becomes the equation for the free energy $\mathcal{G}$ per particle. Fixing T, the right-hand side can be evaluated asymptotically in the limit $L \to \infty$, using the Jacobi theta function transformation (4), which in the special case $z = 0$ and $\tau = it$ ($t \in \mathbb{R}$, with $t > 0$) becomes $\theta_3(0, it) = t^{-1/2}\theta_3(0|it^{-1})$. We find that

$$\exp\Big[-\frac{\mathcal{G}}{kT}\Big] = \Big\{\frac{L}{\lambda_T}\theta_3\Big(0, \exp\Big[-\frac{4\pi L^2}{\lambda_T^2}\Big]\Big) - \frac{1}{2}\Big\}^d \tag{53}$$
$$\sim \Big(\frac{L}{\lambda_T}\Big)^d \quad \text{as } L \to \infty, \tag{54}$$

where for brevity in notation we have introduced the thermal wavelength $\lambda_T = h(2\pi mkT)^{-1/2}$. The single-term approximation (54) is well known [73], but the theta function representations (52) and (53) enable us to compute $\mathcal{G}$ and the associated thermodynamic observables to high precision for any value of L/λ_T.

5.3. Ideal gas of elementary particles

Consider now ideal gases of elementary particles, which may be bosons (such as photons or mesons, for which an arbitrary number of particles can occupy any state) or fermions (such as electrons or neutrinos, for which any state may be occupied by at most one particle). It is more convenient to work with the grand partition function $\mathcal{Q} = \Pi_n\mathcal{Z}_n$, where $\mathcal{Z}_n$ is the canonical partition function for occupancy of the nth state, in which each particle present has energy ε_n and chemical potential μ. Thus we have

$$\mathcal{Z}_n = \sum_{j=0}^{\infty} e^{-j(\varepsilon_n-\mu)/(kT)} = \frac{1}{1 - e^{-(\varepsilon_n-\mu)/(kT)}} \quad \text{(bosons);}$$
$$\mathcal{Z}_n = \sum_{j=0}^{1} e^{-j(\varepsilon_n-\mu)/(kT)} = 1 + e^{-(\varepsilon_n-\mu)/(kT)} \quad \text{(fermions).}$$

If we define the fugacity as usual by $z = \exp[\mu/(kT)]$ we obtain [77]

$$\mathcal{Q}_{\mathrm{B}}(T) = \prod_n \Big[1 - z\exp\Big(-\frac{\varepsilon_n}{kT}\Big)\Big]^{-1} \quad \text{(bosons);}$$
$$\mathcal{Q}_{\mathrm{F}}(T) = \prod_n \Big[1 + z\exp\Big(-\frac{\varepsilon_n}{kT}\Big)\Big] \quad \text{(fermions).}$$

Consider the case of fugacity $z = 1$. If zero point energy is neglected and we write $\varepsilon_n = n\hbar\omega$ as in blackbody radiation, then on writing $x = \exp[-\hbar\omega/(kT)]$ we find [5] that

$$Q_{\mathrm{B}}(T) = \prod_{n=1}^{\infty}(1 - x^n)^{-1} = \frac{1}{\varphi(x)} :$$

$$Q_{\mathrm{F}}(T) = \prod_{n=1}^{\infty}(1 + x^n) = \frac{\prod_{n=1}^{\infty}(1 - x^{2n})}{\prod_{n=1}^{\infty}(1 - x^n)} = \frac{\varphi(x^2)}{\varphi(x)}, \tag{55}$$

where Euler's product $\varphi(x) = \prod_{n=1}^{\infty}(1 - x^n)$, intimately related to the Jacobi theta functions, is discussed in Appendix A and has the transformation formula

$$\varphi(e^{-x}) = \Big(\frac{2\pi}{x}\Big)^{1/2} \exp\Big(\frac{x}{24} - \frac{\pi^2}{6x}\Big)\varphi(e^{-4\pi^2/x}), \tag{56}$$

which is one of the mathematical correspondence principle relations from Table 1. The original expressions for the grand partition function are able to be used for computation for high temperatures, but are useless at low temperatures. However, Eq. (56) gives immediate access to low temperature expansions, since we can eliminate $\varphi(\exp[-\hbar\omega/(kT)])$ in favor of $\varphi(\exp[-4\pi^2 kT/(\hbar\omega)])$. Planat [78] has pursued the implications of the relation between Euler's product and the theory of partitions [79] in the context of the massless Bose gas.

If we retain the zero point energy and consider a general value of the fugacity, we can represent the grand partition functions in terms of the q-Pochhammer function [80]

$$(a;q)_{\infty} = \prod_{n=0}^{\infty}(1 - aq^n) = \sum_{n=0}^{\infty} \frac{(-1)^n q^{n(n-1)/2} a^n}{\prod_{m=1}^{n}(1 - q^m)}. \tag{57}$$

Building on the work of Rogers and Ramanujan [81], there is now an impressive corpus of transformations and identities related to theta functions and q-Pochhammer functions, and they arise frequently in mathematical physics [82].

6. CASIMIR FORCES

The famous prediction of Casimir [83, 84] that the zero temperature energy of interaction of two perfectly conducting plates a distance ℓ apart in vacuum provides an attractive force per unit area $\pi^2\hbar c/(240\ell^4)$ between the plates was a landmark result. Direct experimental verification was challenging, with Sparnaay [85] in 1958 finding that "the attractive interactions do not contradict Casimir's theoretical prediction" (the experiments had problematically large uncertainty). Finally, in 1997, Lamoreaux [86] effectively settled the basic issue [87]: "we have given an unambiguous demonstration of the Casimir force with accuracy of order 5%. Our data is not of sufficient accuracy to demonstrate the finite temperature correction …". (Casimir's original discussion did not address either finite temperature nor limitations on conductivity.) Crudely described, the Casimir effect demonstrates the consequences of geometrically constraining free oscillations of a system (here, the electromagnetic field) compared to the unconstrained state. Entirely classical analogues of the Casimir effect in macroscopic physics have been identified in a maritime context [88], and in an acoustic system suitable for lecture demonstrations [89].

The literature related to the Casimir effect is already voluminous and connections of papers with apparently cognate keywords to physics can be highly tenuous. At one extreme end of the literature [90, 91], since the evaluations of $\zeta(-1)$ and $\zeta(-3)$ are needed in the discussion of the physical Casimir effect (depending on the geometry), the evaluation for $s = -1$ of the analytic continuation of a series of the form $Z(s) = \sum_{\lambda}\lambda^{-s}$, where λ runs through a set of values with an interpretation related to energy levels, has been called the "Casimir energy". The sign of $Z(-1)$ "reflects certain dynamical and arithmetical properties" [91] and formulae related to the so-called Casimir energy can be obtained for compact Riemann surfaces of genus $g \geq 2$.

Of greater physical interest is the embedding of the original Casimir effect in a broader context that admits predictions of interactions between more general classes of matter than perfect conductors at zero temperature [75, 92]. Profoundly important papers by Lifshitz [93, 94, 95] and his subsequent work with Dzyaloshinskii and Pitaevskii [96, 97], also appearing in a later textbook [98], replaced perfect conductors in vacuum by dielectric materials separated by an intervening dielectric material. By permitting the dielectrics to have a frequency-dependent dielectric susceptibility, a wide variety of physical (and even biological) systems could be discussed, and subsequent work of Ninham and Parsegian [99, 100] showed how the required dielectric properties could be determined from spectroscopic data, leading to the ability to make quantitative predictions in experimentally accessible systems. The original Casimir problem arises as an extreme limit of the Lifshitz theory approach, and Lifshitz theory permits the computation of temperature-dependent effects [74, 75, 101]. Experimental validation of the predictions of Lifshitz theory has been obtained in many cases [102].

We focus on the original Casimir problem–plates separated by vacuum–because it exhibits most simply the importance of the mathematical correspondence principle. Let $F(\ell, T)$ denote the free energy of interaction per unit area between two infinite parallel conducting plates separated by a distance ℓ, in vacuum, at finite temperature T. It is instructive to see how extensive use of results of classical analysis such as the Riemann relation for the zeta function and various properties of the gamma function enable the free energy to be expressed in a highly informative way that enables dangerous issues concerning non-uniformity of asymptotic expansions to be dealt with [103]. Where we have already set the dielectric constant of the region between the plates to be unity, then from Lifshitz theory [75] we have

$$F(\ell, T) = \frac{kT}{8\pi\ell^2} \lim_{\Delta\to 1^-} \lim_{\bar{\Delta}\to 1^-} \sum_{n=0}^{\infty}{}' I(\xi_n, \ell), \tag{58}$$

where the prime on the sum indicates that the $n = 0$ term is to be weighted with a factor of 1/2, the parameter ξ_n is defined by

$$\xi_n = \frac{2\pi nkT}{\hbar} \tag{59}$$

and

$$I(\xi_n, \ell) = \Big(\frac{2\xi_n\ell}{c}\Big)^2 \int_1^{\infty} p\Big\{\ln\Big[1 - \Delta^2 \exp\Big(-\frac{2p\xi_n\ell}{c}\Big)\Big] + \ln\Big[1 - \bar{\Delta}^2 \exp\Big(-\frac{2p\xi_n\ell}{c}\Big)\Big]\Big\}dp. \tag{60}$$

To avoid an indeterminacy [104] in the case $n = 0$, we evaluate $I(\xi_n,\ell)$ for small positive real n by use of the change of variables $y = 2p\xi_n\ell/c$ and then take the limit $n \to 0$. The Riemann zeta function first arises from the $n = 0$ contribution from the $s = 3$ case of the integral (15).

For convenience in the asymptotic analysis we write

$$x = \frac{2kT\ell}{\hbar c}. \tag{61}$$

The coupling between the temperature T and the plate spacing ℓ is very important. The limit $x \to 0$ corresponds to the low-temperature limit, provided that the plate separation is constrained, or to the small-spacing limit, provided that the temperature is not too high. The analysis of Ninham and Daicic [103] to this point has

$$F(\ell, T) = \frac{kT}{8\pi\ell^2}\Big\{-\zeta(3) + \sum_{n=1}^{\infty}(2\pi nx)^2 \int_1^{\infty} 2p\ln(1 - e^{-2\pi nxp})dp\Big\}. \tag{62}$$

To evaluate the sum over n we may begin by expanding the logarithm using the series

$$\ln(1 - Z) = -\sum_{m=1}^{\infty} \frac{Z^m}{m} \quad (-1 \le Z < 1).$$

Since Eq. (13) shows that the gamma function is the Mellin transform [43] of the decaying exponential, using the Mellin inversion theorem [Eqs (30) and (31)] we have the contour integral representation

$$e^{-z} = \frac{1}{2\pi i}\int_{\kappa - i\infty}^{\kappa + i\infty} z^{-s}\Gamma(s)ds,$$

with the positive constant κ that places the vertical Bromwich contour $\mathrm{Re}(s) = \kappa$ selected to secure convergence in the subsequent analysis based on this integral ($\kappa > 3$ suffices). We now have

$$\ln(1 - e^{-2\pi nxp}) = -\sum_{m=1}^{\infty}\int_{\kappa - i\infty}^{\kappa + i\infty} \frac{(2\pi mnxp)^{-s}\Gamma(s)ds}{2\pi im},$$

so we can eliminate the logarithm factor from the integrand in Eq. (62), evaluate the resulting elementary integral over p and recognize the sums over m and n as series for the Riemann zeta function [Eq. (12)]. In this way one arrives at the scaled free energy

$$\mathcal{F} = \frac{\hbar c\ell}{(kT)^2}F(\ell, T) \tag{63}$$

$$= -\frac{\zeta(3)}{4\pi x} - \frac{1}{2\pi i}\int_{\kappa - i\infty}^{\kappa + i\infty} \frac{\zeta(s-2)\zeta(s+1)\Gamma(s)ds}{(2\pi x)^{s-1}(s-2)}. \tag{64}$$

The integrand has only four singularities, namely the simple poles at $s = -1, 0, 2$ and 3. To see this, we note that the gamma function has simple poles of residue $(-1)^j/j!$ at $s = -j (j = 0, 1, 2, 3, \ldots)$, while $\zeta(s - 2)\zeta(s + 1)$ has a simple pole at $s = 3$ and simple zeros at $s = -2, -3, -4, \ldots$ (noting that $\zeta(z)$ has a simple pole at $z = 1$ and simple zeros at $z = -2, -4, -6, \ldots$).

The Bromwich contour may be translated an arbitrary finite distance, provided that we account correctly for the residues at poles across which the contour is dragged. If we move the contour to $\mathrm{Re}(s) = 1$, then the term that must be added to account for the pole at $s = 2$ is easily shown to cancel with the first term on the right in Eq. (64). The pole at $s = 3$ leads to a term proportional to $1/x^2$, whose coefficient can be evaluated by recalling that $\zeta(0) = -1/2$ and $\zeta(4) = \pi^4/90$. We find that

$$\mathcal{F} = -\frac{\pi^2}{180x^2} - \frac{1}{2\pi i}\int_{1-i\infty}^{1+i\infty}\frac{\zeta(s-2)\zeta(s+1)\Gamma(s)ds}{(2\pi x)^{s-1}(s-2)}. \tag{65}$$

Since $\zeta(s-2) = -2^{s-2}\pi^{s-3}\sin(\pi s/2)\Gamma(3-s)\zeta(3-s)$ from the Riemann relation (11) and

$$\Gamma(3-s)\Gamma(s) = (2-s)(1-s)\Gamma(1-s)\Gamma(s) = \frac{(2-s)(1-s)\pi}{\sin(\pi s)} \tag{66}$$

we deduce that

$$\mathcal{F} = -\frac{\pi^2}{180x^2} + J(x), \tag{67}$$

where the first term on the right corresponds to the Casimir formula, while

$$J(x) = -\frac{1}{2\pi i}\int_{1-i\infty}^{1+i\infty}\frac{(1-s)\zeta(3-s)\zeta(s+1)ds}{4\pi\cos(\pi s/2)x^{s-1}}. \tag{68}$$

the change of variables $s = 1 + it$ produces

$$\begin{aligned} J(x) &= -\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{t\zeta(2-it)\zeta(2+it)dt}{4\pi\sinh(\pi t/2)x^{it}} \\ &= -\int_0^{\infty}\frac{t\zeta(2-it)\zeta(2+it)\cos[t\ln(x)]dt}{4\pi^2\sinh(\pi t/2)}. \end{aligned} \tag{69}$$

Since $\ln(x) = -\ln(1/x)$, Eq. (69) reveals the remarkable inversion symmetry

$$J(x) = J(x^{-1}). \tag{70}$$

The Casimir term is temperature-independent when one returns to the original variables, but the function $J(x)$ encapsulates genuine temperature dependence. The Riemann relation, one avatar of our central theme summarized in Table 1, has been used to expose the inversion symmetry (a previously observed result [107, 108]), but is also crucial for an efficient extraction of the x dependence. It may be noted in passing that a computer algebra software (such as Mathematica) is very helpful in checking that the intricate manipulations involved are correct, but is presently (and in the foreseeable future may well continue to be) unable to offer much help in guiding the analysis.

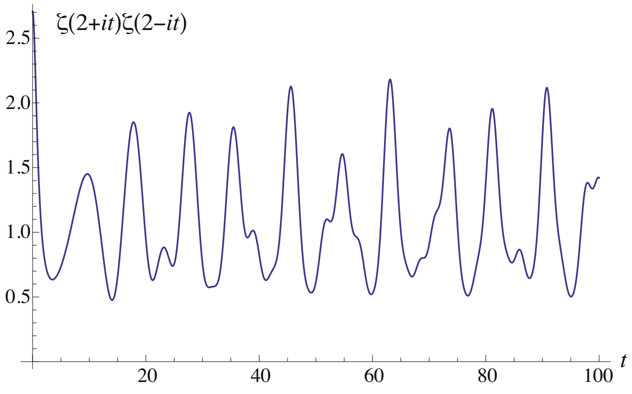


Figure 8. The function $\zeta(2+it)\zeta(2-it)$ in the integrand in Eq. (69): $0 < \zeta(2+it)\zeta(2-it) \le \zeta(2)^2 = \pi^4/36 \approx 2.7$ for all $t \in \mathbb{R}$.

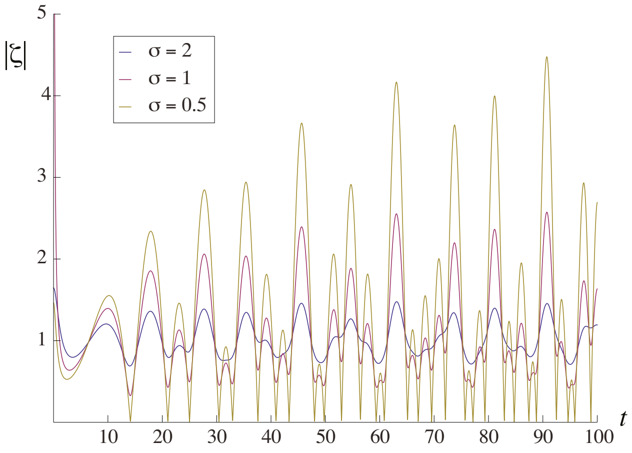


Figure 9. $|\zeta(\sigma+it)| = \sqrt{\zeta(\sigma+it)\zeta(\sigma-it)}$ for $\sigma \in \{1/2, 1, 2\}$, $0 < t < 100$.

We digress for a moment. The function

$$\zeta(2+it)\zeta(2-it)$$

in the integrand in Eq. (69) is easily shown to be real-valued, but is by no means simple in structure: see Fig. 8. In qualitative terms, it appears roughly periodic, but the amplitudes of successive peaks and troughs and their spacing vary in an apparently random manner. More generally, for real σ and t, we have

$$\zeta(\sigma+it)\zeta(\sigma-it) = |\zeta(\sigma+it)|^2. \tag{71}$$

and the strange behavior of $|\zeta(2+it)|^2$ revealed in Fig. 8 arises also for other values of σ. When studying the response to changes in σ it is helpful to consider $|\zeta(\sigma+it)|$ rather than $|\zeta(\sigma+it)|^2$ to reduce the height of the maxima. We plot $|\zeta(\sigma+it)|$ in Fig. 9 for $\sigma = 2$, $\sigma = 1$ and $\sigma = 0.5$. It may be observed that the spacing of the peaks and troughs hardly changes, but the peak heights grow and the trough heights fall as σ decreases. It is relatively easy to prove that $\zeta(\sigma+it)$ is nonzero whenever $\sigma > 1$. That $\zeta(\sigma+it)$ is also never 0 for $\sigma = 1$ is much more difficult to prove. Establishing this was an essential ingredient of the proofs in 1896 by Hadamard [105] and de la Vallée

Poussin [106] that the number of prime numbers less than or equal to n has the asymptotic form $n/\ln n$ as $n \to \infty$. In the context of our discussion, the still-unresolved Riemann hypothesis [17, 19], asserts that $\zeta(\sigma + it) \neq 0$ for $\sigma > 1/2$. It is strangely beautiful that the mathematics of the Casimir effect comes so close to such subtle matters.

Returning to matters more overtly connected to physics, if F_{Casimir} denotes the original single-term Casimir energy prediction for the energy per unit area, we have

$$\frac{F(\ell, T)}{F_{\text{Casimir}}} = 1 + \eta(x), \qquad \eta(x) = \frac{180x^2}{\pi^2} J(x). \tag{72}$$

In assessing the accuracy of the Casimir term, the role of the composite parameter x is crucial. We note that with ℓ measured in metres and T in Kelvin, we have

$$x = \frac{2kT\ell}{\hbar c} \approx 873T\ell. \tag{73}$$

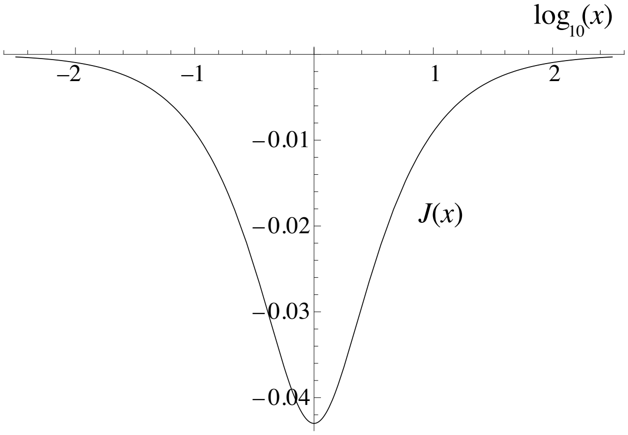


Figure 10. The function $J(x)$ computed from the integral (69) by numerical integration.

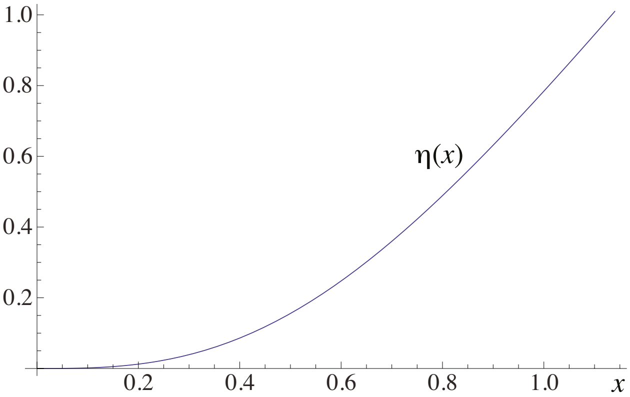


Figure 11. The fractional error $\eta(x)$ in the single-term Casimir formula, defined by Eq. (72), inferred from the numerically evaluated integral (69).

In principle the numerical evaluation of the integral (69) requires some care because of the rapid oscillation of the cosine factor when $x \gg 1$ or $0 < x \ll 1$ and the erratic behavior of the real-valued function $\zeta(2 + it)\zeta(2 - it)$ (see Fig. 8). However, the Riemann–Lebesgue Lemma ensures that $J(x) \to 0$ as $x \to 0$ or as $x \to \infty$ and for the region where $J(x)$ differs perceptibly from zero, Mathematica is up to the task. We show $J(x)$ for $10^{-2} \leq x \leq 10^2$ in Fig. 10. The fractional error η in the one-term Casimir formula, defined in Eq. (72), exceeds 1 for $x > 1.14$ but is less than 5×10^{-5} for $x \leq 0.03$.

In applications of the Casimir formula to experimental situations, relative errors associated with finite conductivity, departure of the real geometry from infinite parallel plates and other practical realities may dominate over the errors arising from the finiteness of the temperature that we have quantified through $\eta(x)$. Having acknowledged that caveat, we note that the case $x \approx 0.5$ arises for atomic dimensions ($\ell \approx 10^{-10}$m) when $T \approx 6 \times 106$ K, within the range of estimated temperatures for the sun ($\approx 4 \times 10^3$K at the surface, 1.6×10^7K at the center [109]).

The analytic structures that have been revealed with the techniques illustrated above have a number of interesting consequences for the Casimir problem in vacuum and for analogous problems involving dielectric or conducting films. Some of these, including connections with nuclear and particle physics, have been explored elsewhere [110, 111]. The point to be made is that viewing physical problems from a mathematical perspective in the spirit of Table 1 leads both to efficient practical analysis and to new insights, though deep and subtle mathematical exotica are seldom far away.

7. CONCLUSIONS

We opened this paper with a reference to the conundrum of the surprising effectiveness of mathematics in physics. Berry has proposed one possible explanation [112].

> *We are beings of finite intelligence in an infinite inscrutable universe. In science, our individual intelligences cooperate, and we can understand more. But still, we are able to comprehend only those structures in the natural world that mirror our mental constructs. And at any stage of humanity's development, the most sophisticated constructs are those of our mathematics. Therefore our deepest penetration into the natural world is limited by our latest mathematics. As mathematics develops, more subtle features of*

the universe become accessible to our understanding... So, 'The unreasonable effectiveness of mathematics in the natural sciences' is not unreasonable at all; on the contrary, it is inevitable.

This is not the only possible explanation, and in some areas, there are credible alternatives. We have shown that what is essentially one grand mathematical idea, which comes expressed in various ways, such as those collected in Table 1, underlies a wide range of apparently disparate physical phenomena. If one is a mathematical platonist–that is, a believer in the existence of abstract mathematical objects that are independent of intelligent agents and their language, thought and practices [113]–it is perhaps not such a leap of faith to conceive of a profound connection between some of these mathematical objects and physical reality.

Amongst a charming collection of pithy quotes and witticisms relevant to science collected by Berry [114] one finds what he calls "three laws of discovery".

1. Discoveries are rarely attributed to the correct person [115].
2. Nothing is ever discovered for the first time [116].
3. Everything of importance has been said before by someone who did not discover it [117].

The existence of a common underlying mathematical theme in many contexts may be an explanation for the applicability of Berry's laws in the sociology of physics.

Most physicists will wisely choose to limit their pondering of metaphysical questions to the bar or coffee shop, but the contemplation of what may be the most natural mathematical framework for physical theory and the pursuit of the implications of the mathematics is a more defensible use of one's office hours. While physical intuition and accumulated conventional wisdom are always worthy of respect, careful analyses with appropriate mathematical insight can yield surprising results. Recently, Lekner [118] has shown that at small separations, charged conducting spheres always attract each other, even when the charges on the spheres are of the same sign, except when the spheres have charges in the ratio that would make them an equipotential surface on contact. This refutation of the rule that "like charges repel" in classical physics is indeed striking. In quantum mechanics, where physical intuition is a more contentious matter (and in the view of many, not even appropriate), accumulated conventional wisdom has still developed, but as noted recently by Ball [119] it is by no means a settled matter that we have either the optimal perspective on the subject or the optimal formulation.

What is the appropriate mathematical training for the modern physicist may be hotly debated, but what we have styled a correspondence principle (embodied in Table 1) and the associated treasures of classical real and complex analysis have legitimate claim for inclusion.

ACKNOWLEDGMENTS

We are grateful to the referees for helpful comments, and for bringing several important references to our attention.

APPENDIX A. JACOBI THETA FUNCTIONS

Where $\mathrm{Im}\{\tau\} > 0$ to secure convergence in the sense of classical analysis and $q = e^{i\pi\tau}$ (so that $|q| < 1$), the four Jacobi theta functions $\theta_k(z|\tau) = \theta_k(z, q)$ are

$$\begin{aligned}
\theta_1(z|\tau) &= 2\sum_{n=0}^{\infty}(-1)^n \exp\{\pi i(n+\tfrac{1}{2})^2\tau\}\sin[(2n+1)z] \\
&= 2\sum_{n=0}^{\infty}(-1)^n q^{(n+1/2)^2}\sin[(2n+1)z] \\
&= 2G(q)q^{1/4}\sin z\prod_{n=1}^{\infty}[1-2q^{2n}\cos(2z)+q^{4n}];
\end{aligned}$$

$$\begin{aligned}
\theta_2(z|\tau) &= 2\sum_{n=0}^{\infty}\exp\{\pi i(n+\tfrac{1}{2})^2\tau\}\cos[(2n+1)z] \\
&= 2\sum_{n=0}^{\infty}q^{(n+1/2)^2}\cos[(2n+1)z] \\
&= 2G(q)q^{1/4}\cos z\prod_{n=1}^{\infty}[1+2q^{2n}\cos(2z)+q^{4n}];
\end{aligned}$$

$$\begin{aligned}
\theta_3(z|\tau) &= \sum_{n=-\infty}^{\infty}\exp\{\pi i n^2\tau+2inz\} = 1+2\sum_{n=1}^{\infty}q^{n^2}\cos(2nz) \\
&= G(q)\prod_{n=1}^{\infty}[1+2q^{2n-1}\cos(2z)+q^{4n-2}];
\end{aligned}$$

$$\begin{aligned}
\theta_4(z|\tau) &= \sum_{n=-\infty}^{\infty}(-1)^n\exp\{\pi i n^2\tau+2niz\} \\
&= 1+2\sum_{n=1}^{\infty}(-1)^n q^{n^2}\cos(2nz) \\
&= G(q)\prod_{n=1}^{\infty}[1-2q^{2n-1}\cos(2z)+q^{4n-2}].
\end{aligned}$$

Here

$$G(q) = \prod_{n=1}^{\infty}(1 - q^{2n}) = \varphi(q^2)$$

and Euler's product $\varphi(x) = \prod_{n=1}^{\infty}(1 - x^n)$ has the remarkable property [5] that

$$\varphi(e^{-x}) = \left(\frac{2\pi}{x}\right)^{1/2} \exp[x/24 - \pi^2/(6x)]\varphi(e^{-4\pi^2/x}).$$

Where $0 < \arg(\tau) < \pi$, the Jacobi transformation formulae are

$$\theta_1(z|\tau) = -i\exp(\frac{i\pi}{4} - \frac{iz^2}{\pi\tau})\tau^{-1/2}\theta_1\left(\frac{z}{\tau}\middle| -\frac{1}{\tau}\right),$$

$$\theta_2(z|\tau) = \exp(\frac{i\pi}{4} - \frac{iz^2}{\pi\tau})\tau^{-1/2}\theta_4\left(\frac{z}{\tau}\middle| -\frac{1}{\tau}\right).$$

$$\theta_3(z|\tau) = \exp(\frac{i\pi}{4} - \frac{iz^2}{\pi\tau})\tau^{-1/2}\theta_3\left(\frac{z}{\tau}\middle| -\frac{1}{\tau}\right),$$

$$\theta_4(z|\tau) = \exp(\frac{i\pi}{4} - \frac{iz^2}{\pi\tau})\tau^{-1/2}\theta_2\left(\frac{z}{\tau}\middle| -\frac{1}{\tau}\right).$$

We observe that in a formal sense, if not within classical analysis (with the limits taken under the restrictions that $|q| < 1$ or $\mathrm{Re}\{\tau\} > 0$, respectively) that

$$\lim_{q\to 1}\theta_3\left(\frac{\pi x}{2L}, q\right) = \lim_{\tau\to 0}\theta_3\left(\frac{\pi x}{2L}\middle|\tau\right) = L\sum_{n=-\infty}^{\infty}\delta(x - 2nL).$$

APPENDIX B. DIFFUSION WITH INCREASING LETHARGY

There is another interesting model for a kind of anomalous diffusion that raises mathematical questions that are at least as subtle as those discussed in Section 3.2 and partly related to them [120]. Consider a random walk in one dimension, for which steps to the left and right are equally likely at each stage, but the length of the nth step is c_n, where $c_n > 0$ and $\sum_{n=1}^{\infty} c_n < \infty$. The walker exhibits increasing lethargy, and asymptotically comes to rest at a random position X relative to the starting location, where $X = \sum_{n=1}^{\infty} \epsilon_n c_n$, the random variables $\{\varepsilon_n\}$ are independent, and $\Pr\{\varepsilon_n = 1\} = \Pr\{\varepsilon_n = -1\} = 1/2$. In the probability literature, the random variable X is described as an "infinite Bernoulli convolution" and it can be proved [121] that provided that $\sum_{n=1}^{\infty} c_n^2 < \infty$, the characteristic function (Fourier transform) associated with the distribution of this random variable exists, and is given by $\mathsf{E}\{\exp(iqX)\} = \prod_{n=1}^{\infty}\cos(c_n q)$.

The case $c_n = \alpha^{n-1}$, where $0 < \alpha < 1$, for which $\mathrm{E}\{\exp(iqX)\} = \prod_{n=0}^{\infty}\cos(\alpha^n q)$, is especially fascinating [122, 123, 124, 125]. This model builds in self-similarity in a way reminiscent of, but different to, Section 3.2. We have $X = \varepsilon_1 + \alpha X_1$, where X and X_1 are identically distributed random variables, while ε_1 and X_1 are independent. Since $\sum_{n=0}^{\infty}\alpha^n = (1 - \alpha)^{-1}$, we know that $-(1 - \alpha)^{-1} \le X \le (1 - \alpha)^{-1}$. For the case $\alpha = 1/2$, it can be proved [127] that X is uniformly distributed on $[-2, 2]$, but for many other values of $\alpha \in (0, 1)$ the distribution of X does not apportion the probability so smoothly.

In general [126], the cumulative distribution function $F(x) = \Pr\{X \le x\}$ of a real random variable X consists of either a single one, or a linear combination of both, of the following components: (i) an "absolutely continuous" component, corresponding to classical probability density function; (ii) a "singular" component, in which the probability all resides on a set of measure zero. For the increasingly lethargic walk with $0 < \alpha < 1/2$, X has a singular distribution: the nonzero probability all resides on a Cantor set of measure zero [123]. The rapid attenuation of the step lengths prevents the walker from exploring the apparent support decently. For $1/2 < \alpha < 1$, it has been proved [122] that the distribution of X for any given value of α is either entirely absolutely continuous or entirely singular. The simplicity of the case $\alpha = 1/2$ might suggest that absolute continuity always prevails for $1/2 < \alpha < 1$, and it has been proven that the set of values of $\alpha \in (1/2, 1)$ for which the distribution is singular has measure zero [128], but a countable number of values of $\alpha \in (1/2, 1)$ for which the distribution is singular were found by Erdös [124] and the search for other anomalous α values continues.

Diffusion with accumulating lethargy also has interesting connections to the discussion of Section 4, where the golden ratio $\tau = (1 + \sqrt{5})/2$ plays a significant role. Hu [129] has shown that for $\alpha = 1/\tau = (\sqrt{5} - 1)/2$ and for $-(1 - \alpha)^{-1} < x < (1 - \alpha)^{-1}$, the local fractal dimension

$$d(x) = \lim_{r\to 0^+} + \log[\Pr\{x - r \le X \le x + r\}]/\log(r)$$

of the distribution of the random variable has maximum value $\log(2)/\log(\tau) \approx 1.4404$ and minimum value $\log(2)/\log(\tau) - 1/2 \approx 0.9404$. For additional related results see Lau and Ngai [130].

REFERENCES

[1] The quotation is from the retirement speech given by Sir Charles Frank (1911–1998) at the University of Bristol in 1976, as recorded by M.V. Berry, "Bristol Anholonomy Calendar", in *Sir Charles Frank OBE FRS, an eightieth birthday tribute*, edited by R.G. Chambers, J.E. Enderby, A. Keller, A.R. Lang, and J.W. Steeds (Adam Hilger, Bristol, 1991), pp. 207-219.

[2] M. Kline, *Mathematics: The Loss of Certainty* (Oxford University Press, New York, 1980).

[3] E.P. Wigner, "The unreasonable effectiveness of mathematics in the natural sciences", Richard Courant lecture in mathematical sciences delivered at New York University, May 11, 1959, Communications on Pure and Applied Mathematics **13**(1), 1–14 (1960).

[4] R.W. Hamming, "The unreasonable effectiveness of mathematics," American Mathematical Monthly **87**(2) 81–90 (1980).

[5] B.W. Ninham, B.D. Hughes, N.E. Frankel, and M.L. Glasser, "Möbius, Mellin, and mathematical physics," Physica A **186**, 441–481 (1992).

[6] N. Bohr, "Über die Serienspektra der Element," Zeitschrift für Physik **2** 423–478 (1920). English translation in L. Rosenfeld and J. Rud Nielsen (editors), *Niels Bohr, Collected Works, Volume 3, The Correspondence Principle (1918–1923)*, pp. 241–282 (North-Holland, Amsterdam, 1976).

[7] W.H. Cropper, *The Quantum Physicists* (Oxford University Press, New York, 1970) notes that 'Although the correspondence principle became increasingly elaborate in later work, it was always based on one simple concept: that when the scale is suitably adjusted, classical physics and quantum physics must merge.'

[8] N. Bohr, "The quantum postulate and the recent development of atomic theory," Nature **121**, 580–590 (1928). Bohr writes (p. 580) '… if in order to make observation possible we permit certain interactions with suitable agencies of measurement, not belonging to the system, an unambiguous description of the state of the system is no longer possible, and there can be no sense of causality in the ordinary sense of the word. The very nature of the quantum theory thus forces us to regard the space-time coordination and the claim of causality, the union of which characterises the classical theories, as complementary but exclusive features of the description…'. See also N. Bohr, "Discussions with Einstein on epistemological problems in atomic physics," in P. Schilpp (editor), *Albert Einstein: Philosopher-Scientist* (Open Court, Chicago, 1949).

[9] M.J. Lighthill, *Introduction to Fourier Analysis and Generalised Functions* (Cambridge University Press, Cambridge, U.K., 1958).

[10] C.G.J. Jacobi, "Theorie der elliptischen Functionen aus den Eigenschaften der Thetareihen abgeleitet", in *Gesammelte Werke*, Band 1, pp. 497–538 (Reimer, Berlin, 1881; reprinted Chelsea, New York, 1969).

[11] E.T. Whittaker and G.N. Watson, *A Course of Modern Analysis*, 4th edition (Cambridge University Press, Cambridge, U.K.,1927).

[12] The wave-particle duality aspect of our central theme can be embedded in a somewhat broader context that we only touch on here. (The authors are grateful to one of the referees for suggesting that we address this, and referring us to relevant literature). Poisson's summation formula, in both the guises (2) and (10), relates spectral information (eigenvalues) to a discrete geometry (a regular lattice). This hints at a class of relations between spectral sums and geometrical or topological sums. Perhaps the simplest manifestation of this is the relation between the ray and mode descriptions of waveguides, which represent complementary perspectives: see C.L. Pekeris, "Ray theory vs normal mode theory in wave propagation problems", Proceedings of Symposia in Applied Mathematics **2**, 71–75 (1950). However, there are many examples, such as: (i) scattering from spheres, with spectral sums over angular momentum vs rays winding different numbers of times round the scattering center (M.V. Berry and K.E. Mount, "Semiclassical approximations in wave mechanics", Reports on Progress in Physics **35**, 315–397 (1972)); (ii) electron diffraction in crystals, with spectral sums over Bloch waves vs classical trajectories winding through the lattice (M.V. Berry, "Diffraction in crystals at high energies", Journal of Physics C **4**, 697–722 (1971)); and (iii) complementary approximate solution methods for the energy eigenvalue problem of quantum mechanics when separation of variables is not available (M.C. Gutzwiller, "Periodic orbits and classical quantization conditions", Journal of Mathematical Physics.**12**, 343–358 (1971)).

[13] A. Córdoba, "La formule sommatoire de Poisson," Comptes rendus de l'Académie des sciences. Série 1, Mathématique **306** (no 8) 373–376 (1988).

[14] A. Córdoba, "Dirac combs", Letters in Mathematical Physics **17**, 191–196 (1989).

[15] See P.L. Butzer, P.J.S.G. Ferreira, G. Schmeisser and R.L. Stens, "The summation formulae of Euler–Maclaurin, Abel–Plana, Poisson, and their interconnections with the approximate sampling formula of signal analysis," *Results in Mathematics* **59**, 359–400 (2012). These authors note that Eq. (10) was first produced by Gauss. The results such as (4) are usually called Jacobi's transformation after their appearance in several of Jacobi's works [C.G.J. Jacobi, "Suite des notices sur les fonctions elliptiques," Journal für die reine und angewandte Mathematik (Crelle's Journal) **3**, 403–404 (1828); C.G.J. Jacobi, "Über die Differentialgleichung, welcher die Reihen $1 \pm 2q \pm 2q^4 \pm 2q^9 +$ etc., $2\sqrt[4]{q} + 2\sqrt[4]{q^9} + 2\sqrt[4]{q^{25}} +$ etc. Genüge leisten," Journal für die reine und angewandte Mathematik (Crelle's Journal) **36**, 97–112 (1848). However if Gauss did not produce them earlier, they are certainly present in work of Poisson, which Jacobi has acknowledged [S.D. Poisson, "Suite du mémoire sur les intégrales définies et sur la sommation des séries, inséré dans les précédens volumes de ce Journal", *Journal de l'École royale polytechnique* **12** (19), 404–509 (1823)].

[16] E.C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, 2nd edition, (Oxford University Press, Oxford, U.K., 1948; reprinted Chelsea, New York, 1986).

[17] B. Riemann, "Ueber die Anzahl der Primzahlen unter einer gegebenen Grösse," Monatsberichte der Königliche Preußische Akademie der Wissenschaften zu Berlin aus der Jahre 1859, 671–680 (1860).

[18] Riemann shows that where $\psi(x) = \sum_{n=1}^{\infty} e^{-n^2\pi x}$ and $\mathrm{Re}\{s\} > 1$, we have $\zeta(s)\pi^{-s/2}\Gamma(s/2) = \int_0^{\infty} x^{s/2-1}\psi(x)dx$. He uses the transformation formula for $\theta_3(0|ix)$ to construct an analytic continuation for $\zeta(s)$ $\pi^{-s/2}\Gamma(s/2)$ that is symmetric under replacement of s by $1 - s$, and the Riemann relation follows.

[19] E.C. Titchmarsh, *The Theory of the Riemann Zeta-function*, 2nd edition, revised by D.R. Heath-Brown (Oxford University Press, Oxford, U.K., 1986).

[20] D. Schumayer and D.A.W. Hutchinson, "Physics of the Riemann hypothesis", Reviews of Modern Physics **83**, 307–330 (2011).

[21] A shorter formal argument based on a binomial expansion of $(n + a)^{-s}$ with $a \in (0, 1)$ and subsequent use of the limit $a \to 1$ yields the formal expression [5]

$$1 - s\zeta(s+1) + \frac{s(s+1)}{2!}\zeta(s+2) - \frac{s(s+1)(s+2)}{3!}\zeta(s+3) + \cdots = 0.$$

from which the same conclusions about simple known values of $\zeta(-m)$ for m a non-negative integer be drawn. The method given in the present paper is classically rigorous. For an alternative rigorous approach, leading to the identity

$$\sum_{q=0}^{\infty} \frac{(s-1)(s)_q}{(q+1)!}[\zeta(s+q) - 1] = 1,$$

see §68 of E. Landau, *Handbuch der Lehre von der Verteilung der Primzahlen* (1st edition, Leipzig and Berlin, Teubner, 1909; 3rd edition, Providence, Rhode Island, AMS Chelsea, 1974).

[22] R. Bellman, *A Brief Introduction to Theta Functions* (Holt, Rinehart and Winston, New York, 1961).

[23] Although we have produced Eq. (2) from Eq. (1), the process is quite reversible. They are essentially equivalent results. J.N. Lyness and B.W. Ninham ["Numerical quadrature and asymptotic expansions", Mathematics of Computation **21**, 162–178 (1967)] have shown how Eq. (1) can be used to deduce a very broad class of quadrature rules for numerical integration on finite intervals, with useful asymptotic expansions for the quadrature error. Further developments of this approach lead to practical techniques for computing analytic continuations of functions defined by parametrized integrals, by direct computation of the integrals even for parameter ranges where the integral diverges. See B.W. Ninham, "Generalised functions and divergent integrals", Numerische Mathematik **8**, 444–457 (1966).

[24] For properties of the modular group and proofs of results discussed in Section 2.2 see the following texts: W. Magnus, *Noneuclidean Tesselations and their Groups* (Academic Press, New York, 1974); J.P. Serre, *A Course in Arithmetic* (Springer-Verlag, New York, 1973); and B. Schoeneberg, *Elliptic Modular Functions* (Springer-Verlag, Berlin, 1974).

[25] E. Fabry, "Sur les pointes singuliers d'une fonction donnée par son développement en série et sur l'impossibilité du prolongement analitique dans les cas très généaux", Annales de 'École normale supérieure (3) **13**, 107–114 (1896).

[26] P. Dienes, *The Taylor Series: an Introduction to the Theory of Functions of a Complex Variable* (Oxford University Press, Oxford, U. K., 1931).

[27] For an easy proof, multiply Eq. (20) by x^2 and integrate by parts twice. We have normalized the total mass to unity, so that $v(x, t)$ can be interpreted as a probability density function.

[28] A. Fick, “Ueber Diffusion”, *Annalen der Physik* **94**, 59–86 (1855), translated as A. Fick, “On liquid diffusion”, *Philosophical Magazine* **10**, 30–39 (1855).

[29] E.W. Montroll and B.J. West, “On an enriched collection of stochastic processes”, in E.W. Montroll and J.L. Lebowitz (editors), *Fluctuation Phenomena*, pp. 61–173 (North-Holland, Amsterdam, 1979).

[30] B.D. Hughes, *Random Walks and Random Environments*, Vol. 1 (Oxford University Press, Oxford, U.K., 1995).

[31] R. Kutner, A. Pękalski, and K. Sznajd-Weron (editors), *Anomalous Diffusion: from Basics to Applications* (Springer, Berlin, 1999).

[32] R. Klages, G. Radons, and I.M. Sokolov (editors), *Anomalous Transport: Foundations and Applications* (Wiley-VCH, Weinheim, Germany, 2008).

[33] B.B. Mandelbrot, *The Fractal Geometry of Nature* (W.H. Freeman, San Francisco, 1982).

[34] B.D. Hughes, M.F. Shlesinger, and E.W. Montroll, “Random walks with self-similar clusters”, Proceedings of the National Academy of Sciences (U.S.A.) **78**, 3287–3291 (1981).

[35] When $\mu > 2$ there is a standard diffusive limit when the time-step t_0 and spatial scale Δ are sent to zero with Δ^2/t_0 held constant. For $\mu = 2$ it is necessary to take $\Delta^2 \ln(1/\Delta)/t_0$ constant to recover diffusion, and for $0 < \mu < 2$ diffusion is never attained.

[36] L.P. Kadanoff, “Scaling laws for Ising models near T_c”, Physics (Long Island City, N.Y.) **2**, 263–272 (1966).

[37] T. Niemeijer and J.M.J. van Leeuwen, (1976). “Renormalization: Ising-like spin systems”, In C. Domb and M.S. Green (editors), *Phase Transitions and Critical Phenomena*, Vol. 6, pp. 425–505. (Academic Press, London, 1976).

[38] For $\mu \leq 1/2$ the series for Q needs to be summed by Abelian means, with a convergence factor $e^{-\delta|m|}$ inserted and the limit $\delta \to 0$ taken after evaluation of the sum.

[39] M.F. Shlesinger and B.D. Hughes, “Analogs of renormalization group transformations in random processes”, Physica A **109**, 597–608 (1981).

[40] W.J. Reed and B.D. Hughes, “On the distribution of family names”, Physica A **319**, 579–590 (2003).

[41] D. Sornette, “Discrete-scale invariance and complex dimensions”, Physics Reports **297** (5), 239–270 (1998).

[42] G.H. Hardy, “Weierstrass’ non-differentiable function”, Transactions of the American Mathematical Society **17**, 301–325 (1916).

[43] For textbook discussions of the Mellin transform techniques used here, see N. Bleistein and R.A. Handelsman, *Asymptotic Expansions of Integrals* (Dover, New York, 1986), B. Davies, *Integral Transforms and Their Applications*, 3rd edition (Springer-Verlag, New York, 2002), or Appendix 2 of Hughes [30].

[44] P. Lévy, *Théorie de l’addition des variables aléatoires* (Gauthier-Villars, Paris, 1937).

[45] Although we have discussed here only the symmetric densities with Fourier transforms $\exp(-c|q|\mu t)$, there is a more general theory of stable densities, containing additional parameters. In this more general context, simple closed form expressions for densities are usually not available, but there have been some interesting developments since 2010. See K.A. Penson and K. Górska, “Exact and explicit probability densities for one-sided Lévy stable distributions”, Physical Review Letters **105**, 210604 (2010); and K. Górska and K.A. Penson, “Lévy stable two-sided distributions: exact and explicit densities for asymmetric case”, Physical Review E **83**, 061125 (2011).

[46] S.A. Fulling and K. S. Güntürk, “Exploring the propagator of a particle in a box,” American Journal of Physics **71**, 55–63 (2003).

[47] B. Gaveau and L.S. Schulman, “Explicit time-dependent Schrödinger propagators,” Journal of Physics A **19** 1833–1846 (1986).

[48] C.G. Darwin, “Free motion in the wave mechanics”, Proc. R. Soc. Lond. A 117 (1927) 258–293.

[49] E.H. Kennard, “Zur Quantenmechanik einfacher Bewegungstypen”, Zeitschrift für Physik 44 (1927) 326-352.

[50] J. Lekner, “Airy wavepacket solution of the Schrödinger equation,” European Journal of Physics **30**, L43–L46 (2009).

[51] W.H. Zachariasen, *Theory of X-Ray Diffraction in Crystals* (Wiley, New York, 1945).

[52] E. Zolotoyabko, *Basic Concepts of Crystallography* (Wiley-VCH, Weinheim, Germany, 2011).

[53] This is a slight over-simplification from the experimental perspective. The amplitude of the Fourier transform can be measured experimentally, but its phase is not directly available.

[54] Let $\mathcal{L}$ be a d-dimensional lattice, that is, a set of points with position vectors $\sum_{j=1}^{d} m_j \mathbf{a}_j$, where $m_j \in \mathbb{Z}$ and the vectors $\{\mathbf{a}_1, \ldots \mathbf{a}_d\}$ are a linearly independent set. The dual lattice or reciprocal lattice consists of these points whose potion vectors $\boldsymbol{\ell}^*$have integer-valued dot products with every position vector $\boldsymbol{\ell}$ of a point in $\mathcal{L}$. Then (see, for example, Theorem 3.2 in Senechal[61]) the Fou-

rier transform of the mass distribution $\rho(\mathbf{r}) = \sum_{\boldsymbol{\ell}\in\mathcal{L}} \delta(\mathbf{r}-\boldsymbol{\ell})$ defined by placing unit mass at each point of $\mathcal{L}$ is given by $\tilde{\rho}(\mathbf{k}) = \sum_{\boldsymbol{\ell}^*\in\mathcal{L}^*} \delta(\mathbf{k}-\boldsymbol{\ell}^*)$. More general results which essentially account for all simple cases have been established by Córdoba [13], who has shown that if μ and ν are two discrete subsets of $\mathbb{R}^d$ (this requires the existence of a nonzero lower bound for the spacing between pairs of points) and $c(\mathbf{s}) > 0$ for all $\mathbf{s} \in \mu$ then the relations

$$\rho(\mathbf{r}) = \sum_{\mathbf{s}\in\mu} \delta(\mathbf{r}-\mathbf{s}), \quad \tilde{\rho}(\mathbf{k}) = \sum_{\mathbf{s}\in\nu} c(\mathbf{s})\delta(\mathbf{k}-\mathbf{s}),$$

between a mass distribution $\rho(\mathbf{r})$ and its Fourier transform

$$\tilde{\rho}(\mathbf{k}) = \int_{\mathbb{R}^d} \exp(2\pi i\mathbf{k}\cdot\mathbf{r}) f(\mathbf{r}) d^d\mathbf{r}$$

can hold simultaneously if and only if the following conditions hold: $c(\mathbf{s}) = 1$ and there exists a linear transformation A of $\mathbb{R}^d$ with determinant 1 such that $\mu = A\mathbb{Z}^d$ and $\nu = (A^{-1})^T\mathbb{Z}^d$, where T denotes the transpose. It may be noted that because the definitions of the Fourier transform and its inverse are simply complex conjugates, the factor of $c(\mathbf{s})$ can be placed in either one of the sums over μ and ν without altering the result. We have used this in rephrasing Córdoba's result to suit our notation and our application of interest, where it is more natural to assert that all masses are the same in physical space and allow position-dependent coefficients in the sum obtained on taking the Fourier transform.

[55] D. Shechtman, I. Blech, D. Gratias, and J. W. Cahn, "Metallic phase with long-range orientational order and no translational symmetry," Physical Review Letters **53**, 1951–1953 (1984).

[56] D. Levine and P. J. Steinhardt, "Quasicrystals: a new class of ordered structures," Physical Review Letters **53**, 2477–2480 (1984).

[57] C. Janot, *Quasicrystals: a Primer*, 2nd edition (Oxford University Press, Oxford, U.K., 1994).

[58] R. Penrose, "The role of aesthetics in pure and applied mathematical research", Bulletin of the Institute of Mathematics and its Applications **10**, 266–271(1974).

[59] A. Hof, "On diffraction on aperiodic structures", Communications in Mathematical Physics **169**, 25–43 (1995).

[60] M. Senechal, "Generalizing crystallography: puzzles and problems in dimension 1," in I. Hargittai (ed.), *Quasicrystals, Networks and Molecules of Fivefold Symmettry* (VCH, Weinheim, Germany, 1990).

[61] M. Senechal, *Quasicrystals and Geometry* (Cambridge University Press, Cambridge U.K., 1995).

[62] M. Senechal and J. Taylor, "Quasicrystals: the view from Les Houches," Mathematical Intelligencer **12** (2), 54–64 (1990).

[63] M. Senechal and J. Taylor, "Quasicrystals: the view from Stockholm," Mathematical Intelligencer **35** (2), 1–9 (2013).

[64] M. Baake and U. Grimm, *Aperiodic Order, Volume 1, A Mathematical Invitation* (Cambridge University Press, Cambridge, U.K., 2013).

[65] In view of Córdoba's observations [54].

[66] B.W. Ninham and S. Lidin, "Some remarks on quasicrystal structure", Acta Crystallographica A **48**, 640–650 (1992) .

[67] These considerations have ancient antecedents among the Greeks and Arabs, and were certainly prominent in renaissance Italy: see, for example, Luca Pacioli, *De divina proportione* (Venice, 1509), available at https://archive.org/details/divinaproportion00paci.

[68] D.W. Thompson, *On Growth and Form*, complete revised edition (New York, Dover, 1992).

[69] A.L. Loeb and W. Varney, "Does the golden spiral exist, and if so, where is its centre?", in *Spiral Symmetry* (edited by I. Hargittai and C.A. Pickover), pp. 47–61 (World Scientific, Singapore, 1992).

[70] M.V. Berry and Z.V. Lewis, "On the Weierstrass–Mandelbrot fractal function", Proceedings of the Royal Society of London, Series A **370**, 459–484 (1980).

[71] M.V. Berry, "A theta-like sum from diffraction physics", Journal of Physics A **32**, L329–L336 (1999).

[72] L.D. Landau and E.M. Lifshitz, *Course of Theoretical Physics: Statistical Physics, Part 1*, 3rd edition, revised by E.M. Lifshitz and L.P. Pitaevski (Pergamon, Oxford, U.K., 1980).

[73] J. Honerkamp, *Statistical Physics*, 2nd edition (Springer, Berlin, 2002).

[74] B.W. Ninham, V.A. Parsegian and G.H. Weiss, "On the macroscopic theory of temperature-dependent van der Waals forces", Journal of Statistical Physics **2**, 323–328 (1970).

[75] J. Mahanty and B.W. Ninham, *Dispersion Forces* (Academic Press, London, 1976).

[76] The approach here is relevant for a semiclassical ideal gas: see Honerkamp [73], pp. 126–127.

[77] In computing thermodynamically interesting functions from ln Q, a multiplicative prefactor

needs to be applied to account for the number of available spin states per particle. We do not address this here.

[78] M. Planat, “From Planck to Ramanujan: a quantum $1/f$ noise in thermal equilibrium”, Journal de Théorie des Nombres de Bordeaux, Université Bordeaux 1, **14** (2002) 585–601 and https://hal.archives-ouvertes.fr/hal-00078140; and M. Planat, “Thermal $1/\ f$ noise from the theory of partitions: application to a quartz resonator”, Physica A **318** (2003) 371–386.

[79] L. Euler, *Introductio in Analysin Infinitorum*, Volume 1, Chapter XVI (Lausanne and Geneva, Marc-Michel Bousquet, 1748); reprinted as *Leonhardi Euleri Opera Omnia*, Series 1, **8** (Leipzig and Berlin,Teubner, 1922).

[80] F.W.J. Olver, D.W. Lozier, R.F. Boisvert and C.W. Clark (editors), *NIST Handbook of Mathematical Functions* (Cambridge University Press, Cambridge, U.K., 2010): see their Eq. (17.5.1). That the series expansion is equivalent to the product is easily deduced from the identity $(1 - a)(aq; q)_\infty = (a; q)_\infty$.

[81] L.J. Rogers, “Second memoir on the expansion of certain infinite products”, Proceedings of the London Mathematical Society (1) **25**, 318–343 (1894); L.J. Rogers, “On two theorems of combinatory analysis and some allied identities”, Proceedings of the London Mathematical Society (2) **16**, 315–336 (1917); S. Ramanujan, *Collected Papers*, pp. 214–215 and pp. 344–346 (Cambridge University Press, Cambridge, U.K., 1927; reprinted AMS Chelsea, Providence, R.I, 1962).

[82] A. Berkovitch and B.M. McCoy, “Rogers–Ramanujan identities: a century of progress from mathematics to physics”, in *Proceedings of the International Congress of Mathematicians*, Vol. 3, pp. 163–172 (Berlin, 1998); available at http://www.mathunion.org/ICM/ICM1998.3/ Main/11/McCoy.MAN.ocr.pdf.

[83] H.B.G. Casimir, “On the attraction between two perfectly conducting plates”, Proceedings of the Koninklijke Nederlandse Academie van Wetenschappen **51**, 793–795 (1948).

[84] H.B.G. Casimir, “Sur les forces van der Waals–London”, in “Colloque sur la theorie de la liaison chimique” (including a response to questions from Magar, Coulson, Prigogine, Pauling and Bauer), Journal de Chimie Physique **46**, 407–410 (1949).

[85] M.J. Sparnaay, “Measurement of the attractive forces between flat plates”, Physica **24** 751–764 (1958).

[86] S.K. Lamoreaux, “Demonstration of the Casimir force in the 0.6 to 6 μm range”, Physical Review Letters **78**, 5–8 (1997).

[87] Actually, Lamoreaux’s experimental system is not one based on parallel plates, but instead uses a flat plate and a sphere, with the force on this system being expressed in terms of the law to be verified for parallel plates using the proximity force theorem of Blocki et al. [J. Blocki, J. Randrup, W.J. Swiatecki, and C.F. Tsang, “Proximity forces”, Annals of Physics (N.Y.) **105**, 427–462 (1977)]. The proximity force theorem is essentially the famous Derjaguin approximation of colloid science [B.V. Derjaguin, “Untersuchungen über die Reibung und Adhäsion. IV. Theorie des Anhaftens kleiner Teilchen”, Kolloid-Zeitschrift **69**, 155–164 (1934)]. For subsequent independent verifications of the Casimir force, with a 1% root-mean-square average deviation between theory and experiment at the smallest measured separations, see: U. Mohideen and A. Roy, “Precision measurement of the Casimir Force from 0.1 to 0.9μmm”, Physical Review Letters **81** (1998) 4549–4552; and A. Roy, C-Y. Lin and U. Mohideen, “Improved precision measurement of the Casimir force”, Physical Review D **60** 111101 (1999).

[88] S.L. Boersma, “A maritime analogy of the Casimir effect”, American Journal of Physics **64**, 539–541 (1996); Boersma writes “The old tales were true. Rolling ships do attract each other. Two ships on a rolling sea attract each other as two atoms do in the sea of vacuum fluctuations.”

[89] A. Larraza, “A demonstration apparatus for an acoustic analog to the Casimir effect”, American Journal of Physics **67**, 1028–1030 (1999).

[90] S. Koyama and N. Kurokawa, “Casimir effects on Riemann surfaces”, Indagationes Mathematicae **13**, 63–75 (2002).

[91] S. Koyama and N. Kurokawa, “Absolute zeta functions, absolute Riemann hypothesis and absolute Casimir energies”, In G. van Dijk and M. Wakayama (editors), *Casimir Force, Casimir Operators and the Riemann Hypothesis: Mathematics for Innovation in Industry and Science* (de Gruyter, Berlin, 2010).

[92] E.M. Lifshitz and L.P. Pitaevski, *Course of Theoretical Physics: Statistical Physics, Part 2* (Pergamon, Oxford, U.K., 1980).

[93] E.M. Лифшиц (E.M. Lifshitz), “Теория молекулярных сил притяжения между конденсированными телами” (“Theory of molecular attraction between condensed bodies”), Doklady Akademii Nauk SSSR **97**, 643–646 (1954).

[94] E.M. Лифшиц (E.M. Lifshitz), “Влияние температуры на молекулярные силы притяжения между конденсиробанными телами” (“Influence of temperature on molecular attraction forces between condensed bodies”), Doklady Akademii Nauk SSSR **100**, 879–881 (1955).

[95] E.M. Lifshitz, “The theory of molecular attractive forces between solids”, Soviet Physics JETP **2**, 73–83 (1956); translation from the Russian original in Zhurnal Eksperimental’noi i Teoreticheskoi Fiziki **29**, 94–110 (1955).

[96] I.E. Dzyaloshinskii, E.M. Lifshitz and L.P. Pitaevskii, “Van der Waals forces in liquid films”, Soviet Physics JETP **10**, 161–170 (1960); translation from the Russian original in Zhurnal Eksperimental’noi i Teoreticheskoi Fiziki **37**, 229–241 (1959).

[97] I.E. Dzyaloshinskii, E.M. Lifshitz and L.P. Pitaevskii, “General theory of van der Waals forces”, Soviet Physics Uspekhi **4**, 153–176 (1961); translation from the Russian original in Uspekhi Fizicheskikh Nauk **73**, 381–422 (1961). The same paper, prepared by a different translator, also appears as “The general theory of van der Waals forces”, Advances in Physics **10** 165–209 (1961).

[98] See §81–82 of Lifshitz and Pitaevski [92].

[99] B.W. Ninham and V.A. Parsegian, “Van der Waals forces across triple-layer films”, Journal of Chemical Physics **52**, 4578–4583 (1970).

[100] B.W. Ninham and V.A. Parsegian, “Van der Waals forces: special characteristics in lipid-water systems and a general method of calculation based on the Lifshitz theory”, Biophysical Journal **10**, 646–663 (1970).

[101] V.A. Parsegian and B.W. Ninham, “Temperature-dependent van der Waals forces”, Biophysical Journal **10**, 664–674 (1970).

[102] E. Elizalde and A. Romeo, “Essentials of the Casimir effect and its computation”, *American Journal of Physics* **59**, 711–719 (1991).

[103] B.W. Ninham and J. Daicic, “Lifshitz theory of Casimir forces at finite temperature”, Physical Review A **57**, 1870–1879 (1998).

[104] It seems that all derivations of the Casimir effect require at least one formal argument to resolve an indeterminacy.

[105] J. Hadamard, “Sur la distribution des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques”, *Bulletin de la Société mathématique de France* **24**, 199–220 (1896).

[106] C. de la Vallée Poussin, “Recherches analytiques sur la théorie des nombres premiers. Première partie: La fonction $\zeta(s)$ de Riemann et les nombres premiers en général”, *Annales de la Société scientifique de Bruxelles* **20** (2), 183–256 (1896).

[107] L.S. Brown and G.J. Maclay, “Vacuum stress between conducting plates: an image solution”, Physical Review **184**, 1272–1279 (1969).

[108] F. Ravndal and D. Tollefsen, “Temperature inversion symmetry in the Casimir effect”, Physical Review D **40**, 4191–4192 (1989).

[109] J. Christensen-Dalsgaard, W. Däppen, S.V. Ajukov, E.R. Anderson, H.M. Antia, S. Basu, V.A. Baturin, G. Berthomieu, B. Chaboyer, S.M. Chitre, A.N. Cox, P. Demarque, J. Donatowicz, W.A. Dziembowski, M. Gabriel, D.O. Gough, D.B. Guenther, J.A. Guzik, J.W. Harvey, F. Hill, G. Houdek, C.A. Iglesias, A.G. Kosovichev, J.W. Leibacher, P. Morel, C.R. Proffitt, J. Provost, J. Reiter, E.J. Rhodes Jr., F.J. Rogers, I.W. Roxburgh, M.J. Thompson and R.K. Ulrich, “The current state of solar modeling”, Science **272**, 1286–1292 (1996).

[110] B.W. Ninham and M. Boström, “Screened Casimir force at finite temperatures: a possible role for nuclear interactions”, Physical Review A **67**, 030701 (2003).

[111] B.W. Ninham, M. Boström, C. Persson, I. Brevik, S.Y. Buhmann and B.E. Sernelius, “Casimir forces in a plasma: possible connections to Yukawa potentials”, European Physical Journal D **68**, 328 (2014).

[112] M.V. Berry, “The arcane in the mundane”, English translation (https://michaelberryphysics.files.wordpress.com/2013/07/berry405.pdf) of a contribution to *Les déchiffreurs: voyage en mathématiques*, edited by J.-F. Dars, A. Lesne and A. Papillault (Éditions Belin, Paris, 2008) pp. 134–135.

[113] Ø. Linnebo, “Platonism in the philosophy of mathematics”, *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/win2013/entries/platonism-mathematics/.

[114] https://michaelberryphysics.wordpress.com/quotations (retrieved January 2015).

[115] Attributed by Berry to Arnold, “implied by statements in his many letters disputing priority, usually in response to what he sees as neglect of Russian mathematicians”. Priority for this observation may, however, be due to S. Stigler, “Stigler’s law of epynomy”, Transactions of the New York Academy of Sciences **39**, 147–158 (1980).

[116] Modestly attributed by Berry to himself, though there is a certain recursively to this assertion. A more modest precursor, viz. that most discoveries of interest have significant precursors, seems worthy of the status of an axiom for human culture.

[117] Quoted by M. Dresden, *H.A. Kramers: between tradition and revolution* (Springer, New York, 1987).

[118] J. Lekner, "Electrostatics of two charged conducting spheres", Proceedings of the Royal Society of London, Series A **468**, 2829–2848 (2012).

[119] P. Ball, "Quantum leaps of faith", Chemistry World (May 2013) http://rsc.org/chemistry-world/2013/04/quantum-classical-mechanics-schrodinger-derivation; and "Quantum quest", Nature **501**, 154–156 (2013).

[120] The material in this appendix has been included as a result of a suggestion made by an anonymous reviewer.

[121] A. Wintner, "On analytic convolutions of Bernoulli distributions", American Journal of Mathematics **56**, 659–663 (1934).

[122] B. Jessen and A. Wintner, "On symmetric Bernoulli convolutions", Transactions of the American Mathematical Society **38**, 48–88 (1935).

[123] R. Kershner and A. Wintner, "On symmetric Bernoulli convolutions", American Journal of Mathematics **57**, 541–548 (1935).

[124] P. Erdös, "On a family of symmetric Bernoulli convolutions", American Journal of Mathematics **61**, 974–976 (1939).

[125] Y. Peres, W. Schlag and B. Solomyak, "Sixty years of Bernoulli convolutions", in C. Bandt, S. Graf and M. Zähle (editors), *Fractal Geometry and Stochastics. II*, pp. 39–65 (Springer, Basel, 1999).

[126] A. Klenke, *Probability Theory* (Springer-Verlag, London, 2008).

[127] Jessen and Wintner [122] prove that X is uniformly distributed on $[-1, 1]$ when $c_n = 2^{-n}$, from which it follows immediately that for $c_n = 2^{1-n}$, the random variable X is uniformly distributed on $[-2, 2]$.

[128] B. Solomyak, "On the random series $\sum \pm \lambda^i$ (an Erdös problem)", Annals of Mathematics **142**, 611–625 (1995).

[129] T.-Y. Hu, "The local dimensions of the Bernoulli convolution associated with the golden number", Transactions of the American Mathematical Society **349**, 2917–2940 (1997).

[130] K.-S. Lau and S.-M. Ngai, "L^q-spectrum of the Bernoulli convolution associated with the golden ratio", Studia Mathematica **131**, 225–251 (1998).

Substantia
An International Journal of the
History of Chemistry

Firenze University Press
www.fupress.com/substantia

Citation: J. Wolfe (2018) From idea to acoustics and back again: the creation and analysis of information in music. *Substantia* 2(1): 77-91. doi: 10.13128/substantia-42

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.

Research Article

From idea to acoustics and back again: the creation and analysis of information in music[1]

Joe Wolfe

University of New South Wales, Sydney, 2052, Australia
E-mail: J.Wolfe@unsw.edu.au

Abstract. The information in musical signals – including recordings, written music, mechanical or electronic storage files and the signal in the auditory nerve – are compared as we trace the information chain that links the minds of composer, performer and listener. The (uncompressed) information content of music increases during stages such as theme, development, orchestration and performance. The analysis of performed music by the ear and brain of a listener may reverse the process: several stages of processing simplify or analyse the content in steps that resemble, in reverse, those used to produce the music. Musical signals have a low algorithmic entropy, and are thus readily compressed. For instance, pitch implies periodicity, which implies redundancy. Physiological analyses of these signals use these and other structures to produce relatively compact codings. At another level, the algorithms whereby themes are developed, harmonised and orchestrated by composers resemble, in reverse, the means whereby complete scores may be coded more compactly and thus understood and remembered. Features used to convey information in music (transients, spectra, pitch and timing) are also used to convey information in speech, which is unsurprising, given the shared hard- and soft-ware used in production and analysis. The coding, however, is different, which may give insight into the way music is understood and appreciated.

Keywords. Information, music, composition, cognition, coding.

INTRODUCTION

Many digital recordings encode microphone signals as 16 bit numbers, which gives a dynamic range (maximum signal range/digitisation step) of 2^{16} = 96 dB. The signal is sampled at 44.1 kHz. This gives a data transmission rate of 706,000 bits per second or 706 kBaud per channel, not counting error correction bits. A traditional compact disc (CD) can store about a thousand megabytes of data: enough to store several hundred novels, or about eighty minutes of uncompressed recorded music. This raises the questions: Where do all these data come from? How much is provided by the composer, by the players and the instruments?

[1] This paper was originally presented and published as a plenary lecture at the Eighth Western Pacific Acoustics Conference, Melbourne, 2003. (C. Don, ed.) Aust. Acoust. Soc., Castlemaine, Australia.

ISSN 1827-9635 (print) | ISSN 1827-9643 (online) | DOI: 10.13128/substantia-42

What happens to that torrent of data when it reaches the listener? The rate delivered by a stereo CD – about one and a half million bits per second – appears to be equivalent to a novel every several seconds. Can our ears and brains cope with such a rate? And finally: Why do we like it? As a composer and physicist, I try here to address these questions from both sides. I suggest some answers, and indicate where research is currently looking for others.

Data compression

Data files can usually be simplified or compressed because they contain much redundancy. For instance, a CD could contain 75 minutes of 1 kHz test tone. This is redundancy on a scale of 1 ms: to a suitably sophisticated receiver, the signal could be sent as the text instruction "p = (1 mPa) sin (2pi*t/ms), 0 < t < 4500 s", which requires only 352 bits in ASCII. For an example of redundancy on a longer scale, consider "house music" in which short sound segments are sampled and repeated many times.

Kolmogorov [1] and Chaitin [2] independently introduced algorithmic entropy to quantify the difference between unpredictable and redundant signals. To paraphrase Chaitin, consider two binary numbers:

10111110010001101010110111011000001101010

and

0101010101010101010101010101010101010101.

The first "looks" random: it was obtained by tossing a coin forty times. The simplest way of transmitting that number is sending the number itself. The second does not "look" random: it can be reconstructed from the instruction "print '01' twenty times". That instruction contains more than forty bits of information, but for a very long predictable number, the reproduction instruction may be rather smaller than the number (e.g. the 208 bit instruction "print '01' a million times" produces a 2 million bit output). The algorithmic entropy is proportional to the number of bits of information in the minimum message needed to reconstruct a signal. (It is thus proportional to the log of the number of permutations and consistent with Gibbs' definition.) The more simple or predictable a signal, the lower its algorithmic entropy and the more it may be compressed. Conversely, the richer in information, the higher the entropy, and the more it resembles a random signal – at least to a receiver that cannot decode it. When sound signals are stored to be heard by humans, they are often compressed using the MPEG (mp3) algorithms. These take advantage of masking in human hearing: one frequency band may mask others, so the masked sounds are omitted. A reconstructed MPEG waveform produces an auditory illusion: its waveform has little resemblance to the original, but it sounds very similar.

Recorded music has relatively small algorithmic entropy. Indeed, its underlying order, at several different levels, is one of its attractions. At the lowest level, there is high redundancy in the waveform. A note with a definite pitch is quasi-periodic: one cycle with the pitch period is followed by many others very like it. Of course, in real, interesting instruments, the periodicity is only approximate: transients and vibrato lead to varying waveforms, as do non-harmonic components in percussion and plucked strings.

Systems of music notation take advantage of this redundancy. In standard (Western) notation, vertical positions of notes on the staff plus accidentals specify pitches and thus, approximately, frequencies. A discrete set of note symbols, plus a few other data (tempo and articulation), specify durations. Some information about the type of waveform, and much else, is contained in a word at the beginning of the music: the name of the instrument that is to play it. From this relatively small data set, performers and instruments construct complete waveforms.

The information content of written music is relatively easy to quantify because written music is digital in pitch and in time: relatively small sets of discrete pitches and durations are used. In contrast, performed music is only approximately digital: musicians make fine adjustments to the durations and timing and, except for keyboard players, adjust the pitch slightly according to context. These adjustments contribute to musical interpretation, to which topic we shall return.

Fig. 2 shows a short example: the first two phrases of the theme of the slow movement in Mozart's clarinet concerto. One way of coding it is to sample the pitch regularly in time. The lowest suitable sampling frequency is the metronome marking times the lowest common multiple of its subdivisions. Most simple themes could be adequately sampled at a rate of order 10 Hz. Five octaves (61 notes) covers the range of most orchestral instruments and can be coded with 6 bits (i.e. $61 < 2^6$), so the notes and rests could be coded at about 60 bits^{-1} (60 Baud).

Most notes are longer than the sampling time, however, so this signal can be compressed by coding for the durations of the notes as well as their pitch. Traditional notation does just this, *inter alia* (Fig. 2b). The bar lines appear to be redundant, but to musicians they also give contextual information relevant to musical expression [3]. They also provide a correction mechanism for accumulated errors in duration decoding.

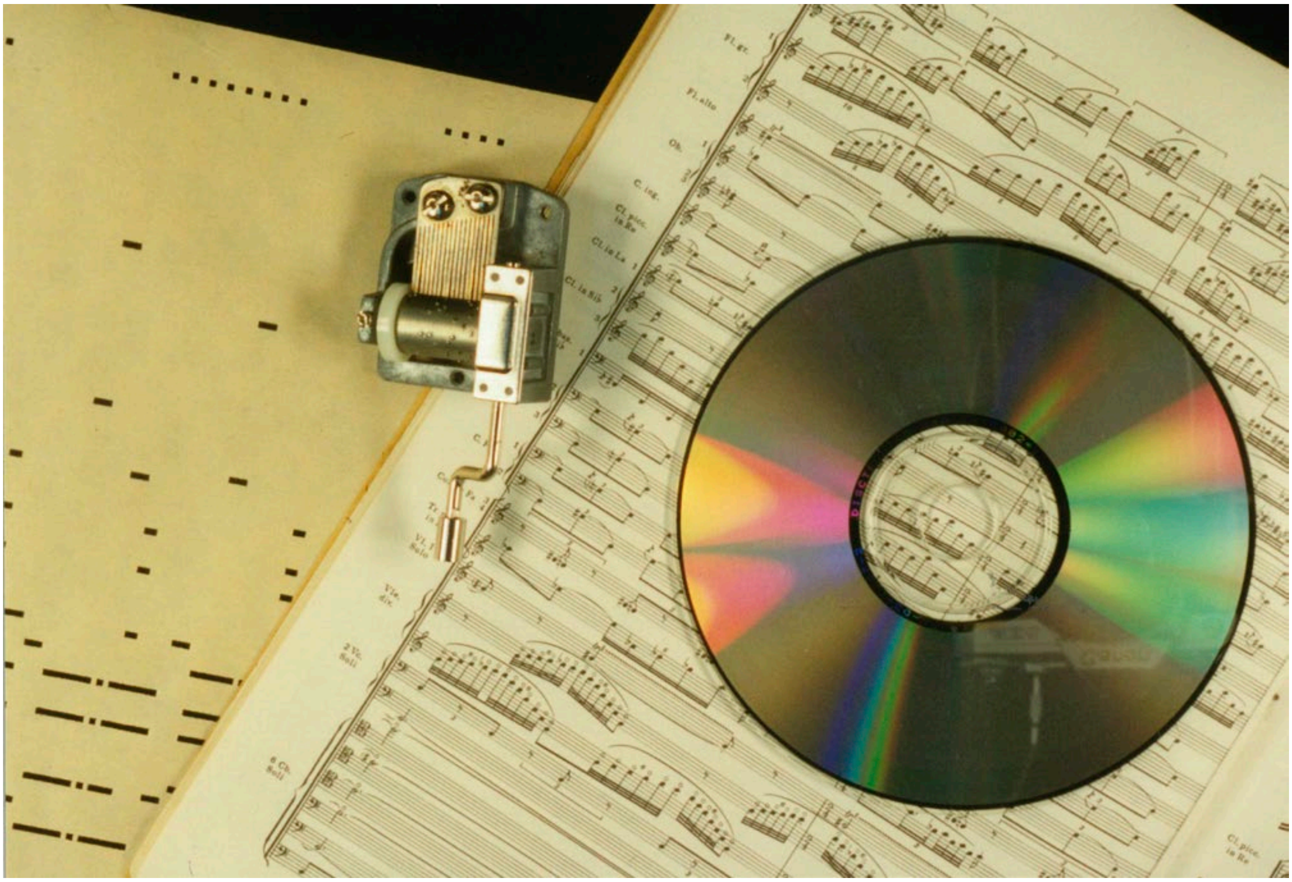


Figure 1. Four digital storage media. (a) The cylinder and comb from a music box play 16 bars from *Lara's Theme* (M. Jarre). The 18 tines of the comb have different masses and thus play different notes when struck by spikes on the cylinder. It has 18 parallel channels – circles round the cylinder. The loudness is binary (spike or no spike, note or no note). The timing is in principle analog, but is here quantised in multiples of 1/12 of a bar. The uncompressed data content of this cylinder is therefore 18 x 12 x 16 = 3456 bits. (b) The pianola roll in the background also has parallel binary channels, but the length of the hole determines the time the strings sound before the damper is replaced. In that sense, both duration and timing could be analogue, but again they are quantised in this example. The uncompressed data content is 35,000 bits per metre. (c) Standard Western music notation is (largely) parallel binary digital coding: each line and space (parallel channels) represents a pitch, though that pitch can be varied by sharps and flats. The time coding is encoded digitally in symbols (see Fig 2). This example (*The Rite of Spring*, I. Stravinsky) has about 30,000 bits on this page, which lasts a few seconds, using a coding somewhat like that in Fig 2c. (d) The CD also carries a binary digital signal ("pit" or "no-pit" in the track) but it is different in all other aspects. The signal is carried in serial rather than in parallel, and it encodes numbers that are proportional to the pressure of a sound wave. This CD records about 5 x 109 bits, not counting error correction bits. The storage efficiencies are approximately: a) 5 x 10^5 bit.kg1, b) 10^6 bit.kg^{-1}, c) 10^7 bit.kg^{-1}, d) 3 x 10^{11} bit.kg^{-1}. The apparatus required for re-creation varies greatly in size: that for (a) is shown (~ 0.01 kg), that for (c) is ~ 10^4 kg.

Figure 2c shows how a simplified binary parallel coding can represent those aspects of traditional notation used here. This example has a data content of 266 bits and, over a duration of about 13 s, a transmission rate of only 20 Baud. No correlation between the quantity of information and its value is implied, of course: many people consider this 266 bit theme more valuable than, say, a Gbyte of white noise!

The encoding used by music sequencers is close to that of music notation. These, the electronic progeny of the musical automata in Fig 1, are computer programs that output signals to synthesisers via a standard Music Industry Digital Interface (MIDI). The MIDI standard transmits data at 31.25 kBaud in serial form. This permits parallel voices and a range of instructions, and its design allowed bandwidth for further developments. Alternative coding protocols have been proposed [4]. More sophisticated representations include expression – variations in loudness, amount of vibrato, fine adjustments to pitch and to timing [5,6].

Another crude but pragmatic way of computing data content is to look at the data files of note proces-

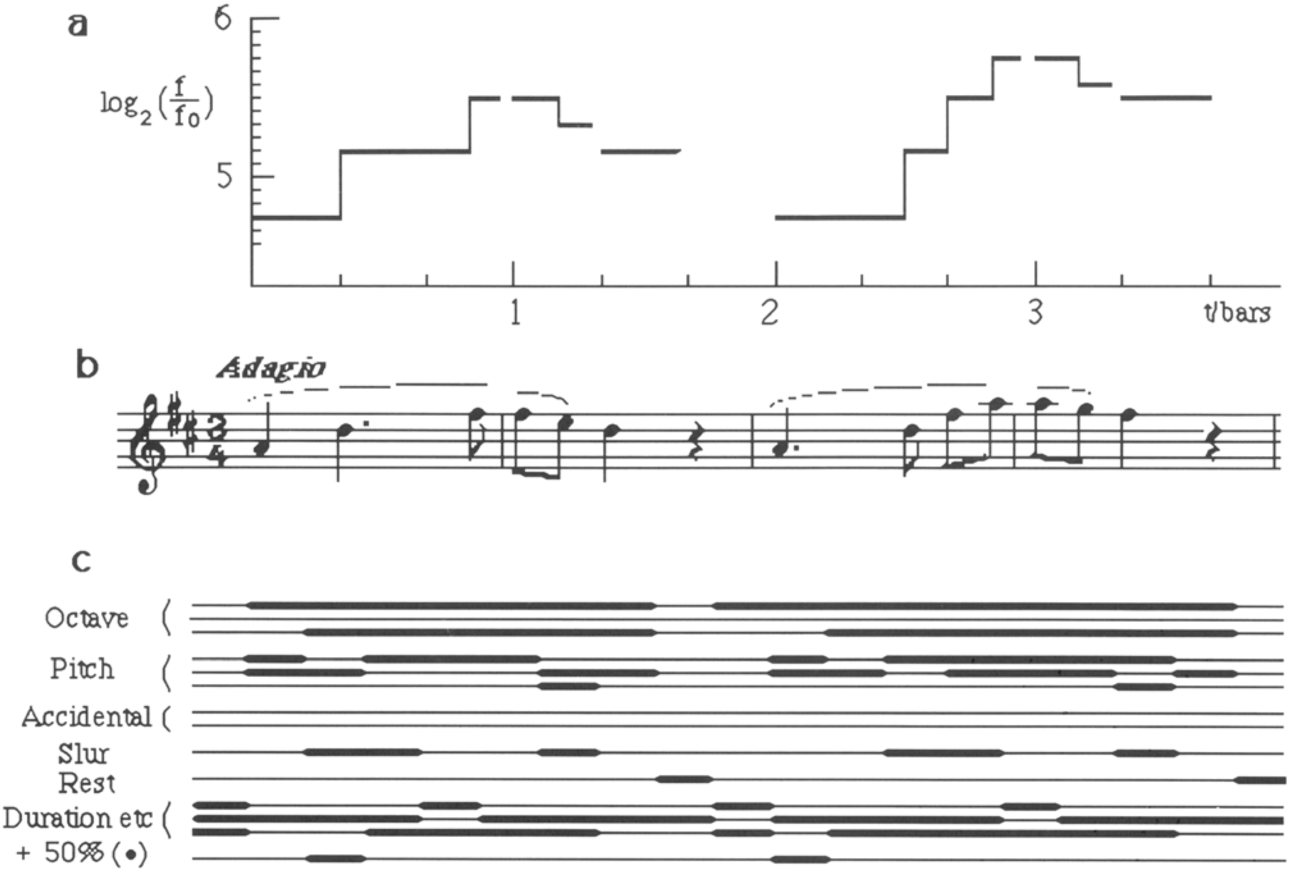


Figure 2. Three ways of coding the first four bars of the theme of the slow movement of Mozart's clarinet concerto. (a) is a semi-log plot of the pitch frequency as a function of time. On the time axis, the larger tics are bars (measures) and the smaller are beats. On the frequency axis, the larger tics are octaves. Notes an octave apart have the same letter name e.g. C5 and C6. The reference frequency is the note called C0, which is currently about 16.3 Hz. The smaller tics are one twelfth of an octave *i.e* frequency ratio of $2^{1/12} \cong 1.059$). These are called equal-tempered semitones: they correspond to the notes on an electronic keyboard. (b) is essentially traditional notation. The vertical and horizontal axes have been adjusted to make it an exactly semi-log plot by varying the spacing between lines, which may represent 3 or 4 semitones. The shapes of notes are a digitised code for duration that has several advantages over the analog time scale used in (a). (c) is a parsimonious parallel binary coding, which is more akin to traditional notation than to (a). The pitches of notes are shown by their octave (top 3 bits) and the note names (next 3 bits) with the most significant bit at the top. The next 2 bits allow for accidentals (sharps, flats and naturals) that are not needed in this example unless the key signature is omitted. The next bit indicates slurs: whether the note is continuous with the preceding one (the curved lines or slurs in (b)). The next bit indicates a rest (silence) of the appropriate length. The next 3 bits show the negative log durations with respect to a whole note. Semibreves, minims, crotchets, quavers and semiquavers (whole, half, quarter, eighth and sixteenth notes) are represented by 000 to 100. 101 is used for a bar line. The final bit allows an increase of 50% in duration (indicated by a dot in (b)). The duration code 111 is reserved as a signal to toggle the coding to text, so that occasional data such as tempo, key signature, expression marks can be added more efficiently. (The unequal spacing of channels is a guide for the eye only).

sors. These are to music what word processors are to text, and are widely used by composers and editors to write and to print music (*Sibelius* and *Finale* are commercial examples). They store written music in digital files that are similar to, but more elaborate than that in Fig 2c. On my hard disc is a 160 kbyte note processor file for a symphonic work. It takes 23 minutes to play, and so its printed score delivers data to the conductor at an average rate of 900 Baud, or 900 bits per second. To achieve the same transmission rate reading this article (not counting figures), one would need to read it at 1100 words per minute. It should be noted that conductors do not absorb all the information in a score in real time.

While comparing written music and written text, it is worthwhile contrasting them as well. One difference is cultural: more people can read text than can read music. Even to those literate in both, however, the aural re-creation is more important in music. Most musicians prefer hearing performances to reading scores, whereas I expect that most text-literate people prefer reading novels (at a rate of several hundred Baud) to hearing them read aloud, at slower rates. In both cases, the auditory transmission contains a great deal more information than does the written version.

THE ORIGIN OF INFORMATION IN MUSIC

Melodic and harmonic structures are good examples of redundancy. In a high information/ high entropy signal, all pitches would occur in approximately equal numbers and it would be impossible to predict the next note: a high information signal sounds or looks random. Music is ordered[2], and this order makes music files compressible.

The generation of information is easy to follow in (Western) concert music because it is usually written down at several different stages, which may be (i) motifs; (ii) their extension to melody, their transformation and development; (iii) the addition of other voices (usually in harmony or polyphony); and iv) orchestration or arranging. In formal music, this results in an orchestral score. In less formal music, analogous processes may lead to a score that is stored in one or more person's memory. In improvised music, the entire "score" may never be stored.

A motif is a characteristic phrase of several notes. The opening four notes of Beethoven's fifth symphony is an example, of which more anon. A motif is usually the origin of a musical composition. Several different pitches over a modest pitch range, and allowing for several different note durations, implies a possible information content of a few hundred bits.

Although the production of this information is difficult to study in detail, textbooks on composition give advice on producing motifs from simpler patterns. Schönberg [7], for example, gives numerous examples of how musically interesting phrases can be constructed from the three notes of a major chord by adding passing notes, repetitions, upbeats, appoggiaturas and alterations of notes. Many composers use comparable techniques to produce melodies.

The processes used by human composers are rarely written down, and are difficult to study explicitly [3]. It may seem prosaic to speculate that they are algorithms (as yet unknown) operating on aspects of the composer's background and stimuli, but to do otherwise seems to lead to Cartesian dualism. A range of explicit automata have been devised to create melodies. A famous example is the dice music attributed to Mozart, in which casting a die decides among several possible subunits. In electronic versions, a random number generator replaces the die. Further, while Mozart's subunits are musical phrases, some composition algorithms start with a scale of notes, some random input and a set of rules. Various automatic composers have thus been devised [8] since Harry Olsen created one in 1951 using rules generalised from the songs of Stephen Foster [9]. Michael Smetanin is an example of a contemporary composer who has used simple rules or algorithms to create musical compositions. It is difficult for an outsider to judge the success of such algorithms *per se*, however, because there is usually some discretionary intervention by a human at the input or output stage. In 'Strange Attractions', Smetanin [10] chose a particular algorithm because it gave melodies that he found attractive. An extreme example of choosing an algorithm and then letting nature take its course is 'White Knight and Beaver' by Martin Wesley-Smith [11], in which the composer assigns a note to each of the four bases of the DNA code, and then notates musically a section of the genome of the bacterium *E. coli*[3]. When other examples are given of tunes created by various algorithms, however, it is usually the case that only the 'best' results are presented – so human decision-making has intervened at the output stage.

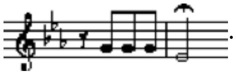
Use of a set of "rules" or fashions to generate combinations of notes and then a decision about which ones to keep is a simple model for the way some human composers work. The "rules" need not be laws (such as "the leading note always rises"[4]) decreed by some authority and observed by composers [12]. Rather they may be habits or tendencies in styles of music. For instance, virtually all composers recognise the octave as the most important and harmonious interval. Even the 'democratisation' of intervals by serialist composers leaves the octave as a very special case [13]. In this case there is a physical explanation: the harmonics of a particular note are a subset of those of the note one octave below, so adding an octave does not, or need not, add any new frequency components. In other cases, the "rules" have more complicated origins: for instance, most compos-

[2] Predictability necessarily implies redundancy. Hearing an unknown piece of tonal music from which some notes had been replaced with obvious blanks, many listeners would be able to guess the missing notes with better than chance scores, just as yo_ cou_d gues_ the _issing lette_s in this sentence.

[3] Does it sound like something that came out of a human colon, one might ask. Well, there are only four notes and they are not discordant. It sounds pleasant and musical, but this listener cannot readily extract a musical meaning.

[4] This rule shows a good example of redundancy: if the leading note were *always* followed by the note above, then an encoding could omit the pitch of the latter, just as one could omit the "u" following "q" in coding English.

ers confine themselves to scales with twelve semitones to the octave. This has a little to do with the physical basis of harmony [14], but it also has to do with what conventional instruments and players can play, what we are used to hearing, and a series of compromises among consonance and keeping the number of notes small. The "rules" for composition in most styles would be difficult to list specifically, but the musical heritage and education of the composer must incline him/her towards some patterns and combinations. Composers have a variety of processes (algorithms) for transforming an old motif into a new one, such as inverting it, changing the rhythm, reversing it, changing one or more intervals [15]. Perhaps the most important stage in producing a good motif is deciding which of many candidates is good. This process, while difficult to analyse, is at least almost universally comprehensible because many music lovers claim an ability to discern a good theme from a bad.

Thus, in one common method of composition, input data and a series of different, often unconscious algorithms generate a short phrase or idea with perhaps some tens or hundreds of bits. This may be developed into a longer melody. In written music, the data content increases in proportion with the length of the melody, but many of the extra data thus produced are redundant, in the scientific sense. The "same" motif may be repeated, transposed, inverted and otherwise transformed to create a much larger work. For one example, note the similarity in the two phrases in Fig. 2. For another, consider the famous opening phrase of Beethoven's fifth symphony: . Much is made of this simple phrase: the motif of three quavers followed by a descent of a third is used dozens of times in the beginning. Simple modifications of it occur in almost every bar of the movement: it is transposed to different positions in the scale, the final interval is changed to a second and sometimes a fourth, the last of the quavers sometimes falls, or the whole phrase is inverted in pitch. Further variants appear in the other movements – a remarkable example of much created from little.

The redundancy or structure that is created by repetition with variation is very common in melodies. In the sixteen bar 'Freude' air of Beethoven's ninth, for example, the phrase of the first four bars is repeated with slight variations in bars five to eight and thirteen to sixteen. This pattern (a,a,b,a) is extremely common, especially in songs. On a larger time-scale, redundancy through explicit repetition is so common that a variety of musical notations exist, including various repeat signs and musical 'goto' statements.

In formal music, there is often a development section in which the original idea is variously transformed: it may appear in different keys, different rhythms, inverted or melodically varied or decorated. The transformed phrase is often sufficiently different that a simple coding cannot easily reduce the length of the simplest representation. The data contained in such sections are thus created by treating the input data (the initial phrase). The existence of important structures with a variety of time scales[5] have made it difficult to formalise or to automate this operation, however. Further, selection among different algorithms and outputs is again an important process. (See the discussions in [3,16].)

Adding harmonies and counter melodies to a principal melodic line adds more data, but in some instances the extra data have relatively great redundancy. A canon is an extreme case, in which the original melody accompanies itself with a time lag, so the only extra information required is the period of the delay. In a fugue, the same or a similar melody enters with a delay, and often a symmetry operation, i.e. transposed or inverted in pitch, with doubled or halved tempo. In these cases, and in polyphony, several parallel channels of melody are of approximately equal importance. In much music however, there is one melody (or foreground) of pre-eminent importance and a harmony or accompaniment (middleground and background).

In many musical styles, the harmony is subject to rules of varying strictness, which to some extent limit the freedom of other voices and thus introduce further redundancy. Students of traditional Western harmony will agree: it often seems that the combination of strict harmony rules and voice ranges, when applied to the melody set in a harmony exercise, allow only a small number of possible 'solutions'. In many styles of music, the second most important line is the bass. If strict harmony rules are applied to a given melody and bass line, the possibilities for further parts is severely limited. Altos and tenors in choirs, or the players of second violin or viola sometimes feel that theirs are the 'left over' notes and that the result is a part that both more difficult and less satisfying than the top or bottom lines. Strict rules are extreme examples [12], but it is rare that harmony or polyphony is without rules, whether formal or informal, rigorous or fuzzy. Thus the generation of the harmony or accompaniment is often aided by the operation of algorithms on the information in the melody [18,19]. Sometimes the harmony is coded in a com-

[5] For example, the use of time-series analysis to predict the next note from the previous several notes may work well for short time scales, but is prone to wander rapidly among keys. Reviewed by Dubnov and Assayag [17].

pact but inexplicit way, such as chord symbols or figured bass. Some of its information (*e.g.* the chord symbol) is sufficiently important that the composer chooses to specify it, but the octave in which the notes occur, or their timing, is left to the performer.

Information other than notes, including articulation, ornamentation and expression marks, may be written above or below the musical staff, to convey information about pitch and duration (*e.g.* trill, staccato etc.) in ways that are more compact and legible than the explicit notation. Others carry information about loudness, articulation and tempo (*pp*, *sfz*, *accel.* etc). Others, particularly in contemporary music, contain instructions about timbre or tone colour [20]. Schönberg proposed the development of Klangfarbenmelodie (tone colour melody) in which changing patterns and structures of timbre would attain a status similar to that of changing pitch in traditional melody. Achievement of this aim might require extra data at a rate of tens or hundreds of bits per second. Some contemporary concert music contains highly specific instructions for performance, sometimes even several instructions per note. Where pitch intervals less than a semitone (microtones) are explicitly required, this is indicated by further qualification (half flat *etc.*). The requirement for slight pitch adjustments is usually implicit: many musicians do not play exactly tempered scales but, according to musical context, make fine adjustments.

One of the most important instructions about timbre is the name of the instrument that plays each part. Orchestration, the process of distributing the parts among the instruments of the orchestra, adds further information. However, there is sometimes a high redundancy when the same notes are played by different instruments.

How many data are stored in an orchestral score? Stravinsky's "The Rite of Spring" [21] provides an example of high content: it is written for a large orchestra and often the parts are relatively independent. In some sections, there are more than 40 distinct musical lines, although of course at any instant there is doubling of notes (Fig 1c). Coding just the notes of this score by sampling in time (*cf* Fig 2a) would require high transmission rates – over 100 kBaud – because of the complicated rhythms. Traditional coding (Fig 2b) is more economical, and requires only several thousand Baud[6].

So a transfer rate of up to several kBaud (equivalent to a few hundred words per second) is available to the conductor of such a work, from the score alone. Not all of this is discernible: if one player in a tutti failed to accent a note, or if the bass clarinet and second bassoon exchanged parts, this would probably pass unnoticed. When one is *not* conducting nor listening to a performance, there is no need to read a score in real time, and one may spend minutes reading carefully a single page of score, which is played in several seconds.

The performer: information input and output

Orchestral players usually read only one line, so they receive and process their written parts at rates of up to a few hundred bits per second. Other visual inputs come from the movements by other musicians, especially the conductor's baton, the leader's bow and the 'body language' of section leaders. Musicians hear the sound around them, and read the gestures and 'body language' of the conductor. This affects their processing of the written information. The interpretation of a dynamic instruction such as *forte* depends on the ensemble loudness at the time. Fine pitch adjustments depend on the prevailing pitch and harmonic context. Players also receive feedback from the interaction with the instrument of their hands, arms and mouths – but this is getting ahead of the logical order, in which the obvious next question is: how much information does the musician put out?

Some instruments have a binary digital component. In keyboard instruments, and in some percussion, the individual pitches are effectively a finite number of parallel pitch channels. In harpsichords and organs, the keys are strictly digital: a key is either depressed or not, and the player's control of the loudness of that note is binary. Bach reportedly said, disingenuously, of his organ playing: "There is nothing remarkable about it. All you have to do is hit the right notes at the right time, and the instrument plays itself" [22]. Bach, who played the viola too, would of course have known that playing a single, beautiful note on such an instrument requires much more than simply starting and stopping at the right time. The exact timing of the depressing and release of keys are analogue parameters of great importance in musical expression. In the piano, another analogue parameter is the momentum with which the hammer strikes the string. In percussion instruments, there are the complications of the position, speed and angle of the strike. Most woodwind and brass instruments have keys and valves used almost always in a binary way: either depressed or not. This does not however restrict the pitch to discrete values because pitch is also controlled by the player's lips and air pressure. In orches-

[6] The example cited is from rehearsal mark 11 in [21]. Demisemiquavers with triplets, quintuplets and septuplets at crotchet = 66 require sampling at 924 Hz. With 6 bits for pitch, the 31 parts require 172 kBaud. Using a code like Fig 2b, but with several more bits of articulation and expression marking, 200-300 notes per bar require several kBaud.

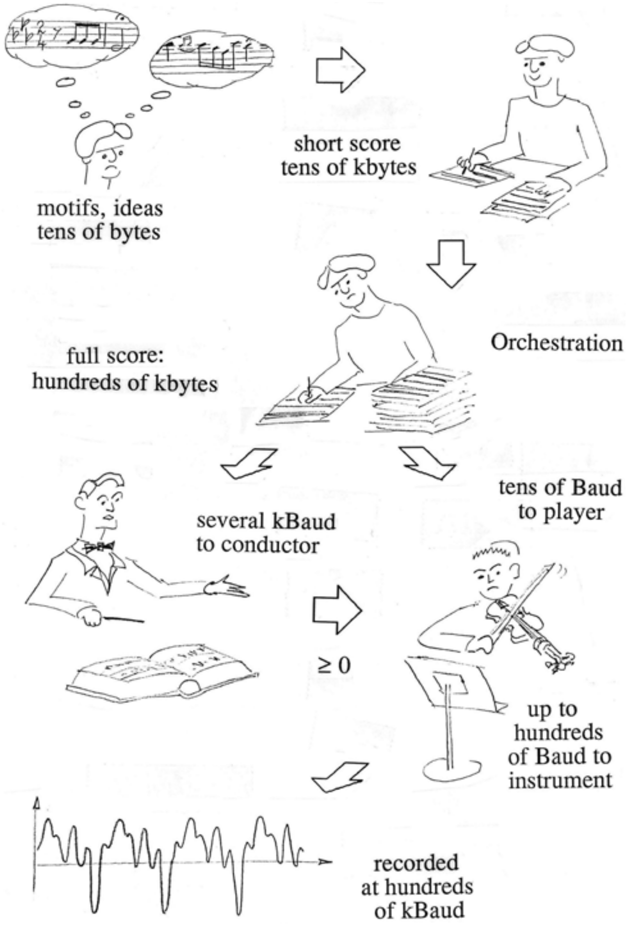


Figure 3. One information chain, from composer's original ideas to performed music. The approximate data content is given in bits and kilobytes (1 kbyte ≅ 8000 bits) and the rate of data transfer is given in Baud (1 Baud = 1 bit per second) and kiloBaud.

tral string instruments, the pitch is controlled by a continuous parameter (position of the finger stopping the string) plus choice of string.

Phrasing and expression are largely supplied by performers. Consciously or unconsciously, musicians decide how to 'shape' the phrase. This includes varying the loudness and amount of vibrato of individual notes, and making slight adjustments to indicated durations. A note judged to be important might be given emphasis by increasing the loudness and vibrato, and by increasing its duration slightly beyond the indicated value. This is one notable – and valuable! – difference between a performance by a musician and one by a primitive music sequencer. To some extent these elements of interpretation are similar among musicians [23] and so they may, to that extent, be codified. Acousticians Friberg, Sundberg and colleagues, in consultation with prominent musicians, have induced and formalised performance rules that add such elements of interpretation to a sequence representing written music [5,24,25]. Their software produces a 'performance' that is much more idiomatic and "musical" than that produced by an ordinary sequencer. These ideas have influenced modern commercial music sequencers[7].

The instrument: input and output

Written music is an incomplete instruction set. To oversimplify, the individual musician reads at typically 100 Baud or less, and outputs time-varying control signals, which may have several times this rate. The instrument outputs an analogue signal. For most monophonic instruments the output spectrum is dominated by approximately harmonic components whose fundamental frequency (which equals the harmonic spacing) determines the pitch. The pitch varies in time (with vibrato and with successive notes) and the amplitudes of the spectral components vary in time. The information required to encode this output depends on the fidelity and dynamic range required. It is at this stage that there is a great increase in the data required for encoding. If the performance is recorded uncompressed on a CD, then it results in the same enormous data transfer rate whether it be the intricate orchestration of 'The Rite of Spring' or one of the much simpler examples given above.

On many instruments, players control several interdependent analogue parameters connected with phrasing, such as vibrato, loudness, and variations in timing and intonation. Performers may also control several parameters that contribute to the timbre. In string instruments these include bow position, speed and force. In wind instruments, they include blowing pressure, several aspects of embouchure (e.g. lip tension, jaw position, position of lips on reed) and the shape of the vocal tract. These parameters may be adjusted several times per second, and each may have several bits of precision. Together they may contribute up to a few hundred Baud.

The instrument, then, is where the data rate increases dramatically. But surely the instrument is not creating information? Rather, we could say that the instrument increases the redundancy – creates redundant data – by a large factor: one period of the note is very similar to the preceding one. This oversimplifies a little: two similar hypothetically identical performances by a player – or even by a music sequencer and synthesiser – will not

[7] These typically have a range of settings for 'expressive' performance, from *meccanico* to *molto espressivo* and *molto rubato*, with varying interpretations including *straight*, *swing*, *Viennese waltz* and *funk.*

produce the same waveform, but the differences are not information to be transmitted from composer and player to listener.

TRANSMISSION AND RADIATION

In performance, instruments radiate sound into the air. These signals, plus background noise, are convoluted by the delays and multiple reflections of the performance venue. This extra information is recognised by listeners who can discern some details about the venue from listening to a recording – the difference between a cathedral and open air is an extreme example. This information contributes feedback to the conductor and players, who in general adapt their performance to the acoustic environment. For instance, they might play more quietly in a room with a low background noise and more slowly and more *marcato* in a room with a long reverberation time.

A performance creates a sound pressure field: the sound pressure p varies with position vector ($\mathbf{r}$) and time (t). It would take a prodigious number of data to record such a field with a resolution in space and time corresponding to the half-wavelength and half-period of the highest audible frequencies (say 30 μs and 1 cm). Of course, the whole field is not sampled by a single listener, who receives just the sound pressure at each ear ($p(\mathbf{r}_1,t)$ and $p(\mathbf{r}_2,t)$), although the positions of the ears may vary in time as the listener moves his/her head. So each ear receives an analogue signal which, if the level of background noise is sufficiently low, may have the same dynamic and frequency range as the sum of the signals from the instruments.

Our imaginary composer, orchestrator, musicians, conductor and performance venue have now delivered to the ear the great data rate mentioned in the introduction. Because of the high signal redundancy, the information rate or algorithmic entropy rate is considerably lower, but still perhaps impressive. The information has been generated by mental processes of the composer and performers, which we may consider as algorithms – subtle and in many cases not understood – processing inputs from memory, education and culture. The instrument has turned this information into the radiated signal, which has been filtered and convolved by the acoustic environment. It's now time to follow the signal into the listener's head.

THE ANALYSIS OF INFORMATION

The outer and middle ear are, for our purposes, primarily acoustic and mechanical impedance transformers that overcome the mismatch between the air of the radiation field and the cochlear fluid in the inner ear. (They are also filters, transmitting some frequencies more effectively than others.) The qualitative change occurs in the cochlea of the inner ear in which the input signal – single channel analog – is actively filtered, compressed and converted to parallel digital electrical signals in the auditory nerve.

Because of the position-dependent mechanical properties of the basilar membrane, pitch is in part coded by channel: only low frequency waves reach the apical end of the membrane, so nerve fibres from this region carry information about low frequencies. It is also partly coded in rate of firing, at low frequencies at least, because the hair cells are stimulated at the frequency of the motion[8]. Signal amplitude is also partly coded by channel (some fibres only respond to large signals) and partly by (analog) signal firing rate: overall, larger stimuli produce higher firing rates. The minimum firing rate is not however zero: most neurones have a 'background firing rate' – a rate at which they fire in the absence of any signal. This makes a neuron capable of carrying a "negative" signal: if the cell is inhibited by a neighbour, its firing rate falls below the background rate. Lateral inhibition among neighbouring cells is useful in amplifying small simultaneous differences. Nerves also become less sensitive with continued stimulation, so a changing signal usually has a greater effect than a steady one. For more detail, the reader is referred to reviews of perception and neurobiology [27,28,29,30,31,32].

Coding in the auditory nerve

The pulses in the nerve fibres, called action potentials[9], are binary – either the stimulus is strong enough produce an action potential, which travels along the nerve fibre, or else nothing happens. As in electronics, the advantage of digital signals is their immunity to noise and distortion. Nerve fibres are very lossy coaxial cables, so an unamplified signal is substantially lost after transmission of a few millimetres. Many stages of amplification and pulse shaping are conducted by the nerve membrane where it is exposed at the nodes of Ranvier.

What is the data transfer rate at this stage? There are about 30,000 nerve fibres or channels, each capable of

[8] Experiments with implanted electrodes show that, at low stimulation rates, perceived pitch depends approximately logarithmically on the stimulation rate but also linearly on the electrode position [26].

[9] The voltage inside biological cells is usually tens of mV negative. When nerve cells are stimulated (by briefly making their membrane "insulation" leaky), the internal voltage rises ~100 mV before returning to the resting value.

transmitting a few hundred action potentials per second. If the coding were strictly digital, the data transfer rate would surpass that of a CD. The practical rate is much less, because of redundancy: in part because nearby fibres carry highly correlated signals. What happens to this signal in the brain is difficult to follow directly. The experimental observations of psychophysics include integration, sampling and signal treatment at higher levels.

Effects including the active filtering in the basilar membrane give rise to the masking of weak signals by strong signals in nearby frequency bands. There are only roughly 30 critical bands so, instead of 30,000 parallel frequency channels, perception effectively involves only of the order of 30. For an unmasked tone, the just noticeable difference (JND) in sound level is roughly 1 dB. Over a short term dynamic range of 60 dB, this gives about 60 perceptible loudness levels (requiring 6 bits). The JND for frequency may be as small as tenths of a percent for sustained signals, but in our calculation the maximum frequency resolution is limited over most of the range by the temporal sampling rate. The greatest perceptual resolution in time is a few tens of milliseconds. At this rate, the number of different frequency percepts is about 1000 (10 bits). So there are about 16 bits, sampled at up to 30 times per second, in 30 channels. The product gives data transmission rate of 16 kBaud: a considerable overestimate because the JNDs increase towards the ends of the frequency range and as sampling rate and number of simultaneous stimuli increases[10]. Whatever the actual maximum rate, to achieve it would require a signal that, at the perceptual level, had no redundancy or order: a signal that sounded random. Not music.

Processing – sorting into notes

It is easier to perceive notes (which usually include several or many separate frequency components) than to perceive the individual frequency components of its spectrum. With practice and careful listening, one can distinguish some spectral components in notes in some circumstances[11]. That naïve listeners rarely do so suggests that we have either a very well-learned or an inbuilt mechanism for combining the various frequency components of a note together and perceiving it as a whole. This capacity is partly explained in terms of two general properties attributed to the nervous system: that change is more noticeable than lack of change, and that things that change in the same way are often grouped together. Consider a note comprising several harmonics: if the pitch of the note changes (either melodically or due to vibrato), then the pitches of all its components change in exact proportion; if the loudness changes, then the loudness of the harmonics also changes. Evidently we possess signal processors that group these separate, but similarly changing elements together and identify them as a single note. Instrumental and operatic soloists make use of vibrato to make their notes identifiable against the sound of the orchestra[12].

The system works especially well for notes whose spectral components are approximately harmonic, which we identify as having a definite pitch. This capacity may have been important in the evolution of human audition. Many human vocal sounds (the vowels in speech, but also inarticulate cries and screams, whether sung or spoken) have at any instant a definite pitch and spectral components which fall in the harmonic series. It is likely that we have evolved hard- and soft-ware capable of identifying vocalised sounds among other sounds that do not have harmonic structure, such as wind noise. The system works so well that we hear missing fundamentals and Tartini tones.

Analysis in time

The shortest time scale of interest in music is the period of the vibration. This ranges from about 50 μs to 50 ms. For low pitches, the auditory nerve carries some information about pressure variation on this time scale, but while we are aware of pitch, we are rarely aware of the variation in pressure that gives rise to that pitch[13].

The next time scale is that of transients. When an instrument begins to play a note, there is a short time (tens of milliseconds) over which the amplitudes of the various components vary considerably before ‘settling down’ to establish a relatively unvarying spectrum. These transients are so important to the timbre of a note that different wind instruments are readily confused if the initial and final transients are removed [34]. Transients in musical notes are analogous to plosive conso-

[10] There are further complications such as feedback loops and other control signals which come “downwards” from the brain to the ear, and these affect the “upwards” signals to the brain [33].

[11] Or, conversely, a small number of harmonics may be made sufficiently louder than the rest that they can be identified as separate notes, as in harmonic singing.

[12] This effect is especially useful if some of the harmonics of the soloist occur in a frequency range where the accompanying sounds have relatively low level – if we can hear one component clearly, it seems that we can track other components which have the same vibrato and phrasing.

[13] A contrabassoon can play Bb$_0$ at 29 Hz. When this note is played loudly, we can just detect a periodic variation as the reed opens and closes 29 times per second. Most of the sound we hear, however, is in the higher harmonics rather than the fundamental.

nants (d, t, g, k, b, p) in speech or singing. In both cases we are capable of concentrating and hearing them with some clarity, but under most circumstances these details are analysed subconsciously.

The third time scale (several tens of milliseconds and longer) is that of notes [35,36]. It is at this level that we sense pitch and timing: the basic elements of melody. With little concentration, we can readily be conscious of the rhythm and the pitch, and also of the timbre of the instrument playing it. It is, however, difficult to introspect much beyond this: although our ears and their associated low-level processing have coded the various component frequencies and how they vary on the scale of tens of milliseconds, we are usually aware of the signal at a higher level: that of pitches, rhythms and timbres.

A changing signal is less redundant than a constant one, and our senses reflect this. After a while we no longer notice the sound of the wind, the weight of our clothes, the strange colour of artificial lighting; but we do notice sudden changes in them – changes over time. Similarly, we notice sharp boundaries in a visual image rather than a gradual change between two colours or shades – changes in space or channel. Changes in time are enhanced by the property of nerves to fire more rapidly when first excited than they do during a steady stimulus. Differences in space or channel number are enhanced by neural circuits that effectively subtract the signals from adjacent nerves using lateral inhibition [37].

Pitch sensitivity provides a good example. A single note without vibrato is a steady signal, which is probably carried at all times by the same nerve fibres. A note with vibrato is a varying signal, which is probably carried at different times by different nerve fibres. Vibrato makes notes more noticeable, and also makes it easier to identify a single instrument in an ensemble. Timing sensitivity provides another example. We are usually less conscious of the duration and end of the note than the beginning: a variation in the timing of the end of each note is noticed as a change in articulation – some notes more staccato than others; a variation in the timing of the beginning is noticed as a variation in the rhythm, and is more noticeable.

Symmetries: the ear and the instrument

In this sense, our ears and their associated low-level processing perform a role that is the reverse of that of the instrument: the player controls the note's pitch, duration and often the timbre; the instrument converts the player's partly digital, partly analogue parallel signal into a complicated vibration, or equivalently a set of simple (usually harmonic) vibrations in a mechanical oscillator (string or air column). These vibrations, often via an impedance transformer (bridge and body of string instruments, bells of brass instruments) cause a pressure wave that is a single analogue signal: p(t).

The ear receives a wave p'(t) and, via impedance transformers (the outer and middle ear) this causes a complicated vibration, or equivalently a set of simple, often harmonic, vibrations in a mechanical oscillator (the basilar membrane). These vibrations are sensed and processed, and we perceive the note's timing, pitch, duration and timbre.

The perception of notes is subject to categorisation (*i.e.* digitisation): when fine differences in pitch are presented, listeners, especially those with musical training, tend to sort them into the discrete notes in a scale [38]. Thus the perception of pitch is partly digital and partly analogue – we perceive a note, but may remark that it was a little sharp or flat.

More symmetries: the listener and the composer

On time-scales larger than those discussed above, listeners are capable of perceiving structures and features in music: we may identify (whether consciously or otherwise) themes, harmonies, orchestration etc. This article gives no more than some pointers to research in this area. Sloboda [3] compares the analysis of linguistic structure by Chomsky with the analysis of musical structure by Schenker, which uses hierarchies of note groupings and their functions. Some seem general, while others are specific to certain cultures. One way of studying this level of structure is by proposing plausible models and comparing their performance with that of human subjects [39-41].

These processes complete the communication symmetry. To the extent that the listener hears melodic patterns, repeats and transformations of thematic material, s/he reverses the process of composition and may leave the concert hall humming the themes or ideas that began the whole process.

The information transmitted between the minds of composer and listener may differ in detail, but the coding is physiologically similar in the two minds, in that it involves many parallel digital signals in neurones. Between the two, however, the information passes through a coding totally foreign to the operation of the brain – a data-rich, serial, analogue signal. The interpreters for this foreign signal are the musical instrument in one direction and the ear in the other, whose symmetry is discussed above. The performing musicians direct and supervise translation at one end. The listener has an interpretive role that may be the reverse of those of player and composer, depending on training and attitude. A discussion of this is beyond our current aim.

MUSICAL COMMUNICATION

To a communications engineer, music might seem inefficient and unreliable. Different listeners may extract different messages from the same signal. Listeners may differ with the composer over the question "what is it about?" This does not mean, of course, that it is without meaning or value: the signal is rich in information often input by different people (composer, performers, conductor) so it is not surprising that different people extract different subsets of that information, or interpret it differently. To quote Aaron Copland: "'Is there a meaning to music?' My answer to that would be, 'Yes'. And 'Can you state in so many words what the meaning is?' My answer to that would be, 'No'." [42]. Researchers are however quantifying aspects of the meaning. Schubert [43], for instance, measures emotional responses to music in a two-parameter space and finds reasonably consistent responses, with a resolution of a few bits in each direction and a time resolution of seconds. This gives a Baud rate not far below that of text being read.

In the context of musical enjoyment, the processes of encoding and decoding may be at least as important as any part of the communication. But why do we so enjoy this encoding and decoding? Why have we evolved the capacity for this sophisticated, complicated but imprecise method of communication of abstract ideas? Does musical ability confer survival advantages on individuals possessing it? Why can such abstract communication have powerful emotional effects? These questions are invitations for speculation, but it is interesting to look at them with regard to information coding.

Music and speech: similarities and complementarities

The physiological hardware used for listening to music and speech is the same, and some of the software may be shared too. Most speech sounds involve vibration of the vocal folds. The time scale of these vibrations is shorter than that of nerve or muscle response, so any given vibration is very similar to its predecessor, so the sound produced is usually quasi-periodic. These periodic speech sounds (as well as screams, cries, and moans) have harmonic spectra. The ability to discern a set of harmonic frequency components as an entity, and to track simultaneous changes in that set, is an ability to discern one voice or cry from background sound. It is also much of the ability to follow a melody.

On the other hand, the signal codings of speech and music are different. Oversimplifying for the sake of the argument, we could say that they are almost complementary, especially with regard to digitisation. Speech coding is digital in that it uses a discrete set of speech sounds (phonemes). In alphabetic languages, (a subset of) these are all that is recorded, as letters, in the text or transcribed form. Further, they are digitised in perception (*i.e.* they are perceived categorically [44]). Phonemes are encoded by features of the sound spectrum (formants and formant trajectories) and by transients. But in music, transients (especially the way notes start) and features of the spectrum are together what we call timbre. Most of the 'text' of music is notes: digital representation of pitch and timing. These are also perceived digitally (categorically) in music [38]. In speech, however, these features are prosody and (except in tonal languages such as Mandarin and Thai) they are analog variables, which are not notated. So the texts of music and speech use the acoustical features and digitisation in almost complementary ways, as the table shows. I discuss this in greater detail elsewhere [45].

Why music?

The capacity to communicate using sound, whether by speech or more primitive articulations, may have been sufficiently important to select for a suitable capacity for sound analysis. This explains (at the evolutionary level) why we have the mechanisms that we use for analysing music. But why do we so use those mechanisms? Why do parents sing to infants? Why do we like and make music? Perhaps signal processing can provide part of the answer.

Those who write or use automatic speech recognition software know that it is non-trivial to extract the spectral features, envelope and pitch that carry information in both speech and music, especially in the presence of background noise. In some cases, however, it may be easier in music. Consider an unaccompanied melody, sung or played by a single instrument, which might be an example of music from our early pre-history. This signal has frequencies that are usually stable during a note, compared with the rapid, continuous (*i.e.* analogue) pitch changes in speech. Rhythms in music are also more regular in music than in speech. In instrumental music and in *vocalise* (singing without words), the spectral features change less, and in a more regular way, than they do speech. When we sing to babies [46], is it possible that we are using the reduction-

Table 1. Acoustical features of music and speech signals show complementary coding. (Reproduced from [45]).

Acoustical feature	**Music without words**	**Speech**
Fundamental frequency (when quasi periodic)	*pitch component of melody* categorised notated precision possible	*pitch component of prosody* not categorised not notated variability common
Temporal regularities and quantisation on a longer time scale	*rhythmic component of melody* categorised notated precision possible	*rhythmic component of prosody* not categorised not notated variability common
Short silences	*articulation* sometimes notated	*parts of plosive phonemes* implicitly notated
Steady formants	*components of instrumental timbre* not notated not categorised *per se*	*components of sustained phonemes* notated categorised
Varying formants	*not widely used* —	*components of plosive phonemes* categorised notated
Transient spectral details	*components of timbre* not categorised sometimes notated	*components of consonants* categorised notated

ist method to teach them how to listen, developing the skills necessary to understand speech?

Could music be a game for the ear? Games are often described as models of social behaviour, that develop useful mental and physical skills. Games develop reflexes, co-ordination and muscular strength that may confer evolutionary advantages. Intellectual and socialising games develop skills that could also confer survival or mating advantages. If speech and signal processing skills enhanced our ancestors' chances of survival or mating, the game of music may have been selected, whether it were transferred between generations by genetics or culture.

The basic skills of sound analysis are subtle and beyond introspection, but that is true of many games: we are no more conscious of how we analyse sounds than we are of the muscular control we used to catch a ball. What we do with these skills is sometimes elaborate, but that is also true of games such as cricket and chess. In games and in music, our enjoyment of neurological exercise and challenges seems to require successively more complicated games as our capacities develop.

Speech carries the meaning of the words spoken, but it also carries information in the way in which the words are spoken. The rhythms and tempi, subtle pauses and variations in articulation and loudness, the overall register and the changing pitch – all carry information. Information of this latter type gives subtle shades to the meaning conveyed by the words, and it often tells of the speaker's emotional state. The ability to convey this information distinguishes a good actor from someone who just reads the words. Music also carries expressive information in subtle variations in rhythm and phrasing [24,47], coded in a comparable way [48].

However, an important vehicle for affective information in speech is prosody. These features, completely omitted in the text of speech, are the dominant features of music, whereas the features used to encode the explicit information in speech are used, in music, for timbre and are often varied little. I end by inviting the reader to wonder, as I do, whether this may one of the reasons for the attraction and emotional power of music, this peculiarly coded, abstract method of communication.

ACKNOWLEDGMENT

I thank Sten Ternström and Emery Schubert for helpful comments.

REFERENCES

1. Kolmogorov, A.N. "Three approaches to the definition of the concept "amount of information"." (1965) Problemy Peredachi Informatsii, **1**, 3-11 (Russian, cited by Chaitin, *ibid.*)

2. Chaitin, G.J. “Information, Randomness and Incompleteness”. (World Science, Singapore, 1987).
3. Sloboda, J. “The Musical Mind” Oxford: Clarendon Press, (1985).
4. Garnett, G. “Music, Signals, and Representations: a Survey” in “Representations of Musical Signals”, de Poli, Piccialii and Roads, eds., (MIT Press, Cambridge Mass, 1991)
5. Sundberg, J. “Musical performance: a synthesis-by-rule approach”. Computer Music J. **7**, 37-43 (1987).
6. Friberg, A. “Generative rules for music performance: a formal description of a rule system”. Computer Music J., **15**, 49-55 (1991).
7. Schönberg, A. “Fundamentals of Musical Composition” (Faber, London, 1967).
8. Schwanauer, S.M. and Levitt, D.A. eds. “Machine Models of Music”, (MIT Press, Cambridge MA, 1993).
9. Cope, D. “Experiments in musical intelligence (EMI): non-linear linguistic-based composition”, Interface, **18**, 117-139 (1989).
10. Smetanin, M. “Strange Attractions”, (Sounds Australian, Sydney, 1990).
11. Wesley-Smith, M. “White Knight and Beaver”. (Sounds Australian, Sydney, 1984).
12. Masson, C. “Nouveau Traité des Règles pour la Composition de la Musique” (1705). (Facsimile edition, Minkoff, Geneva, 1971).
13. Leibowitz, R. “Introduction à la musique de douze sons”. (L’Arche, Paris, 1949)
14. Helmholtz, H.L.F. “On the Sensations of Tone as a Physiological Basis for the Theory of Music”, (1877) English translation by A.J. Ellis, (Dover, N.Y. 1954).
15. Stravinsky, I “The Rite of Spring: sketches 1911-1913. Facsimile reproductions with commentary by R. Craft”. (Boosey & Hawkes, London, 1969).
16. Dirst, M. and Weigend, A.S. “Baroque forecasting: on completing J.S. Bach’s last fugue”, in “Time Series Prediction: Forecasting the Future and Understanding the Past” A.S. Weigend and N.A. Gershenfeld, eds. (Addison-Wesley, Reading MA 1993).
17. Dubnov, S. and Assayag, G. “Universal pediction applied to stylistic music generation”, in “Mathematics and Music” G.Assayag, H.G.Feichtinger and J.F.Rodrigues, eds. (Springer, Berlin, 2002).
18. Maxwell, H.J. “An expert system for harmonizing analysis of tonal music” in “Understanding Music with AI: Perspectives on Music Cognition” M. Balaban, K. Ebcioglu and O.Laske, eds.pp 335-353. (MIT Press, Cambridge, MA, 1992).
19. Hild, H., Feulner, J. and Menzel, W. “HARMONET: a neural net for harmonizing chorals in the style of J.S. Bach” in “Advances in Neural Information Proceessing Systems, J.E. Moody, S.J. Hanso and R.P. Lippman, eds, 4:267-274. Morgan Kauffman, San Mateao, CA (1992).
20. Stone, K. “Music Notation in the Twentieth Century”. (Norton, New York, 1980).
21. Stravinsky, I. “The Rite of Spring” (1921). The example cited is from rehearsal mark 11. Boosey & Hawkes, London (1967).
22. Köhler, J.F. “Historia Scholarum Lipsiensium” (1776), quoted by David, H.T. and Mendel, A. “The Bach Reader”, (Norton, NY, 1972)
23. Repp, B.H. “A constraint on the expressive timing of a melodic gesture: evidence from performance and aesthetic judgment”, Music Perception, **10**, 22-242 (1992).
24. Sundberg, J. Fribert, A. and Fryden, L. “Threshold and preference quantities of rules for music performance”, Music Perception, **9**, 71-92 (1991).
25. Juslin, P.N; Friberg, A., Bresin, R. “Toward a computational model of expression in music performance: The GERM model.” Musicae Scientiae. Spec Issue, 2001-2002, 63-122 (2002).
26. Fearn, R., Carter, P. and Wolfe, J. “The perception of pitch by users of cochlear implants: possible significance for rate and place theories of pitch” Acoustics Australia, 27, 41-43 (1999).
27. Barlow, H.B. in “Physics and mathematics of the nervous system” (Conrad, M, Güttinger, W. and Dal Cin, M., eds) (Springer-Verlag, Berlin, 1974).
28. Møller, A.R. “Auditory Physiology”. (Academic, NY, 1983).
29. Fletcher, N.H. “The physical bases of perception”, Interdisciplinary Sci. Rev., **9**, 6-13 (1984)
30. Kandel, E.R. and Schwartz, J.H. Principles of Neural Science, (Elsevier, 1985).
31. Altschuler, R.A., Bobbin, R.P., Clopton, B.M. and Hoffman, D.W. “Neurobiology of Hearing: the Central Auditory System”. (Raven, NY, 1991).
32. Yates, G.K. “The Ear as an Acoustical Transducer”, Acoustics Australia, **21**, 77-81 (1993).
33. Spangler, K.M. and Warr, W.B. “The descending auditory system” in “Neurobiology of Hearing” R.A. Altschuler et al, eds, pp 27-45, (Raven, NY, 1991).
34. Berger, K.W. Some factors in the recognition of timbre, J. Acoust. Soc. Am. **36**, 1888 (1963).
35. Warren, R.M., Gardner, D.A., Brubaker, B.S. and Bashford, J.A. “Melodic and nonmelodic sequences of tones: effects of duration on perception”, Music Perception, **8**, 277-290 (1991).
36. Warren, R.M. “La perception des séquences acoustiques: intégration globale ou résolution tempo-

relle?" *in* "Penser les Sons. Psychologie Cognitive de l'Audition" McAdams, S. and Bigand E., eds., Presses Universitaires de France (1994).

37. Shepard, G.M. "Neurobiology", (Oxford Uni. Press, 1988).
38. Locke, S. and Kellar, L. "Categorical perception in a non-linguistic mode" Cortex, **9**, 355-369 (1973).
39. Lischka, C. "Understanding Music Cognition: A Connectionist View" in "Representations of Musical Signals", de Poli, Piccialii and Roads, eds., (MIT, Cambridge Mass, 1991).
40. Longuet-Higgins, H.C. "Artificial intelligence and musical cognition", Phil. Trans. R. Soc. Lond. A **349**, 103-113 (1994).
41. Longuet-Higgins, H.C. and Lisle, E.R. "Modelling musical cognition", Contemporary Music Review, **3**, 15-27 (1989).
42. Copland, A., "What to listen for in music". (New American Library, NY, 1967).
43. Schubert, E. "Continuous measurement of self-report emotional response to music" in "Music and emotion: Theory and research. Series in affective science." Juslin, P.N. (Ed); Sloboda, J.A. (Ed), eds. pp 393-414. (Oxford University Press, London, 2000).
44. Clark, J. and Yallop, C. "An Introduction to Phonetics and Phonology" (Blackwell, Oxford, 1990).
45. Wolfe, J. "Speech and music, acoustics and coding, and what music might be 'for'". International Conference on Music Perception and Cognition, Sydney, 2002, K Stevens, D. Burnham, G. McPherson, E. Schubert, J. Renwick, eds. pp 10-13 (2002). www.phys.unsw.edu.au/~jw/ICMPC.pdf
46. Gérard, C and Auxiette, C. "The processing of musical prosody by musical and nonmusical children" Music perception, **10**, 93-126 (1992).
47. Mersenne, M. "Harmonie Universelle, contenant la Théorie et la Pratique de la Musique" (1636). (Facsimile edition, CNRS, Paris, 1975).
48. Banse, R. and Scherer, K.R. "Acoustic profiles in vocal emotion and expression" J. Personality and Social Psychology, **70**, 614-636 (1996).

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

Historical Article

Snapshots of chemical practices in Ancient Egypt

Jehane Ragai

Emeritus Professor of Chemistry, The American University in Cairo, 37, Sedley Taylor road, CB28PN, Cambridge, UK
E-mail: jragain@aucegypt.edu

OPEN ACCESS

Citation: J. Ragai (2018) Snapshots of chemical practices in Ancient Egypt. *Substantia* 2(1): 93-101. doi: 10.13128/substantia-43

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. This article gives a historical overview of a number of chemical practices carried out by the Ancient Egyptians and shows that beyond being purely empirical, in more than one instance their methods suggest an understanding of the rudiments of modern day chemistry. A close analysis of some of their preparations indicates that Ancient Egyptians were familiar with the principles of oxidation and reduction, could control the pH of a solution and were successful in preparing novel compounds through a controlled technology of chemical synthesis. In the latter endeavor it is shown that these Ancient people embraced the scientific method, preceding Aristotle's rejection in Ancient Greece of a purely deductive approach to scientific enquiry. Egyptian Blue, the only pigment synthesized by the Ancient Egyptians is also discussed, and attention is drawn to its potential future contributions to modern high-tech applications.

Keywords. Ancient egyptians, chemical synthesis, Egyptian Blue, Kohl, scientific method**.**

> "..*It appears that Egyptians have developed a technology of chemical synthesis in solution that allowed preparations of original compounds..*"
> *Phillip Walter* [1]

INTRODUCTION

The embryonic stage of modern chemistry "*Alchemy*" can be traced back to Ancient Egypt, where Hermes Trismegistus[2] said to be a contemporary of Moses, founded the art of *Alchemy* often dubbed the *Hermetic art*[a]. To many the words *Alchemy* and *Chemistry* are linked to "*Khema*" or "*Chemi"*[3] which referred to the ancient name for Egypt meaning the black land.

On the other hand Plutarch attributes the name "*Alchemy*" to the Ancient Egyptian activities, referring to their skills in the extraction of metals, the preparation of alloys, and the working of gold[4-6] all of which contributed to the practical part of Alchemy.

Today an observer reflecting on the achievements of the Ancient Egyptians, would certainly recognize a flurry of activities that could be referred to as '*Chemical*', which not only served their *Religious beliefs* but also had *utilitarian*, *aesthetic,* and *symbolic* connotations.

It is generally believed that such accomplishments resulted from purely empirical observations. However the question remains: did the Ancient Egyptians at any point grasp the chemical significance of some of these practices?

In what follows snapshots are provided of some of the most impressive '*Chemical*' achievements of the Ancient Egyptians, the origins of which can in many cases be traced back to their religious convictions.

RELIGIOUS BELIEFS AS CATALYSTS FOR CHEMICAL ACTIVITIES

In Ancient Egypt an almost obsessive horror of death and extinction was reconciled with an absolute faith in immortality.

To ensure eternal life, it was essential that the body be preserved in a good condition and that the tomb of the deceased be equipped with implements, stuffed animals, donations, jewelry (*in case of the rich and powerful*)… that would serve the deceased in the afterlife. Corpses as early as the third millennium BC were preserved by a special technique of embalming (*referred to as* **mummification)** where the chemical process of *osmosis* played a crucial role. The main purpose of mummification was the dehydration of the body so as to prevent anaerobic bacteria from living on its tissues, causing their putrefaction and decay.

It is very probable that the Ancient Egyptians did not understand the chemistry behind the phenomenon of osmosis but must have been aware, on a purely *empirical* basis, of the special role of natron (*a mixture of sodium carbonate, bicarbonate with very small amounts of sodium sulfate and sodium chloride*) in this dehydration process. It is of significance that Herodotus and Diodorus used the same word for preserved fish as that for mummy, considering that even in pre-mummification times, salt was used to dry fish.[7]

With bodies placed on a slanting board and covered with dry natron for forty days, fluids flowed readily by osmosis from inside the body through the skin and to the outer high concentration of natron, resulting in total dessication. Bodies were preserved by such a chemical process and satisfied the Ancient Egyptians dreams of immortality as well as their strong belief in the great beyond[8]. The precise methods of mummification varied from period to period, and also within the same period depending on the social status of the dead person.

THE TOMB: A PROMISE OF ETERNAL LIFE

With the onset of the Dynastic period (~3300BC) Egyptians built elaborate **tombs** which housed, protected, and equipped their dead for the afterlife. Initially built as a flat-roofed, rectangular structure: the '*Mastaba*', which included a shaft that led to an underground burial chamber, would soon give way to the pyramidal structure of the Giza pyramids (erected around ~2600-2500 BC). A special **mortar** was used as a binder to stabilize the heavy limestone blocks that formed the core and outer layers of these massive constructions.

Alfred Lucas[7] in his pioneering work on Ancient Egyptian mortar, asserted that before Graeco-Roman times, the mortar employed for stone in Ancient Egypt was mainly 'gypsum'.

Some writers on Ancient Egypt have described Ancient Egyptian mortar as burnt lime, however, chemical examination by Lucas[7] has shown that Ancient Egyptians never used lime until the Roman period[b]. Such results were later corroborated by Coppola and co-workers [9] who analysed mortars belonging to the Ramesside era and found that they all had a gypsum based binder.

When heated at temperatures as low as 110°C-160°C gypsum loses water to produce the powder, plaster of Paris ($CaSO_4 \cdot ½H_2O$) according to the reaction:

$$2CaSO_4 \cdot 2H_2O \xrightarrow{heat} 2CaSO_4 \cdot ½H_2O + 3H_2O$$

and when water is added to the powder of plaster of Paris it rehydrates (absorbs water) and hardens rapidly.

$$2CaSO_4 \cdot ½H_2O + 3H_2O \rightarrow 2CaSO_4 \cdot 2H_2O$$

According to Coppola, Ancient Egyptian workers seemed to be conscious of the fact that the quality of the raw materials and methods of firing influenced the nature of the final product[9].

There is no doubt that the Ancient Egyptians recognised on *a purely experimental basis* the deceptively simple chemical reactions involved in the preparation of gypsum. They also most likely understood that *lime mortar* entailed the formation of calcium oxide (CaO or quicklime) with the subsequent formation of $Ca(OH)_2$ (slaked lime) to give mortar.

$$CaCO_3 + heat \rightarrow CaO + CO_2 \text{ and}$$

$$CaO + H_2O \rightarrow Ca(OH)_2$$

Even though lime mortar could have been used in their constructions, they probably realized that the prep-

Figure 1. The Kephren Pyramid in which gypsum mortar was used.

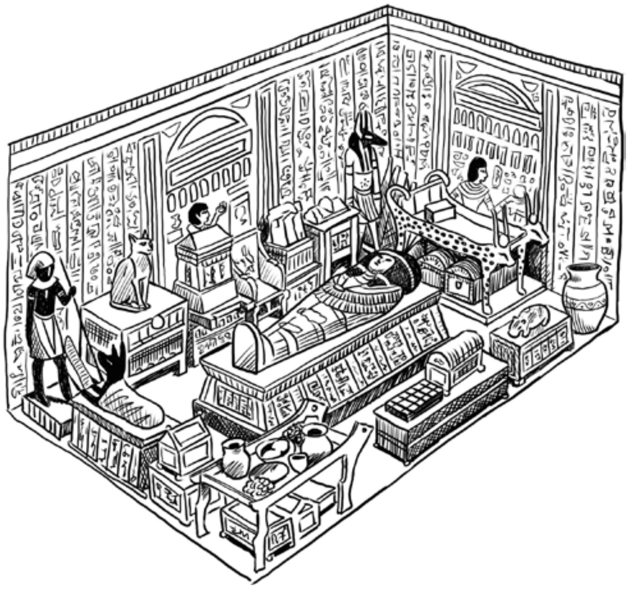


Figure 2. Interior of a tomb.

aration of CaO, requiring a heating temperature close to 900 °C, was not an ideal choice. According to Lucas the scarcity of fuel in Ancient Egypt and the low temperature processing of gypsum undoubtably must have been the reason why Ancient Egyptians preferred gypsum over lime.

It is very probable that the process of preparing a suitable mortar had to be learnt by multiple trials, the results being gauged by the nature of the final product.

Members of the ruling class and nobles were particularly meticulous in preparing their tombs and strived to ensure that all their needs for the afterlife were addressed. Metallic implements, small copper and bronze statuettes representing a variety of deities, colored faience, glass amulets and shawabtis (*small statuettes generally 365, that would serve the deceased every day of the year*), intricate jewelry… all had to be scrupulously prepared and securely buried with the deceased.

The skill for the production of burial implements and accessories, was generally attained through experimental observations, but may have necessitated a certain degree of primitive chemical knowledge.

COPPER AND BRONZE: THE BEGINNING OF A STRONG TRADITION OF METAL WORKING

It is only through heating the copper ore in the presence of carbon (*charcoal*) as reducing agent that, as early as the Old Kingdom (~2600 BC), a successful extraction of copper could be achieved in Ancient Egypt.

The ore basically in the form of malachite (basic copper carbonate) was crushed into small pieces and heated, in the presence of charcoal, to temperatures beyond 1000^0C (*reached at by means of blow pipes or foot bellows*) changing at first to copper oxide and then to molten copper.

$$CuCO_3{\cdot}Cu(OH)_2 \rightarrow 2\ CuO + CO_2 + H_2O$$

$$C + CuO \rightarrow Cu + CO$$

It would take a thousand years after that discovery for bronze (*an alloy of copper and tin*) to enter into extensive use in Ancient Egypt. A wide spectrum of statues, implements, weapons made out of this alloy were all buried in the tombs. As there are no tin ores in Egypt, it has been suggested that the tin was probably imported from Persia and possibly brought in by Phoenician traders[10].

The Egyptians would soon realize in their preparation of bronze, that the melting point of copper could be lowered by alloying it with tin, an observation which would prove handy in many instances.

Both the extraction of copper and the preparation of bronze seem to have been achieved through trial and

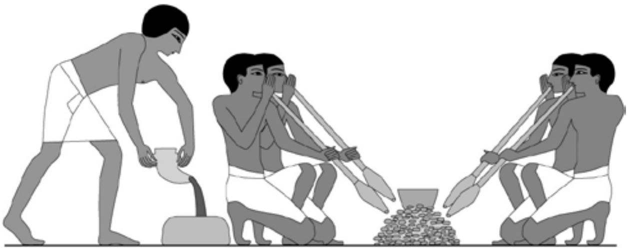


Figure 3. Extraction of Copper (malachite mixed with charcoal heated with blowpipes and then pouring of the molten copper as shown on the left).

Figure 4. Bronze statue of the Goddess Bastet.

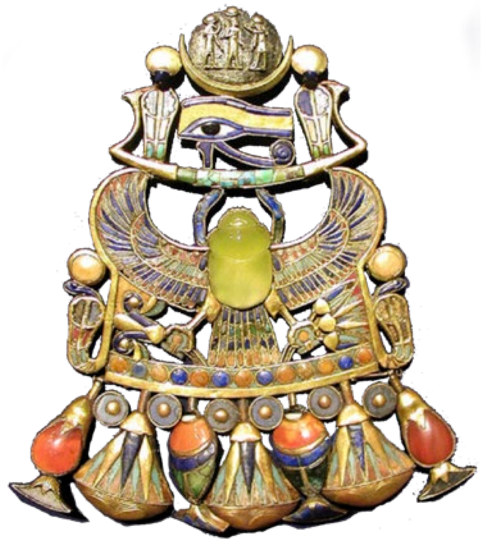


Figure 5. Gold Pectoral (parts soldered by colloidal hard soldering).

error with the results embodied in empirical observations which were methodically recorded and strictly followed.

GOLD: FROM *RELIGIOUS STATUARY* TO LEGENDARY *JEWELRY*

Identified with the sun god Ra and with the dazzling solar light, gold occupied a very special place amongst Ancient Egyptian metals. Its shine and glitter also made it quite attractive for use in ornaments and jewelry.

The successful preparation of bronze would pave the way for the progress Egyptians would attain in working gold, namely in relation to the successful soldering of intricate pieces of gold jewelry.

As judiciously pointed out by Cyril Aldred:

> *There seems little doubt that ancients soldered their goldwork by the process known as colloidal hard soldering. In colloidal hard soldering, ground copper carbonate, probably in ancient Egypt in the form of powdered malachite so commonly used as an eye-cosmetic, is mixed with gum or glue and this adhesive is employed to stick the grains or wire into place, or to coat the adjacent edges of the parts to be joined…*[11]

After heating and gradually increasing the temperature from 100˚C to 880 ˚C,

> *… at about 880˚C … the gold in contact with the copper melts to form a welded joint, whereas both gold and copper melt at nearly the same temperature well above this point, viz. 1083˚C and 1063˚C respectively.* [11]

This lowering of the melting point of both Copper and Gold was, as mentioned earlier, certainly inspired by bronze making. There is no recorded evidence, however, that the Egyptians had grasped the underlying chemical principle that governs such a phenomenon (*lowering of the chemical potential etc.….)* as we understand it today.

EGYPTIAN BLUE: A REMINDER OF SUBLIME JUSTICE AND PERFECTION.

Tomb walls were generally adorned with colored representations, mostly associated with deities, and symbolizing the sacred over the profane. Many of these depictions were painted in blue as to the Ancient Egyptians the blue color had a special significance. Worn on the breast plates of Egyptian priests, it was regarded as the color of *Divine Truth*. The blue colored *war* crown became very popular during the New Kingdom (~1500 BC) and was believed to confer upon its wearers, special protection from mysterious hostile forces.[7]

Lapis lazuli (most important component is *lazulite* of formula $Na,Ca)_8(AlSiO_4)_6(S,SO_4,Cl)_{1-2}$.) and the naturally occurring blue pigment Azurite ($Cu_3(CO_3)_2(OH)_2$ *a basic copper carbonate*) were both known to the Ancient Egyptians. However, the rare occurrence of lapis lazuli and the pale blue color of the azurite pigment, encouraged the Egyptians to produce their first synthetic pigment, the well known 'Egyptian blue'[b]

The earliest recorded use of Egyptian Blue is in the Old Kingdom (~ 2600-2100 BC) and its preparation continued into the Greco-Roman Period (330BC-400AD)[10] [c]

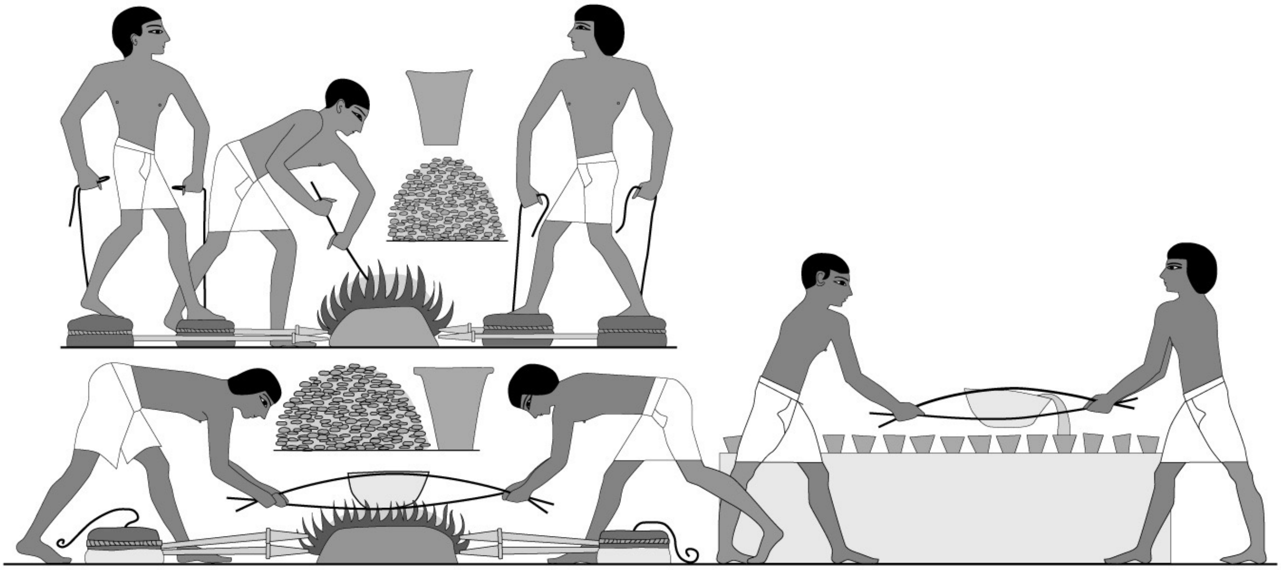


Figure 6. Preparation of Bronze (heating by means of foot bellows).

The secret of its manufacture was lost in the fourth century AD, and rediscovered only in the nineteenth.

The nature of this pigment has been extensively investigated[12-15] and was found to be a calcium-copper tetrasilicate with the formula $CaCuSi_4O_{10}$ or ($CaO.CuO.4SiO_2$) with a definite composition and crystal structure. Building on previous work and through their own attempts at preparing Egyptian Blue, Wiedemann and Bayer[13] concluded that the raw materials had to be close to the stoichiometric composition 1 CaO, 1 CuO, $4SiO_2$ and that small amounts of fluxes (borax, salt) were needed to catalyse the reaction and to yield a better crystalline structure. It was also observed that in order to achieve a bright blue color the synthesis had to be carried out in an oxidizing atmosphere and at a temperature lower than 1000^0C.

Even though it may never have been explicitly indicated, these results suggest that the Egyptians must have also been familiar with the rudiments of oxidation and reduction.

Today *Egyptian Blue* is another important and fascinating legacy spawn by this Ancient Egyptian civilization with the recent observation that when irradiated with visible light, it fluoresces with exceptional strength in the near infrared region of the electromagnetic spectrum [16-17]. Such a property appears to have an important future in modern high-tech applications, ranging from special fibre optical systems for telecommunications[18] , state-of-the-art high resolution biomedical imaging [19-22] luminescent fingerprinting dusting powder [23]and security ink technology[24].

COLORED AMULETS: BESTOWING PROTECTION, HEALTH AND GOOD LUCK

To match their magico-religeous beliefs the Ancient Egyptians fashioned small and beautifully colored objects (*notably scarabs, amulets, ushabtis*) made of a ware far better adapted that the rough clay, the so-called *Egyptian Faience*. Such a ware composed of a body(core) coated with an alkali-based glaze, gave the Ancient Egyptians the opportunity to produce a wide spectrum of colors.

Until the XVIIIth dynasty (1550 – 1295 BC), blue faience was produced from the thermal decomposition of malachite or azurite during the manufacture of the glaze (the color was mainly due to copper *Cu* in the

Figure 7. Hippo Goddess (Taweret, goddess of childbirth) in blue faience.

form of copper oxide CuO). Ancient Egyptians would soon become aware that a prolonged exposure to high temperature would favor the green color over the blue.[25] Caution therefore had to be exercised in the heating of the glaze and it was only during the Middle Kingdom (2055–1650 BC) that blue became common in faience.

During the XVIIIth dynasty(c. 1550-c. 1292 BC), cobalt in the form of cobalt oxide (CoO) was the principal colorant for blue faience and was generally accompanied by a significant amount of copper. An intense blue color was obtained with concentrations of CoO as low as 0.05 per cent which turned to violet or indigo when concentration was increased to 0.2 percent.

In the coloring of the glaze, the Egyptians realised, again probably on a purely empirical basis, that the colors exhibited by metal oxides in minerals could not always be transferred to the vitreous state. In view of the ligand field and crystal field effects, rarely does a transition metal in glass have the same chemical environment as in a mineral.

Amulets in the form of the Ankh sign (*the key of life*), the Eye of Horus, or the scarab, were often depicted in blue and were worn by the Egyptians and also buried with their mummies providing protection and prosperity.

Here to the symbolism of color (blue) was added the symbolism of form (amulet).

EYE MAKEUP: A VEHICLE TO GOOD HEALTH AND IMMORTALITY

*Makeu*p occupied a primary position at all levels of Ancient Egyptian society and played an important role in funerary rites for the purification of the body[26]. Beauty symbolised holiness – a key to the attainment of eternal life – and eye makeup in particular had an important standing in the Ancient Egyptians' collection of cosmetic elements.

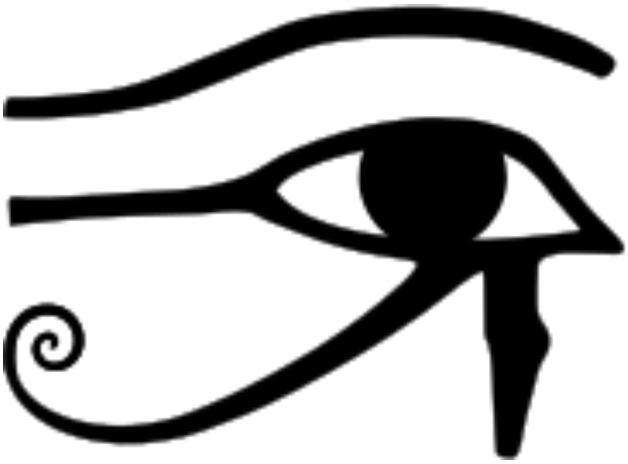


Figure 8. The eye of Horus.

A close connection was perceived between the *madeup eye*, the lunar cyclical renewal and the clash between the gods Horus and Set[26]. According to one myth, Set gouged out Horus's eye in a battle, an event perceived by the Egyptians as an interruption of the usual lunar cycle and a threat to the return of the new moon[27]. For the reestablishment of the cosmic order the eye had to be reconstituted and cured – a task successfully achieved by *Thot* the God of writing.

Philip Walter referring to the rehabilitation of Horus's eye points out that..

> *...The eye of the God should... be completed, reconstituted with makeup and unguents to ensure by the beneficial power of cosmetics the integrity and the health of the Divine eyes, and the victory of the Light.* [28]

The *Eye of Horus adorned with makeup* came to symbolise the moon with all its powerful and protective connotations. Ancient Egyptian religious texts attest to its primary symbolic role and importance:

> *Take two eyes of Horus, the black and the white, take them to your forehead that they may illuminate your face...*[29]

Such beliefs led the Egyptians to regularly use eye makeup during their lifetime. They also ensured that upon their death and as a vehicle to good health and immortality, containers holding cosmetic powders would be included in their burial surroundings .

KOHL AND THE PRACTICE OF WET CHEMISTRY

The 1798 Napoleonic expedition to Egypt brought back a large number of these powders preserved in alabaster, ceramic, wood or reed jars dating from 2000 B.C., the latter have been kept in the storage rooms of Louvre's laboratories.[26,30] Amongst these were two forms of eye makeup used since predynastic times: the green eye paint prepared from the mineral malachite which was usually applied to the lower eyelids and the black makeup, known as *Kohl*, generally used for the upper lids.

In 1995, a group of French scientists led by chemist Philip Walter started researching these Ancient Egyptian cosmetics through a collaborative partnership between the CNRS (National Center for Scientific Research), the Louvre Department of Egyptian Antiquities and the Scientific laboratories of l'Oreal. The collaboration lasted close to seventeen years and part of the work entailed the analysis of black ***Kohl***[26].

Using scanning electron microscopy in conjunction with X-ray diffraction for structural characterisation and phase identification, Walter and co-workers identified two ores of lead namely galena and cerussite (PbS and $PbCO_3$), but to their great surprise the analyses also revealed the presence of copious amounts of laurionite (Pb(OH)Cl) and phosgenite ($Pb_2Cl_2CO_3$)[31].

In view of their very rare presence in nature and their copious amounts in the cosmetic vials, it was concluded that these minerals must have been artificially produced. The excellent state of preservation of the containers ruled out any possibility of weathering or alteration effects.

For comparative purposes Walter and co-workers prepared these lead compounds by stirring lead oxide (PbO litharge) with rock salt (NaCl) in carbonated free water to give *laurionite* (Pb(OH)Cl) and by adding to the mixture natron (mainly Na_2CO_3 and $NaHCO_3$) to obtain *phosgenite* ($Pb_2Cl_2CO_3$). In both cases the prepared minerals were very close in composition and texture to the archaeological compounds[32].

To avoid the undesirable formation of hydroxides the reactions taking place (equations 1 and 2), seemingly quite simple, had to be closely monitored as a neutral pH needed to be maintained.

$$PbO + H_2O + NaCl \rightarrow Pb(OH)Cl + NaOH \quad [1]$$

$$PbO + H_2O + NaCl + 1/2Na_2CO_3 \rightarrow 1/2\ Pb_2Cl_2CO_3 + 2\ NaOH \quad [2]$$

Such a preparation therefore entailed the repeated addition of fresh water and sodium chloride and the continuous removal of the supernatant liquid. The process required several weeks to reach completion[31].

As pointed out by Patricia Pineau, director of research communication for the cosmetics giant L'Oreal:

> *Without knowing much chemistry, how did they have the foresight to know that a chemical reaction started on one day would produce such and such a result after several weeks?*[30]

This discovery led to the astonishing revelation that the Ancient Egyptians were in fact quite versed in the rudiments of ***wet chemistry***, a practice which enabled them to synthesise original compounds in solution. The question remained: why did the Ancient Egyptian add these preparations to their eye makeup?

Old manuscripts indicated that eye cosmetics "*... were essential remedies for treating eye illness and skin ailments...*"[32] and the Ancient Egyptian Ebers medical papyrus mentioned Kohl for the treatment of a plethora of eye diseases[33].

Intrigued by this situation, the French scientists Amatore, Walter and co-workers embarked on a project to find out if lead compounds had indeed any therapeutic effects. Using ultramicroelectrodes they showed that submicromolar concentrations of Pb^{+2} generated by the partial solubility of *laurionite* (Pb(OH)Cl) and added to human skin cells led to the production of NO, a molecule which played a role in the body's immune response.[32] Commenting on such findings Martin Oliver from McGill university suggested that the released nitric oxide could either stimulate the immune cells present in the eye or alternatively kill the disease-forming bacteria close to the eye[34] .

It is therefore possible that the Ancient Egyptians realised on a purely empirical basis, that whenever a white paste (identified today as laurionite or phosgenite) was present in the eye makeup preparation, it would have a therapeutic effect on its bearers and would give them greater immunity. This observation may have been the driving force behind this specific synthesis[32] .

According to Bernstein such an activity "*....remains the first known example of a large scale chemical process*" in Ancient Egypt[35] !

SOME REFLECTIVE COMMENTS

Having dwelt at length upon some of the fascinating '*Chemical*' accomplishments of the ancient Egyptians, it is now perhaps prudent to examine more closely the role of *empirical probing* as opposed to *rational thinking* and *quantitative speculation* in the chemical endeavors of these remarkable people.

Almost a century ago, L.E. Warren referring to the Ancient Egyptians chemical practices expressed the following opinion:

> *It should be understood that the Egyptians in general did not possess an inquiring mind and that ordinarily they would not conduct experiments merely for the purpose of satisfying curiosity or gaining knowledge...* [36]

More recently Wledemann and Berke[14], with regard to Egyptian Blue and other Ancient Egyptian chemical activities, suggested that:

> *...Man-made blue pigments required sophisticated chemical and technological developments... Ancient chemical achievements could not be based on atomic or molecular grounds. Therefore any progress was established by long and tedious processes of empirical probing*[14]

There is no doubt that many of the Egyptians accomplishments, some of which are described in this text, must have come as a result of long and laborious experimental scrutiny and elaborate processes of empirical trials.

However when one considers Egyptian blue, the rigorous stoichiometric requirements *and* the specific conditions for its preparation (*oxidizing atmosphere, addition of a small amount of catalytic flux, temperature control*)[13], must have no doubt involved some degree of *quantitative speculation* and *rational thinking*. Furthermore, the *idea* in itself of preparing one of the first known synthesized pigments, suggests the Ancient Egyptians' ability to *innovate* and *think creative*ly.

The same is true with regard to the synthesis of the new compounds *laurionite* and *phosgenite* using "*wet chemistry*" in which the acidity had to be controlled over several weeks, and the alkaline supernatant liquid continuously removed with the attendant addition of salted water. This certainly entailed quite an elaborate empirical process but also reflects an activity which is implicitly intermingled with sound knowledge *and* which is not totally devoid of any *rational speculation*.

With regard to the synthesis of phosgenite, Philip Walter suggested :

> *We might presume that the observation of natural phenomena may have enabled them to develop and invent such a science. Due to the regular flooding of the Nile and the presence of the desert, Egypt is a country that offers opportunities to observe a large number of mineral formations of exceptional character, especially around the salt lakes of the Wadi Natrun which supplied the natron so necessary to mummification. These carbonates of sodium are produced by chemical reactions between the salt water of the lake and the limestone substrate of the lake bottom, following very similar mechanisms to those involved in the making of the synthetic constituents of cosmetics…*[28]

It can be therefore be surmised that the Ancient Egyptians in their synthesis of *phosgenite* applied a version of our modern day scientific method. Assuming that they *observed* as suggested by Walter the natural formation of this compound, this would then have led them to *hypothesise,* as an informed guess, the needed conditions for a successful preparation, followed by *testing* through carefully controlled and replicable experiments (*control of the acidity through continuous washing…)* and ultimately *verifying* the validity of their hypothesis and *checking* whether or not it needed modification (*obtaining a white compound with immunological effects..*).

This should not come as a surprise to us since according to the Edwin Smith papyrus there are indeed indications that the Ancient Egyptians had a rational approach to medicine in which they applied the present day *scientific method*[37,38].

The chemical practices of the Egyptians stand in partial contrast to Plato's deductive mode of thinking where pure reasoning was the only route to knowledge at the total exclusion of experimental verification. There is no doubt that the Egyptians' manufacture of these artificial lead–based compounds reflects an inductive approach very much in keeping with our modern scientific mode of inquiry!

NOTES

a) "Hermeticism, also called Hermetism is a religious, philosophical, and esoteric tradition based primarily upon writings attributed to Hermes Trismegistus[39]. According to the 'Hermetic view' man can share in divinity and is therefore at least potentially in constant communication with God. The notion of a mystical ascent to the good acts as a unifying theme in 'Hermetism'.

b) Other pigments used in Ancient Egypt were mostly natural minerals. When working in 1980 as a chemical consultant to' The Sphinx Project' at the American Research Center in Egypt (ARCE), I analysed by X-ray Diffraction some blue pigments which were extracted from a cache in the front paws of the Sphinx. These were identified as Egyptian blue (Jehane Ragai: Special report to ARCE, 1982).

c) My own analysis of a series of Ancient Egyptian mortars extracted from the Great Giza pyramid, the second Giza pyramid and the Sphinx revealed the predominant presence of a Gypsum based binder.

REFERENCES

1. P. H. Walter, *Cosmetic and Therapeutic Chemicals* **2003**, 1.
2. Herodotus, *Euterpe,* Vol. 1, LXXXII. 381,Earle, Philadelphia, p. 1814.
3. *The Alchemy Reader: From Hermes Trismegistus to Isaac Newton*, (Ed.: S. J. Linden), Cambridge, **2014**, p. 5.
4. L. E. Warren, *J. Chem. Educ.*, DOI 10.1021/ed011p146.
5. *A History of metallurgy*, (Ed.: F. Habashi), Quebec, **1994**, pp. 11-42.
6. J. Ogden in *Ancient Egyptian Materials and Technologies*, (Eds.: P. T. Nicholson, I. Shaw), Cambridge, **2000**, pp. 148-173.

7. A. Lucas, J. Harris, *Ancient Egyptian Materials and Industries*, Dover Publications, **1999**, p. 283.
8. J. Ragai, G. De Young in *Encyclopedia of the History of Science, Technology and Medicine in Non-Western Cultures***,** (Ed.: H. Seline), Springer publications, **2008**, pp. 749-750.
9. M. Coppola, M. G. Taccia, C. Tedeschi, *Proceedings of Conference Built Heritage 2013, Monitoring Conservation and Management* **2013**, 1382.
10. H. A. Ead, **2011**, http://www.touregypt.net/science.htm
11. C. Aldred, *Jewels of the Pharaohs: Egyptian Jewelry of the dynastic Period*, Thames and Hudson Ltd, **1971**, p. 99.
12. W. T. Chase, *Science and archaeology*, Cambridge, Mass, **1971**, pp. 80-90.
13. H. G. Wiedemann, G. Bayer, *Anal. Chem.* **1982**, 54(4), 619.
14. H. G. Wiedemann, H. Berke, *Proceedings of the conference : The polychromy of Antique Sculpture and the Terracota Army of the first Chinese Emperor: Studies on Materials and Techniques conservation* **2001**, 158.
15. M. S. Tite, M. Bimson, M. R. Cowell in *Archaeological Chemistry III*, (Ed.: J. B. Lambert), Washington, DC, **1984**, pp. 215-242.
16. G. Accorsi, G. Verri, M. Bolognesi, N. Armaroli, C. Clementi, C. Miliani, A. Romani, *Chem. Commun.* **2009**, 23, 3392-4.
17. R. Brazil, *Chemistry World* **2017**, 14(6), 19.
18. D. Johnson-McDaniel, C. A. Barrett, A. Sharafi, T. T. Salguero, *J. Am. Chem. Soc.*, DOI 10.1021/ja310587c.
19. G. Pozza, D. Ajò, G. Chiari, F. De Zuane, M. Favaro, *J. Cult. Heritage* **2000**, 1(4): 393.
20. G. M. Davies, R. J. Aarons, G. R. Motson, J. C. Jeffery, H. Adams, S. Faulkner, M. D. Ward, *Dalton Trans.*, DOI 10.1039/B400992D.
21. N. M. Shavaleev, L. P. Moorcraft, S. J. A. Pope, Z. R. Bell, S. Faulkner, M. D. Ward, *Chem. Eur. J.* **2003**, 9, 5283.
22. M. H. V. Werts, R. H. Woudenberg, P. G. Emmerink, R. van Gassel, J. W. Hofstraatand, J. W. Verhoeven, *Angew. Chem.* **2000**, 39, 4542.
23. B. Errington, G. Lawson, S. W. Lewis, G. D. Smith, *Dyes and Pigments* **2016**, 132, 310.
24. C. Q. Choi, *Scientific American* **2013**.
25. A. Kaczmarczy, R. E. M. Hedges, *Ancient Egyptian Faience*, Aris and Phillips Ltd., London, **1983**.
26. P. H.Walter, F. Cardinali, *L'art-chimie: enquête dans le laboratoire des artistes*, Michel de Maule, **2013**.
27. G. Pinch, *Egyptian Mythology: A Guide to the Gods, Goddesses, and Traditions of Ancient Egypt*, Oxford University Press, **2004**, pp. 131-132.
28. P. H. Walter, *Molecular and Structural Archaeology: Cosmetic and Therapeutic Chemicals* **2003**, 1.
29. R. O. Faulkner, *The Ancient Egyptian Pyramid Texts*, Clarendon Press, Oxford, **1969**.
30. B. Thorson, *The Independent*, **1999**, http://www.independent.co.uk/arts-entertainment/science-the-art-of-ancient-egypt-1110924.html
31. *P.* Walter, P. Martinetto, G. Tsoucaris, R. Brniaux, M. A. Lefebvre, G. Richard, J. Talabot, E. Dooryhée, *Nature* **1999**, 397(6719), 483.
32. *I.* Tapsoba, S. Arbault, P. Walter, C. Amatore, *Anal. Chem.*, DOI 10.1021/ac902348g.
33. R. Kreston, *Discover magazine*, **2012**, http://blogs.discovermagazine.com/bodyhorrors/2012/04/20/ophthalmology-of-the-pharaohs/#.WbqO_8Zx3IU
34. W. R. Corliss, *Science Frontiers Online*, **1999**, http://www.science-frontiers.com/sf123/sf123p01.htm.
35. M. Bernstein, M. Woods, *ACS*, **2010**, https://www.acs.org/content/acs/en/pressroom/newsreleases/2010/january/ancient-egyptian-cosmetics.html.
36. L. E. Warren, *J. Chem. Educ.*, **1934**, 11(3), 146.
37. D. K. Mak, A. T. Mak, A. B. Mak, *Solving everyday problems with the scientific method: thinking like a scientist*, World Scientific, **2016.**
38. M. Stiefel, A. Shaner, S. D. Schaefer, *The Laryngoscope*, **2006**, 116(2), 182.
39. R. Audi, *The Cambridge Dictionary of Philosophy (2nd ed.),* Cambridge University Press, **1999**.

PICTURE CREDITS

All pictures are from the public domain except for Figures 2, 3 and 4 were drawn by Fadia Badrawi.

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

OPEN ACCESS

Citation: F.M. Rubino (2018) The "*Bitul B'shishim (one part in sixty)*": is a Jewish conditional prohibition of the Talmud the oldest-known testimony of quantitative analytical chemistry?. *Substantia* 2(1): 103-119. doi: 10.13128/substantia-44

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.

Historical Article

The "*Bitul B'shishim (one part in sixty)*": is a Jewish conditional prohibition of the Talmud the oldest-known testimony of quantitative analytical chemistry?

Federico Maria Rubino

Università degli Studi di Milano, Department of Health Sciences, Ospedale San Paolo, v. A. di Rudinì 8, I-20142 Milano (Italy).
E-mail: federico.rubino@unimi.it

Abstract. Accomplishments of Hellenistic science and technology in some fields, such as mathematics, physical cosmology and engineering, has recently been re-evaluated and can be considered as of the same level that the scientific revolution in Western Europe reached at the beginning of the XVII century CE. Information on the level of chemical science is scanty; however, independent ancient sources such as the Jewish Talmud can yield significant clues. The still existing dietary laws include a practice to assess the acceptability of food mixtures with two complementary assessment techniques. One enforces a specific minimum mixing ratio (1:60) of unacceptable-to-acceptable ingredients, the other uses a sensory assessment to exclude the presence of a tasty unacceptable ingredient. This practice is likely the first historical example of quantitative analytical chemistry. This article collects clues that this approach is rooted in the implicit acceptance by Hellenistic chemical science of an atomic paradigm and on the awareness that interaction of different matter yields product that are different from the starting ones. Quantitative assessment of the presence of unacceptable ingredients by sensorial assessment or by mixing ratio likely points to a forgotten practice of Hellenistic experimental pharmacology and physiology to test the efficacy of drugs and poisons, that was performed in animals, with the use of a control group, and on human subjects.

Keywords. Dilution, food contamination, halacha, Hellenism, Jews, kasherut, mixture, Talmud.

1. MUCH EARLIER ANALYTICAL CHEMISTRY THAN ANTICIPATED?

Humankind practiced empirical chemistry since the farthest of times to produce food, materials and market goods,[1] and documental sources report a good deal of recipes since ancient Near Eastern civilizations.[2] Studies of paleo-chemistry and ancient chemistry however suffer from several sources of difficulty. One is the fragmentation and obscurity of documental sources, and the inherent limitation in reconstructing technological achievements of different cultures over time. However, when the contemporary researcher re-considers and follows in detail ancient recipes, such as those reported in Early Modern "alchemical" documents, the obtained results closely approach the descriptions given by the original Authors, as has been recently documented.[3-7]

Paleo-chemistry (and contemporary chemistry) was more often concerned with manufacturing goods, such as metals and alloys, dyes, medicinal drugs and poisons,[1] rather than with the intellectual effort to understand the properties of matter. Miners often performed assays for metals in ores and alloys, and the composition of ancient pharmaceutical preparations manufactured from mixtures of individual ingredients (each of which needed to be authenticated, especially when it came from remote locations) is often reported in quantitative terms [8]. However, very little is known on the assays, if any, that were employed to identify raw materials and to check their adequacy to specific purposes, such as the integrity of metals, the composition of alloys, of food, drugs, dyes and perfume, an activity that tantamount to analytical chemistry. Among the few reported "analytical" methods is the assay of metal ores by cupellation. This technique allows to concentrate precious metals, such as gold and silver, from ores by a solid-liquid extraction into low-melting lead, followed by recovering the precious metal(s) from the latter, more easily oxidized metal, by ashing the metal button in air.[9] The described "analytical" method thus essentially corresponds to a small-scale preparation, and weighting in a scale is the final method of measurement.[10]

Seldom are a very small number of material documents, such as the residues of old-time equipment, vessels, raw materials and preparations found in archeological studies and only very recently some could be compositionally characterized.[11-13] The only clue that in the Antiquity an assay was used to detect a specific component in a mixture is the detection of Iron in biological fluids with the use of the extract of oak gall.[14,15]

The lack of information on other chemical analyses is not surprising, since also the contemporary discipline of analytical chemistry achieved a distinctive status within chemical sciences much later than the traditional branches of mineral (inorganic) and organic chemistry, and of *materia medica*, the forerunner of pharmacology, toxicology and medicinal chemistry. Methods for chemical analysis of minerals, *i.e.*, to distinguish the different simple constituents and to identify new chemical elements were published as early as the early XVII century, even in the lack of a consistent theory on the composition of matter and of an operative definition of what a "simple" chemical body, *i.e.*, a chemical element is. It is only in 1861 that the renowned independent analyst Carl Remigius Fresenius founded the first scientific journal specifically dedicated to analytical chemistry as an independent discipline.[16] Only in the 1940s the most authoritative of contemporary scholarly journals of analytical chemistry, the Analytical Chemistry of the American Chemical Society, gained an independent status, formerly being since 1929 a supplementary issue of the Society's magazine of industrial chemistry.[17]

2. A REAPPRAISAL OF HELLENISM: THE BOOM AND THE DOOM.

Hellenism is the period of Mediterranean history that stands between Alexander the Great's death in 323 BCE and the battle of Actium in 31 BCE that ended the Ptolomean rule of Egypt. Historians have long considered this as a ripe age with little real intellectual achievement, when compared, on arbitrary terms, with the previous Classical period of Greek history that established as paradigm of Western European culture. Among prejudices on this period is that according which abundance of human slave workforce caused a limited interest in mechanization of work, and the consequent lack of a developed production economy. The Old Mediterranean and Greek-Romans thus failed to understand and exploit the natural world. The missed opportunity to develop the budding knowledge into a "modern" framework also led to the withdrawal of some intellectual achievements of early Hellenism that actually foreran those of the Early Modern age. In particular, this is exemplified by the fact that Ptolemaic model of geocentric universe overcame the Aristarchus' heliocentric one, Galenic medicine mostly cancelled Herasistratus' physiology, and Archimedean mechanics only found limited exploitation in devices, such as those later described by Heron.[18]

On the contrary, a very recent re-interpretation of the surviving Hellenistic and later texts and of material artefacts indicates that the intellectual development, scientific and technological advancement at the peak of that period was at the same level that Modern Western Europe only reached in late Renaissance and Early Modern age. One main scholar to initiate this innovative interpretation, Lucio Russo, reconstructed some parts of the "lost knowledge", mainly in the fields of geographical physics, cosmology and astronomic navigation.[18-20]

A real intellectual *Boom* occurred between the III and II centuries BCE with the establishment in the recently founded Alexandria of the Museum and Library by Tolomy II Eupator.[21] Other Hellenistic kingdoms followed, such as the Attalids in Pergamon (to tackle an *ante-litteram* embargo to the export of papyrus as the substrate for writing, the use of parchment was locally boosted), the Seleucids in Syria and Babylon, and elsewhere in the *koiné*, for which information is much more limited.

The abrupt geo-political *Doom* of the Hellenistic civilization occurred in the mid-II century BCE, with the almost contemporary destruction of Corinth and of Cartago in 146 BCE. Just a few years later, the anti-Greek Alexandria *pogrom* sponsored by Ptolemy III Euergetes in 137 BCE caused the migration of the luckiest scholars to safer remote places. The wholesome destruction of the main intellectual centers of the time determined a break in the transmission of knowledge, surviving scholars relocating to safer areas in the Eastern Mediterranean, in Syria and further eastwards to Bactria and Maurya India. Scholars in the quieter I century CE and later strove to revive the mostly interrupted intellectual activity, but were no more able to recover the loss, and most scientific advancements in applied mathematics, astronomy, geography and navigation, in natural sciences and medicine faded into oblivion, due to the inability to reconstruct the underlying methodological and theoretical framework.[22]

Ancient knowledge in "empirical sciences", such as in medicine and chemistry, cannot be as easily reconstructed, due to the loss of most original sources, and to the corruption of residual information that could not be understood any more in *post-Doom* times.[22] In the field of chemistry, Greek and Hellenistic scholars had conceived a rationally based precursor of the current atomic-molecular model of the composition of matter as early as the V century BCE, building on the first intuitions of Democritus, and progressively developed by Aristotle, the Epicurean school, Crisippus and the Stoic school.[22] After the *Doom*, the tenets on which the budding theory of matter composition had been developed were abandoned, purportedly for their sheer materialistic content. As in other disciplines, such as the drift of Hellenistic physical astronomy into astrology, the remnants of that knowledge merged with other philosophical and religious traditions, and evolved into alchemy.[23,24] Likely, the practical contents that dealt with manufacturing high-tech materials, such as imitation gems and gold-looking alloys, dyes, pigments, perfumes, pharmaceutical drugs and poisons, faded into the practical recipes which artisans transmitted through oral tradition, in a social environment that now was well detached from the shrinking population of educated scholars.

However, it is conceivable that some "fossil" knowledge of the *Hellenistic boom*, and especially its quantitative applications, was already embedded as common discourse in sources that have been so far untapped, and that their exploration can yield new insight on their knowledge in the other fields.

3. "FOSSIL" INFORMATION FROM AN OLD MEDITERRANEAN PEOPLE.

One possible, and so far little examined, source of information is the Talmud, a written compilation of discussions in the application of Hebrew religious Law (*halacha*) to everyday affairs that was elaborated in Roman Palestine (*Talmud Jerushalmi*) and in Parthian Babylonia (*Talmud Bavli*)[25] from the III to the VI century CE. *Halacha* developed from the normative books of the "written Torah" (*Torah she-bi-khtav*, the Pentateuch of the Old Testament) and on the "oral Torah" (*Torah Sheba'al Peh*), the sources of which were rooted at least four centuries earlier.[26,27] The exploration of Talmud unveils earlier knowledge science and technology, encompassing – in contemporary terms – animal and human anatomy and physiology, chemical technologies,[28,29,30] statistics[31,32] (including the earliest-known description of random sampling: Chullin 4a), risk prevention and management.

One peculiar aspect of *halacha* is the enforcement of several alimentary *taboos*, some of which – such as abstinence from pig – are so well known as to become symbolic and even a synecdoche of the Jewish identity to the other peoples. The main primary sources of information for the Jewish dietary rules are the Torah (in particular, Deut. 14:1-26 contains the well-known compilation of allowed and forbidden animals for food), the Talmud and a much later compilation, the XVI century CE *Shulchan Aruch.*[33] that summarized halachic rules as enforced by Sephardi Jewry in the Mediterranean area. Several treatises of the Talmud report the very complicated rules on inacceptable (*issur*) and acceptable (*heter*) food and on mixtures, such as in *Chullin* (most *loci* that are especially pertinent to this essay are between 82a and 98b; *v. infra*), and related information occurs in treatises that discuss other sources of material impurity. The topic is of a great importance to practicing Jews down to present times.[34]

Among the lesser-known food regulations are discarding the sciatic nerve (*gid hanasheh*) from the thigh of ritually slaughtered animals, completely removing fat from the slaughtered carcass, and completely draining meat from blood, the ban to mixing meat and milk in food (*basar be chalav*). In particular, the Torah states thrice the *basar be chalav* prohibition (Es. XXIII, 19 and XXXIV, 26; Deut. XIV, 21), that was unknown before

Moses' Covenant was established (in the Book of Exodus), since Abraham offers to the three Visitors meat cooked in sour milk (Gen. XVIII, 7-8).

> Es. XXIII, 19. "Bring the best of the firstfruits of your soil to the house of the LORD your God. "Do not cook a young goat in its mother's milk.
> Es. XXXIV, 26. "Bring the best of the firstfruits of your soil to the house of the LORD your God. "Do not cook a young goat in its mother's milk."
> Deut. XIV, 21. [21] Do not eat anything you find already dead. […] Do not cook a young goat in its mother's milk.
> Gen. XVIII, 7-8. [7] Then he ran to the herd and selected a choice, tender calf and gave it to a servant, who hurried to prepare it. [8] He then brought some curds and milk and the calf that had been prepared, and set these before them.

The *gid hanasheh* prohibition comes earlier in the Torah, since it links to Jacob-Israel's fight with the Angel, who left him lame for life (Gen. XXXII, 22-31).

> Gen. XXXII, 22-31. . […] [25] When the man saw that he could not overpower him, he touched the socket of Jacob's hip so that his hip was wrenched as he wrestled with the man. […] and he was limping because of his hip. […]

The *basar be chalav* prohibition had no apparent explanation since the most ancient times of Jewish culture, and therefore the religious authorities expanded its application in order not to infringe the ban.[35] In general, to avoid cross-contamination of mutually incompatible food ingredients, separate sets of pots are used, and ritual cleaning with water or on fire is performed.[34]

These food-mixing bans, however, admit an exception to thrashing the forbidden food mixture, in the case mixing of forbidden ingredients occurred by accident. In this case, to test whether the mixture is still admissible as food, two complementary routes are available. One states that if the contaminating ingredient is present in a proportion that is less than one-sixtieth of the main one (*bitul b'shishim*: one part in sixty), the food is still *kosher* (ritually acceptable). Another possibility is that if the contaminant does not impart to the mixture its distinctive properties of taste (*ta'am k'ikar*: the taste is equivalent to the substance), the food is still acceptable.[34]

Both approaches are so familiar to a present-day regulatory chemist or toxicologist, as to remind other similarly "modern" Hellenistic accomplishments that are considered as "anticipating" contemporary views in mathematics, in astronomy, in mechanics, in natural sciences.

The *bitul b'shishim* "*one part in sixty*" approach closely resembles the contemporary practice of toxicological risk assessment, whereby the presence of a contaminant is compared to an enforced lower limit, and decision upon acceptability is taken consequently.

The *ta'am k'ikar* "*the taste is equivalent to the substance*" approach corresponds to the use of a sensorial assay, and is similar in principle and setup to what is nowadays performed for similar applications. In particular, it is requested that an unaware, extraneous assessor (the *akum*, a Gentile) taste the mixture for absence/presence of the undesired ingredient (in modern terms, a *blind test*). The use of a *biological response as endpoint* foreruns the now abandoned approach to limit setting for airborne industrial solvents that was in use in the former Soviet Union, based on the measurement of evoked electrophysiological potentials triggered by body exposure to exogenous substances.[36] The criterion whereby acceptability comes when the unacceptable component is *no longer perceived* is again a forerunner of the ALARA (As Low As Reasonably Achievable) principle that is adopted in radio-protection and in the management of environmental and occupational risk from carcinogenic chemicals.[37] *Nihil sub sole novi* (Qohelet, 1,9).

The Jewish normative texts (several *loci* in *Chullin*, *e.g.* especially between 89b and 120a; *Shulchan Aruch's Yoreh De'ah*, several *loci*; *see* Appendix) describe in much detail the transmission of the off-flavor of unacceptable substances from a contained liquid or solid-in-liquid to the container, or from an unexpected and not allowed contaminant to the bulk of food. Both ancient texts correctly identify as the determinants of the process: temperature, contact time, the nature of the liquid, that of the material and the surface-to-bulk ratio of the immersed solid and the material of the container, and enforce consequently the halachic rules. Those reported below are just a few examples that describe both assessment strategies.

> **Ch. 89b chap. VII:** […] If a thigh was cooked together with the sciatic nerve it is forbidden if it imparts a taste [into the thigh]! (Note 11: I.e., if the thigh that was cooked was not sixty times greater than the forbidden nerve; for the Rabbis have estimated that if there were more than sixty parts of permitted matter as against one part prohibited, the latter cannot impart a flavor unto the former.) […]
> **Ch. 96b chap. VII:** MISHNAH. IF A THIGH WAS COOKED TOGETHER WITH THE SCIATIC NERVE AND THERE WAS SO MUCH [OF THE NERVE] AS TO IMPART A FLAVOUR [TO THE THIGH], IT IS FORBIDDEN. HOW DOES ONE MEASURE THIS? AS IF IT WERE MEAT [COOKED] WITH TURNIPS (Note 7: It is estimated by the Rabbis that meat cannot impart its taste to any substance that is cooked with it if the latter is sixty times as large in bulk as the meat.).

> **Ch. 97a,b:** […] the Rabbis ruled that one may rely upon a [gentile] cook, and yet [in other cases] the Rabbis ruled that the test is sixty [to one]. Therefore we say, where substances of different kinds, each kind being permitted by itself, were mixed together, the test is whether or not one imparts a flavor to the other; 1 and if one of the substances was forbidden 2 then we rely upon the opinion of a gentile cook. […]
> **Ch. 97b:** […] R. Nahman said: The [sciatic] nerve [is neutralized] in sixty-fold, but the nerve itself is not to be included to make up this number. (Note 14: I.e., there must be sixty times the volume of the forbidden nerve.) The udder is neutralized in sixtyfold, but the udder itself is to be included. (Note 15: If an udder which was not emptied of its milk was cooked together with meat, the entire contents of the pot would be forbidden unless there was in the pot sixty times as much as the milk of the udder. (The quantity of milk in the udder is regarded as equal to the volume of the udder). Now the udder can also be included to make up this sixty-fold since it is not the udder that is forbidden but only the milk contained in it. In other words, there must be in the pot fifty-nine times the quantity of the udder; v. infra ***109a***.) An egg (Note 16: Of an unclean bird which was boiled with eggs of clean birds.) is neutralized in sixty-fold, but the egg itself is not to be included.
>
> …………..

It is apparent from the reported excerpts that both assessment practices: *bitul b'shishim* (one part in sixty), and *ta'am k'ikar* (the taste is equivalent to the substance) correspond to the likely earliest example of a quantitative analytical chemical assay. It is worth considering that no information on a corresponding quantitative approach is anywhere found in survived texts of Greek, Hellenistic and Greco-Roman technology.

The particular dilution factor most commonly considered as upper limit for halachic acceptance, 1/60 (*approx.* 1.6%, or 98.4% pure), matches that which is also nowadays a useful threshold for the presence in a "technical grade" product of undesired contaminants or off-products, which are devoid of particular concern such as toxicological or microbiological health risk. This level is also close to the minimum detectable amount of some tests designed for the detection of adulterants in food, such as the late XIX-century Villavecchia-Fabris test that discriminates edible from industrial vegetable oils purposely adulterated with 5% sesame oil.[38]

It may thus be of an interest to understand whether this approach to quantitative chemical analysis might come into the Talmud deriving from Hellenistic chemical conceptions, to fulfill a specific halachic task. In a complementary way, this notion, which is contained in an early-CE text and the roots of which may well extend several centuries before, may be a clue to reverse-understand the nature and level of Hellenistic concepts on the composition of matter and on the relationship with other fields of natural science. As such, this interpretation suffers from lack of sufficient internal evidence, and may generate a circular argument. However, even if the practice of "alternative history" may quickly lead to fictionalized accounts, nevertheless by adopting some rules to control the construction of scenarios the voids in documentation can be credibly filled to re-create plausible descriptions of past events.[39]

4. ALEXANDRIA, ANTIOCHIA AND PERGAMON: MEDITERRANEAN BRIDGES BETWEEN HELLENISM AND JUDAISM?

A significant cultural interaction of Greeks and Jews developed starting in the IV century BCE, encompassed the early and late Hellenism, the Roman suzerainty and final conquer of Palestine, the Diaspora and continued after the fall of the Roman Eastern and Byzantine Empire to the Parthians and the Sassanids.[40,41] Briefly, and to the aim of this reconstruction, it is conceivable that some elements of Hellenistic knowledge in natural science and medicine outpoured into Hebrew halakhic discourse that developed in the "oral phase", even before the discussions started to be registered in writing as the Mishnah and Gemara, at the beginning of the III century CE. The Talmud and its earlier Judaic sources in fact contain several items that have long been recognized as of a likely Hellenistic origin.[42] Some Talmudic knowledge may thus represent one of the few remnants of lost Hellenistic science and technology that developed and was commonplace *before the Doom* in Ptolemaic Egypt, in Seleucid and Hasmonean Palestine, and in the Hellenistic kingdoms of Anatolia, mainly those of Pergamon and of Bythinia, and did not survive in the transmitted body of text of the Greco-Roman world.

As for the plausibility of this scenario, it is widely accepted now that, in several fields, Jews who were in contact with the Greek and Hellenistic environment reinterpreted Greek and Hellenistic knowledge, or just the cultural suggestions that their neighbors spread.[43]

> *There were one thousand young men in my father's house, five hundred of whom studied the Law, while the other five hundred studied Greek wisdom.*

To further appreciate the possible degree of interaction of Palestinian Jews who had a role in setting the *halacha* with the contemporary Hellenistic culture, we may recall a famed Talmud episode (Shabbat, 31a). As known, the rigorous Talmud Master Shammai […] *repulsed him* (the curious Gentile who sought for instant information on the intricacies of the Jewish faith) *with the builder's cubit, which was in his hand* […]. Shammai was a wealthy architect in I century CE Palestine, the cubit was in fact a measuring ruler, a rather sophisticated professional device akin to those in use by technicians and calculators as far as the 1980s, and his professional training very likely included elements of knowledge that was of Hellenistic derivation.

A clue that Greek and Hellenistic philosophy was at least known to educated, if observant Jews of the early Talmudic era is the use of the term "*apikoros*" to grossly indicate a secular thinker who negates most or all the tenets of Judaism, or of any revealed religion for the good. This word first occurs in rabbinic literature in the Mishnah (Sanh. 10:1),[44] and derives from the IV century BCE Greek philosopher Epicurus, who advocated a materialistic explanation of the world and the pursuit of a quiet happiness through vegetarianism and abstention from greed and violence.[45]

> Q. Horatius Flaccus, Epist., I, 4, 10. […] Me pinguem et nitidum bene curata cute vises, / cum ridere voles, Epicuri de grege porcum. (*If you ask of myself, you will find me, whenever you want something to laugh at, in good case, fat and sleek, a true hog of Epicurus' herd*)
> D. Aligheri, Commedia, Inf. X 13-15. Suo cimitero da questa parte hanno / con Epicuro tutti suoi seguaci, / che l'anima col corpo morta fanno. (*In this place Epicurus and all his followers are entombed, who say the soul dies with the body.*)

Epicurus is the only Greek philosopher who is explicitly mentioned in the Talmud, while there is no mention of the competing Stoic school.

> **Sanh. 10:1.** […] But the following have no portion in the World to Come: He who says that resurrection is not a Torah doctrine, the Torah is not from Heaven, and an apikoros [who denigrates Torah and Torah scholars] . Rabbi Akiva adds: One who reads from heretical book […] http://www.emishnah.com/PDFs/Sanhedrin%2010.pdf

During the Hellenistic period, there was an increased opportunity for Jews to spread, especially in Anatolia, where Attalus III and Mithridates VI favored the transferred of a large body of Jewish settlers.[46] It is conceivable that some immigrants belonged to socially educated strata[47] and may transmit knowledge and suggestions to the still vital Jerusalem center, possibly in the occasion of pilgrimages to the Temple.[40] Given the advancement of agricultural, pharmacological and toxicological studies of natural substances in the Anatolian Hellenistic kingdoms, in Seleucid Babylon and in Ptolemaic Egypt,[48,49,50,51,52,53] it is conceivable that such information may reach Palestine through multiple routes.

5. AT THE CORE OF THE ISSUE: FROM DRUG TITRATION TO ISSUR ESTIMATION … AND BACKWARDS.

That the late Egyptian Pharaoh's court was likely interested in experimentation at large is witnessed by the Greek historian Herodotus (?485-425 BCE). As reported (*Historiai*, Part 1, Book 2, paragraph 2) Psammetichus I (664-610 BCE) had a baby raised without hearing any spoken language, in the earliest recorded psychological experiment, to determine whether human beings have an innate capacity for speech, and if so, which particular language is innate.

Medical studies flourished in Hellenistic Alexandria and in other cities, and eventually developed into "research-oriented" anatomical and physiological studies.[54] Reportedly, Herasistratus was the first able to differentiate motor from sensor nerves by experiments that would not be possible to perform in animals or in dead human bodies, but only in living humans.[55,56] The argument of their cruelty was used by early Christian polemists, such as Tertullian, against paganism;[57] however, the *querelle* over whether Herasistratus really performed such experiments continues.[58]

A little exploited information to support the likeliness of the information comes from the Talmud treatise on womanly issues (Niddah, 30b), which reports that Cleopatra VII of Egypt performed systematic experiments on human fertilization, likely including the use of contraceptive drugs, forced timed intercourse mating and surgical abortion.

> **Niddah 30b.** […] A story is told of Cleopatra the queen of Alexandria that when her handmaids were sentenced to death by royal decree they were subjected to a test (note 23: Fertilization and subsequent operation) and it was found that both [a male and a female embryo] were fully fashioned on the forty-first day. […] They were made to drink (note 31: Before they were experimented on), a scattering drug (note 32: i.e., destroying the semen in the womb) […]. A story is told of Cleopatra the Grecian queen, that when her handmaids were sentenced to death under a government order they were subjected to a test and it was found that a male embryo was fully fashioned on the forty-first day and a female embryo on the eighty-first day.

As for the availability of the needed experimental tools, detailed knowledge of human fertility is well documented in Pharaohs' Egypt. The Berlin Papyrus of 1.350 BCE witnesses the knowledge and application of tests to assess pregnancy, based on the stimulating effect of pregnancy hormones excreted in urine on the germination of corn and barley seeds,[59,60] and a later one (papyrus Kahun[61,62]) uses swamp canes to the same purpose. There is clue that those ancient claims are even supported by empirical evidence.[63,64]

Due to the use of deliberate poisoning as a weapon in political struggle in the Hellenistic period (and later), efficacy studies on drugs and poisons were much developed, especially in the kingdoms of Bithynia (by Mithridates IV) and Pergamon (by Attalus III).[48-54] The high level of knowledge on toxic poisons and their antidotes calls for the use of systematic experiments, reportedly performed even on humans, such as slaves and convicted criminals, as witnessed by several nearly contemporary testimonies, especially regarding the former character.[39] It is conceivable that an assessment of the desired level of activity of concoctions (deadly toxic, or sub-toxic, "*Mithridatic*") should be performed, if the preparations were to be reliably used to their intended purpose. The II century CE pseudo-Galen text *Theriaca ad Pisonem* contains an important testimony to this practice, whereby the efficacy of a preparation against venomous animals is tested on animals, with use of a control group, but there is no clear indication that specific doses of the *pharmakon* were administered.

> 10R […] we being unable to test it on men do the same on certain other living beings and try to arrive at a true verdict on the drug. So we take cocks – not those that live with us under the same roof, but rather wild ones, and with a rather dry constitution, and we put poisonous beasts among them, and those who have not drunk theriac die immediately, but those who have drunk it are strong and stay alive after being bitten. […] (*Theriaca ad Pisonem*, ch. 1;10)[8]

It is here that the knowledge embedded in the Talmud's *bitul b'shishin* likely comes forward as a neutral witness.

It is conceivable that activity titration of *pharmaka* was possibly performed by testing the effects of progressive ("scalar") dilutions according to definite proportions,[65,66] an approach that would eventually re-surface in the XIII-XIV century CE, when the Montpellier medical school re-discovered the same principle from the al-Kindi treatise *Quia primos*.[67]

The measurement systems in the ancient world are difficult to reconstruct, since there were differences among regions and over time.

As witnessed also in the Talmud, several different scales were simultaneously in use, in particular the sexagesimal, decimal and binary ones. The binary, harmonic or Pythagorean scale ($1/2^n$, *i.e.* denominators in the sequence 1:2:4:8:16:32:64, and so on; Figure 1) is used still today for the same purpose. Dilutions with this scale are very easy to prepare and ensure a tight control over the concentration of the proband substance, since variation occurs by halving the preceding dilution. Moreover, such relative scale allows comparing the strength of different preparations, even when the actual quantity of the active material in a complex mixture is unknown and consistent units of measure may not be available. In fact, this is the case of natural extracts, and until the very recent advent of physico-chemical techniques for separation, identification and quantification of complex mixtures, this was the way to titrate biological drugs such as insulin.[68]

The actual correspondence to contemporary standards of the Talmudic units of measure and their mutual conversion is a matter of current controversy, since it occurs not only the realm of antiquary sciences and archeology, but also has a value for the enforcement of *halacha*. Not only the names and size of units changed over time and varied between different places, but also the scales used to build multiples were heterogeneous and often ill defined. The Talmud reports different and partially overlapping measures of volume for liquids and for dry (grain).[69] The *kav* (around 1,22 L; used for both dry and liquid) is the basic unit from which others are derived. The *log* (around 0.306 L; Lev. 14:10)

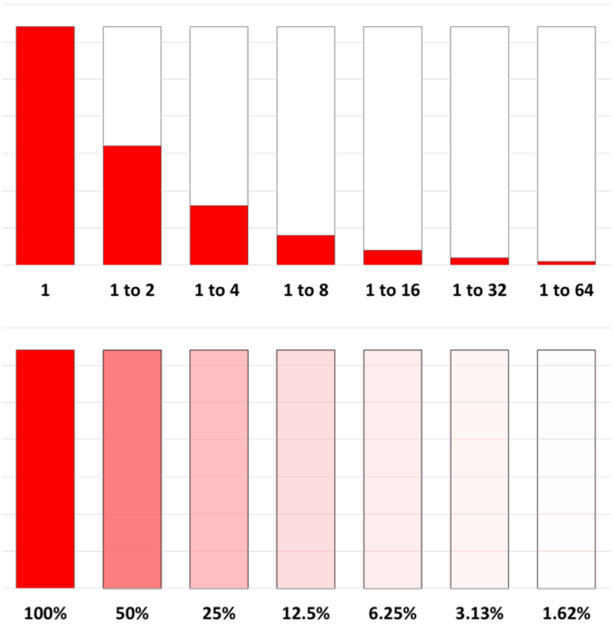


Figure 1. Scalar dilution of a concentrate (red) active substance in a solvent (white). Seven progressive dilutions are shown, starting from the mother liquor and ending with the sixth (1/64). The panel above shows the exponential decrease of the concentration form the arbitrary level of "one unit" to that corresponding to the sixth dilution. The fading of color of a solution subject to the progressive 1:2 dilution is exemplified in the panel below.

corresponds to the Babylonian *mina* and the Talmud mentions half-logs and quarter-logs, as well as eighths, sixteenths, and sixty-fourths of a log (a *kortov*). Liquid measures include a *hin* (around 3.67 L), ½ *hin*, 1/3 *hin*, ¼ *hin*.[70]

As for the dilution factor of one to sixty that is reported in the Talmud (only a slight difference, in real terms, from 1/64), it may be recalled that the number 60 is the base of the Babylonian sexagesimal system, the one that is still tenaciously in use to divide time and to measure angles. The Babylonian *maris* has multiples by the factors of 12, 24, 60, 72 (60 + 12), 120, 720, and a sub-multiple, the *shekel*, as 1/60 of a *mina*.

The number 60 has prime factors 2^2 * 3 *5 and 12 integer divisors; it is also the product of the numbers of the fundamental Pythagorean triplet (*i.e.*, 3 * 4 * 5 = 60, and $3^2 + 4^2 = 5^2$). A 3:4:5 Pythagorean triangle has a perimeter of (3+4+5=12) units, and the sides differ from one another by one unit (5-4 = 4-3 = 1). Early Babylonian calculators knew the arithmetic properties that make this device among the most useful for field measurement. It is held that Pythagoras only reported and possibly demonstrated as a theorem what amounted to a long known empirical practice that had found wide application in land allotment and building (Figure 2).[71]

Given the possibility of Hellenistic technology to build finely machined devices, such as the Antikythera astronomical clock,[72] it is as well conceivable that a Pythagorean triplet might be used as reference to manufacture or carve a matching pair of containers, the larger of which exactly contains 60 times as much as the smaller one, as illustrated in Figure 3.

What may be conceived from the specific value of "*sixty*" for the denominator in the Talmudic criteria of *bitul b'shishim* is that, in the early Ptolemaic times, some Alexandrian experimental physicians performed pharmacological activity tests of medicinal preparations through scalar dilution in the geometrical proportion.[65] Some of Herophilus' disciples and followers, such as Mantias and Apollonius Mys, were reportedly pharmacologists,[54] and Galen recognized the former as being the father of the "compound drug" tradition, while the latter is the main source of Galen's *On the Composition of Drugs according to Places* (XII Kuhn). Dioskorides Phakas, allegedly a relative of Cleopatra VII Philopator, is credited as the first to use a color test to detect iron in biological fluids.[15]

At some time, *poskim* (Jewish assessors) who were acquainted with this method started to apply it to the assessment of food according to *halacha*. They came to determine that an eight-fold scalar dilution ($1/2^6 = 1/64$, or » 1,6%) was sufficient to lose the taste of some tasty, yet unacceptable ingredient, such as *chalav* (milk) in *basar* (meat) and vice-versa, and *gid hanasheh* (sciatic nerve in halachically slaughtered calves; *see* above).

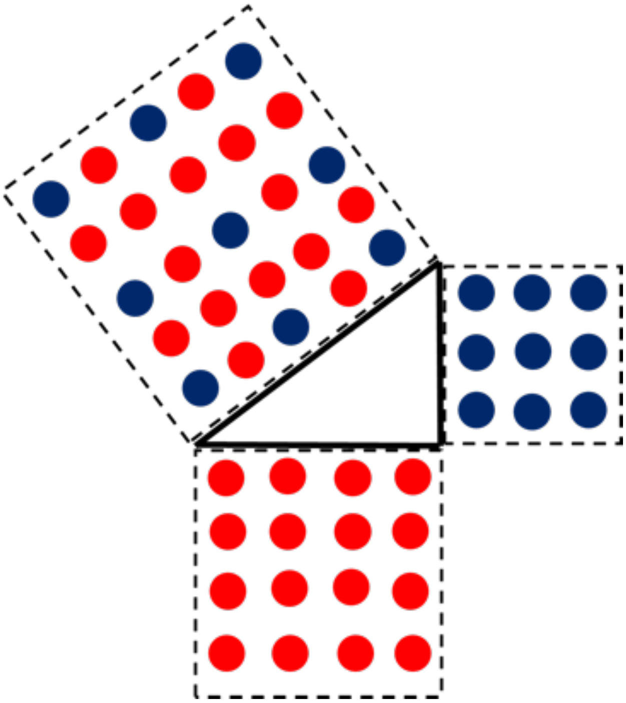


Figure 2. A chessboard that shows the empyrical derivation of the so-called Pytagora's theorem on right triangles and the existence of the Pytagoric triplet 3,4,5.

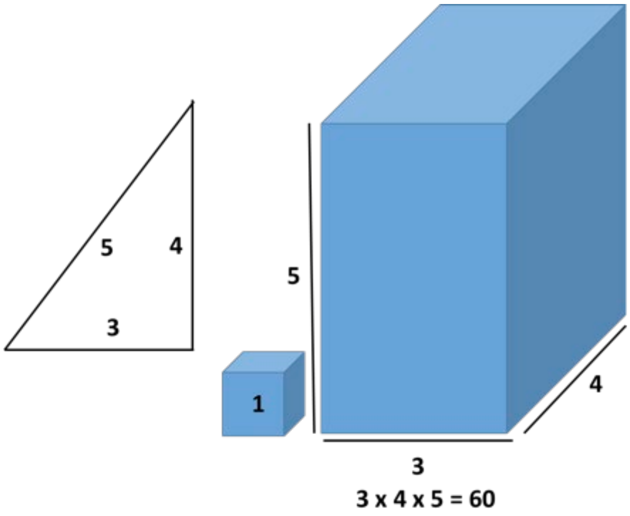


Figure 3. Relationship of two solids (containers), the larger exactly 60 times as much as the smaller, built as a unit-side cube (the smaller, acting as the unit-volume) and as a rectangular prism with sides in the 3:4:5 ratio of the fundamental Pythagorean triplet (the larger).

The detailed intellectual basis of this method might go lost at the time of the *Hellenistic Doom*, or even later, due to the loss of the trans-generational continuity of knowledge transmission in science and technology. What possibly resisted within the halachic tradition was the recollection of the use of a sensorial assessment to test for unacceptable food mixtures (*taam k'ikur*), and an approximate appreciation of a generic acceptance threshold value (*bitul b'shishim*). The one-to-sixty ratio was the closest when the 1:2 geometrical scalar dilution and the Babylonian system of multiples of the *mina* were compared, and so this specific value was consolidated some time during the formation of the Talmud. Such consolidation might occur in Palestine or in Babilonia, due to the exchange of scholars since the most ancient Hillel and Shammai time [73], but only in the *Bavli* there is a *Chullin* treatise concerning food. The Jewish community of Alexandria did not develop an autonomous halachic tradition and referred to Palestine (Niddah 69b) until its dissolution following the 115-117 CE revolt under Trajan. However, they shared with the Palestinian Masters several intellectual skills in the field of textual interpretation[74] [*e.g.* p.66] and in mechanical technology[74] [*e.g.* p.66].

> **Niddah 69b:** Our Rabbis taught: Twelve questions did the Alexandrians address to R. Joshua b. Hananiah. (Note 33) Three were of a scientific nature, (Note 34: Lit., 'the way of the earth', worldly affairs) three were matters of aggada, three were mere nonsense and three were matters of conduct.

This explanation can be strengthened by considering that the Talmud Masters did not always accept as such any of the two possibilities to enforce *kasherut* on mixtures, and in particular highlighted the difference between mild-tasting prohibited ingredients (the *basar b'chalav*) and strongly tasting ones, such as spices (*tavlin*), some of which may be halachically prohibited[34] [*passim*].

6. SENSORIAL ASSESSMENT AND SENSORIAL THEORY: AN IMPLICIT ACCEPTANCE OF ATOMISM?

The criteria of *bitul b'shishim* (*one part in sixty*) and of *taam k'ikar* (*the taste imparts identity to the substance*) represent two coordinate and complementary strategies for halachic food assessment that closely resemble contemporary approaches to regulatory food toxicology. As highlighted above, the first option is more akin to "*assessment by modelling*" in that estimation and calculation is employed to decide the *heter vs. issur* issue. The second option necessarily resorts to "*assessment by measurement*" and employs a biological sensor as the measuring device of a physico-chemical phenomenon: such is, in fact, the action of tasting. That chemical analysis owes to sensorial assays its beginnings is apparent from the etymology itself of the terms employed to describe its operations. *Test*, *saggio* (in Italian), *assay*, all derive from the sensorial assessment performed through the mouth: to taste, *assaggiare*; the wise (il *saggio*) is the man who knows by personal appreciation, who has learnt to distinguish the taste of salt, il *sapore di sale*, from that of other substances.

The very possibility to use taste as an assessment technique, as in the case of food *halacha* by *taam k'ikar* is not only rooted in human common sense, but is implicitly founded on a sensorial theory, and in turn on an underlying theory of matter. It is here that *halacha* shows its likely derivation from previous (*i.e.*, Old Mediterranean) methods of assessment of the strength of preparations with sensorial or biological activity, such as of tasty spices or pharmacologically active preparations. Information in the Talmud may thus shed light on the prevailing theory of matter in Hellenistic time, and possibly on its evolution or transformation over time, down to the Late Middle Age.

It is generally accepted that non-atomistic theories developed by Aristotle and his successor Theophrastus, were largely prevalent, while Democritean atomism did not develop beyond the imaginative level depicted in the Epicurus-Lucretius tradition. The Aristotelian four-element, four-quality theory was fully exploited in several branches of knowledge, in particular as the foundation of the four-humor physiological model that was the mainstay of the Galenic medical thought for the best of the following fifteen centuries.[75,76] In turn, its application to *materia medica* classified remedies, such as spices and herbal medicines, according to their purported qualities (hot-cold, humid-dry) and attempted to assign to each herb the nature of the qualities and a semi-quantitative appreciation of the respective strength, as "grades" in a cardinal scale.[4]

The overall Aristotle-Theophrastus sensorial theory (*de Sensu*; *de Sensibus*)[77,78,79] classifies the five senses into three groups. The first group of senses includes sight and hearing, and, in both cases, no physical contact occurs between the sensing organ and the object from which the stimulus originates. The second group includes touch, for which the interaction of the receptor with the stimulus is of direct physical contact. The third group includes smell and taste, the nature of which is somewhat ambiguous, and the interpretation of which by Aristotle and Theophrastus is uncertain, but likely foreruns the contemporary, physiologically correct receptor-agonist theory and identifies the transporters of the stimuli of taste and smell as material entities that diffuse from the object to the organs of sense [80,81,82]. Although Aristotle and Theophrastus did not adhere to atomism, nevertheless, the sensorial theory of smell and taste as such is only justified as long as the transporters of the stimulus are of a material constitution (and so are the matching receptors).[83,84] Furthermore, a law according to which the strength of the stimulus is directly proportional to the quantity of the transporters of the stimulus holds, and by consequence the weaker the stimulus, the weaker is the presence of the substance.

It is thus conceivable that the logical bases for halachic rules such as *bitul bitushin* stem by a more or less direct intellectual route from familiarity with, and the acceptance of Greek or Hellenistic knowledge by the Masters of the Talmud, in a field that we may now consider at the merge of physical chemistry and physiology. Several *loci* of the Talmud where medical knowledge is instrumental to solving halachic issues show that Greek-Hellenistic medical theory was pragmatically accepted and re-elaborated according to need.[85,86,87]

Another strong clue that the theory of matter implicitly accepted by the Masters of the Talmud is corpuscular and atomistic is the strong attention that the Talmud gives to the regulation of the halachic status of (food) mixtures. This includes the transport of the status of substances to vessels and through vessels to other (food) substances, caused by absorption-desorption phenomena (*ta'am balua*, absorbed taste). In particular, the transmission of taste is fastidiously related to parameters of time (one day, *ben yomo*, or more than one day, *eino ben yomo*), temperature (sparking hot, *libbun gamur*), physical condition and comminution of the interacting substances (*chaticot*, pieces or *lach*, liquid mixture) and the material of the containers (essentially pottery, metal or glass). To get rid of the prohibited *ta'am*, decontamination with either hot water (*hagala*) or with fire (*libbun*) is prescribed, according to the nature of the *issur* and of the container.[34 [*passim*]]

The XVI century CE *Shulkan Aruch* treatise summarized most food-related *halacha* in the *Yoreh De'ah* section (*see* Supplementary), with reference to the original Talmudic source and to the later elaboration of the main Middle Age commentators (according to the Hebrew time scan, the Geonic and Rishonic period). The closest earliest similar observation to that reported in the XVI century *Yoreh De'ah* occurs in Vannoccio Biringuccio's treatise on metals (*De la pirotechnia*, 1540), with just the plain observation that bulk metal takes more time to dissolve than comminuted one. More strikingly, the phenomenological basis and fundamental laws of mass-transfer and chemical kinetics only became apparent in the late XIX century (CE!), when investigation on catalysis in organic chemistry started and the laws of absorption-desorption, enzyme kinetics and dose-response were rationalized.

There is an even more sophisticated issue in the Talmud that may hide a fossil notion of a Hellenistic theory of matter transformation (*i.e.*, of modern chemistry).

Cooking entails the (irreversible) transformation of some components of food into others with different characteristics, such as taste, smell or texture, and halachic status, as well. In particular, a regulation considers that two different (food) substances, each with its own halachic status, can combine to produce another, with an individual, specific halachic status, usually from allowed to forbidden (*chaticha na'aset neveila*) [34, *passim*].

The Avodah Zarah of the Mishnah (2:6) and of the Talmud (37b) contain a regulation, *bishul Yisrael*, according to which Jews can only eat food that is prepared by Jews (or under supervision of a Jew).

> **Avodah Zarah 37b.** [...] A comparison is to be drawn with water — as only water which has undergone no change [is permitted to Jews] so also must the food have undergone no change [at the hand of heathens]. [...] ears of corn should also be prohibited when roasted by them [...] wheat should be prohibited when milled by them [...]

However this law applies only to those foods that, according to the Talmud, are "fit for a king's table" (*oleh al shulchan melachim*) and/or are not usually consumed raw (AZ 38a).

> **Avodah Zarah 38a.** [...] Whatever is eaten raw does not come within [the law of what is prohibited] on account of having been cooked by heathens. [...] Whatever is not brought upon the table of kings to serve as a relish with bread does not come within [the law of what is prohibited] on account of having been cooked by heathens. [...]

It is thus conceivable that its formulation by the Masters of the Talmud reflects their knowledge of, and the agreement on, a theory of matter according to which forms of matter irreversibly become (transform) into a different one (*davar hadash*) by natural or voluntary human means.[88] The same is likely to apply to the Talmud commentators and *halacha* regulators who lived in the West European Middle and early Modern Age, since no substantial theoretical change in the theory of matter occurred until truly atomistic theories developed in the late XVII century CE.[89].

Thus, if this chain of reasoning could be retrieved from the ancient Jewish sources, it would strengthen the possibility that the Hellenistic theory of matter already distinguished physical mixtures (*myxis*) from the product(s) of chemical reactions (*krasis*) that occur between mixture components.[20 [pp. 157-167]] Such processes were already common in the antiquity and led to new materials that do not exist in nature (such as soft glass), to dying products (the reversible air-induced reduction-oxidation reaction of indigo and of purple mussel extracts that originates the two famous colors), and to other artificial goods.

As a recent example, the halachic status of mono- and di-glyceride emulsifiers that derive from natural fats, has been assessed by considering their relationship to the starting products. Some contemporary Jewish religious authorities in the USA decided that the meaning of “*fit for a king's table*” is that the product should stay edible throughout the process that transforms a raw, inedible (or halachically forbidden) food into the finite product. Since one of the preparation steps of glyceride emulsifiers entails the formation of an inedible, even caustic, mixture of fat with strong acid, this event breaks the edibility chain, since now the concoction is no more *oleh al shulchan melachim* (or for anyone's mouth, really).[90]

7. CONCLUSIONS, AND A ROADMAP FOR FURTHER STUDY.

The several, however far from exhaustive, nuggets of Talmudic information reported here suggest that the elaboration and incorporation of Greek, Hellenistic and later knowledge into the Hebrew Talmud occurred very early, and continued through the centuries. Such cultural event was likely occasioned by the proximity of Jewish scholars, who also were the earliest Talmud Masters, to Gentiles who practiced Hellenistic science and technology, especially in large cosmopolitan cities. The cultural melting pot of Alexandria is one likely candidate, yet in several other areas of the Hellenistic and Roman world, and in Parthian Babylonia, occasions of interaction between learned individuals may have played a similar role. Such interaction, direct or mediated as it might develop, would induce Hellenistic advances in science and technology into the halachic discussion of the time, and this embedded knowledge survived the *Hellenistic Doom* and was preserved as a component of Talmudic knowledge even when its Hellenistic roots had been severed.

This cultural phenomenon may not be unique in the history of Western Judaism. As Ptolemaic Alexandria might be the cradle of the Hellenistic-Judaic interaction in the III to II century BCE, as well Islamo-Judaic *al-Andalus-Sepharad* visited by Christian post-docs in the X-XII centuries CE was where quantitative studies of the pharmacology of simple and mixed *remedia*[65,67] could be re-appraised through the inherently quantitative approach of *halacha*. The intellectual pathway towards quantitative pharmacology traced by the al-Kindi – Gerard of Cremona – Arnald Villanova – Jordan de Turre connection of the Montpellier medical school[66] in fact developed at the same time of the flourishing Catalan-Provençal Geonim. Even the converted Jew who took part in the Gospel *vs.* Talmud polemic debates at least until the Paris Talmud fires of 1240 CE might play a yet unconsidered role in highlighting to the Christian scholarly world some unsuspected sides of Talmudic thought.[91,92]

Later, throughout the Humanism, Renaissance and until the Counter-Reformation, there was a surge of interest among educated Gentiles on studying Hebrew to meddle into their texts, such as the Hebrew source of the Septuaginta and the *Kabbala,*[93] as a source of *prisca sapientia*. In addition, “court Jews” were among the first to get involved in chemical manufacturing of high-tech commodities, such as dyes, drugs, gunpowder.[94] Due to the perceived complexity of the operations, chemical manufacturing was long known as “practical alchemy” or “white witchcraft”, and a then current (and still lasting) prejudice[23,40,93,95] considered Jews in general as particularly suited, for the good and for the evil, in the “esoteric” science of transformation of matters.

In all cases, geopolitical events beyond the pale of individuals would close the short “windows of opportunity”, and in general, the Jewish intellectual world contributed much less than deserved to the development of Western culture at large. The recently started preparation of a version of the Talmud in Italian[69] is expected to facilitate data mining of this huge text by means of Information Technology, independent from halachic studies and from knowledge of the specific Hebrew language. Furthermore, availability of an increasing volume of related information through computer-aided translation will prompt investigations on the possible (and even likely) transmission of Ancient Mediterranean knowledge in the Hebrew Talmud, and through this text, its comments and other Hebrew sources into that of the reawakening Western Europe: *if not now, when?*

REFERENCES

1. S. C. Rasmussen (Ed), *Chemical Technology in Antiquity. ACS Symposium Series Vol. 1211* **2015**, American Chemical Society.

2. M. Levey, *Osiris* **1956**, 12: 376-389.
3. L. M. Principe, *Ambix* **1987**, 34: 21-30.
4. L. M. Principe, *The Secrets of Alchemy*. University of Chicago Press, 2013.
5. H. Fors, L. M. Principe, H. O. Sibum, *Ambix* **2016** May; 63(2): 85-97.
6. L. M. Principe, *Ambix* **2016** May; 63(2): 118-44.
7. H. Robertson, *Ambix* **2016** May; 63(2): 145-61.
8. Galen, Kuhn Ed. vol XII, XIII, XIV at: https: //archive.org/details/hapantaoperaomni13galeuoft; an English translation of *Theriaca ad Pisonem* (Kuhn, XIV) is: Robert Adam Leigh, On Theriac to Piso, Attributed to Galen. PhD thesis, University of Exeter, 2013: https: //ore.exeter.ac.uk/repository/bitstream/handle/10871/13641/LeighR.pdf?sequence=1; (access 2017).
9. F. Téreygeol, N. Thomas, *Revue d'Archéométrie* **2003** 27, 171-181.
10. B. Vannoccio, Tr. Eng. *The Pirotechnia of Vannoccio Biringuccio* / translated from the Italian with an introduction and notes by Cyril Stanley Smith & Martha Teach Gnudi. **1942** The American institute of mining and metallurgical engineers, New York.
11. M. Ciaraldi, *Veget Hist Archeobot*. **2000** 9, 91-98.
12. M. R. Guasch-Jane, M. Ibern-Gomez, C. Andres-Lacueva, O. Jauregui, R. M. Lamuela-Raventos, *Anal. Chem*. **2004** 76: 1672-1677.
13. D. Namdar, A. Gilboa, R. Neumann, I. Finkelstein, S. Weiner, *Mediterranean Archaeology and Archaeometry* **2013** (12)3: 1-19.
14. M. Nierenstein, *Isis* **1931** 16,2 (Nov.), 439-446.
15. M. Nierenstein, *Bristol Med Chir J (1883)* **1946** Summer; 63(226): 75-81.
16. W. Fresenius, *Fresenius J Anal Chem* **2001** 71: 1041-1042)
17. R. Murray, *Anal. Chem*. **2011** 83(23): 8825.
18. L. Russo. *La Rivoluzione dimenticata*. 2° ed. **2013** Feltrinelli, Milano,; chapp. 4, 9
19. L. Russo. *L'America dimenticata*. **2013** Mondadori, Milano,
20. L. Russo. *Stelle, atomi e velieri. Percorsi di storia della scienza*. **2015** Mondadori Scuola, Milano,
21. P. M. Fraser. *Ptolemaic Alexandria*. 3 vols. **1972** New York: Oxford University Press.. Ch. xx
22. L. Russo. *La Rivoluzione dimenticata*. 2° ed. **2013** Feltrinelli, Milano; chap. 10.8
23. R. Patai. *Jewish Alchemists: A History and Source Book*. **1995** Princeton University Press.
24. L. M. Principe, **2013** *cit.* chap. 1, 2
25. Anon. (III-VI sec. CE) *Talmud Bavli*. English translation (only Mishnah and Gemara): www. https: //ia800501.us.archive.org/1/items/TheBabylonianTalmudcompleteSoncinoEnglishTranslation/The-Babylonian-Talmud-Complete-Soncino-English-Translation.pdf (last accession 2017); some treatises also at: http: //www.come-and-hear.com/talmud/ (last accession 2017)
26. A. Steinsaltz, *The Essential Talmud*. **2006** Basic Books (New York, USA).
27. H. Freedman, *The Talmud: A Biography*. Tr.It. Storia del Talmud. **2016** Bollati Boringhieri.
28. A. Steinsaltz, *Isis* **1977** 68(1): 104-105.
29. N.L. Rabinovitch, G. Nàdor, *Isis* **1976** 67: 103-105; *ibid.* **1977** 68: 104.
30. M. E. Singer, *J. Halacha Contemp*. Soc. 42, Sukkot 2001.
31. M. Hasofer, *Biometrika* **1967** 54(1/2): 316-321.
32. N.L. Rabinovitch. *Biometrika* **1969** 56(2): 437-441.
33. Caro I. (XVI sec. C.E.) *Shulkan Aruch (Yoreh Deah)*. Tr. Eng. (partial) in: https: //en.wikisource.org/wiki/Translation: Shulchan_Aruch/Yoreh_Deah; *see* also http: //www.yonanewman.org/kizzur/ (last accession 2017).
34. D. Brodsky (Rav). *The Laws of Kosherut*. **2017** At: http: //etzion.org.il/en/topics/laws-kashrut (access 2017).
35. J. M. Sasson, *in:* U. Hübner, E. Axel Knauf (Eds.) *Kein Land für sich Allein: Studien zum Kulturkontakt in Kanaan, Israel/Palestina und Ebirnâri für Manfred Weippert zum 65 Geburtstag*, **2002** Orbis Biblicus et Orientalis 186 Freiburg Universitätsverlag 294-308.
36. V.A. Ryazanov, *Arch Environ Health*. **1962** 5: 480-94.
37. S. J. Walker, *Permissible Dose: A History of Radiation Protection in the Twentieth Century*. 1st Ed., **2000** University of California Press. Review: A. Colombi, F.M. Rubino, *Med. Lav. (Milano)* **2003** 94(3): 334-336.
38. V. Villavecchia, *Trattato di Chimica Analitica Applicata*. 2°ed., **1922** vol. II. U. Hoepli Ed. (Milano).
39. Mayor A. *The Poison King. The life and legend of Mithridates, Rome's deadliest Enemy.* **2009** Princeton University Press. Tr. it. *Il Re Veleno*, **2010** Einaudi, Torino, and references therein.
40. V. Tcherikover, *Hellenistic Civilization and the Jews* (translated by S. Applebaum) **1959** Jewish Publication Society of America.
41. S. T. Katz (Ed.), *The Cambridge History of Judaism, Volume 4. The Late Roman-Rabbinic Period* (**2008**).
42. D. Boyarin, in: *The Cambridge companion to the Talmud and rabbinic literature*. **2007** Charlotte Elisheva Fonrobert, Martin S. Jaffee (Eds.), Cambridge University Press, 336-364.
43. S. Lieberman, *Greek in Jewish Palestine*. (5702-**1942**) New York, The Jewish Theological Seminary of

America.

44. J. R. Labendz, *Hebrew Union College Annual*, **2003** 74: 175-21
45. A. Keimpe (Ed.), *The Cambridge History of Hellenistic Philosophy* **2008** Cambridge Histories Online. Cambridge University Press.
46. M. Rostovtsev, *The American Historical Review* **1921** 26(2): 203-224; p.221.
47. R. Buitenwerf, *Book III of the Sibylline oracles and its social setting*. With an Introduction, Translation, and Commentary. **2003** BRILL, Leiden 304-320 (https: //books.google.it/books?isbn=9004128611).
48. J. Scarborough, *Studies on Ancient Medicine* **2008** 138-156.
49. L. M. V. Totelin, *Early Science and Medicine* **2004** 9: 1-19.
50. S. Norton, *Molecular Interventions* **2006** 6: 60-66.
51. L. Totelin, *Phoenix* **2012** 66(1-2) 122-144.
52. P. Schneider, *Phoenix* **2012** 66(3-4) 272-297.
53. A. J. Marshall, Phoenix 29(2): 139-154 (1975)
54. H. von Staden, *Herophilus: The Art of Medicine in Early Alexandria* **1989** Cambridge.
55. H. von Staden, in: *Alexandria and Alexandrianism*. Kenneth Hamma (Ed.) The J. Paul Getty Museum 85-106 (1996)
56. L. Russo, *La Rivoluzione dimenticata*. 2° ed. **2013** Feltrinelli, Milano; chap. 5.
57. Quintus Septimius Florens Tertullianus, (155-230 CE) http: //www.tertullian.org/anf/anf03/anf03-22.htm (access: 2017)
58. J. Ganz, *Istoriâ mediciny* **2014** 4(4): 5-12.
59. H. P. Bayon, *Proc R Soc Med.* **1939** 32(11): 1527-1538.
60. P. Ghalioungui, S. Khalil, A. R. Ammar, *Med Hist* **1963** 7: 241-46.
61. J. M. Stevens, *Med J Aust*. **1975** 2(25-26): 949-52.
62. L. Smith, *J Fam Plann Reprod Health Care* **2011** 37(1): 54-5.
63. R. Haimov-Kochman , Y. Sciaky-Tamir, A. Hurwitz, *Eur J Obstet Gynecol Reprod Biol*. **2005** 123(1): 3-8.
64. http: //gifh.wordpress.com/2012/02/19/la-travagliata-storia-del-test-digravidanza/; ; http: //www.laprovinciadicomo.it/stories/cultura-e-spettacoli/292109_test_di_gravidanza_egizio_funziona_dopo_4_mila_anni/ (last access 2017)
65. M. R. McVaugh, *Bull Hist Medicine* **2017** 91(2): 183-209.
66. J. Ricordel, *Revue d'histoire de la pharmacie*, **2001** 89^{e} année, n°330, 135-148.
67. Arnaldus de Villanova. *Aphorismi de gradibus*. Edidit et praefatione et commentariis anglicis instruxit Michael R. McVaugh. **1975** Granada; Barcelona.
68. C. Sinding, *Bull Hist Med*. **2002** 76(2): 231-70.
69. R.S. Di Segni (Ed.) *Talmud Babilonese - Trattato Rosh haShanà*. **2016** Giuntina, Milano.
70. *Encyclopedia Judaica. Units of measure*. At: http: //www.torahcalc.com/info/biblical-units/; https: //www.sil.org/system/files/reapdata/.../silewp2014_003.pd; http: //www.ajdler.com/jjajdler/Talmudic3.pdf; *also:* http: //www.jewishencyclopedia.com/articles/14821-weights-and-measures; www.dafyomi.co.il/general/info/units-of-measurement.pdf; dafyomi.shemayisrael.co.il/general/.../units-of-measurement.pdf (last access 2017)
71. L. Russo. *La Rivoluzione dimenticata*. 2° ed. **2013** Feltrinelli, Milano, ch. 2, 52-53.
72. *Antikythera Mechanism Research Project* at: https: //www.antikythera-mechanism.gr/ (Access 2017)
73. Y. Z. Eliav, in: *The Archaeology and Material Culture of the Babylonian Talmud*. **2015** (Markham J Geller, Ed.), Brill, Leiden 186-224.
74. S. Lieberman. *Hellenism in Jewish Palestine*. (5722-**1962**) New York, The Jewish Theological Seminary of America; *passim*; p.66; p.94.
75. E. Nicholas. *The Alphabet of Galen: Pharmacy from Antiquity to the Middle Ages: A Critical Edition of the Latin Text*. **2012** Univ. of Toronto Press.
76. P. Holmes, *J Amer Herbalist Guild* **2002** Spring/Summer: 7-18.
77. Aristotle. *De Sensu et sensibilibus*. Tr. It. A. L. Carbone in: *L'anima e il corpo, Parva Naturalia*. II Ed. **2015** Bompiani (Milano).
78. Aristotle. *De anima*. Tr. It. G. Movia in: *L'anima*. VII Ed. **2015** Bompiani (Milano).
79. Theophrastus of Erebus. *De Sensibus*. Tr. It. L. Torraca in: *I Dossografi Greci*. **1961** CEDAM (Padova).
80. R. Sorabji, *The Philosophical Review* **1971** 80(1): 55-79.
81. T. K. Johansen, *Phronesis* **1996** XLI/1, 1-19.
82. J. Ellis, *Phronesis* **1990** XXXV/3, 290-302.
83. J. Mansfeld, *Phronesis* **1996** XLI/2, 158-188.
84. B. Stansfield Eastwood, *Rheinisches Museum für Philologie, Neue Folge*, 124. Bd., H. 3/4 268-290 (1981)
85. S. T. Newmyer, *Aufstieg und Niedergang der römischen Welt II* **1996** 17(3): 2895-2911.
86. S. S. Kottek, *Aufstieg und Niedergang der römischen Welt II* **1996** 17(3): 2912-2932.
87. F. Rosner, *Aufstieg und Niedergang der römischen Welt II* **1996** 37(3): 2866-2887.
88. K. Abelson, M. Rabinowitz (Rabbis), *Proceedings of the Committee on Jewish Law and Standards* I **1980-1985** 187-190. Rabbinical Assembly, **1988** http: //www.rabbinicalassembly.org/yoreh-deah (last access 2017)

89. B. T. Moran, *Distilling Knowledge: Alchemy, Chemistry, and the Scientific Revolution*. **2005** Harvard University Press.
90. K. Abelson (Rabbi), *Proceedings of the Committee on Jewish Law and Standards* I **1980-1985** 181-185. Rabbinical Assembly, **1988** http: //www.rabbinicalassembly.org/yoreh-deah (last access 2017)
91. A. Fidora, *J. of Transcultural Medieval Studies* **2014** 1(2): 337-342.
92. A. Fidora, *J. of Transcultural Medieval Studies* **2014** 2(1): 63-78.
93. G. Scholem. *Kabbala e Alchimia* [tr.it.] **2015** SE.
94. D. Jutte, *Isis* **2012** 103: 668-686.
95. G. Ferrario, in: *Chymia: Science and Nature in Medieval and Early Modern Europa*. **2010** M. Lopez Perez, D. Kahn, M. Rey Bueno (Eds.) Cambridge Scholars Publishing, 19-29.

Supplementary Materials

Shulchan Aruch Part II: Yoreh De’ah (http://www.torah.org/advanced/shulchan-aruch/archives.html)

	Loci
Yoreh De’ah Chapter 7a - ABSORPTION	
If forbidden food (or an object that has absorbed forbidden food within the past 24 hours) is in contact even momentarily with hot liquid in a utensil that has been on a fire, or with salty liquid for 18 minutes, or with any liquid for 24 hours, permissible food that was in contact with the liquid for that period of time becomes forbidden unless the forbidden components are less than 1/60 of the total.	(69:1,9,11,15,18;70:6;98:4; 104:1-2;105:1-3)
In estimating 1/60 only food below the surface of the liquid is considered	(see 94:1; 98:4; 99:1,4; 105:1).
If the utensil has not been on a fire, but a hot component comes from such a utensil	(see 68:10-11, 13,15),
the surface of the permitted food must be peeled off where it comes in contact with the liquid	(105:3).
Even if the components are hot, if they do not come from utensils that have been on a fire and the liquid is not salty	(see 69:9 and 91:7)
and the permitted food remains in it for less than 24 hours, the food need only be washed off and the liquid is permitted	(91:1-4;105:2-3).
These and the following rules also apply to contact between milk and meat products	(87:10;91:4-6; see 92:1,4-6).
On absorption from a forbidden egg that is still in its shell	see 86:5-6.
If the components are near a fire and hot but not in contact with liquid, and an object that has absorbed forbidden food touches permitted food, it becomes forbidden unless the forbidden components are less than 1/60 of the total; and the same is true if forbidden food of a type that sometimes contains fat touches permitted food, even if the forbidden food is absorbed in other food; but if the forbidden food is of a type that never contains fat, the places where it touched the permitted food need only be removed to a depth of a fingerbreadth, and this must be done in any case if the places where it touched are known	(see 68:4,9;105:4-5,7-8).
If the components are not near a fire and a hot component that has been on a fire is added, the places where the permitted food touches a forbidden component need only be peeled	(68:10-11,15;92:7;94:8;105:6).
If no component has been on a fire only washing is needed, and nothing need be done if the components are dry	(91:1-4).
If the components are not near a fire, but a forbidden food component is heavily salted	(see 69:8;70:6)
and not entirely dry	(see 91:5;95:7),
the places where a permitted component touches it even momentarily must be peeled;	(see 69:8,18;70:6;105:1)
and if the forbidden food is of a type that sometimes contains fat, the permitted food that touches it becomes forbidden unless the forbidden components are less than 1/60 of the total, and it must also be peeled if the places where it touched are known	(64:16,20; 65:1; 69:9,16,18,20;70:1-4; 72:2;105:9-11).
Salt causes absorption even in an object	(see 69:16-17),
but it does not draw out forbidden food that has been absorbed in an object	(69:16;70:2;105:12-13).
If the forbidden food is meat from which the salt is drawing blood, and the salt is also drawing juices out of the permitted food, or the permitted food is on top, it does not absorb the blood and need only be washed	(70:1-4);
or if the permitted food is meat that still contains blood, any blood that it absorbs in this way can be extracted together with its own blood, though other blood that it absorbs cannot	(69:2;70:2,6;72:2).
Similarly, the blood that a fire draws from meat is not absorbed in other meat that is near the fire, but other blood is	(69:4,20;76:1-2;77:1).
Yoreh De’ah Chapter 7b - ABSORPTION cont’d	
Some substances absorb more easily or less easily than others; for examples	see 64:18-19;96:5;121:1.
Pressure (as in cutting with a knife or grinding in a mortar) increases the depth of absorption;	see 94:7 and 96:1-3 as well as 10:1-3;64:16;89:4.
Even in cases where the forbidden component is less than 1/60 of the total, if it can be recognized or separated it must be removed; and if it is attached to or first entered a permitted component, that component is forbidden and must be removed if it can be recognized	(69:11;72:2-3; 73:6;90:1;92:2-4;94:3;98:4; 106:1-2).
When a permitted food component becomes forbidden because of thorough mixing (see Ch.8a) or absorption it is regarded as entirely forbidden even if it absorbed an amount smaller than its volume	(92:4;98:5;99:3,5;106:1;107:2),
but if it absorbs meat or milk it is not regarded as being entirely meat or milk	(94:6).
If an object absorbs an unknown amount of forbidden food it is regarded as entirely forbidden	(see 94:2)
unless the absorption was of a type that requires only peeling	(98:4).

	Loci
If it absorbs a known amount of forbidden food it is not regarded as entirely forbidden unless it is made of pottery or it has also absorbed an unknown amount of permitted food	(98:5; see 92:5-7 and 94:6).
Permitted and forbidden foods should not be heated together in an enclosed space (such as an oven) unless one of them is covered or both of them are in containers and the oven is not completely enclosed, but if this was done the food remains permitted if the oven is not completely enclosed unless one of the foods has a sharp taste or unless a mixture containing even a tiny quantity of the forbidden food would be forbidden	(90:2;97:3;108:1-2; see Ch.8a).
Some foods absorb odors even if the source is covered;	see 108:4.
Similar laws apply to heating them one after the other if the first one causes steam to form in the oven	(108:1; see also 92:7-8;93:1; 105:3).
Tasting forbidden foods even without swallowing them is forbidden	(108:5),
but smelling them is not forbidden unless it is forbidden to derive benefit from them	(108:7).
An object that was in contact only with cold, unsalted forbidden food can be cleaned by thorough washing	(121:1),
but if it has absorbed forbidden food it should not be used even with cold, unsalted permitted food even after it has been washed unless it is earthenware	(see 69:16; 94:7;121:5).
If an object made of metal, wood or stone absorbed forbidden food in the presence of hot liquid, the absorbed food can be removed from it by immersing it in boiling water at least 24 hours after the food was absorbed in it	(91:5;108:3;121:2).
If the absorption was in the presence of heavy salt or of hot liquid that is no longer in a utensil that has been on the fire, it is necessary only to scrape off the object's surface where the food or liquid touched it;	see 92:9
. If it absorbed forbidden food by heating in the absence of liquid (this includes frying) the absorbed food can be burnt out of it by heating it to a high temperature	(97:2;121:4-6).
If it is a knife it may be used with cold food after thoroughly cleaning or grinding it down; to use it with hot food it must be heated to a high temperature or ground down and immersed in boiling water	(see 10:1-3; 64:17;69:20; 89:4;94:7;121:7).
These laws are also treated in Orach Chayim 55:1-2; see 121:3. On the procedures for cleaning utensils that were used with forbidden wine see Ch.10b.	Orach Chayim 55:1-2; see 121:3
Yoreh De'ah Chapter 8a - MIXTURES OF FOOD cont'd	
If forbidden and permitted foods are mixed together thoroughly the mixture is permitted if no one forbidden component is more than 1/60 of the total	(98:1,6,9; see 99:1-2,4).
In defining a component, things that have the same name are regarded as the same whether or not they taste the same;	see 98:2.
For some types of forbidden foods amounts different from 1/60 are required; for other types any amount makes a mixture forbidden	(see 98:7-8).
If an intrinsically forbidden component can be detected by its taste or by its effect on the mixture	(e.g., 87:11; 102:1),
or if a forbidden component can be recognized but cannot be removed	(104:1,3),
the mixture is forbidden even if the component is less than 1/60	(98:8; 105:14).
It is forbidden to mix forbidden food with permitted food to produce a permitted mixture; if this was done, the person who did it or for whom it was done is forbidden to derive benefit from the result	(94:5-6;101:6).
It is forbidden to use a utensil that has absorbed forbidden food if the utensil is sometimes used for less than 60 times as much permitted food	(99:7;122:5).
If a mixture contains less than 1/60 of a forbidden component, and more of that component is added so that the total reaches 1/60, the mixture becomes forbidden; but if a mixture contains less than 1/60 of meat (or milk) it does not become forbidden even if milk (or meat) is added to it afterwards	(99:6).
If an entire (dead) creature or (named) body part that has always been forbidden is mixed with any amount of permitted food the mixture is forbidden, but if the forbidden component can be recognized and removed the remaining mixture is permitted if the forbidden component was less than 1/60 of it	(100:1-3).
Similarly, if a portion of food that is intrinsically forbidden and is large enough to serve to guests in its present condition is mixed with any amount of permitted food, the mixture is forbidden as long as the portion may have remained intact	(69:14;81:2; 92:3; 101:1-7; 105:9;106:1).
If food that is only temporarily forbidden or that can be made permitted without much effort is mixed with any amount of permitted food of the same type, the mixture is forbidden until the forbidden component becomes permitted; but if it is mixed with permitted food of a different type, or is not intrinsically forbidden, or became forbidden only after it was mixed, or can be recognized and removed, the mixture is permitted if the forbidden component is less than 1/60 of the total	(102:1-4).

	Loci
Yoreh De’ah Chapter 8b - MIXTURES OF FOOD cont’d	
If forbidden food is tasteless or gives a mixture a permanent bad taste (or if it is a creature: itself has a bad taste) it does not make the mixture forbidden unless it is the majority ingredient, but it should still be removed from the mixture if possible;	see 81:8;95:4;100:2;103:1-4; 104:1-3;107:2;122:1;123:25.
Food absorbed in an object loses its taste after 24 hours and no prohibition results if it is reabsorbed in other food afterwards	(93:1;94:4;95:2;103:5,7;122:4,6-7),
but food adhering to the surface of an object does not lose its taste, and in any case if an object has absorbed forbidden food it should not be used with permitted food even after 24 hours until the absorbed food is removed from it	(122:2-3; see Ch.7b).
Milk or meat absorbed in an object and reabsorbed in meat or milk within 24 hours results in a prohibition, but if it is first reabsorbed in something else it becomes a “second-order” taste and can no longer result in a prohibition	(94:5,9;95:1-3).
In strong-tasting food, even absorbed tastes that are 24 hours old or second-order are not permitted	(95:2;96:1-5;103:6;122:3).
Precautions should be taken to avoid the possibility of forbidden and permitted things becoming mixed up;	see 101:8-9;110:10;123:23.
If a piece of forbidden food that is not large enough to serve to guests becomes mixed up with two or more pieces of permitted food of the same type the pieces are all regarded as permitted, but one person should not eat all of them	(109:1, and see Ch.9).
If they were cooked together the result is forbidden unless the forbidden food is less than 1/60 of the total of the food that is in doubt	(109:2, and see 111:7;
a person is allowed to add permitted food to the mixture before cooking it to ensure that the forbidden portion is less than 1/60).	
If a piece of forbidden food becomes mixed up with pieces of permitted food of a different type and cannot be distinguished, it is not regarded as permitted unless it is less than 1/60 of the mixture	(109:1);
but if it is more than 1/60 the mixture is not regarded as entirely forbidden, and if more permitted food is added to it until the forbidden portion becomes less than 1/60 the mixture becomes permitted	(92:4).
If an object that has absorbed forbidden food becomes mixed up with other objects they are all permitted	(102:3; 122:8).
If an “important” forbidden thing (for example, a living creature or anything that is counted rather than measured) becomes mixed up with any number of permitted things of the same kind the mixture is forbidden	(e.g. 86:3),
but if the things lose their importance (for example, the living creatures are slaughtered and are not large enough to serve to guests) and this was not done deliberately the mixture becomes permitted	(110:1-2),
and if one of the things is accidentally destroyed the others become permitted because we assume that the forbidden one was destroyed (this also applies to creatures, large portions, and things that are only temporarily forbidden), but they should be eaten two at a time by more than one person and they should not all be eaten at once	(110:7).
See Ch.11 on the case where the things are forbidden because of idolatry.	

Substantia
An International Journal of the History of Chemistry

Firenze University Press
www.fupress.com/substantia

OPEN ACCESS

Citation: F. Bagnoli and R. Livi (2018) Michael Faraday: a virtuous life dedicated to science. *Substantia* 2(1): 121-134. doi: 10.13128/substantia-45

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.

Historical Article

Michael Faraday: a virtuous life dedicated to science

Franco Bagnoli and Roberto Livi

Department of Physics and Astronomy and CSDC, University of Florence
E-mail: franco.bagnoli@unifi.it

Abstract. We review the main aspects of the life of Michael Faraday and some of his main scientific discoveries. Although these aspects are well known and covered in many extensive treatises, we try to illustrate in a concise way the two main “wonders” of Faraday’s life: that the son of a poor blacksmith in the Victorian age was able to become the director the Royal Institution and member of the Royal Society, still keeping a honest and “virtuous” moral conduct, and that Faraday’s approach to many topics, but mainly to electrochemistry and electrodynamics, has paved the way to the modern (atomistic and field-based) view of physics, only relying on experiments and intuition. We included many excerpts from Faraday’s letters and laboratory notes in order to let the readers have a direct contact with this scientist.

Keywords. Faraday, history of science, biography.

> *I have far more confidence in the one man who works mentally and bodily at a matter than in the six who merely talk about it [1]*

LIFE

Michael Faraday was born in 1791 in the village of Newington, now part of the urban area of London. His father worked as a blacksmiths and his family was able to allow him but the most basic education. At the age of 14 he had to find a job to ease the burden on his family.

He was hired in a London bookshop, run by Mr. George Riebau (Figure 1). Here he had to carry out the activities of an apprentice bookbinder and seller of books and newspapers for a period of seven years.

This work enabled Michael Faraday to read many books, that passed through the bookshop of Mr. Riebau. In particular, as he wrote remembering his early experiences, he was greatly impressed by reading the book “Conversations in Chemistry” by Prof. Marcet and the treatise on electricity that appeared in the *Encyclopaedia Britannica* [3].

These readings inspired him the idea of simple electrochemical experiments, building elementary electric generators (batteries). These simple

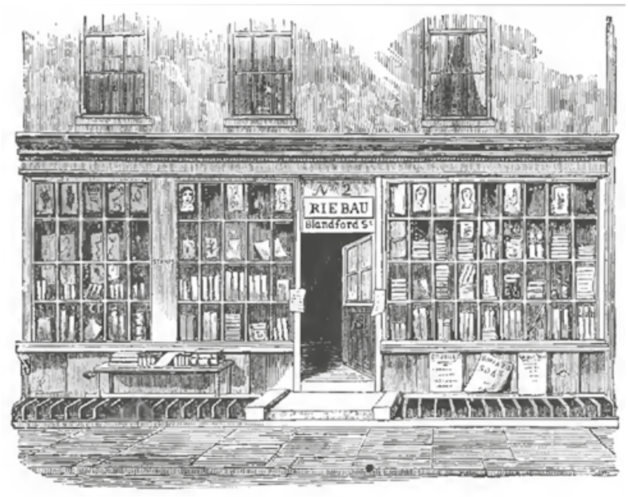


Figure 1. The bookshop of Mr. Riebau in London [2].

experiments were carried out by Faraday almost simultaneously with those conducted in Italy by Alessandro Volta and Luigi Galvani, who are recognized as the first discoverers of the electrodynamic effects (the Voltaic pile and experiments on animal electricity, respectively). In particular Faraday, at the age of 21, managed to discover the principles of electrolysis, which will be the basis of the great electrochemical developments in the nineteenth century.

Among the visitors of the library of Mr. Riebau there was one of the greatest British scientists of that time, Sir. Humphry Davy, the director of the Royal Institution of Great Britain (Figure 2).

This British scientific institution was founded in 1799 after the initiative of Lord Rumford (Benjamin Thomson, industrial, scientist and adventurer) with the aim of improving the technical skills of English scientists, still too much limited by the traditional academic-oriented training provided by the main British universities such as Cambridge and Oxford. Lord Rumford was deeply concerned by the gap accumulated by the English science with respect to the French one, in the decades between the Revolution and the rise to power of Napoleon. The French scientific institutions experienced a deep cultural revolution, no less impetuous than the political revolution, leading to a synthesis of modern science and technology. In order to find a comparable event in the western history, one has to go back to the Hellenistic period, when a huge number of scholars, among which one can mention Archimedes, Anaximander and Ptolemy, built the foundations of our scientific knowledge.

Figure 2. Sir Humphry Davy [4].

The institutional and the social importance attributed to the science in the post-revolutionary French society is definitely a factor of absolute modernity, that influenced, through the spirit of admiration and imitation, the entire European continent and also the Great Britain, mainly for a justified fear of its principal political and military competitor.

The defeat of Napoleon and the subsequent restoration contributed to lowering the tension. The directorship of the Royal Institution passed to H. Davy, who modified the ambitious program of that institution, which since the very beginning had been devoted to the training of high quality technicians (corresponding to nowadays engineers). One of his goals was to present science as an object of amusement for the educated public, which he amazed with spectacular experiments prepared in his laboratory.

This notwithstanding, Davy remained a great scientist and, once paid the fee of being a brilliant disseminator appreciated in all the cultural clubs of the capital, he devoted his time with rigor and continuity to the activities of a true researcher and teacher. Faraday, during his apprenticeship, was admitted to attend the lessons by Davy. Actually Mr. Riebau recommended him to the great British chemist, who was impressed by Faraday's ability as a student, not only in preparing the notes of his lectures, but also in decorating them with useful illustrations and comments.

Obviously Faraday, who was a skillful bookbinder, bound the notes in a very elegant fashion, and offered them as a present to Davy. All these facts convinced

Figure 3. Michael Faraday in his thirties [5].

Davy that Faraday was such a brilliant young man so fond of science that he hired him, at the end of his apprenticeship with Mr. Riebau, in the Royal Institution. One of the members of the Advisory Board suggested Davy to propose Faraday being hired as a cleaner of the laboratory glassware. Faraday accepted the offer, and this was the beginning of his scientific career in the Royal Institution, in 1813. There he spent the remaining 45 years of his scientific life, first as a chemical assistant of Davy, then as his first collaborator and finally as his successor at the head of this institution.

We want to point out that only the extraordinary scientific achievements in his career enabled Faraday to reach career milestones and a social position otherwise unattainable for the son of a poor blacksmith, moreover member of the Sandemanian religious confraternity, which did not recognize the religious authority of the Anglican church (Sandemanians considered themselves as a part of the Reformed Church of Scotland). Just to understand the uniqueness of Faraday's life, it should be noted that only Anglicans were allowed access to the great British universities, such as Oxford and Cambridge, and any academic member of these universities was necessarily required to be also a member of the Anglican Church.[1]

Figure 4. Alessandro Volta presents his battery (Voltaic pile) to Napoleon [6].

The first scientific experience of Faraday outside England came soon in his life. In 1814, he went along with H. Davy and his wife in a trip on the Continent, first in France and then in Switzerland, Italy and finally in Belgium. The entire journey lasted almost one year and a half.

In these years the war between France and England was raging. Nevertheless, Napoleon invited Davy to Paris and honoured him, for his researches on electricity and magnetism, with the scientific prize previously attributed to Alessandro Volta (Figure 4). Napoleon's decision should be understood not only for its political significance, i.e., demonstrating to the world that France acknowledged and honoured the scientific genius even from an enemy, but also for the genuine interest that the French emperor had for scientific discoveries, that would have shaken the future world. Already in 1804, deeply convinced of the social impact of the new scientific discoveries, Napoleon reformed the system of *Grandes Ecoles*, bringing them under the strict control of the state and reorganizing them as military schools.

Napoleon provided Davy and his companions with a special safe-conduct to reach Paris and then move freely throughout Europe, in the areas under the French control. As an illustration of the social rigidity of those times, Faraday had to carry out the role of valet of Davy's spouse, because of his humble origins, despite his mentor was persuaded of the outstanding scientific quality of his young assistant. But Lady Davy could not tolerate that a young son of a blacksmith could participate in their social life, other than with a well-defined subordinate role.

[1] Isaac Newton was against trinitarianism and therefore a sort of heretic. Being a fellow in Cambridge, he was asked to take a vow of celibacy and become a member of the Church of England. While the first requirement was not a burden to him, he considered stopping his studies in order to avoid the ordination. He was finally dispensed with this duty.

Figure 5. Michael Faraday and Sarah Barnard [7].

Figure 6. André-Marie Ampère [8].

The impossibility of finding a valet in France, imposed Faraday to play this role for the entire duration of their journey, although he had previously agreed with Davy to accept it only for the period necessary to reach Paris from London.

This situation contributed to making the relationship between Davy and his protégé very tense. At some moment, during their stay in Paris, Faraday was at the point of going back home. We can conjecture that two main factors made Faraday change his mind. On one hand this European tour with Davy represented an exceptional opportunity for him to come into direct contact with the scientific world of the continent. On the other hand, we should consider that Faraday's strong religious education and his deep respect for the precepts of humility and human solidarity, that animated his confraternity, allowed him to endure the psychological pressure of such a conflictual situation.

This is a side of Faraday's character, which played a key role not only in his private life, but also in his professional activity. In fact, one of the foundations of his beliefs was the tolerance towards his fellows, regardless of their religious confession (many of his closest friends, in fact, did not share his own religious beliefs) and their social status. Moreover, his scientific work was pervaded by his strong conviction that trying to understand the laws of nature was a direct way to approach the knowledge of God's work in creation [9].

His life was marked by a personal commitment in support of poor people, not only through the actions promoted by his confraternity. For Faraday, this was the most natural behaviour, but also an aspect of his life that made no sense to brag about: his privacy, to that effect, was total. Faraday married in 1821 Sarah Barnard, a member of his own confraternity (Figure 5). He expressly recognized that his wife was a crucial support throughout all of his personal and professional life.

Faraday's scientific fame grew over the years, not only at home, due to his amazing discoveries in the field of electromagnetism, that we shall discuss later. Faraday maintained contacts with many of his contemporary scientists. His exchange of letters with the French scientist André-Marie Ampère (Figure 6) on the nature of electromagnetic forces (see the debate between action at

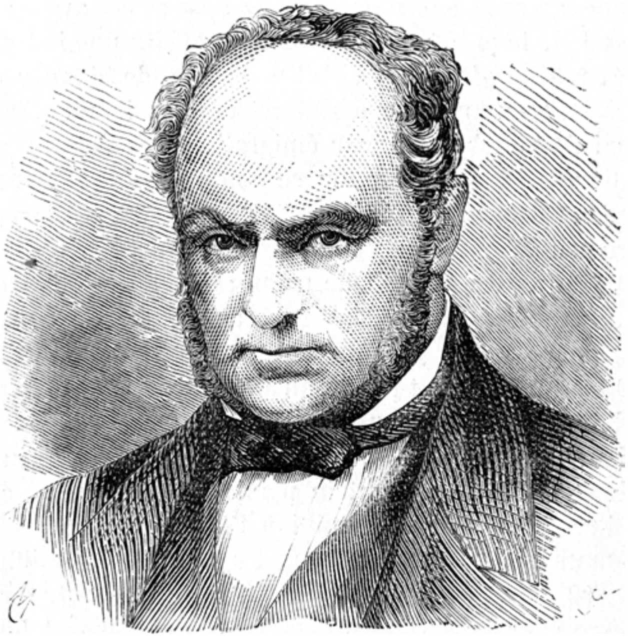


Figure 7. Carlo Matteucci [10].

Figure 8. Ottaviano Fabrizio Mossotti [11].

distance and force fields) was very popular and strongly inspired the forthcoming scientific debate.

Faraday was also in contact with some Italian physicists. As a matter of fact Faraday learnt some Italian to exchange letters with Carlo Matteucci (Figure 7), a physiologist of international fame, well known in the scientific community for his research on electrophysiology. Faraday personally met also Ottaviano Fabrizio Mossotti (Figure 8). In fact, Mossotti had been living for some time in London and developed a theory of dielectrics. His scientific work deeply attracted the attention of Faraday and significantly influenced the work of James Clerk Maxwell, mainly in relation to the discovery of the so-called "displacement current".

Mossotti and Matteucci, after several personal misfortunes also linked to their political and military involvement in the Italian *Risorgimento*, met again at the *Scuola Normale* in Pisa, where they managed to organize one of the first experimental physics laboratories in Italy.

Faraday was a tireless worker and a great science disseminator (Figure 10), with an uncommon skill for physical phenomena. He succeeded, as we shall see, in giving birth to original and almost revolutionary ideas, without any specific training in Mathematics, that would have presumably led him to translate his intuitions and numerous experimental findings in a complete general theory of electromagnetic phenomena.

Fortunately for him and all of us, a young Scottish scientist, James Clerk Maxwell (Figure 9), succeeded in such a task. It should be pointed out that Maxwell explicitly recognized that his fundamental work, on the theory of electromagnetic phenomena, was an almost direct deduction from the titanic experimental framework achieved by Faraday.

It is worth to be noticed that Maxwell, brilliant mathematician and theoretical physicist, was a rather poor lecturer, while the self-made-experimenter Faraday was a worthy heir of Sir Davy (Figure 10).

Because of his tireless activity, in 1839 Faraday had a nervous breakdown, from which he recovered with difficulties, and thanks to his wife's care. He managed to return to his studies on electromagnetism, but he had to refrain from the exhausting working periods he was used to, and slowed down his personal involvement as director of the Royal Institution (Figure 11).

The achievements already attained and those that followed in the second period of his scientific life granted him such a scientific reputation that he was worldwide renown in the middle of the XIX centuty. As a matter of fact Faraday was elected as a member of the most prestigious scientific academies of that time, but he always remained faithful to the principle of avoiding awards and honours at home.

When, during the Crimean War (1853-1856), the British government asked him to contribute to the production of chemical weapons, Faraday strongly opposed his ethical principles. With equal firmness he refused the social and economic benefits of becoming a knight,

Figure 9. James Clerk Maxwell [12].

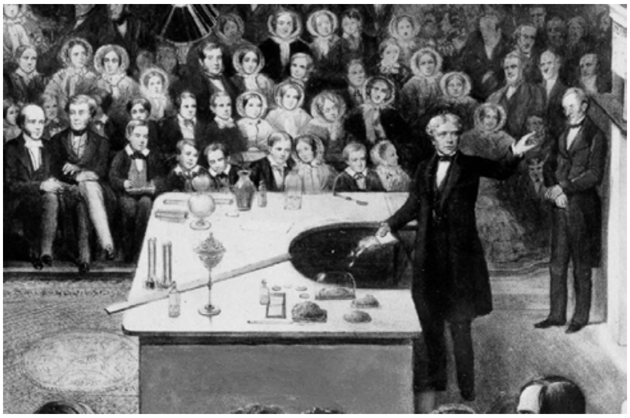


Figure 10. M. Faraday while holding a lecture at the Royal Institution in London [13].

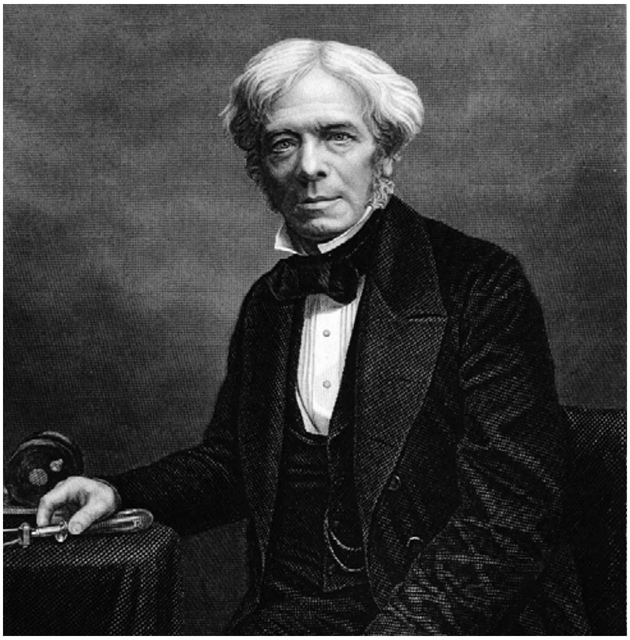


Figure 11. Michael Faraday in his sixties [14].

and, twice, the appointment as president of the Royal Society. Only in 1848 he accepted the house at Hampton Court from the British Royal Family, in recognition of his scientific achievements.

In 1858 he moved there, where he died in 1867, at the age of 75. The great English scientist was buried in the Highgate Cemetery in London, because he had also refused the privilege of being buried in the English pantheon, the Westminster Abbey in London. As a Sandemanian, he decided that even after death he would have never entered an Anglican church, least of all lay there until the day of judgement. In Westminster Abbey only a simple plaque nearby the tomb of Isaac Newton remembers the greatest British experimental scientist of all times.

SCIENTIFIC DISCOVERIES

History of Faraday's scientific discoveries

- 1810-1820 First Electrochemical Experiments
- 1820-1830 Electrical conduction experiments
- 1831 Law of electromagnetic induction
- 1832-1833 Laws of electrolysis
- 1837-39 Dielectric materials
- 1845-1846 Diamagnetism and Faraday effect
- 1855 Studies on paramagnetism

At the age of twenty, one year before the end of his apprenticeship at Mr. Riebau's bookshop, Faraday performed the first of the many important experiments of his scientific career.

As he stated in a letter to his friend Benjamin Abbott, Faraday set up an electric battery (Figure 12), using zinc and copper disks with the size of a half penny, separated by pieces of paper soaked in "Muriate of Soda".

Faraday discovered that the electrical power of this rudimentary stack was capable of separating the components of a magnesium sulphate solution. He submerged in the solution the copper wires connected to the poles

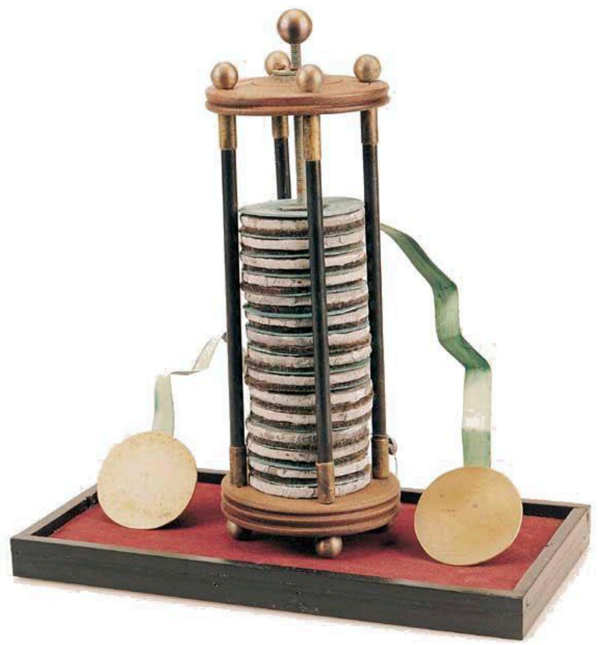


Figure 12. Alessandro Volta's pile (modern reconstruction) [15].

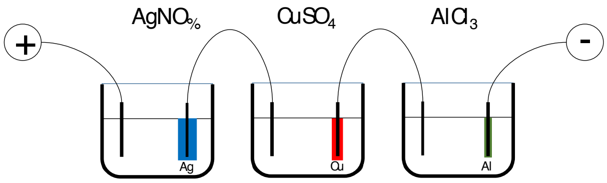


Figure 13. Demonstration of the experimental laws of Faraday on electrolysis (1834). Solutions of silver nitrate, copper sulphate and aluminium chloride. The electric current flowing through the cells is the same and the metal ions are deposited on the negative electrode in an amount proportional to the current flowing through the electrolytic cells (first Faraday's Law). Furthermore, if the amount of silver deposited is 108 g (atomic weight of Ag) that of copper is equal to 31.7 g (about half the atomic weight of Cu) and that of aluminium is 9 g (one third of the atomic weight of aluminium): this indicates that the Cu and Al ions, respectively, carry a double and triple charge with respect to that of the Ag ion (second Faraday's Law) [16].

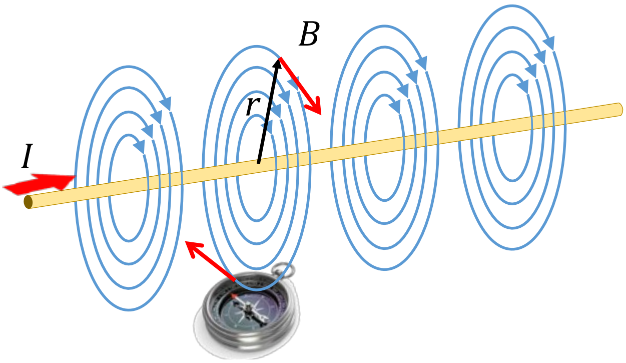


Figure 14. Magnetic field force lines generated by a rectilinear conductor carrying a constant electric current [16].

of the pile. In a few hours the solution became muddy, because of the formation of a suspension of magnesia. In this way Faraday discovered the phenomenon of electrolysis. This was the beginning of an intense line of systematic and experimental research studies in the years to come. During his stay at the Royal Institution, Faraday managed to obtain the two fundamental laws of electrolysis (Figure 13).

First LAW: For a given solution, the quantity of matter that is deposited on the electrodes is proportional to the amount of charge which passes through the solution.

This implies that the ions carrying the charge through the solution have a well-defined electric charge.

Second LAW: The monovalent ions of different substances carry an equal quantity of electric charge, while the bi- or tri-valent ones carry a correspondingly higher charge.

This law implies the existence of an elementary charge unit that at the Faraday's times was attributed to individual atoms, but we now know to be that of the electron.

The results of this research were published by Faraday in 1834, almost at same time as Carlo Matteucci, who formulated the same laws of electrolysis but using completely independent methods.

Faraday was the author of numerous publications in scientific journals, but we can say with no doubt that his main contributions are collected in his Laboratory Journal, which he hold regularly from 1820 until 1862. This diary was published in 1932 by the Royal Institution in seven large volumes, totalling 3,236 pages accompanied by nearly a thousand illustrations of instruments and experimental setups. In these pages Faraday also describes his first experiments on electrical conduction (1820-1830) in which he confirmed his talent as experimenter, sensing and describing crucial phenomena, such as the fact that a conductor wire carrying an electric current exerts a force on the poles of a bar needle, showing that this force is the same along a circle concentric to the wire (Figure 14).

Similarly, a current-carrying wire can be put in rotation around a magnetic bar (Figure 15). The observation of this phenomenon can be considered as the first step of Faraday towards the design of an electric motor. This setup, far more refined than the ones used in the preliminary experiments carried out by the Danish physicist

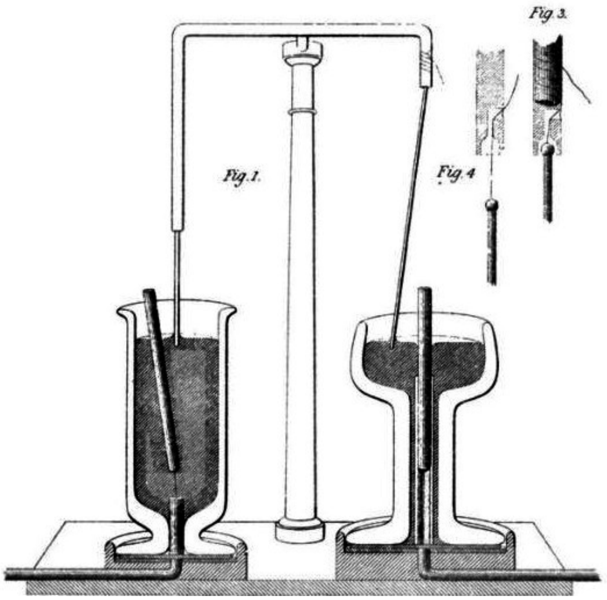


Figure 15. Left: ampoule filled with mercury which is anchored to a movable magnetic rod which rotates to the passage of an electric current along the conductor. Right: ampoule filled with mercury, which contains a fixed magnetic bar around which a conductive wire is crossed by an electric current [17].

H.C. Orsted, was built by Faraday after discussing the idea with H. Davy and W.H. Wollastone.

Faraday published the results of his research without mentioning Davy's and Wollastone's contributions (who, despite their efforts, were never able to build any device functioning as an electric motor) and this opened such a hard dispute, that at the end Faraday was forced to officially give up with his studies on electromagnetism. It is interesting to read from Faraday's words how he justified his act to the managers of the Royal Institution:

> *I hear every day more and more those sounds which though only whispers to me are I suspect spoken aloud amongst scientific men and which as they in part affect my honour and honesty I am anxious to do away with or at least to prove erroneous in those parts which are dishonourable to me. You know perfectly well what distress the very unexpected reception of my paper on Magnetism in public has caused me and you will not therefore be surprised at my anxiety to get out of it though I give trouble to you and others of my friends in doing so.*
> *I understand I am charged 1. with not acknowledging the information I received in assisting Sir H. Davy in his experiments on this subject; 2. with concealing the theory and views of Dr Wollaston; 3. with taking the subject whilst Dr Wollaston was at work on it; and 4. with dishonourably taking Dr Wollaston's thoughts and pursuing them without acknowledging the results I have brought out [18].*

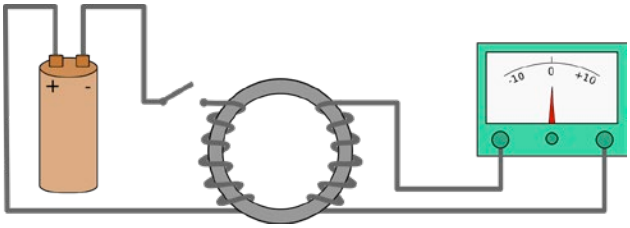


Figure 16. Scheme of the experiment on electromagnetic induction by Faraday, August 29, 1831. When the circuit on the left is closed, the presence of an induced current in the circuit on the right is observed, as long as the first does not reach the steady state value when the current in the second circuit disappears. In fact, the magnetic field flux inside the soft iron ring around which the two filaments are wrapped changes over time after closing the first circuit. The same phenomenon is observed after opening again the left circuit, in which case the flux of the magnetic field decreases. Therefore, induction, differently from the electrostatic interaction, is a dynamic phenomenon [19].

It is no exaggeration to affirm that the ability of Faraday had too much shaded that of his mentor, who decided to significantly limit the scientific autonomy of his assistant. Officially, Faraday was allowed only to continue his studies on electrochemistry, and was eventually able to come back formally on those about electromagnetism immediately after the death of Davy (1829), when he took over the direction of the Royal Institution. In two years of intense experimentation, he managed to understand the phenomenon of the magnetic induction (already highlighted by the abbot Francesco Zantedeschi).

Faraday's laboratory notebook reports his fundamental discovery of the electromagnetic induction on 29th August 1831. The page is entitled: "Experiments on the production of electricity from magnetism". In his notebook Faraday provided a detailed description of his experimental apparatus (Figure 16).

Note that also the French physicist André-Marie Ampere in 1822 had conjectured the possibility of such a phenomenon, based on the analogy with electrostatics. More precisely, Ampere believed that if a circuit had been traversed by a current, in a second circuit in its vicinity a current should be observed.

Actually Ampere did notice that something happened in the second circuit for a very short time, when the first circuit was moved with respect to the first, but he did not attribute any importance to this transient phenomenon. Faraday instead highlighted this effect in the experiment of 29th of August 1831 and only two months later, on the 17th of October, he was able to obtain the confirmation of the dynamic nature of the electromagnetic induction phenomenon, performing the experiment depicted in Figure 17.

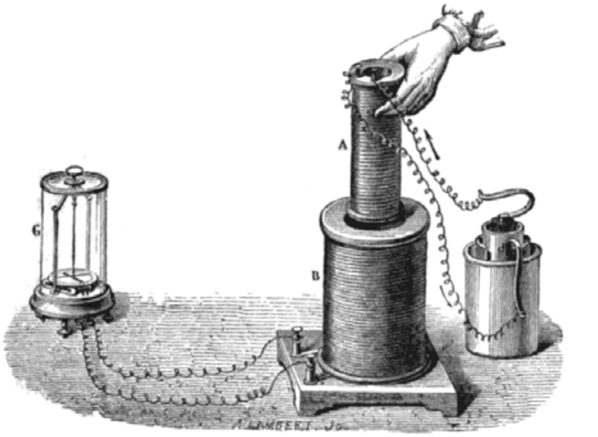


Figure 17. Scheme of the experiment of electromagnetic induction by Faraday, October 17th, 1831. By vertically moving the smallest coil within the larger one the magnetic field flux concatenated to the latter is varied and a passage of current in it is observed. This is a confirmation of the dynamic nature of the phenomenon of the electromagnetic induction (1892) [20].

This time he built a tight coil crossed by an electric current and therefore capable of generating a magnetic field directed along its axis, and inserted it inside a larger coil, moving the former vertically. This operation allowed him to vary the flux of the magnetic field linked with the largest coil in a controlled manner, not relying upon the too rapid variation produced by turning on and off the current in the primary circuit of the first experiment.

He observed that in this way a current was produced: it was circulating in the second coil as long as the magnetic field flux continued to change over time. These experiments provided Faraday the final confirmation that the electromagnetic phenomena were represented with lines of force that pervaded the surrounding space and that made such phenomena much easier to illustrate with respect to the law of the action at distance, borrowed by the principles of Newtonian mechanics.

It should be stressed that, at that time, "action at distance" was a scientific cornerstone, that had been substantiated by the work of the French physicists, such as Coulomb and Ampère, strongly linked to the rationalist tradition of French science, the guardian of the modern formulation of Newtonian mechanics, rewritten in the language of rational mechanics.

Faraday, fortunately for us, was an experimental scientist of great talent, very little inclined to be fascinated by metaphysical paradigms. His conception of lines of force constituted a rather different point of view, in some ways antithetical to that of the French rationalists.

His position was definitely not free of romantic influences, which tended to question the scientific rationalism of the time, but certainly his intellectual honesty and his intuition led him to strongly support the idea of a "fluid dynamics" approach to electromagnetic forces, which, in this first level of understanding, had to occur through the presence of a material medium interposed in the space where such phenomena were observed. Since his early scientific experiences Faraday had strongly supported these ideas, even discussing them in a scientific correspondence with Ampere, who recognized in the English scientist a correspondent worthy of great consideration, although their ideas on electromagnetic phenomena were indeed quite antithetical.

Indeed, Faraday was probably intimidated by his lack of mathematical background, so instead of quarrelling with Ampère, he preferred to stick to his experimental observations:

> *I am unfortunate in a want of mathematical knowledge, and the power of entering with facility into abstract reasoning. I am obliged to feel my way by facts closely placed together, so that it often happens I am left behind in the progress of a branch of science not merely from the want of attention but from the incapability I lay under of following it, notwithstanding all my exertions. It is so just now, I am ashamed to say, with your refined researches in electromagnetism or electrodynamics.*
> *On reading your papers and letters I have no difficulty in following the reasoning but still at last I seem to want something more on which to steady the conclusions. I fancy the habit I got into of attending too closely to experiment has somewhat fettered my powers of reasoning and chains me down and I cannot help now and then comparing myself to a timid ignorant navigator who though he might boldly and safely steer across a bay or an ocean by the aid of a compass which in its actions and principles is infallible, is afraid to leave sight of the shore because he understands not the power of the instrument that is to guide him [21].*

However, it is worth stressing that Faraday's approach was quite more modern than Ampère's view. By systematically using iron filings, he was able to mentally visualize the existence of the magnetic field, thus separating Ampère's interactions among circuit in two steps: the generation of the magnetic field (which can originate also from a magnet) and the effect of the magnetic field on a circuit (or another magnet).

After the formulation of the laws of electrolysis, that we already mentioned, Faraday, between 1837 and 1839, devoted himself to the study of dielectric materials (the very name "dielectric" is due to him), which are in fact insulators.

These studies aimed at testing how the electromagnetic fields act on matter through the lines of force. Faraday also introduced the concept of dielectric constant,

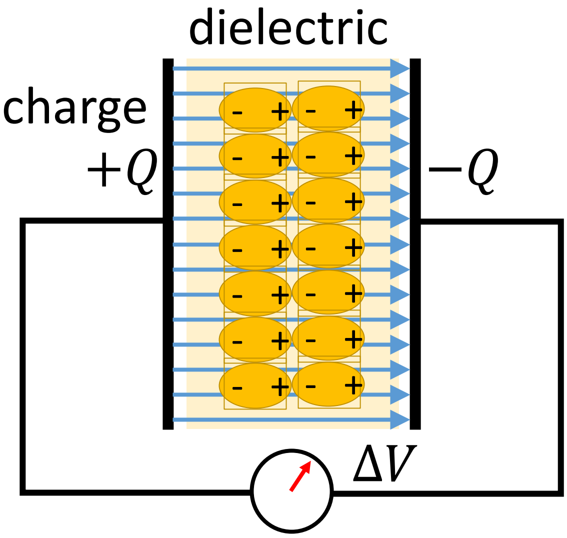


Figure 18. A measure of the relative dielectric constant K can be obtained by interposing the dielectric material (insulator) between the plates of a charged capacitor and measuring the potential difference in the absence and in the presence of the dielectric: K =ΔV0/ΔV > 1 [16].

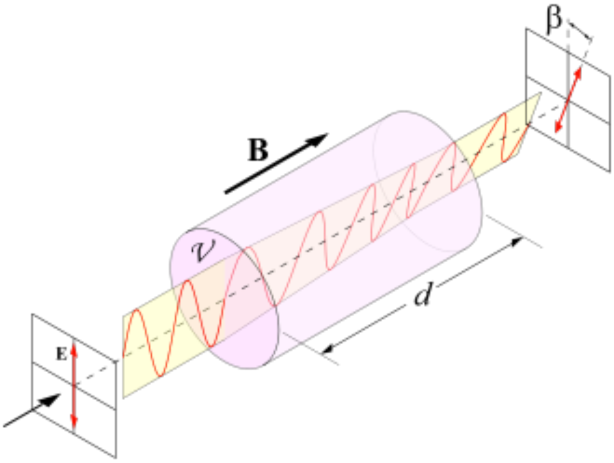


Figure 19. The Faraday effect. Faraday observed the rotation of the polarization plane of the light when this crosses a glassy material (lead silicoborate) subjected to an intense magnetic field: this effect is due to interaction of the light (electromagnetic wave) with the electrons present in the glass atoms. The phenomenon is the basis of the microwave technology and of optical insulators used in modern communications technologies [22].

while providing an operational definition for its measurement (Figure 18).

After a period of inactivity of almost four years in 1845 Faraday returned to his experiments on magnetic materials, classifying them in diamagnetic and paramagnetic, depending on their propensity to oppose or promote the penetration of the magnetic field.

These early studies paved the way for the important developments in research on the magnetic properties of matter in the following decades. Faraday was already aware of the complementary nature of electric and magnetic phenomena and continued investigating this topic, by addressing specific questions about the nature of the optical effects of magnetic fields.

As usual, he faced the problem with his brilliant experimental intuition and in 1845 he discovered the phenomenon known as “Faraday rotation”. After a series of unsuccessful experiments, he succeeded to develop an experimental apparatus capable of detecting a rotation of the polarization plane of light when this crossed a material (glass) merged into a strong magnetic field, directed in the direction of propagation of light (Figure 19).

The last years of the scientific life of Faraday were devoted to improve his understanding of paramagnetic materials and to perform some unsuccessful attempts to find a relationship between electromagnetic and gravitational forces. It is hard not to be amazed by facing this last great Faraday’s intuition, that anticipates some of the most recent discoveries in physics, related to the possibility of representing a unified theory of all force fields (sub-nuclear, nuclear, electromagnetic and gravitational).

In 1849 he wrote in his laboratory diary:

> *Surely this force [gravity] must be capable of an experimental relation to Electricity, Magnetism and the other forces, so as to bind it up with them in reciprocal action and equivalent effect...What in Gravity answers to the dual or antithetical nature of the forms of force in Electricity and Magnetism? Perhaps the to and fro, that is, the ceding to the force or approach of Gravitating bodies, and the effectual reversion of the force or separation of the bodies, quiescence being the neutral condition. Try the question experimentally on these grounds [23].*

Faraday began a long series of experiments that did not provide him any evidence of his expectations. He concluded this stage of his research on the diary by noting

> *[The negative results] do not shake my strong feeling of the existence of a relation between gravity and electricity, though they give no proof that such a relation exists [24].*

This shows that although Faraday was an empirical scientist, however he did not consider only the results of his experiments. In fact, due to his deep philosophical beliefs he was convinced that the laws of nature must have a unique common origin, as a manifestation of something that deeply and mysteriously permeates the universe.

Figure 20. Statue of Michael Faraday in Savoy Place, London [26].

In the case of Faraday, we can certainly say that this was intrinsic to his ethical and religious convictions. But we have also to make clear that Faraday had no sanctimonious attitude, nor he was motivated by any superstition or dogmatic belief.

Some historians believe, with some justification, that his self-taught training and his free spirit played a crucial role in providing him with a unique talent as a creator of crucial experiments, while escaping the cultural traps of a traditional mathematical background, that, in his time, was attributed the value of incontrovertible truths. George Gamow (the famous Russian scientist of the twentieth century) wrote about Faraday

> *Impressive as Faraday's experimental discoveries were, they are matched by his theoretical ideas. Having had very little education and having known practically no mathematics, Faraday could not be what is usually called a theoretical physicist. But the fact is that, for conceiving a theoretical picture of a puzzling physical phenomenon, a knowledge of intricate mathematics is often quite unnecessary and sometimes even harmful. The explorer may easily be lost in the jungles of complicated formulas and, as a Russian proverb says, "cannot see the forest for the trees" [25].*

Figure 21. A 20-pound banknote with the Faraday effigy [27].

Thanks to his tireless and amazing work as experimentalist, famous during his life and after his death, James Clerk Maxwell developed his theory of electromagnetic field, which is one of the cornerstones of the building of modern physics. It can be said, without doubt, that Maxwell's scientific path would hardly have come to fulfilment without Faraday having largely prepared it.

Considering his shyness, it is possible that Faraday would not be pleased by the huge number of studies dedicated to him, nor by the present public recognition of his character (Figures 20 and 21).

SOME EXCERPTS BY FARADAY

> *It may be asked, what lines of force are there in nature, which are fitted to convey such an action, and supply for the vibrating theory the place of the ether? I do not pretend to answer this question with any confidence; all I can say is, that I do not perceive in any part of space, whether (to use the common phrase) vacant or filled with matter, anything but forces and the lines in which they are exerted. The lines of weight or gravitating force are, certainly, extensive enough to answer in this respect any demand made upon them by radiant phenomena; and so, probably, are the lines of magnetic force: and then, who can forget that Mossotti has shown that gravitation, aggregation, electric force, and electro-chemical action may all have one common connexion or origin; and so, in their actions at a distance, may have in common that infinite scope which some of these actions are known to possess?*
>
> *The view which I am so bold as to put forth considers, therefore, radiation as a high species of vibration in the lines of force which are known to connect particles and also masses of matter together. It endeavours to dismiss the ether, but not the vibrations. The kind of vibration which, I believe, can alone account for the wonderful, varied, and beautiful phenomena of polarization, is not the same as that which occurs on the surface of disturbed water, or the waves of sound in gases or liquids, for the vibrations in*

these cases are direct, or to and from the centre of action, whereas the former are lateral. It seems to me, that the resultant of two or more lines of force is in an apt condition for that action which may be considered as equivalent to a lateral vibration; whereas a uniform medium, like the ether, does not appear apt, or more apt than air or water. The occurrence of a change at one end of a line of force easily suggests a consequent change at the other. The propagation of light, and therefore probably of all radiant action, occupies time; and, that a vibration of the line of force should account for the phenomena of radiation, it is necessary that such vibration should occupy time also. I am not aware whether there are any data by which it has been, or could be ascertained, whether such a power as gravitation acts without occupying time, or whether lines of force being already in existence, such a lateral disturbance of them at one end as I have suggested above, would require time, or must of necessity be felt instantly at the other end [28].

Years ago I believed that electrolytes could conduct electricity by a conduction proper; that has also been denied by many through long time: though I believed myself right, yet circumstances have induced me to pay that respect to criticism as to reinvestigate the subject, and I have the pleasure of thinking that nature confirms my original conclusions. So though evidence may appear to preponderate extremely in favour of a certain decision, it is wise and proper to hear a counter-statement. You can have no idea how often and how much, under such an impression, I have desired that the marvellous descriptions which have reached me might prove, in some points, correct; and how frequently I have submitted myself to hot fires, to friction with magnets, to the passes of hands, &c., lest I should be shutting out discovery;—encouraging the strong desire that something might be true, and that I might aid in the development of a new force of nature [29].

Magnetic lines of force convey a far better and purer idea than the phrase magnetic current or magnetic flood: it avoids the assumption of a current or of two currents and also of fluids or a fluid, yet conveys a full and useful pictorial idea to the mind [30].

All your names I and my friend approve of or nearly all as to sense & expression, but I am frightened by their length & sound when compounded. As you will see I have taken dioxide and skaiode because they agree best with my natural standard East and West. I like Anode & Cathode better as to sound, but all to whom I have shown them have supposed at first that by Anode I meant No way[2] *[31].*

Although we know nothing of what an atom is, yet we cannot resist forming some idea of a small particle, which represents it to the mind … there is an immensity of facts which justify us in believing that the atoms of matter are in some way endowed or associated with electrical powers, to which they owe their most striking qualities, and amongst them their mutual chemical affinity [32].

I require a term to express those bodies which can pass to the electrodes, or, as they are usually called, the poles. Substances are frequently spoken of as being electro-negative, or electro-positive, according as they go under the supposed influence of a direct attraction to the positive or negative pole. But these terms are much too significant for the use to which I should have to put them; for though the meanings are perhaps right, they are only hypothetical, and may be wrong; and then, through a very imperceptible, but still very dangerous, because continual, influence, they do great injury to science, by contracting and limiting the habitual view of those engaged in pursuing it. I propose to distinguish these bodies by calling those anions which go to the anode of the decomposing body; and those passing to the cathode, cations; and when I have occasion to speak of these together, I shall call them ions [33].

I wanted some new names to express my facts in Electrical science without involving more theory than I could help & applied to a friend Dr Nicholl [his doctor], who has given me some that I intend to adopt for instance, a body decomposable by the passage of the Electric current, I call an 'electrolyte' and instead of saying that water is electro chemically decomposed I say it is 'electrolyzed'. The intensity above which a body is decomposed beneath which it conducts without decomposition I call the 'Electrolyte intensity' &c &c. What have been called: the poles of the battery I call the electrodes they are not merely surfaces of metal, but even of water & air, to which the term poles could hardly apply without receiving a new sense. Electrolytes must consist of two parts which during the electrolization, are determined the one in the one direction, and the other towards the poles where they are evolved; these evolved substances I call zetodes, which are therefore the direct constituents of electrolytes [34].

I have taken your advice, and the names used are anode cathode anions cations and ions; the last I shall have but little occasion for. I had some hot objections made to them here and found myself very much in the condition of the man with his son and ass who tried to please every body; but when I held up the shield of your authority, it was wonderful to observe how the tone of objection melted away [35].

[The new term] Physicist is both to my mouth and ears so awkward that I think I shall never use it. The equivalent of three separate sounds of i in one word is too much [36].

[2] Here "No way" is presumably not an idiomatic exclamation, but a misinterpretation from the Greek prefix, *-a* "not" or "away from," and *hodos* meaning "way." The Greek ἄνοδος (*anodos*) means "way up" or "ascent."

PUBLICATIONS BY FARADAY AND RELEVANT BIBLIOGRAPHY

Michael Faraday, *Chemical Manipulation, Being Instructions to Students in Chemistry, on the methods of performing experiments of demonstration or of research, with accuracy and success*, John Murray, London 1830. https://archive.org/details/chemicalmanipula-00fararich

Michael Faraday, John Tyndall, *Experimental Researches in Electricity*, Dent, London 1922; Dutton, New York 1914. https://archive.org/details/experimentalrese-00faraiala

Michael Faraday, *Experimental Researches in Chemistry and Physics*, Taylor and Francis, London 1859. https://archive.org/details/experimentalrese00fararich

Michael Faraday, *A Course of Six Lectures on the Chemical History of a Candle*, W. Crookes, ed., Harper & Brothers, New York 1861. https://archive.org/details/acoursesixlectu01croogoog

Michael Faraday, *On the Various Forces Of Nature And Their Relations To Each Other*, W. Crookes, ed.,The Viking Press, New York 1960. https://archive.org/details/onthevariousforc007796mbp

Faraday, Michael, *Faraday's diary: being the various philosophical notes of experimental investigation*, T. Martin, ed., G. Bell and Sons, LTD, London 1932. https://archive.org/details/faradaysdiarybei00fara_1. See also the 2009 publication of Faraday's diary. http://www.faradaysdiary.com/

Michael Faraday, *Curiosity Perfectly Satisfyed: Faraday's Travels in Europe 1813–1815*, B. Bowers and L. Symons, ed., Institution of Electrical Engineers, P. Peregrinus, London 1991.

Michael Faraday, T*he Correspondence of Michael Faraday*, F. James, ed., Institution of Engineering and Technology 1991-2011. http://www.rigb.org/our-history/michael-faraday/michael-faraday-correspondence

Michael Faraday , *Michael Faraday's Mental Exercises: An Artisan Essay Circle in Regency London,* Alice Jenkins, ed. Liverpool University Press, Liverpool (2008).

Michael Faraday, *Course of six lectures on the various forces of matter, and their relations to each other,* R. Griffin, London; Glasgow 1860. https://archive.org/details/courseofsixlectu00fararich

Michael Faraday and Michael Northmore, *The Liquefaction of Gases*, W. F. Clay, Edinburgh 1896. https://archive.org/details/liquefactionofga00fararich

Michael Faraday and Christian Friedirich Schoenbein, *The letters of Faraday and Schoenbein 1836–1862. With notes, comments and references to contemporary letters* , Williams & Norgate, London 1899. https://archive.org/details/lettersoffaraday00fararich

James Hamilton, *Life of Discovery: Michael Faraday, Giant of the Scientific Revolution*, Random House, New York 2004.

Michael Faraday, *Faraday's Experimental Researches in Electricity: Guide to a First Reading*, Howard J. Fisher, ed., Green Lion Press, Santa Fe, New Mexico 2001.

Olivier Darrigol, *Electrodynamics, from Ampere to Einstein,* Oxford University Press, Oxford 2002.

Sydney Ross, *Nineteenth-Century Attitudes: Men of Science* volume 13: *Chemists and Chemistry,* Springer Science+Business Media B.V., Dordrecht 1991.

See also the Wikipedia entry on Faraday with many references and biographies [37].

REFERENCES

1. M. Faraday, *Letter to John Tyndall* (19 April 1851); letter 2411, The correspondence of Michael Faraday, Volume 4, Frank A. J. L. James, ed., Institution of Engineering and Technology (1999) p. 281.
2. Henry Bence Jones, *The Life and Letters of Michael Faraday*, Band 1, S. 9. https://commons.wikimedia.org/wiki/File:Faraday-riebaus_shop.png
3. James Hamilton, *Life of Discovery: Michael Faraday, Giant of the Scientific Revolution*, Random House, New York 2004 p. 8.
4. Unknown author from Popular Science Monthly **XIV** p. 696 1878-1879. https://archive.org/stream/popularsciencemon14newy
5. Engraving by John Cochran after a painting Painted by H.W. Pickersgill (1826), https://commons.wikimedia.org/wiki/File:Faraday_Cochran_Pickersgill.jpg
6. Fresco by Gaspero Martellini, after the drawing of Nicola Cianfanelli, Tribuna di Galileo (Florence), 1841. http://www.internetculturale.it/opencms/opencms/upload/exhibits3d/tribuna/Desc/images/007.jpg]
7. Photograph by Henry Dixon & Sons Ltd. About 1850, From Wellcome Image Library. https://wellcomecollection.org/works/pz2uwk3q
8. Unknown author from *Practical Physics*, Millikan and Gale, 1920. https://commons.wikimedia.org/wiki/File:Andre-marie-ampere2.jpg
9. Ian H. Hutchinson, *The Genius and Faith of Faraday and Maxwell*, The New Atlantis **41**, Winter 2014, pp. 81-99. https://www.thenewatlantis.com/publications/the-genius-and-faith-of-faraday-and-maxwell

10. Maria Chenu, L'Illustration, Journal Universel, juin 1862, p. 372. https://commons.wikimedia.org/wiki/File:L%27Illustration_1862_gravure_ministre_Matteucci.jpg
11. https://it.wikipedia.org/wiki/File:Ottaviano_Fabrizio_Mossotti_2.jpg
12. Drawing by unknown, Popular Science Monthly **17** back of cover (1880). https://archive.org/details/popularsciencemo17newy
13. Detail of a lithograph by Alexander Blaikley (1856). https://upload.wikimedia.org/wikipedia/commons/a/a0/Faraday_Michael_Christmas_lecture_detail.jpg]
14. From Wellcome Image Library (cropped), https://wellcomecollection.org/works/xm6ap73v?query=faraday
15. http://ppp.unipv.it/Mostra/Pagine/Frame%20S4/FrameS46.htm
16. Image by the authors.
17. Drawing by Michael Faraday, *Experimental Researches in Electricity,* Dent, London 1922; Dutton, New York 1914, volume 2, plate 4. https://en.wikipedia.org/wiki/Homopolar_motor#/media/File:Faraday_magnetic_rotation.jpg
18. Faraday to Managers of the Royal Institution, 11 October 1861; L. Pearce Williams, Rosemary Fitzgerald and Oliver Stallybrass, *The Selected Correspondence of Michael Faraday* Cambridge University Press, New York, 1971, no. 785.
19. https://en.wikipedia.org/wiki/Faraday%27s_law_of_induction#/media/File:Faraday_emf_experiment.svg
20. J. Lambert (1892) https://en.wikipedia.org/wiki/Faraday%27s_law_of_induction#/media/File:Induction_experiment.png
21. Corr. No. 369 bis, pp. 928-931.Faraday to Ampere, 3 September 1822; *The selected correspondence of Michael Faraday,* L. P. Williams, ed., Cambridge University Press, New York, 1971, Vol. 1, pp. 134-135
22. https://en.wikipedia.org/wiki/Faraday_effect#/media/File:Faraday-effect.svg
23. Michael Faraday, *Experimental Researches in Electricity,* Dover, New York 1965, Vol. 3, p. 2717.
24. Michael Faraday, *Diary V*, pp. 10018-10019 paraphrased in Experimental Researches in Electricity vol. III, Dent, London 1922; Dutton, New York 1914, p. 2703. Cf. *Diary V*, pp. 10934-10935.
25. G. Gamow, *The Great Physicists from Galileo to Einstein*, Dover, New York 1988 p. 149.
26. https://it.wikipedia.org/wiki/Michael_Faraday#/media/File:Michael_Faraday_statue.jpg
27. http://www.bankofengland.co.uk/banknotes/Documents/withdrawnrefguide.pdf
28. Michael Faraday, *Thoughts on Ray-vibrations. To Richard Phillips, Esq.,* Philosophical Magazine ***xxiv*** 1844, p. 136—or *Experimental Researches in Electricity*, vol. **ii**. Dent, London 1922; Dutton, New York 1914 p. 284.
29. Michael Faraday, *Observations on Mental Education, Lectures on Education*, Parker and Son 1855.
30. Michael Faraday, *Diary,* Entry for 10 Sep 1854. In Thomas Martin, ed., *Faraday's Diary: Being the Various Philosophical Notes of Experimental Investigation*. G. Bell, London 1932, Vol. 6, 315.
31. Michael Faraday, *Letter to William Whewell*, who coined the terms (3 May 1834). In Frank A. J. L. James, ed., *The Correspondence of Michael Faraday*, L. P. Williams, ed., Cambridge University Press, New York, 1971, Vol. 2, p. 181.
32. Faraday, M.: *Experimental Researches in Electricity*, Dent, London 1922; Dutton, New York 1914, section 852.
33. Michael Faraday, Philosophical Transactions of the Royal Society of London **124**, 79 (1834).
34. Michael Faraday, *Letter to William Whewell* (24 Apr 1834), In Frank A. J. L. James, ed., *The Correspondence of Michael Faraday*, L. P. Williams, ed., Cambridge University Press, New York, 1971, Vol. 2, p. 176.
35. Michael Faraday, *Letter to William Whewell* (15 May 1834), In Frank A. J. L. James, ed., *The Correspondence of Michael Faraday*, L. P. Williams, ed., Cambridge University Press, New York, 1971, Vol. 2, p. 186.
36. Micheal Faraday, Quoted in Sydney Ross, *Nineteenth-Century Attitudes: Men of Science* volume 13: *Chemists and Chemistry,* Springer Science+Business Media B.V., Dordrecht 1991, p. 10.
37. Faraday entry on Wikipedia. https://en.wikipedia.org/wiki/Michael_Faraday

Finito di stampare da
Logo s.r.l. – Borgoricco (PD) – Italia

March 2018 No walls. Just bridges

Substantia

An International Journal of the History of Chemistry

Vol. 2 – n. 1

Table of contents

Stig E. Friberg
Emulsion Thermodynamics - In from the Cold 7

Hans-Jürgen Apell
Finding Na,K-ATPase: I - From Cell to Molecule 17

Louis Caruana SJ
Mechanistic Trends in Chemistry 29

F. Tito Arecchi
Cognition and Reality 41

Barry D. Hughes and Barry W. Ninham
A Correspondence Principle 51

Joe Wolfe
From idea to acoustics and back again: the creation and analysis of information in music 77

Jehane Ragai
Snapshots of chemical practices in Ancient Egypt 93

Federico Maria Rubino
The "*Bitul B'shishim (one part in sixty)*": is a Jewish conditional prohibition of the Talmud the oldest-known testimony of quantitative analytical chemistry? 103

Franco Bagnoli and Roberto Livi
Michael Faraday: a virtuous life dedicated to science 121