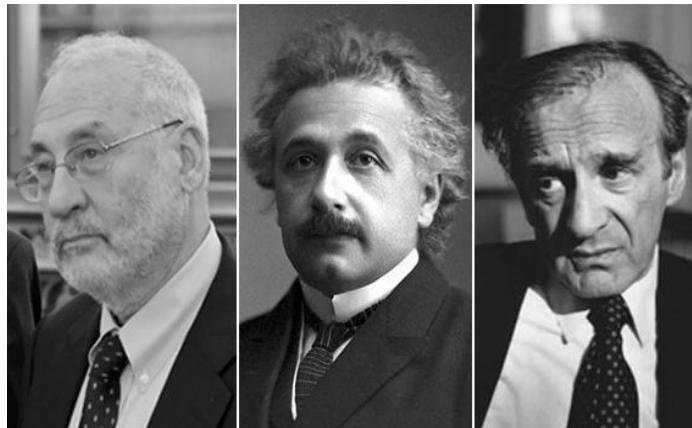


Jewish Nobel Lectures



Jews or Jewish people are an ethnoreligious group and a nation, originating from the Israelites and Hebrews of historical Israel and Judah. Jews make up 0.2% of the world's population, yet they have won over 20% of the Nobel prizes.

Contents

1. Robert John Aumann

War and Peace

2. Hans Albrecht Bethe

Energy Production in Stars

3. Konrad Emil Bloch

The biological synthesis of cholesterol

4. Felix Bloch

The Principle of Nuclear Induction

5. Niels Henrik David Bohr

The structure of the atom

6. Max Born

The Statistical Interpretations of Quantum Mechanics

7. Sir Ernst Boris Chain

The chemical structure of the penicillins

8. Albert Einstein

Fundamental ideas and problems of the theory of relativity

9. Richard Phillips "Dick" Feynman

The development of the space-time view of quantum electrodynamics

10. James Franck

Transformations of kinetic energy of free electrons into excitation energy of atoms by impacts

11. Vitaly Ginzburg

On Superconductivity and Superfluidity

12. Fritz Haber

The synthesis of ammonia from its elements

13. Bernard Katz

On the quantal mechanism of neural transmitter release

14. Fritz Albert Lipmann

Development of the acetylation problem: a personal account

15. Otto Fritz Meyerhof

Energy conversions in muscle

16. Albert A. Michelson

Recent advances in spectroscopy

17. Paul Ehrlich

Partial cell functions

18. Wolfgang Ernst Pauli

Exclusion principle and quantum mechanics

19. Otto Stern

The method of molecular rays

20. Eugene Wigner

Events, Laws of Nature, and Invariance Principles

21. Richard Martin Willstätter

On plant pigments

22. Julius Axelrod

Noradrenaline: fate and control of its biosynthesis

23. Gerty Theresa Cori

Polysaccharide phosphorylase

24. George de Hevesy

Some applications of isotopic indicators

25. Rita Levi-Montalcini

The nerve growth factor: thirty-five years later

26. Emilio Gino Segrè

Properties of antinucleons

27. Selman Abraham Waksman

Streptomycin: background, isolation, properties, and utilization

28. Otto Wallach

Alicyclic compounds

WAR AND PEACE

Prize Lecture¹, December 8, 2005

by

ROBERT J. AUMANN

Center for the Study of Rationality, and Department of Mathematics, The Hebrew University, Jerusalem, Israel.

“Wars and other conflicts are among the main sources of human misery.” Thus begins the *Advanced Information* announcement of the 2005 Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel, awarded for Game Theory Analysis of Conflict and Cooperation. So it is appropriate to devote this lecture to one of the most pressing and profound issues that confront humanity: that of War and Peace.

I would like to suggest that we should perhaps change direction in our efforts to bring about world peace. Up to now all the effort has been put into resolving specific conflicts: India–Pakistan, North–South Ireland, various African wars, Balkan wars, Russia–Chechnya, Israel–Arab, etc., etc. I’d like to suggest that we should shift emphasis and study war in general.

Let me make a comparison. There are two approaches to cancer. One is clinical. You have, say, breast cancer. What should you do? Surgery? Radiation? Chemotherapy? Which chemotherapy? How much radiation? Do you cut out the lymph nodes? The answers are based on clinical tests, simply on what works best. You treat each case on its own, using your best information. And your aim is to cure the disease, or to ameliorate it, in the specific patient before you.

And, there is another approach. You don’t do surgery, you don’t do radiation, you don’t do chemotherapy, you don’t look at statistics, you don’t look at the patient at all. You just try to understand what happens in a cancerous cell. Does it have anything to do with the DNA? What happens? What is the process like? *Don’t* try to cure it. Just try to *understand* it. You work with mice, not people. You try to make them sick, not cure them.

Louis Pasteur was a physician. It was important to him to treat people, to cure them. But Robert Koch was not a physician, he didn’t try to cure people. He just wanted to know how infectious disease works. And eventually, his work became tremendously important in treating and curing disease.

War has been with us ever since the dawn of civilization. Nothing has been more constant in history than war. It’s a phenomenon, it’s not a series of iso-

¹ A very lightly edited version of the 40-minute lecture actually delivered at the Royal Swedish Academy of Sciences in Stockholm. We are grateful to Professor Nicolaus Tideman for pointing out an error in a previous version.

lated events. The efforts to resolve specific conflicts are certainly laudable, and sometimes they really bear fruit. But there's also another way of going about it – studying war as a general phenomenon, studying its general, defining characteristics, what the common denominators are, what the differences are. Historically, sociologically, psychologically, and – yes – *rationally*. Why does *homo economicus* – rational man – go to war?

What do I mean by "rationality"? It is this:

A person's behavior is rational if it is in his best interests, given his information.

With this definition, can war be rational? Unfortunately, the answer is yes; it can be. In one of the greatest speeches of all time – his second inaugural – Abraham Lincoln said: "Both parties deprecated war; but one would make war rather than let the nation survive; and the other would accept war rather than let it perish. And the war came."

It is a big mistake to say that war is irrational. We take all the ills of the world – wars, strikes, racial discrimination – and dismiss them by calling them irrational. They are not necessarily irrational. Though it hurts, they may be rational. If war is rational, once we understand that it is, we can at least somehow address the problem. If we simply dismiss it as irrational, we can't address the problem.

Many years ago, I was present at a meeting of students at Yale University. Jim Tobin, who later was awarded the Prize in Economic Sciences in Memory of Alfred Nobel, was also there. The discussion was freewheeling, and one question that came up was: Can one sum up economics in one word? Tobin's answer was "yes"; the word is *incentives*. Economics is all about incentives.

So, what I'd like to do is an economic analysis of war. Now this does *not* mean what it sounds like. I'm not talking about how to finance a war, or how to rebuild after a war, or anything like that. I'm talking about the *incentives* that lead to war, and about building incentives that prevent war.

Let me give an example. Economics teaches us that things are not always as they appear. For example, suppose you want to raise revenue from taxes. To do that, obviously you should raise the tax rates, right? No, wrong. You might want to *lower* the tax rates. To give people an incentive to work, or to reduce avoidance and evasion of taxes, or to heat up the economy, or whatever. That's just one example; there are thousands like it. An economy is a game: the incentives of the players interact in complex ways, and lead to surprising, often counter-intuitive results. But as it turns out, the economy really works that way.

So now, let's get back to war, and how *homo economicus* – rational man – fits into the picture. An example, in the spirit of the previous item, is this. You want to prevent war. To do that, obviously you should disarm, lower the level of armaments. Right? No, wrong. You might want to do the exact opposite. In the long years of the cold war between the US and the Soviet Union, what prevented "hot" war was that bombers carrying nuclear weapons were in the air 24 hours a day, 365 days a year. Disarming would have led to war.

The bottom line is – again – that we should start studying war, from all viewpoints, for its own sake. Try to understand what makes it happen. Pure, basic science. *That* may lead, eventually, to peace. The piecemeal, case-based approach has not worked too well up to now.

Now I would like to get to some of my own basic contributions, some of those that were cited by the Prize Committee. Specifically, let's discuss repeated games, and how they relate to war, and to other conflicts, like strikes, and indeed to all interactive situations.

Repeated games model long-term interaction. The theory of repeated games is able to account for phenomena such as altruism, cooperation, trust, loyalty, revenge, threats (self-destructive or otherwise) – phenomena that may at first seem irrational – in terms of the “selfish” utility-maximizing paradigm of game theory and neoclassical economics.

That it “accounts” for such phenomena does not mean that people deliberately choose to take revenge, or to act generously, out of consciously self-serving, rational motives. Rather, over the millennia, people have evolved norms of behavior that are by and large successful, indeed optimal. Such evolution may actually be biological, genetic. Or, it may be “memetic”; this word derives from the word “meme,” a term coined by the biologist Richard Dawkins to parallel the term “gene,” but to express social, rather than biological, heredity and evolution.

One of the great discoveries of game theory came in the early seventies, when the biologists John Maynard Smith and George Price realized that strategic equilibrium in games and population equilibrium in the living world are defined by the same equations. Evolution – be it genetic or memetic – leads to strategic equilibrium. So what we are saying is that in *repeated* games, strategic equilibrium expresses phenomena such as altruism, cooperation, trust, loyalty, revenge, threats, and so on. Let us see how that works out.

What do I mean by “strategic equilibrium”? Very roughly, the players in a game are said to be in *strategic equilibrium* (or simply *equilibrium*) when their play is *mutually optimal*: when the actions and plans of each player are rational in the given strategic environment – i.e., when each knows the actions and plans of the others.

For formulating and developing the concept of strategic equilibrium, John Nash was awarded the 1994 Prize in Economics Sciences in Memory of Alfred Nobel, on the fiftieth anniversary of the publication of John von Neumann and Oskar Morgenstern’s *Theory of Games and Economic Behavior*. Sharing that Prize were John Harsanyi, for formulating and developing the concept of *Bayesian* equilibrium, i.e., strategic equilibrium in games of incomplete information; and Reinhard Selten, for formulating and developing the concept of *perfect* equilibrium, a refinement of Nash’s concept, on which we will say more below. Along with the concepts of *correlated* equilibrium (Aumann 1974, 1987), and *strong* equilibrium (Aumann 1959), both of which were cited in the 2005 Prize announcement, the above three fundamental concepts constitute the theoretical cornerstones of noncooperative game theory.

Subsequent to the 1994 prize, two Prizes in Economic Sciences in Memory of Alfred Nobel were awarded for *applications* of these fundamental concepts. The first was in 1996, when William Vickrey was awarded the Prize posthumously for his work on auctions. (Vickrey died between the time of the Prize announcement and that of the ceremony.) The design of auctions and of bidding strategies are among the prime practical applications of game theory; a good – though somewhat dated – survey is Wilson 1992.

The second came this year – 2005. Professor Schelling will, of course, speak and write for himself. As for your humble servant, he received the prize for applying the fundamental equilibrium concepts mentioned above to *repeated* games. That is, suppose you are playing the same game G , with the same players, year after year. One can look at this situation as a single big game – the so-called *supergame* of G , denoted G^∞ – whose rules are, “play G every year.” The idea is to apply the above equilibrium concepts to the supergame G^∞ , rather than to the one-shot game G , and to see what one gets.

The theory of repeated games that emerges from this process is extremely rich and deep (good – though somewhat dated – surveys are Sorin 1992, Zamir 1992, and Forges 1992). In the few minutes that are available to me, I can barely scratch its surface. Let me nevertheless try. I will briefly discuss just one aspect: the *cooperative*. Very roughly, the conclusion is that

Repetition Enables Cooperation.

Let us flesh this out a little. We use the term *cooperative* to describe any possible outcome of a game, as long as no player can *guarantee* a better outcome for himself. It is important to emphasize that in general, a cooperative outcome is *not* in equilibrium; it's the result of an agreement. For example, in the well-known “prisoner's dilemma” game, the outcome in which neither prisoner confesses is a cooperative outcome; it is in neither player's best interests, though it is better for both than the unique equilibrium.

An even simpler example is the following game H : There are two players, Rowena and Colin. Rowena must decide whether both she and Colin will receive the same amount – namely 10 – or whether she will receive ten times more, and Colin will receive ten times less. Simultaneously, Colin must decide whether or not to take a punitive action, which will harm both Rowena and himself; if he does so, the division is cancelled, and instead, each player gets nothing. The game matrix is

		Acquiesce	Punish
		10	0
Divide Evenly	10	0	
	100	0	
Divide Greedily	1	0	
	100	0	

The outcome **(E,A)**, yielding 10 to each player, is a cooperative outcome, as no player can guarantee more for himself; but like in the prisoner's dilemma, it is not achievable in equilibrium.

Why are cooperative outcomes interesting, even though they are not achievable in equilibrium? The reason is that they are achievable by contract – by agreement – in those contexts in which *contracts are enforceable*. And there are many such contexts; for example, a national context, with a court system. The Talmud (Avot 3, 2) says,

הוי מתפלל בשלום של מלכות, שאלמלא מורה, איש את רעה חיים בלעו.

“Pray for the welfare of the government, for without its authority, man would swallow man alive.” If contracts are enforceable, Rowena and Colin can achieve the cooperative outcome **(E,A)** by agreement; if not, **(E,A)** is for practical purposes unachievable.

The cooperative theory of games that has grown from these considerations predates the work of Nash by about a decade (von Neumann and Morgenstern 1944). It is very rich and fruitful, and in my opinion, has yielded *the* central insights of game theory. However, we will not discuss these insights here; they are for another Prize in Economic Sciences in Memory of Alfred Nobel, in the future.

What I do wish to discuss here is the relation of cooperative game theory to repeated games. The fundamental insight is that repetition is like an enforcement mechanism, which enables the emergence of cooperative outcomes *in equilibrium* – when everybody is acting in his own best interests.

Intuitively, this is well-known and understood. People are much more cooperative in a long-term relationship. They know that there is a tomorrow, that inappropriate behavior will be punished in the future. A businessman who cheats his customers may make a short-term profit, but he will not stay in business long.

Let's illustrate this with the game H . If the game is played just once, then Rowena is clearly better off by dividing Greedily, and Colin by Acquiescing. (Indeed, these strategies are *dominant*.) Colin will not like this very much – he is getting nothing – but there is not much that he can do about it. Technically, the *only* equilibrium is **(G,A)**.

But in the supergame H^∞ , there *is* something that Colin can do. He can threaten to Punish Rowena for ever afterwards if she ever divides Greedily. So it will not be worthwhile for her to divide greedily. Indeed, in H^∞ this is actually an equilibrium in the sense of Nash. Rowena's strategy is “play E for ever”; Colin's strategy is “play A as long as Rowena plays E; if she ever plays G, play P for ever afterwards.”

Let's be quite clear about this. What is maintaining the equilibrium in these games is the *threat of punishment*. If you like, call it “MAD” – mutually assured destruction, the motto of the cold war.

One caveat is necessary to make this work. The discount rate must not be too high. Even if it is anything over 10% – if \$1 in a year is worth less than 90

cents today – then cooperation is impossible, because it's still worthwhile for Rowena to be greedy. The reason is that even if Colin punishes her – and himself! – for ever afterwards, then when evaluated today, the entire eternal punishment is worth less than \$90, which is all that Rowena gains today by dividing greedily rather than evenly.

I don't mean just the monetary discount rate, what you get in the bank. I mean the personal, subjective discount rate. For repetition to engender cooperation, the players must not be too eager for immediate results. The present, the now, must not be too important. If you want peace now, you may well never get peace. But if you have time – if you can wait – that changes the whole picture; *then* you may get peace now. It's one of those paradoxical, upside-down insights of game theory, and indeed of much of science. Just a week or two ago, I learned that global warming may cause a cooling of Europe, because it may cause a change in the direction of the Gulf Stream. Warming may bring about cooling. Wanting peace now may cause you never to get it – not now, and not in the future. But if you can wait, maybe you will get it now.

The reason is as above: The strategies that achieve cooperation in an equilibrium of the supergame involve punishments in subsequent stages if cooperation is not forthcoming in the current stage. If the discount rates are too high, then the players are more interested in the present than in the future, and a one-time coup now may more than make up for losses in the sequel. This vitiates the threat to punish in future stages.

To summarize: In the supergame H^∞ of the game H , the cooperative outcome (\mathbf{E}, \mathbf{A}) is achievable in equilibrium. This is a special case of a much more general principle, known as the *Folk Theorem*, which says that *any* cooperative outcome of *any* game G is achievable as a strategic equilibrium outcome of its supergame G^∞ – even if that outcome is not an equilibrium outcome of G . Conversely, every strategic equilibrium outcome of G^∞ is a cooperative outcome of G . In brief, for any game G , we have

THE FOLK THEOREM: The cooperative outcomes of G coincide with the equilibrium outcomes of its supergame G^∞ .

Differently put, repetition acts as an enforcement mechanism: It makes cooperation achievable when it is not achievable in the one-shot game. Of course, the above caveat continues to apply: In order for this to work, the discount rates of all agents must be low; they must not be too interested in the present as compared with the future.

There is another point to be made, and it again relates back to the 1994 Prize. John Nash got the Prize for his development of equilibrium. Reinhard Selten got the Prize for his development of *perfect* equilibrium. Perfect equilibrium means, roughly, that the threat of punishment is *credible*; that *if* you have to go to a punishment, then after you punish, you are still in equilibrium – you do not have an incentive to deviate.

That certainly is *not* the case for the equilibrium we have described in the supergame H^∞ of the game H . If Rowena plays \mathbf{G} in spite of Colin's threat,

then it is *not* in Colin's best interest to punish forever. That raises the question: In the repeated game, can (\mathbf{E}, \mathbf{A}) be maintained not only in strategic equilibrium, but also in *perfect* equilibrium?

The answer is yes. In 1976, Lloyd Shapley – whom I consider to be the greatest game theorist of all time – and I proved what is known as the *Perfect Folk Theorem*; a similar result was established by Ariel Rubinstein, independently and simultaneously. Both results were published only much later (Aumann and Shapley 1994, Rubinstein 1994). The Perfect Folk Theorem says that in the supergame G^∞ of any game G , any cooperative outcome of G is achievable as a *perfect* equilibrium outcome of G^∞ – again, even if that outcome is not an equilibrium outcome of G . The converse of course also holds. In brief, for any game G , we have

THE PERFECT FOLK THEOREM: **The cooperative outcomes of G coincide with the perfect equilibrium outcomes of its supergame G^∞ .**

So again, repetition acts as an enforcement mechanism: It makes cooperation achievable when it is not achievable in the one-shot game, even when one replaces strategic equilibrium as the criterion for achievability by the more stringent requirement of *perfect* equilibrium. Again, the caveat about discount rates applies: In order for this to work, the discount rates of all agents must be low; they must not be too interested in the present as compared with the future.

The proof of the Perfect Folk Theorem is quite interesting, and I will illustrate it very sketchily in the game H , for the cooperative outcome (\mathbf{E}, \mathbf{A}) . In the first instance, the equilibrium directs playing (\mathbf{E}, \mathbf{A}) all the time. If Rowena deviates by dividing **Greedy**, then Colin punishes her – plays **P**. He does not, however, do this forever, but only until Rowena's deviation becomes unprofitable. This in itself is still not enough, though; there must be something that motivates Colin to carry out the punishment. And here comes the central idea of the proof: If Colin does not punish Rowena, then Rowena must punish Colin – by playing **G** – for not punishing Rowena. Moreover, the process continues – any player who does not carry out a prescribed punishment is punished by the other player for not doing so.

Much of society is held together by this kind of reasoning. If you are stopped by a policeman for speeding, you do not offer him a bribe, because you are afraid that he will turn you in for offering a bribe. But why should he not accept the bribe? Because he is afraid that you will turn him in for accepting it. But why would you turn him in? Because if you don't, he might turn you in for not turning him in. And so on.

This brings us to our last item. Cooperative game theory consists not only of delineating all the possible cooperative outcomes, but also of choosing among them. There are various ways of doing this, but perhaps best known is the notion of *core*, developed by Lloyd Shapley in the early fifties of the last century. An outcome x of a game is said to be in its “core” if no set S of players can *improve* upon it – i.e., assure to each player in S an outcome that is bet-

ter for him than what he gets at x . Inter alia, the concept of core plays a central role in applications of game theory to economics; specifically, the core outcomes of an economy with many individually insignificant agents are the same as the competitive (a.k.a. Walrasian) outcomes – those defined by a system of prices for which the supply of each good matches its demand (see, e.g., Debreu and Scarf 1963, Aumann 1964). Another prominent application of the core is to *matching* markets (see, e.g., Gale and Shapley 1962, Roth and Sotomayor 1990). The core also has many other applications (for surveys, see Anderson 1992, Gabszewicz and Shitovitz 1992, Kannai 1992, Kurz 1994, and Young 1994).

Here again, there is a strong connection with equilibrium in repeated games. When the players in a game are in (strategic) equilibrium, it is not worthwhile for any one of them to deviate to a different strategy. A *strong* equilibrium is defined similarly, except that there it is not worthwhile for any set of players to deviate – at least one of the deviating players will not gain from the deviation. We then have the following

THEOREM (AUMANN 1959): The core outcomes of G coincide with the strong equilibrium outcomes of its supergame G^∞ .

In his 1950 thesis, where he developed the notion of strategic equilibrium for which he got the Prize in Economic Sciences in Memory of Alfred Nobel in 1994, John Nash also proposed what has come to be called the *Nash Program* – expressing the notions of cooperative game theory in terms of some appropriately defined noncooperative game; building a bridge between cooperative and noncooperative game theory. The three theorems presented above show that repetition constitutes precisely such a bridge – it is a realization of the Nash Program.

We end with a passage from the prophet Isaiah (2, 2–4):

והיה באחרית הימים, נכון יהיה ה' בראש ההרים, ונישא מגבעות, ונהרו אליו כל הגוים.
והלכו עמים רבים ואמר, לכו ועולה אל ה' יי', אל בית אלהי יעקב, וירונו מדרביו, ונלכה באורחותיו; כי
מצין תצא תורה, ודבר יי' מירושלם. ושפט בין הגוים, והוכיה עמים רבים, וכיתתו הרבותם לאיתם,
וחניתותיהם למזרמות; לא ישא גוי חרב, ולא ילמדו עוד מלחמה.

“And it shall come to pass ... that ... many people shall go and say, ... let us go up to the mountain of the Lord, ... and He will teach us of His ways, and we will walk in His paths. ... And He shall judge among the nations, and shall rebuke many people; and they shall beat their swords into ploughshares, and their spears into pruning hooks; nation shall not lift up sword against nation, neither shall they learn war any more.”

Isaiah is saying that the nations can beat their swords into ploughshares when there is a central government – a Lord, recognized by all. In the absence of that, one *can* perhaps have peace – no nation lifting up its sword against another. But the swords must continue to be there – they cannot be beaten into ploughshares – and the nations must continue to *learn* war, in order *not* to fight!

REFERENCES

- Anderson, R. M., 1992, "The Core in Perfectly Competitive Economies," in Aumann and Hart 1992, 413–457.
- Aumann, R. J., 1959, "Acceptable Points in General Cooperative n -Person Games," in *Contributions to the Theory of Games IV*, Annals of Mathematics Study 40, edited by A. W. Tucker and R. D. Luce, Princeton: at the University Press, 287–324.
- Aumann, R. J., 1964, "Markets with a Continuum of Traders," *Econometrica* 32, 39–50.
- Aumann, R. J., 1974, "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics* 1, 67–96.
- Aumann, R. J., 1987, "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica* 55, 1–18.
- Aumann, R. J. and Hart, S. (eds.), 1992, 1994, 2002, *Handbook of Game Theory, with economic applications*, Vols. 1, 2, 3, Elsevier, Amsterdam.
- Aumann, R. J. and Shapley, L. S., 1994, "Long-Term Competition: A Game-Theoretic Analysis," in *Essays in Game Theory in Honor of Michael Maschler*, edited by N. Megiddo, Springer, New York, 1–15.
- Debreu, G. and Scarf, H., 1963, "A Limit Theorem on the Core of an Economy," *International Economic Review* 4, 235–246.
- Forges, F., 1992, "Repeated Games of Incomplete Information: Non-Zero-Sum," in Aumann and Hart 1992, 155–177.
- Gabszewicz, J. J. and Shitovitz, B., 1992, "The Core in Imperfectly Competitive Economies," in Aumann and Hart 1992, 459–483.
- Gale, D. and Shapley, L. S., 1962, "College Admissions and the Stability of Marriage," *American Mathematical Monthly* 69, 9–15.
- Kannai, Y., 1992, "The Core and Balancedness," in Aumann and Hart 1992, 355–395.
- Kurz, M., 1994, "Game Theory and Public Economics," in Aumann and Hart 1994, 1153–1192.
- Peleg, B., 1992, "Axiomatizations of the Core," in Aumann and Hart 1992, 397–412.
- Roth, A. and Sotomayor, M., 1990, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Econometric Society Monograph Series, Cambridge: at the University Press.
- Rubinstein, A., 1994, "Equilibrium in Supergames," in *Essays in Game Theory in Honor of Michael Maschler*, edited by N. Megiddo, Springer, New York, 17–28.
- Sorin, S., 1992, "Repeated Games with Complete Information," in Aumann and Hart 1992, 71–107.
- von Neumann, J., and Morgenstern, O., 1944, *Theory of Games and Economic Behavior*, Princeton: at the University Press.
- Wilson, R., 1992, "Strategic Analysis of Auctions," in Aumann and Hart 1992, 227–279.
- Young, H. P., 1994, "Cost Allocation," in Aumann and Hart 1994, 1193–1236.
- Zamir, S., 1992, "Repeated Games of Incomplete Information: Zero-Sum," in Aumann and Hart 1992, 109–154.

Portrait photo of Robert J. Aumann by photographer D. Porges.

H. A. BETHE

Energy production in stars

Nobel Lecture, December 11, 1967

History

From time immemorial people must have been curious to know what keeps the sun shining. The first scientific attempt at an explanation was by Helmholtz about one hundred years ago, and was based on the force most familiar to physicists at the time, gravitation. When a gram of matter falls to the sun's surface it gets a potential energy

$$E_{\text{pot}} = -GM/R = -1.91 \cdot 10^{15} \text{ erg/g} \quad (1)$$

where $M = 1.99 \cdot 10^{33}$ g is the sun's mass, $R = 6.96 \cdot 10^{10}$ cm its radius, and $G = 6.67 \cdot 10^{-8}$ the gravitational constant. A similar energy was set free when the sun was assembled from interstellar gas or dust in the dim past; actually somewhat more, because most of the sun's material is located closer to its center, and therefore has a numerically larger potential energy. One-half of the energy set free is transformed into kinetic energy according to the well-known virial theorem of mechanics. This will permit us later to estimate the temperature in the sun. The other half of the potential energy is radiated away. We know that at present the sun radiates

$$\epsilon = 1.96 \text{ erg/g sec} \quad (2)$$

Therefore, if gravitation supplies the energy, there is enough energy available to supply the radiation for about 10^{15} sec which is about 30 million years.

This was long enough for nineteenth century physicists, and certainly a great deal longer than man's recorded history. It was not long enough for the biologists of the time. Darwin's theory of evolution had just become popular, and biologists argued with Helmholtz that evolution would require a longer time than 30 million years, and that therefore his energy source for the sun was insufficient. They were right.

At the end of the 19th century, radioactivity was discovered by Becquerel and the two Curie's who received one of the first Nobel prizes for this discovery. Radioactivity permitted a determination of the age of the earth, and more

recently, of meteorites which indicate the time at which matter in the solar system solidified. On the basis of such measurements the age of the sun is estimated to be 5 milliards of years, within about 10%. So gravitation is not sufficient to supply its energy over the ages.

Eddington, in the 1920's, investigated very thoroughly the interior constitution of the sun and other stars, and was much concerned about the sources of stellar energy. His favorite hypothesis was the complete annihilation of matter, changing nuclei and electrons into radiation. The energy which was to be set free by such a process, if it could occur, is given by the Einstein relation between mass and energy and is

$$c^2 = 9 \cdot 10^{20} \text{ erg/g} \quad (3)$$

This would be enough to supply the sun's radiation for 1500 milliards of years. However nobody has ever observed the complete annihilation of matter. From experiments on earth we know that protons and electrons do not annihilate each other in 10^{30} years. It is hard to believe that the situation would be different at a temperature of some 10 million degrees such as prevails in the stars, and Eddington appreciated this difficulty quite well.

From the early 1930's it was generally assumed that the stellar energy is produced by nuclear reactions. Already in 1929, Atkinson and Houtermans¹ concluded that at the high temperatures in the interior of a star, the nuclei in the star could penetrate into other nuclei and cause nuclear reactions, releasing energy. In 1933, particle accelerators began to operate in which such nuclear reactions were actually observed. They were found to obey very closely the theory of Gamow, Condon and Gurney, on the penetration of charged particles through potential barriers. In early 1938, Gamow and Teller² revised the theory of Atkinson and Houtermans on the rate of « thermonuclear » reactions, *i.e.* nuclear reactions occurring at high temperature. At the same time, von Weizsäcker³ speculated on the reactions which actually might take place in the stars.

In April 1938, Gamow assembled a small conference of physicists and astrophysicists in Washington, D. C. This conference was sponsored by the Department of Terrestrial Magnetism of the Carnegie Institution. At this Conference, the astrophysicists told us physicists what they knew about the internal constitution of the stars. This was quite a lot, and all their results had been derived without knowledge of the specific source of energy. The only assumption they made was that most of the energy was produced « near » the center of the star.

Properties of Stars

The most easily observable properties of a star are its total luminosity and its surface temperature. In relatively few cases of nearby stars, the mass of the star can also be determined.

Fig. 1 shows the customary Hertzsprung-Russell diagram. The luminosity, expressed in terms of that of the sun, is plotted against the surface temperature, both on a logarithmic scale. Conspicuous is the main sequence, going from upper left to lower right, i.e. from hot and luminous stars to cool and faint ones. Most stars lie on this sequence. In the upper right are the Red Giants, cool but brilliant stars. In the lower left are the White Dwarfs, hot but faint. We shall be mainly concerned with the main sequence. After being assembled, by gravitation, stars spend the most part of their life on the main sequence, then develop into red giants, and in the end, probably into white dwarfs.

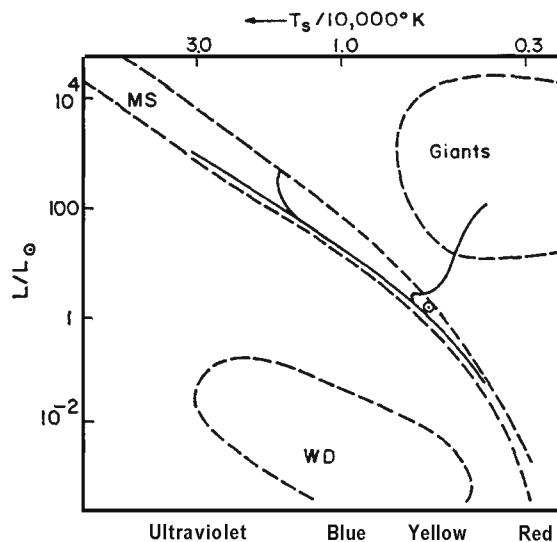


Fig.1. Hertzsprung-Russell diagram. From E.E.Salpeter, in *Apollo and the Universe*, Science Foundation for Physics, University of Sydney, Australia, 1967.

The figure shows that typical surface temperatures are of the order of 10^4 °K. Fig. 2 gives the relation between mass and luminosity in the main sequence. At the upper end, beyond about 15 sun masses, the mass determinations are uncertain. It is clear, however, that luminosity increases rapidly with mass.

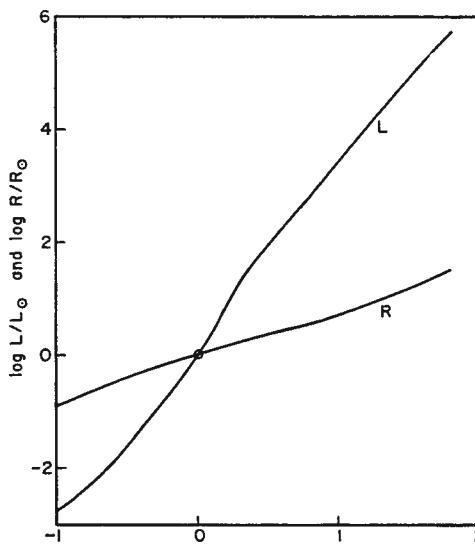


Fig. 2 Luminosity and radius of stars vs. mass. Abscissa is $\log M/M_\odot$. Data from C. W. Allen, *Astrophysical Quantities*, Athlone Press, 1963, p. 203. The curve for $\log L/L_\odot$ holds for all stars, that for R/R_\odot only for the stars in the main sequence. The symbol \odot refers to the sun.

For a factor of 10 in mass, the luminosity increases by a factor of about 3000, hence the energy production per gram is about 300 times larger.

To obtain information on the interior constitution of the stars, astrophysicists integrate two fundamental equations. Pioneers in this work have been Eddington, Chandrasekhar and Strömgren. The first equation is that of hydrostatic equilibrium

$$\frac{dP}{dr} = -GM(r)\frac{\varrho(r)}{r^2} \quad (4)$$

in which P is the pressure at distance r from the center, ϱ is the density and $M(r)$ is the total mass inside r . The second equation is that of radiation transport

$$\frac{1}{\varkappa\varrho} \frac{d}{dr}(\frac{1}{3}\alpha c T^4) = -\frac{L(r)}{4\pi r^2} \quad (5)$$

Here \varkappa is the opacity of the stellar material for black-body radiation of the local temperature T , α is the Stefan-Boltzmann constant, and $L(r)$ is the flux of radiation at r . The value of L at the surface R of the star is the luminosity.

In the stars we shall discuss, the gas obeys the equation of state

$$P = R \varrho / \mu \quad (6)$$

where R is the gas constant, while μ is the mean molecular weight of the stellar material. If X , Y and Z are respectively concentrations by mass of hydrogen, helium and all heavier elements, and if all gases are fully ionized, then

$$\mu^{-1} = 2X + \frac{3}{4}Y + \frac{1}{2}Z \quad (7)$$

In all stars except the very oldest ones, it is believed that Z is between 0.02 and 0.04; in the sun at present, X is about 0.65, hence $Y=0.33$ and $\mu=0.65$. In many stars the chemical composition, especially X and Y , vary with position r . The opacity is a complicated function of Z and T , but in many cases it behaves like

$$\kappa = C \varrho^{-3/4} \quad (8)$$

where C is a constant.

The integration of (4) and (5) in general requires computers. However an estimate of the central temperature may be made from the virial theorem which we mentioned in the beginning. According to this, the average thermal energy per unit mass of the star is one-half of the average potential energy. This leads to the estimate of the thermal energy per particle at the center of the star,

$$k T_c = \alpha \mu G H M/R \quad (9)$$

in which H is the mass of the hydrogen atom, and α is a constant whose magnitude depends on the specific model of the star but is usually about 1 for main sequence stars. Using this value, and (9), we find for the central temperature of the sun

$$T_{c6} = 14 \quad (10)$$

where T_c denotes the temperature in millions of degrees, here and in the following. A more careful integration of the equations of equilibrium by Demarque and Percy⁴ gives

$$T_{c6} = 15.7; \varrho_c = 158 \text{ g/cm}^3 \quad (11)$$

Originally Eddington had assumed that the stars contain mainly heavy elements, from carbon on up. In this case $\mu = 2$ and the central temperature is increased by a factor of 3, to about 40 million degrees; this led to contradictions with the equation of radiation flow, (5), if the theoretical value of the

opacity was used. Strömgren pointed out that these contradictions can be resolved by assuming the star to consist mainly of hydrogen, which is also in agreement with stellar spectra. In modern calculations the three quantities X , Y , Z , indicating the chemical composition of the star, are taken to be parameters to be fixed so as to fit all equations of stellar equilibrium.

Thermonuclear Reactions

All nuclei in a normal star are positively charged. In order for them to react they must penetrate each others Coulomb potential barrier. The wave mechanical theory of this shows that in the absence of resonances, the cross section has the form

$$a(E) = \frac{S(E)}{E} \exp\left(-\sqrt{\frac{E_G}{E}}\right) \quad (12)$$

where E is the energy of the relative motion of the two colliding particles, $S(E)$ is a coefficient characteristic of the nuclear reaction involved and

$$E_G = 2 M (\pi Z_0 Z_1 e^2 / \hbar)^2 = (2\pi Z_0 Z_1)^2 E_{\text{Bohr}} \quad (13)$$

Here M is the reduced mass of the two particles, Z_0 and Z_1 their charges, and E_{Bohr} is the Bohr energy for mass M and charge 1. (13) can be evaluated to give

$$E_G = 0.979 W \text{ MeV} \quad (14)$$

with

$$W = A Z_0^2 Z_1^2 \quad (14a)$$

$$A = A_0 A_1 / (A_0 + A_1) \quad (14b)$$

in which A_0 , A_1 are the atomic weights of the two colliding particles. For most nuclear reactions $S(E)$ is between 10 MeV-barns and 1 keV-barn.

The gas at a given r in the star has a given temperature so that the particles have a Boltzmann energy distribution. The rate of nuclear reactions is then proportional to

$$(8/\pi M)^{1/2} (k T)^{-3/2} \int \sigma(E) E \exp(-E/kT) dE \quad (15)$$

It is most convenient⁵ to write for the rate of disappearance of one of the reactants

$$dX_0/dt = -[o_1] X_0 X_1 \quad (16)$$

where X_o and X_i are the concentrations of the reactants by mass, and

$$[\text{OI}] = 7.8 \cdot 10^{11} (Z_0 Z_i / A)^{1/3} S_{\text{eff}} \rho T_6^{-2/3} e^{-\tau} \quad (17)$$

$$\tau = 42.487 (W/T_6)^{1/3} \quad (17a)$$

Since the reaction cross section (12) increases rapidly with energy, the main contribution to the reaction comes from particles which have an energy many times the average thermal energy. Indeed the most important energy is

$$E_0 = (\tau/3) k T \quad (18)$$

For $T = 13$ which is an average for the interior of the sun, we have

$$\begin{aligned} \tau/3 &= 4.7 \text{ for the reaction H + H} \\ &= 19 \text{ for the reaction C + H} \\ &= 25 \text{ for the reaction N + H} \end{aligned} \quad (19)$$

It is also easy to see from (17) that the temperature dependence of the reaction rate is

$$\frac{d \ln [\text{OI}]}{d \ln T} = \frac{\tau - 2}{3} \quad (20)$$

Nuclear Reactions in Main Sequence Stars

Evidently, at a given temperature and under otherwise equal conditions, the reactions which can occur most easily are those which have the smallest possible value of W (14a). This means that at least one of the interacting nuclei should be a proton, $A_o = Z_o = 1$. Thus we may examine the reactions involving protons.

The simplest of all possible reactions is

$$H + H = D + \varepsilon^+ + \nu \quad (21)$$

(ε^+ = positron, ν = neutrino).

This was first suggested by von Weizsäcker³, and calculated by Critchfield and Bethe⁶. The reaction is of course exceedingly slow because it involves the beta disintegration. Indeed the characteristic factor S is

$$S(E) = 3.36 \cdot 10^{-25} \text{ MeV-barns} \quad (22)$$

This has been derived on purely theoretical grounds, using the known coupling constant of beta disintegration; the value is believed to be accurate to 20% or better. There is no chance of observing such a slow reaction on earth, but

in the stars we have almost unlimited time, and a large supply of protons of high energy. As we shall see presently, the rate of energy production by this simple reaction fits the observed energy production in the sun very well.

The deuterons formed in (21) will quickly react further, and the end product is ${}^4\text{He}$. We shall discuss the reactions in more detail later on.

The proton-proton reaction (21), although it predicts the correct energy production in the sun, has a rather weak dependence on temperature. According to (19), (20), it behaves about as T^4 . Since central temperatures change only little from the sun to more massive stars, the energy production by this reaction does likewise. However as we have seen in Fig. 2, the observed energy production increases dramatically with increasing mass. Therefore there must exist nuclear reactions which are more strongly dependent on temperature; these must involve heavier nuclei.

Stimulated by the Washington Conference of April 1938, and following the argument just mentioned, I examined⁷ the reactions between protons and other nuclei, going up in the periodic system. Reactions between H and ${}^4\text{He}$ lead nowhere, there being no stable nucleus of mass 5. Reactions of H with Li, Be and B, as well as with deuterons, are all very fast at the central temperature of the sun, but just this speed of the reaction rules them out: the partner of H is very quickly used up in the process. In fact, and just because of this reason, all the elements mentioned, from deuterium to boron, are extremely rare on earth and in the stars, and can therefore not be important sources of energy.

The next element, carbon, behaves quite differently. In the first place, it is an abundant element, probably making up about 1% by mass of any newly formed star. Secondly, in a gas of stellar temperature, it undergoes a cycle of reactions, as follows



Reactions a, c, and d are radiative captures; the proton is captured by the nucleus and the energy emitted in the form of gamma rays; these are then quickly converted into thermal energy of the gas. For reactions of this type, $S(E)$ is of the order of 1 keV-barn. Reactions b and e are simply spontaneous beta decays, with lifetimes of 10 and 2 min respectively, negligible in com-

parison with stellar times. Reaction f is the most common type of nuclear reaction, with 2 nuclei resulting from the collision; $S(E)$ for such reactions is commonly of the order of MeV- barns.

Reaction f is in a way the most interesting because it closes the cycle: we reproduce the ^{12}C which we started from. In other words, carbon is only used as a catalyst; the result of the reaction is a combination of 4 protons and 2 electrons⁸ to form one ^4He nucleus. In this process two neutrinos are emitted, taking away about 2 MeV energy together. The rest of the energy, about 25 MeV per cycle, is released usefully to keep the sun warm.

Making reasonable assumptions of the reaction strength $S(E)$, on the basis of general nuclear physics, I found in 1938 that the carbon-nitrogen cycle gives about the correct energy production in the sun. Since it involves nuclei of relatively high charge, it has a strong temperature dependence, as given in (19). The reaction with ^{14}N is the slowest of the cycle and therefore determines the rate of energy production; it goes about as T^{24} near solar temperature. This is amply sufficient to explain the high rate of energy production in massive stars⁹.

Experimental Results

To put the theory on a firm basis, it is important to determine the strength factor $S(E)$ for each reaction by experiment. This has been done under the leadership of W.A. Fowler¹⁰ of the California Institute of Technology in a monumental series of papers extending over a quarter of a century. Not only have all the reactions in (23) been observed, but in all cases $S(E)$ has been accurately determined.

The main difficulty in this work is due to the resonances which commonly occur in nuclear reactions. Fig. 3 shows the cross section of the first reactions (23a), as a function of energy. The measured cross sections extend over a factor of 10^7 in magnitude; the smallest ones are 10^{-11} barns = 10^{-35} cm^2 and therefore clearly very difficult to observe. The curve shows a resonance at 460 keV. The solid curve is determined from nuclear reaction theory, on the basis of the existence of that resonance. The fit of the observed points to the calculated curve is impressive. Similar results have been obtained on the other three proton-capture reactions in (23).

On the basis of Fig. 3 we can confidently extrapolate the measurements to lower energy. As we mentioned in (18) the most important energy contribut-

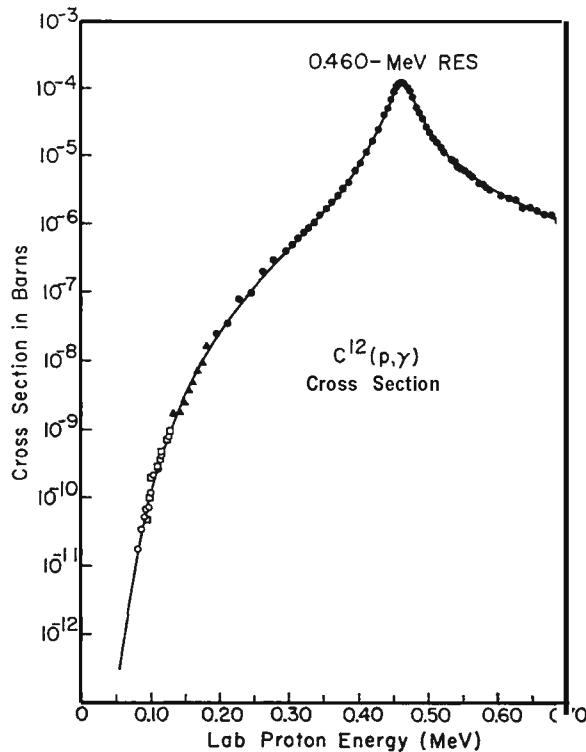
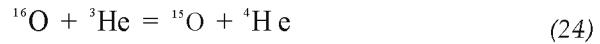


Fig. 3. Cross section for the reaction $^{12}\text{C} + \text{H}$, as a function of the proton energy. From Fowler, Caughlan and Zimmerman⁵.

ing to the reaction rate is about $20 kT$. For $T_e = 13$, we have $kT = 1.1 \text{ keV}$; so we are most interested in the cross section around 20 keV. This is much too low an energy to observe the cross section in the laboratory; even at 100 keV, the cross section is barely observable. So quite a long extrapolation is required. This can be done with confidence provided there are no resonances close to $E = 0$. Therefore a great deal of experimental work has gone into the search for such resonances.

The resonances exist of course in the compound nucleus, *i.e.* the nucleus obtained by adding the two initial reactants. To find resonances near the threshold of the reactions (23), it is necessary to produce the same compound nucleus from other initial nuclei, e.g., in the reaction between ^{14}N and H, the compound nucleus ^{15}O is formed. To investigate its levels Hensley¹¹ at CalTech studied the reaction



He found indeed a resonance 20 keV below the threshold for $^{14}\text{N} + \text{H}$ which in principle might enhance the process (23d). However the state in ^{15}O was found to have a spin $J= 7/2$. Therefore, even though ^{14}H has $J= 1$ and the proton has a spin of $1/2$, we need at least an orbital momentum $\lambda= 2$ to reach this resonant state in ^{15}O . The cross section for such a high orbital momentum is reduced by at least a factor 10^4 , compared to $\lambda=0$, so that the near-resonance does not in fact enhance the cross section $^{14}\text{N}+\text{H}$ appreciably. This cross section can then be calculated by theoretical extrapolation from the measured range of proton energies, and the same is true for the other reactions in the cycle (23).

On this basis, Fowler and others have calculated the rate of reactions in the CN cycle. A convenient tabulation has been given by Reeves¹²; his results are plotted in Fig. 4. This figure gives the energy production per gram per second as a function of temperature. We have assumed $X= 0.5$, $Z= 0.02$.

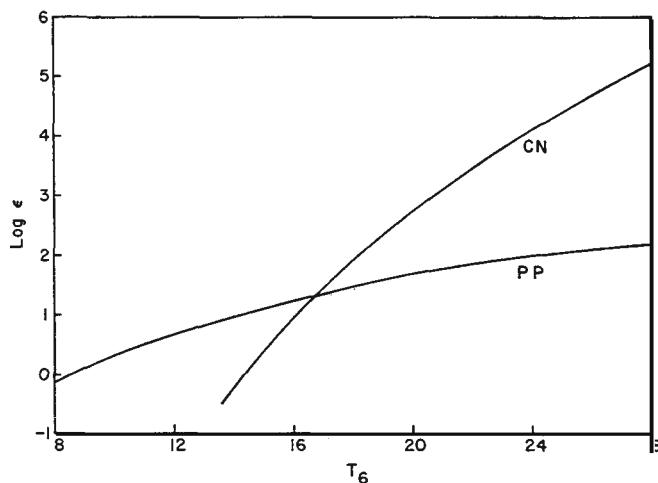
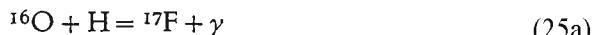


Fig. 4. The energy production, in erg/g sec as a function of the temperature in millions of degrees. For the proton-proton reaction (PP) and the carbon-nitrogen cycle (CN). Concentrations assumed $X= Y= 0.5$, $Z= 0.02$. Calculated from Tables 8 and 9 of Reeves¹².

The figure shows that at low temperature the $\text{H} + \text{H}$ reaction dominates, at high temperatures the $\text{C} + \text{N}$ cycle; the crossing point is at $T_6= 13$; here the energy production is 7 erg/g sec. The average over the entire sun is obviously smaller, and the result is compatible with an average production of 2 erg/g sec.

The energy production in the main sequence can thus be considered as well understood.

An additional point should be mentioned. Especially at higher temperature, when the CN cycle prevails, there is also a substantial probability for the reaction chain



This chain is not cyclic but feeds into the CN cycle. It is customary to speak of the whole set of reactions as the CNO bi-cycle. The effect of reactions (25) is that ^{16}O initially present will also contribute to the reactants available, and thus increase the reaction rate of the CN cycle somewhat. This has been taken into account in Fig. 5.

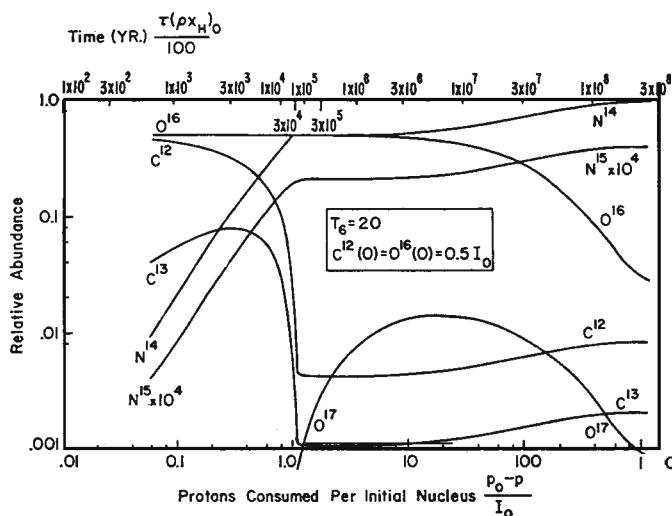


Fig. 5. Variation with time of the abundances of various elements involved in the CNO cycle. It is assumed that initially ^{12}C and ^{16}O have the same abundance while that of ^{14}N is small. From G.R. Caughlan, *Astrophys. J.* (1967).

If equilibrium is established in the CNO bi-cycle, eventually most of the nuclei involved will end up as ^{14}N because this nucleus has by far the longest lifetime against nuclear reactions. There is no observable evidence for this; in fact wherever the abundance can be observed, C and O tend to be at least

as abundant as N. However this is probably due to the fact that the interior of a star stays well separated from its surface; there is very little mixing. Astrophysicists have investigated the circumstances when mixing is to be expected, and have found that surface abundances are quite compatible with these expectations. In the interstellar material which is used to form stars, we have reason to believe that C and O are abundant and N is rare. This will be discussed later.

The Completion of the Proton-Proton Chain

The initial reaction (21) is followed almost immediately by



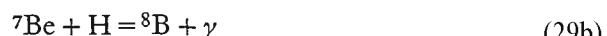
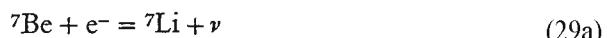
The fate of ^3He depends on the temperature. Below about $T_c = 15$, the ^3He builds up sufficiently so that such nuclei react with each other according to



This reaction has an unusually high $S(E) = 5$ MeV- barns⁵. At higher temperature, the reaction



competes favorably with (27). The ^7Be thus formed may again react in one of two ways



At about $T_c = 20$, reaction (29b) begins to dominate over (29a). (29b) is followed by (29c) which emits neutrinos of very high energy. Davies¹³, at Brookhaven, is attempting to observe these neutrinos.

Evolution of a Star

A main sequence star uses up its hydrogen preferentially near its center where nuclear reactions proceed most rapidly. After a while, the center has lost al-

most all its hydrogen. For stars of about twice the luminosity of the sun, this happens in less than 10^{10} years which is approximately the age of the universe, and also the age of stars in the globular clusters. We shall now discuss what happens to a star after it has used up the hydrogen at the center. Of course, in the outside regions hydrogen is still abundant.

This evolution of a star was first calculated by Schwarzschild¹⁴ who has been followed by many others; we shall use recent calculations by Iben¹⁵. When hydrogen gets depleted, not enough energy is produced near the center to sustain the pressure of the outside layers of the star. Hence gravitation will cause the center to collapse. Thereby, higher temperatures and densities are achieved. The temperature also increases farther out where there is still hydrogen left, and this region now begins to burn. After a relatively short time, a shell of H , away from the center, produces most of the energy; this shell gradually moves outward and gets progressively thinner as time goes on.

At the same time, the region of the star outside the burning shell expands. This result follows clearly from all the many numerical computations on this subject. The physical reason is not clear. One hypothesis is that it is due to the discontinuity in mean molecular weight: Inside the shell, there is mostly helium, of $\gamma = 4/3$, outside we have mostly hydrogen, and $\gamma = 0.65$. Another suggestion is that the flow of radiation is made difficult by the small radius

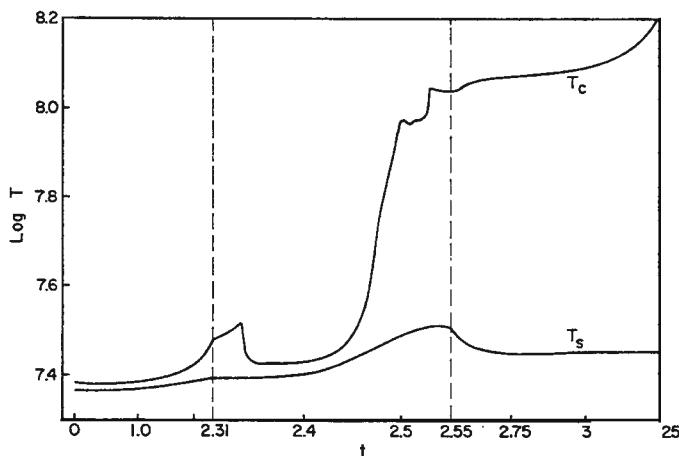


Fig. 6. Evolution of a star of 3 solar masses, according to I. Iben, *Astrophys.J.*, 142 (1965) 1447. Abscissa is time in units of 10^9 years (note the breaks in scale at $t = 2.31$ and 2.55). I. Temperature (on logarithmic scale) : T_c = temperature at center of star, T_s = same at mid-point of source of energy generation, which, after $t = 2.48$ is a thin shell. T_c increases enormously, T_s stays almost constant.

of the energy source, and that this has to be compensated by lower density just outside the source.

By this expansion the star develops into a red giant. Indeed, in globular clusters (which, as I mentioned, are made up of very old stars), all the more luminous stars are red giants. In the outer portion of these stars, radiative transport is no longer sufficient to carry the energy flow; therefore convection of material sets in in these outer regions. This convection can occupy as much as the outer 80% of the mass of the star; it leads to intimate mixing of the material in the convection zone.

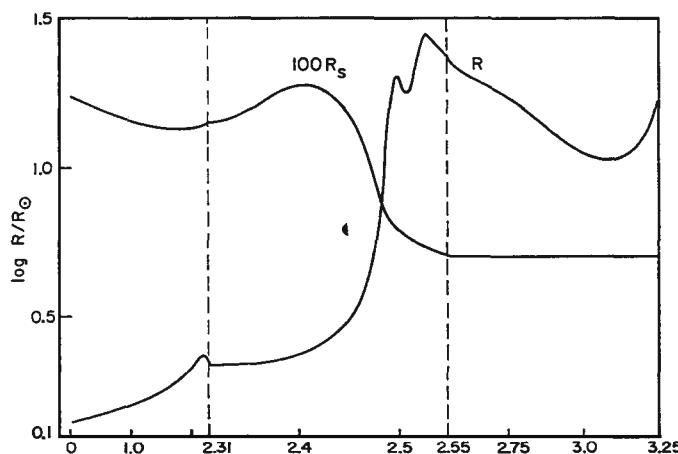


Fig. 7. Evolution of a star, (see caption to Fig. 6). II Radius, in units of that of the sun, on logarithmic scale. R = total radius, $100 R_s$ = 100 times the radius of mid-point of energy source. R increases tremendously, while R_s shrinks somewhat.

Iben¹⁵ has discussed a nice observational confirmation of this convective mixing. The star Capella is a double star, each component having a mass of about 3 solar masses, and each being a red giant. The somewhat lighter star, «Capella F» (its spectral type is F) shows noticeable amounts of Li in its spectrum, while the somewhat heavier Capella G shows at least 100 times less Li. It should be expected that G, being heavier, is farther advanced in its evolution. Iben now gives arguments that the deep-reaching convection and mixing which we just discussed, will occur just between the evolution phases F and G. By convection, material from the interior of the star will be carried to the surface; this material has been very hot and has therefore burned up its Li. Before deep convection sets in (in star F) the surface Li never sees high temperature and thus is preserved.

Following the calculations of Iben we have plotted in Figs. 6-9 the development of various important quantities in the history of a star of mass= 3 solar masses. The time is in units of 10^8 years. Since the developments go at very variable speed, the time scale has been broken twice, at $t= 2.31$ and $t= 2.55$. In between is the period during which the shell source develops.

During this period the central temperature rises spectacularly (Fig. 6) from about $T_c = 25$ to $T_c = 100$. At the same time the radius increases from about 2 to 30 solar radii; subsequently, it decreases again to about 15 (Fig. 7). The central density, starting at about 40, increases in the same period to about $5 \cdot 10^4$ (Fig. 8). The luminosity (Fig. 9) does not change spectacularly, staying always between 100 and 300 times that of the sun.

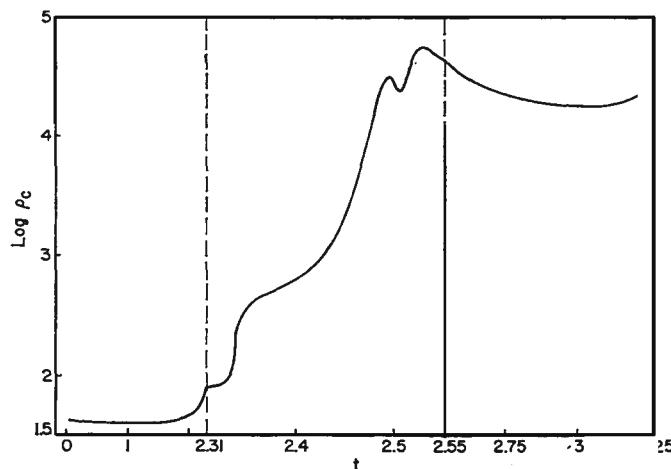


Fig. 8. Evolution of a star (see caption to Fig. 6).111. Density, on logarithmic scale, at the center of the star. This quantity increases about 1000-fold.

While the inside and the outside of the star undergo such spectacular changes, the shell in which the hydrogen is actually burning, does not change very much. Fig. 9 shows m , the fraction of the mass of the star enclosed by the burning shell. Even at the end of the calculation, $t= 3.25$, this is only $m= 0.2$. This means that only 20% of the hydrogen in the star has burned after all this development. Fig. 6, curve T_s , shows the temperature in the burning shell which stays near 25 million degrees all the time. Fig. 7, curve R_s , shows the radius of the shell, in units of the solar radius; during the critical time when the shell is formed this radius drops from about 0.15 to 0.07. This is of course the mechanism by which the shell is kept at the temperature which originally prevailed at the center.

In the meantime, the temperature at the center increases steadily. When it reaches about $T_e = 100$, the ^4He which is abundant at the center, can undergo nuclear reactions. The first of these, which occurs at the lowest temperature (about $T_e = 90$) is

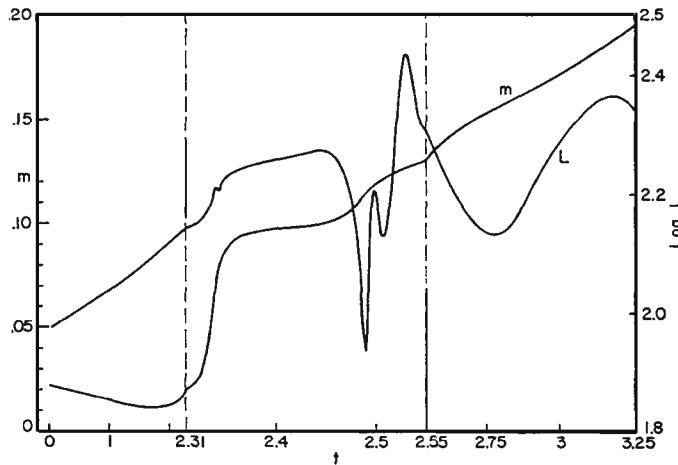


Fig. 9. Evolution of a star (see caption to Fig. 6). IV. Curve L , luminosity relative to that of the sun, on logarithmic scale. This quantity does not change very much during the life of the star. Curve m , fraction of the mass of the star enclosed by energy-producing shell, on linear scale. This fraction increases slowly with time.

While this reaction goes on, the central temperature remains fairly constant. However, there is not much ^{14}N so the reaction soon stops (after about $0.02 \cdot 10^8$ years), and the center contracts further.

The next reaction makes use entirely of the abundant ^4He , *viz.*



This reaction has the handicap of requiring a simultaneous collision of 3 alpha particles. This would be extremely unlikely were it not for the fact that it is favored by a *double* resonance. Two alpha particles have nearly the same energy as the unstable nucleus ^9Be , and further $^9\text{Be} + ^4\text{He}$ has almost the same energy as an excited state of ^{12}C . This reaction can of course not be observed in the laboratory but the two contributing resonances can be. The importance of the first resonance was first suggested by Salpeter¹⁶, the second by Hoyle¹⁷. Recent data indicate that (31) requires a temperature of about $T_e = 110$, at the

central densities corresponding to $t=2.5$, i.e. $\rho_c > 10^4$. Once this reaction sets in, the central temperature does not rise very fast any more.

Reaction (31) is most important for the buildup of elements. Early investigators^{3,7} had great trouble with bridging the gap between ${}^4\text{He}$ and ${}^{12}\text{C}$. Two nuclei in this gap, mass 5 and mass 8, are completely unstable, the rest disintegrate in a very short time under stellar conditions. Reaction (31) however leads to stable ${}^{12}\text{C}$. This nucleus can now capture a further alpha particle



the temperatures required for this are about the same as for (31). There is also some capture of alpha particles by ${}^{16}\text{O}$ leading to ${}^{20}\text{Ne}$, but the next step, ${}^{20}\text{Ne} \rightarrow {}^{24}\text{Mg}$, cannot occur appreciably at these temperatures; instead, the helium gets used up in forming ${}^{12}\text{C}$, ${}^{16}\text{O}$ and some ${}^{20}\text{Ne}$.

Helium is depleted first in the center, and now the same process repeats which previously took place with hydrogen. A shell of burning He is formed, at a smaller radius than the H shell, and of course at a higher temperature. The center of the star now contracts further by gravitation and reaches still higher temperatures.

Buildup and Dispersal of Elements

The further developments of a massive star are more speculative. However the theory of Hoyle and collaborator¹⁸ is likely to be correct.

The center of the star heats up until the newly formed carbon nuclei can react with each other. This happens at a temperature of roughly 10°degrees. Nuclei like ${}^{24}\text{Mg}$ or ${}^{28}\text{Si}$ can be formed. There are also more complicated mechanisms in which we first have a capture reaction with emission of a gamma ray, followed by capture of this gamma ray in another nucleus which releases ${}^4\text{He}$. This ${}^4\text{He}$ can then enter further nuclei and build up the entire chain of stable nuclei up to the most stable Fe. Not much energy is released in all of these processes.

The center of the star contracts further and gets still hotter. At very high temperatures, several milliards of degrees, thermal equilibrium is no longer strongly in favor of nuclei of the greatest binding energy. Instead, endothermic processes can take place which destroy some of the stable nuclei already formed. In the process, alpha particles, protons and even neutrons may be released. This permits the buildup of elements beyond Fe, up to the top of the

periodic table. Because of the high temperatures involved all this probably goes fairly fast, perhaps in thousands of years.

During this stage, nuclear processes tend to consume rather than release energy. Therefore they no longer oppose the gravitational contraction so that contraction continues unchecked. It is believed that this will lead to an unstable situation. Just as the first contraction, at the formation of the H shell source, led to an expansion of the outer envelope of the star, a similar outward expansion is expected now. But time scales are now short, and this expansion may easily be an explosion. Hoyle *et al.*¹⁸ have suggested this as the mechanism for a supernova.

In a supernova explosion much of the material of the star is ejected into interstellar space. We see this, e.g., in the Crab Nebula. The ejected material probably contains the heavy elements which have been formed in the interior of the massive star. Thus heavy elements get into the interstellar gas, and can then be collected again by newly forming stars. It is believed that this is the way how stars get their heavy elements. This means that most of the stars we see, including our sun, are at least second generation stars, which have collected the debris of earlier stars which have suffered a supernova explosion.

To clinch this argument it must be shown that heavy elements cannot be produced in other ways. This has indeed been shown by Fowler¹⁹. He has investigated the behavior of the enormous gas cloud involved in the original « Big Bang », and its development with time. He has shown that temperatures and densities, as functions of time, are such that heavy elements beginning with C cannot be produced. The only element which can be produced in the big bang is ^4He .

If all this is true, stars have a life cycle much like animals. They get born, they grow, they go through a definite internal development, and finally they die, to give back the material of which they are made so that new stars may live.

I am very grateful to Professor E. E. Salpeter for his extensive help in preparing this paper.

1. R. d'E. Atkinson and F. G. Houtermans, *Z. Physik*, 54 (1929) 656.
2. G. Gamow and E. Teller, *Phys. Rev.*, 53 (1938) 608.
3. C.F. von Weizsäcker, *Physik. Z.*, 38 (1937) 176.

4. P.R. Demarque and J.R. Percy, *Astrophys. J.*, 140 (1964) 541.
5. W. A. Fowler, G. R. Caughlan and B.A. Zimmerman, *Ann. Rev. Astron. Astrophys.*, 5 (1967) 525 .
6. H. A. Bethe and C. L. Critchfield, *Phys. Rev.*, 54 (1938) 248.
7. H.A. Bethe, *Phys. Rev.*, 55 (1939) 436.
8. The electrons are used to annihilate the positrons emitted in reactions b and e.
9. The carbon-nitrogen cycle was also discovered independently by C. F. von Weizsäcker, *Physik. Z.*, 39 (1938) 633, who recognized that this cycle consumes only the most abundant element, hydrogen. But he did not investigate the rate of energy production or its temperature dependence.
10. W. A. Fowler, many papers in *Phys. Rev., Astrophys. J.* and other publications. Some of this work is summarized in ref. 5.
11. D. C. Hensley, *Astrophys. J.*, 147 (1967) 818.
12. H. Reeves, in *Stellar Structure*, Vol. 8 of G.P. Kuiper (Ed.), *Stars and Stellar Systems*, University of Chicago Press, Chicago, Ill., 1965, especially Tables 8 and 9.
13. R. Davies Jr., *Phys. Rev. Letters*, 12 (1964) 303.
14. M. Schwarzschild, *Structure and Evolution of the Stars*, Princeton University Press, 1958.
15. I. Iben Jr., *Astrophys. J.*, 141 (1965) 993, 142 (1965) 1447, 143 (1966) 483.
16. E.E. Salpeter, *Phys. Rev.*, 88 (1952) 547.
17. F. Hoyle, *Astrophys. J.*, Suppl. 1 (1954) 121.
18. E. M. Burbidge, G. R. Burbidge, W. A. Fowler and F. Hoyle, *Rev. Mod. Phys.*, 29 (1957) 547.
19. W.A. Fowler, *Internat. Ass. Geochem. Cosmochem., 1st Meeting, Paris*, 1967.

KONRAD BLOCH

The biological synthesis of cholesterol

Nobel Lecture, December 11, 1964

In the early 1930's after decades of effort, the structural elucidation of cholesterol had reached the stage of completion and with this achievement one of the most brilliant chapters of organic chemistry came to a close. To chemists and biochemists alike the structure of cholesterol at once presented a biosynthetic problem of exceptional challenge. At first sight the molecular architecture of cholesterol seemed enigmatic and devoid of any clues as to how this complex molecule might be constructed from the smaller molecules available in the cell. It is, therefore, all the more remarkable that much of the early speculative thinking came so close to predicting some of the general principles that were eventually shown to operate in sterol biosynthesis. The early proposals were mainly concerned with the problem of ring formation and quite uniformly they envisioned an origin of the tetracyclic steroid ring system from an appropriately folded long-chain precursor. This intuition proved to be correct. All of the speculative schemes have in some way influenced the research that was to take place later, but none equaled in perspicacity L. Ruzicka's unifying hypothesis on the common origin of terpenes and steroids and the suggestion by Sir Robert Robinson that cholesterol might arise by cyclization of the hydrocarbon squalene. These bold and appealing ideas carried perhaps more weight because they had some experimental evidence to support them. Already in 1926 Shannon, following a proposal by Heilbron, Kamm and Owens¹, had shown that squalene fed to animals increases the cholesterol content of the tissues².

The phase of continuous research on cholesterol biosynthesis begins in 1937 with two independent and remarkably complementary investigations. In the course of their pioneering studies on intermediary metabolism with the aid of stable isotopes, Rittenberg and Schoenheimer at Columbia arrived at the conclusion that the process of cholesterol formation involves the coupling of smaller molecules, "possibly those which have been postulated to be intermediates in the fat and carbohydrate metabolism"³. During the same year Sonderhoff and Thomas noting an incorporation of trideuterioacetate into the unsaponifiable materials of yeast, stated: "Es lässt sich darauf schliessen

dass die Sterine der Hefe auf einem recht unmittelbaren Wege aus der Essigsäure entstehen"⁴.

As a graduate student under Hans T. Clarke at Columbia and later as an assistant of R. Schoenheimer, I was introduced to the new and powerful tool of isotopic tracers and shared the intellectual excitement of applying it to problems of biosynthesis, a previously closed area of metabolic investigation. Schoenheimer, a pathologist and biochemist, had long been interested in the origin and metabolism of cholesterol and planned to pursue this problem actively when his brilliant career came to an untimely end in 1941. Taking up the promising leads which already existed, Rittenberg and I began a systematic study of the utilization of labeled acetic acid for cholesterol synthesis in animal tissues and could soon show that acetate contributes in a major way to the synthesis of both the aliphatic side chain and of the tetracyclic moiety of the sterol molecules^{5,6}. After my move to the University of Chicago these studies were continued with the objective of establishing the origin of all carbon atoms of the cholesterol skeleton and with the hope that the pattern of distribution would be informative. Individual positions of the isooctyl side chain and the angular methyl groups were relatively accessible to chemical degradation and analysis, and with information available for only a few carbon atoms, we ventured the prediction that a two-carbon metabolite of acetate is the principal if not the sole building block of cholesterol⁷ (Fig. 1).

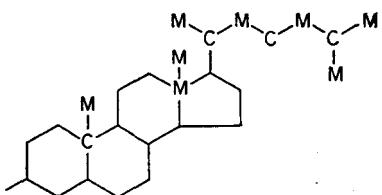


Fig. 1. Origin of some of the carbon atoms of cholesterol from methyl (M) and carboxyl (C) carbon atoms of acetic acid. (Wuersch, Huang and Bloch, 1952)

The arguments were largely deductive and therefore more convincing evidence for the exclusive origin of the sterol molecule from acetate was sought. A mutant of *Neurospora crassa*, which E. L. Tatum had isolated, served this purpose admirably because a deficiency in pyruvate metabolism made the growth of the mutant dependent on exogenous acetate. Cells of the mutant strain grown on labeled acetate produced ergosterol essentially without dilution of the isotope and this proved that no other carbon source contributed significantly to the synthesis of the sterol skeleton⁸.

One might ask why the biochemical mutant technique which has contributed so decisively to the charting of biosynthetic pathways has not been employed more often in studying sterol biosynthesis. Attempt to use it may have been made but apparently without success. Sterol-less mutants of molds or fungi have never been reported and any search for bacterial mutants would have been in vain because the pathways of sterol biogenesis do not exist in the bacterial phylum. Nevertheless, mutant strains of bacteria have in one instance been exceedingly important, if inadvertently, for the discovery of a key sterol precursor.

The thinking in our laboratory about modes for constructing larger molecules from acetate units was greatly influenced by information which had come from studies on rubber biosynthesis. Bonner and Arreguin had demonstrated the utilization of acetate for the biosynthesis of this isoprene polymer and they speculated how three acetate molecules might combine to form the requisite isoprenoid subunits for the macromolecule by way of acetoacetate and β -methylcrotonic acid⁹ (Fig. 2). Assuming that their scheme was also applicable to the construction of cholesterol, we could predict how isotopes from the two carbons of acetate should be arranged in certain portions of the molecule. For the isoctyl side chain of cholesterol the distribution pattern could be checked experimentally¹⁰. Prediction and experimental fact agreed satisfactorily, and therefore the view evolved quite naturally that cholesterol like many other natural substances was derived from a polyisoprenoid intermediate (Fig. 3). For this view to take hold, the ground was, of course, well prepared by Robinson's hypothesis¹¹, according to which cholesterol was formed by the cyclization of squalene, a polyisoprenoid hydrocarbon*. At this stage

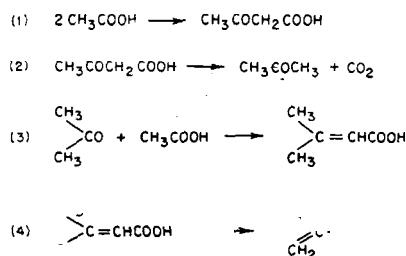


Fig. 2. Proposed reactions for the formation of isoprene from acetic acid. (Bonner and Arreguin 1949)

* As Ruzicka points out¹² a possible relation between the triterpenes and the sterols was discussed in his laboratory as early as 1925: "The hypothesis may be formulated that the steroids and the triterpenes have at least partially a common origin" (E. A. Rudolph, Ph. D., thesis, Eidgenössische Technische Hochschule, Zürich, 1925).

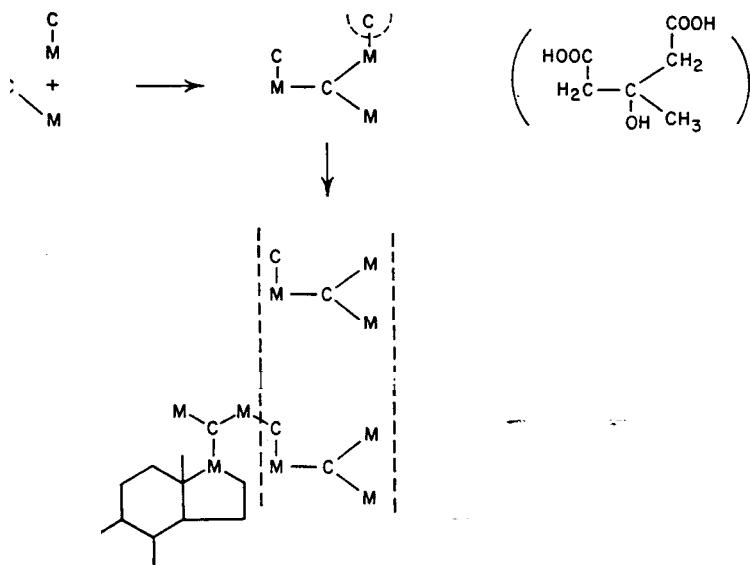


Fig. 3. Predicted and experimental distribution of acetate carbon atoms in cholesterol formed from isoprenoid precursors. Shown in the upper right corner, β -hydroxy- β -methylglutaric acid; see p. 86.

an outline of the major stages of the overall process emerged: acetate \rightarrow isoprenoid intermediate squalene \rightarrow cyclization product \rightarrow cholesterol.

In order to verify the general hypothesis, I attempted to demonstrate the formation of squalene from labeled acetate in the shark, an animal species which, for reasons yet unknown, accumulates this hydrocarbon in unusually large amounts. The project involved an excursion into comparative biochemistry for which the Biological Station in Bermuda was chosen. Experimentation with intact sharks or with the lipid-rich shark liver posed, however, considerable technical difficulties and I failed in the desired objective. Fortunately, the more realistic decision to study squalene synthesis in the liver of the rat succeeded in the laboratory at Chicago in the skillful hands of R. G. Langdon¹³. Once available in labeled form, squalene could readily be shown to serve as a precursor of cholesterol in the intact animal¹⁴. Only much later (in 1958) was the squalene-sterol conversion rigorously proven with authentic, chemically synthesized [¹³C] all-*trans* squalene¹⁵.

The available information thus left little doubt about the key role of squalene in sterol biosynthesis, but whether the cyclization proceeded by folding of the hydrocarbon chain as Robinson had specified could not yet be answered. It was obviously necessary to know whether the arrangement of

acetate carbons in the steroid skeleton conformed with prediction as it did in the sterol side chain. This involved the formidable chemical task of a complete carbon by carbon dissection of the cholesterol nucleus, eventually achieved by the elegant and definitive experiments of Cornforth and Popják¹⁶⁻¹⁸ between 1953 and 1957 (Fig.4). In the meantime the Robinson hypothesis needed to be examined more closely for different reasons. Lanosterol, the sterol from wool fat, had been known for some time to combine in its structure features of both the sterols and of the pentacyclic triterpenes, but it was not until 1952 that Ruzicka, Jeger and their collaborators furnished unequivocal proof for the 4,4,14-trimethylcholestane structure of lanosterol¹⁹. I learned of this development in the fall of 1952 after a lecture I presented to the Chemistry Department at Harvard. In a discussion that followed - joined by Professor E. R. H. Jones - Professor R. B. Woodward argued incisively for a novel type of squalene folding (Fig. 5). The appeal of his proposal was that it uniquely rationalized the structure of lanosterol. It also placed lanosterol in the pathway between squalene and cholesterol, at least by inference. One major and immediately verifiable consequence of the mechanism suggested by Woodward was that it predicted an alternate arrangement of acetate carbons in the steroid molecule differing from the pattern called for by Robinson's scheme in four of the twenty-seven positions, (at C₇, C₈, C₁₂ and C₁₃) (Fig. 6). Returning to the Chicago laboratory I was able to confirm the origin of C₁₃ from a methyl group of acetate in accordance with prediction. This welcome result encouraged us to commit ourselves to the new and much more plausible scheme for the cyclization of squalene²⁰. A similar proposal was made independently by Dauben *et. al.*²¹ shortly thereafter. The "correct" origin of C₇, another of the critical carbon atoms, was demonstrated²² later* and the evidence became

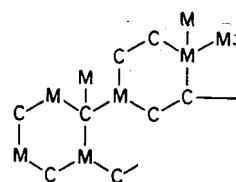


Fig. 4. Distribution of methyl (M) and carboxyl (C) carbon atoms of acetic acid in the nucleus of cholesterol. (Cornforth and Popják, 1953-1957)

*I made this finding while on Sabbatical leave at the Eidgenössische Technische Hochschule, Zürich, Switzerland, in 1953. During this stay I had the benefit of numerous stimulating discussions on terpene-sterol relationships with Professors Ruzicka, Prelog, Jeger and their associates. Their wisdom and counsel were invaluable for the later progress of our research.

CYCLIZATION OF SQUALENE

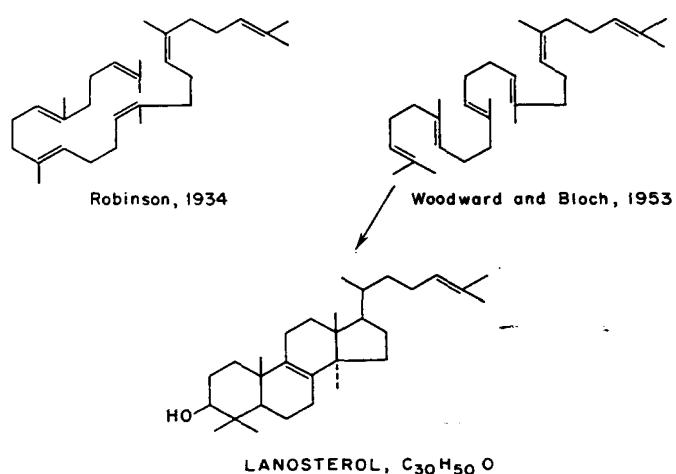


Fig. 5. Schemes for the cyclization of squalene and the relation of squalene to lanosterol.
(Jeger *et al.*, 1952)

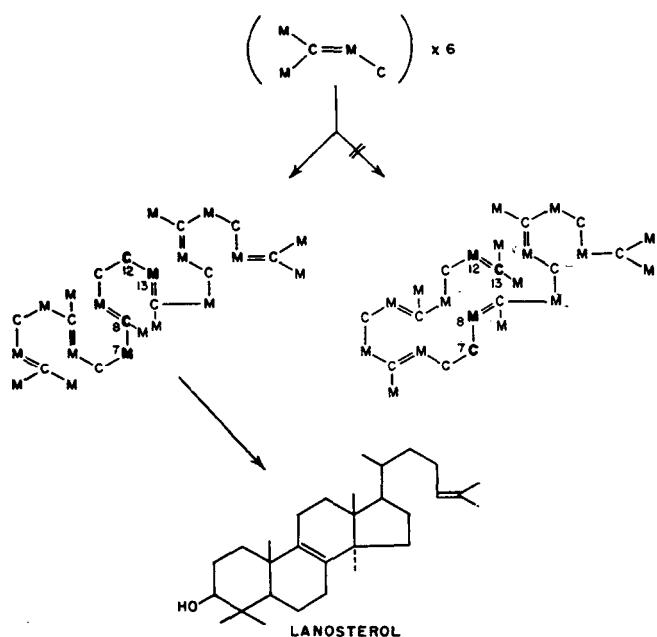


Fig. 6. Alternate arrangements of acetate carbons in sterol precursors predicted by the two cyclization schemes. The four affected positions are shown in bold type.

complete when Cornforth and Popják identified C₈ and C₁₂ in the course of their comprehensive degradation of the cholesterol nucleus. The same investigators greatly strengthened the, as yet, loose and speculative fabric by showing that squalene derived from methyl-labeled acetate was indeed isotopically constituted as the hypothesis had predicted^{2,3}. The essential demonstration of the conversion of squalene to lanosterol^{2,4} and of lanosterol to cholesterol came in due course²⁵. In these experiments we followed the directions of N. L. R. Bucher for preparing liver homogenates²⁶ a technique which proved invaluable in all subsequent studies on sterol biogenesis.

In 1953, Ruzicka and his colleagues published the first of two classical papers on the origin of terpenes and sterols and on the biogenetic relationships which the various structures suggested. Numerous natural products, some clearly related, others related in a much less obvious way, were now connected by plausible mechanisms and revealed as stemming from the main trunk or from the branches of a single biogenetic tree¹². The views of the Zürich school were developed further in 1955 in a publication²⁷ which Cornforth²⁸ has aptly referred to as "the apotheosis of the isoprene rule". The stereochemical arguments developed here led, *inter alia*, to a greatly refined formulation for the mechanism by which squalene cyclizes to lanosterol (Fig. 7). Three of their

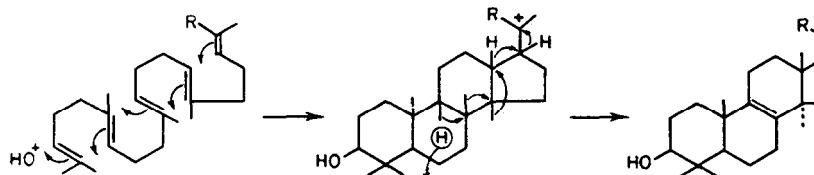


Fig. 7. Mechanism for the cyclization of squalene to lanosterol (refs. 12, 27).

main postulates were of special interest to us: (1) cyclization is initiated by a formal cation OH⁺ attacking the position of squalene which becomes C₃ of the sterol nucleus, (2) cyclization results from a series of interactions of electrophilic centers with electrons of suitably disposed double bonds, affording lanosterol without stabilization of products at any intermediary stage, and (3) once the tenacyclic ring system is established, two hydrogen atoms and two methyl groups migrate to neighboring positions in a concerted rearrangement that terminates in lanosterol. Submitting these postulates to experimental test we obtained results which were in each instance in accordance with theory. T. T. Tchen found that molecular oxygen, the formal equivalent of OH⁺, and not water is the source of the 3-hydroxyl group of lanosterol.

Furthermore, his experiments showed that the enzymatic cyclization of squalene to lanosterol in a D₂O medium proceeds without attachment of deuterium to carbon as it should if any (but the last) in the series of presumptive carbonium ions fail to stabilize²⁹.

Another predictable feature of the squalene-lanosterol transformation was the relocation of two branched methyl groups in the squalene skeleton. This could occur by a single 1,3 methyl shift or more likely by the migration of two methyl groups to their adjacent carbons atoms. The problem was resolved in favor of two 1,2 methyl shifts (from C₈ to C₁₄ and from C₁₄ to C₁₃) by Cornforth and Popják and their associates³⁰, and independently by us¹⁵. The two groups followed separate approaches but both exploited the fact that mass spectrometry enables one to distinguish the masses of molecules differing by one or more isotopic atoms. In our laboratory the problem of methyl migration was studied with synthetic all-*trans* squalene³¹ suitably labeled with ¹³C. This was transformed enzymatically into lanosterol and the product exhaustively oxidized to acetic acid which in turn was converted to ethylene for mass spectrometric analysis. Briefly, the results demonstrated that the methyl group at C₁₄ of lanosterol was not the same group that had originally been attached to the corresponding position of squalene but that it had come there from a neighboring position by a 1,2 methyl shift. It logically followed that the second of the shifting methyl groups also migrates to an adjacent position. The English workers used for the same purpose [¹³C] mevalonic acid which they converted to cholesterol. Again the acetic acid obtained by oxidation was analyzed, the results permitting the additional conclusion that the methyl migration from C₁₄ to C₁₃ is intra- and not intermolecular. The successful outcome of the studies on squalene cyclization has been particularly gratifying. It demonstrates exceptionally well that the organic chemist and the biochemist must interact closely if they wish to solve the problems of biochemical reaction mechanisms.

Continuing a historical rather than a systematic description of cholesterol biosynthesis and returning to the earlier stages of the process, we proceed to the discovery of mevalonic acid, the substance that proved to be the link between acetic acid and the biological isoprene unit. Since 1952 several laboratories had searched for branched-chain C₅ or C₆ intermediates without discovering any compounds that showed the requisite biological activity. β -Hydroxy- β -methyl glutaric acid, a substance known at that time to occur in plants³², seemed structurally attractive as a condensation product of three moles of acetate and as a precursor of a branched C₅ unit³³, but its biological

activity was disappointing. (Later, it was recognized that hydroxymethylglutarate is in fact an intermediate, but only in the form of the mono-CoA derivative). The break through came in 1956 with the isolation of mevalonic acid by Wright, Folkers and associates at the Merck, Sharp and Dohme laboratories³⁴. The original purpose of these investigations was to isolate and characterize a factor which was exceptionally active as a substitute for acetate in the nutrition of acetate-requiring strains of *Lactobacillus acidophilus*³⁵.

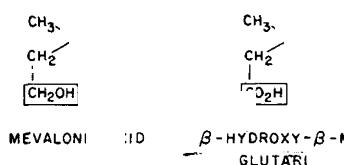


Fig. 8. Structural relation of mevalonic acid and β -hydroxy- β -methylglutaric acid.
(Merck, Sharpe & Dohme Laboratories, 1965)

Noting the structural resemblance of mevalonic acid and hydroxymethylglutarate-their carbon skeletons are identical (Fig. 8) - Tavormina, Gibbs and Huff tested the bacterial growth factor and found it to be remarkably active as a precursor of squalene and sterol³⁶. Conversion was essentially quantitative assuming that only one of the enantiomorphs was active. The discovery of mevalonic acid as a key intermediate in terpene and sterol biosynthesis is one of the examples of serendipity on which the progress of science depends so critically. The example is the more remarkable because up to the present the function of mevalonic acid in the *Lactobacilli* which require it has remained unknown. Sterols play no essential role in the metabolism of bacteria nor is mevalonic acid required by *Lactobacilli* as a precursor of known polyisoprenoid derivatives.

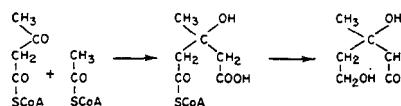


Fig.9. Biosynthesis of mevalonic acid. (Rudney, Lynen; 1957-1958)

The enzymatic link between acetyl-CoA and mevalonate by way of acetoacetyl-CoA and hydroxymethylglutaryl-CoA was soon established in the laboratories of Rudney^{37,38} and of Lynen^{39,40} (Fig. 9). Clearly, Lynen's fundamental discovery of acetyl-CoA in 1951 paved the way for studying the

early steps in sterol biosynthesis as it did for the understanding of all phases of fatty acid metabolism. One important issue that remains to be resolved is the origin of the acetoacetyl-CoA that enters into the synthesis of hydroxymethylglutarate. Conceivably the initial condensation to the four-carbon level involves an acetyl unit and a malonyl unit⁴¹ as in the synthesis of long-chain fatty acids. Energetically this mechanism is certainly more favorable than reversal of the β -ketothiolase reaction. We have observed, however, that in biotin-deficient yeast sterol synthesis from acetate proceeds normally, whereas fatty acid synthesis is greatly impaired⁴². This finding would seem to argue against any role of biotin-CO₂ and therefore of malonyl units in sterol synthesis from acetate, at least in some organisms. The purposes of metabolic regulation would also be served better if the acetoacetyl precursors for sterols and fatty acids were furnished by independent mechanisms. This would allow for control of the entry of acetyl-CoA into the two pathways at the earliest possible stage.

Mevalonic acid has the same oxidation level as isoprene and is formally convertible to isoprene by decarboxylation and by the loss of two molecules of water. The possibility was, therefore, attractive that the subunits condensing to the presumptive mono- and sesquiterpenoid intermediates might be closely related to isoprene itself. In our first experiments, designed to test this view, we studied the retention of tritium in the enzymatic conversion of [5-di-³H] - mevalonate to squalene. We found that in this transformation only a small fraction of the twelve hydrogen atoms attached to C₅ of the six participating mevalonate molecules was removed⁴³. Thus one of the two bond-forming centers (C₅) seemed to remain in the reduced state during the process of carbon-carbon bond formation. Later experiments with D₂O and 5-D₂-mevalionic acid strengthened this conclusion and allowed us to make the same deduction for carbon atom 2 of mevalonate. Still lacking any direct knowledge of the nature and number of intermediates, we were nevertheless in a position to interpret these results in terms of a general reaction mechanism for the coupling of C₅ or C₆ subunits⁴⁴ (Fig. 10). Bond formation had to occur without loss or re-introduction of hydrogen at the reacting centers, *i.e.* by interaction of mevalonic acid derivatives containing -CH₂-groups at both the C₂ and C₅ positions. From the same experiments it could be inferred further that the removal of the tertiary hydroxyl group and the loss of the carboxyl function of mevalonic acid proceed concertedly to a C₅ compound bearing a methylene group. Possible structures for the reactive condensing unit were thereby limited to isoprene itself or a derivative of isopentene⁴⁴.

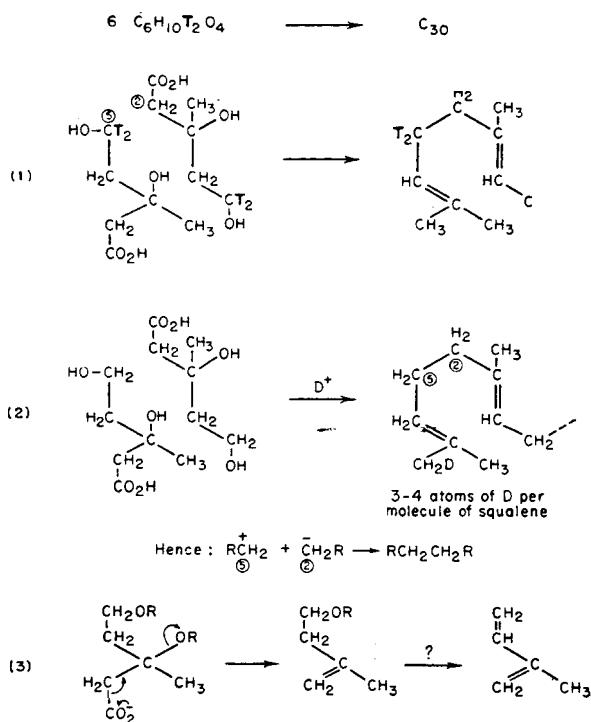


Fig. 10. Studies with heavy hydrogen pertaining to the structure of the biological isoprene unit.

Concurrently we investigated the requirements for cofactors in squalene synthesis from mevalonic acid in yeast extracts and found that both ATP and TPNH had to be supplied to the enzyme system⁴³. Suspecting phosphorylated intermediates, T. T. Tchen isolated a stable monophosphate of mevalonic acid and the requisite kinase enzyme⁴⁵⁻⁴⁶. The derivative was the 5 - monophosphate ester, as Lynen was able to show⁴⁰. Commenting on the mevalionic kinase reaction, Tchen pointed out that a phosphate ester grouping would be an appropriate leaving anion for hydroxyl elimination and thus aid in forming the anticipated terminal methylene group of the biological isoprene unit. As subsequent developments showed, this guess needed to be modified only slightly.

5-Phosphomevalonic acid was an excellent substrate in squalene synthesis, but since it was not converted unless once again ATP was provided for the reaction, additional phosphorylation steps were indicated. At this time (May 1958), investigators in the field met in London at a CIBA Symposium on

Terpene and Sterol Biogenesis. It was clear that three laboratories were on the trail of several phosphorylated derivatives of mevalonic acid, Lynen's group and ours working with yeast extracts and the group of Popják and Cornforth with liver preparations. We reported the isolation of two new derivatives of mevalonic acid, one containing two atoms of phosphorus per molecule and cleaved by mild acid hydrolysis to orthophosphate and mevalonic monophosphate, and another, yielding orthophosphate and Δ^3 -isopentenol when digested with snake venom phosphatase. Our first and premature structural assignments for these substances were 3,5 - diphosphomevalonate and isopentenylmonophosphate⁴⁷. These structures were soon revised when additional experiments showed that the five-carbon compound contained two atoms of phosphorus and therefore had to be isopentenyl pyrophosphate^{48,49}. On the basis of this assignment, the logical structure for the six-carbon compound was mevalonic acid-5-pyrophosphate⁴⁹ (Fig. 11). Lynen and his associates came to the same conclusion and, moreover, provided firm proof for the identity of isopentenylpyrophosphate by chemical synthesis⁵⁰.

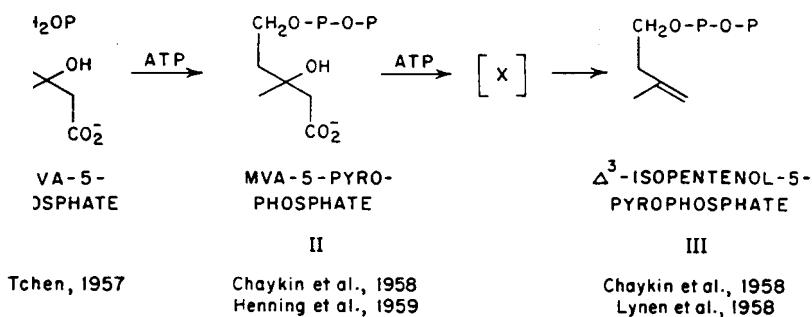


Fig. 11. Phosphorylated derivatives of mevalonic acid.

Chemically the most interesting of the reactions in the sequence from mevalonic acid to isopentenylpyrophosphate is the ATP-facilitated decarboxylative β -elimination of mevalonic acid 5-pyrophosphate, the step that generates the biological isoprene unit. The concerted nature of this reaction had been anticipated by the results of our earlier deuterium studies⁴⁴. It was, therefore, gratifying to find that purified "anhydrodecarboxylase" catalyzed the coordinated removal of the carboxyl group and of the tertiary hydroxy function as postulated⁴⁴. Data obtained with ^{18}O suggest that 3-phosphomevalonic-5-pyrophosphate is a transitory intermediate⁵¹, ATP serving as the phosphorylating agent for the tertiary hydroxyl group thereby promoting its elimination (Fig. 12).

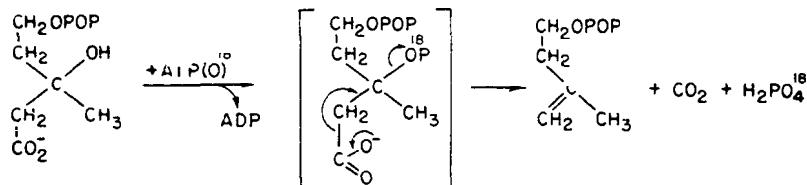


Fig.12. Decarboxylative β -elimination of mevalonic acid-5-pyrophosphate to isopentenylpyrophosphate. (De Waard, Phillips and Bloch, 1959; Lindberg, Yuan, De Waard and Bloch, 1962)

In the symmetrical squalene molecule there are two terminal isopropylidene groups, and since the "biological isoprene unit" has the isopropenyl structure, two of the six isopentenyl groups must isomerize at some stage of squalene synthesis. One step that might appropriately initiate the condensation of C_5 units is an interaction of an isopropenyl ester with enzyme in a double displacement reaction to yield pyrophosphate and isopentenyl enzyme, the latter isomerizing subsequently to the dimethylallyl structure⁴⁷. This possibility now seems unlikely. Lynen proposed an alternative mecha-

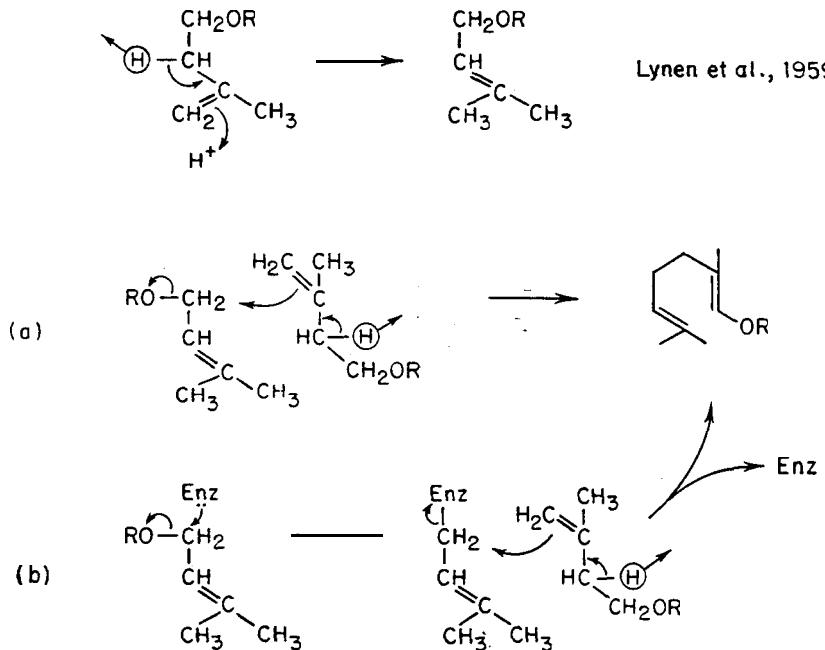


Fig. 13. Isomerization of isopentenylpyrophosphate and mechanisms for the condensation of isopentenyl units.

nism (Fig. 13) involving the isomerization of free isopentenylpyrophosphate to dimethylallyl pyrophosphate prior to condensation⁵⁰. Support and proof for this mechanism was provided by Lynen and his associates when they isolated the requisite isomerase and showed that this enzyme is an essential component in the coupling system when isopentenylpyrophosphate is the sole substrate⁵². The discovery of geranylpyrophosphate⁵³ and of farnesylpyrophosphate⁵⁰ in Lynen's laboratory completed the description of the biosynthetic pathway and of each enzymatic step up to the sesquiterpene stage (Fig. 14). As Popják was able to show, the same intermediates participate in the synthesis of squalene by liver enzymes^{54,55}.

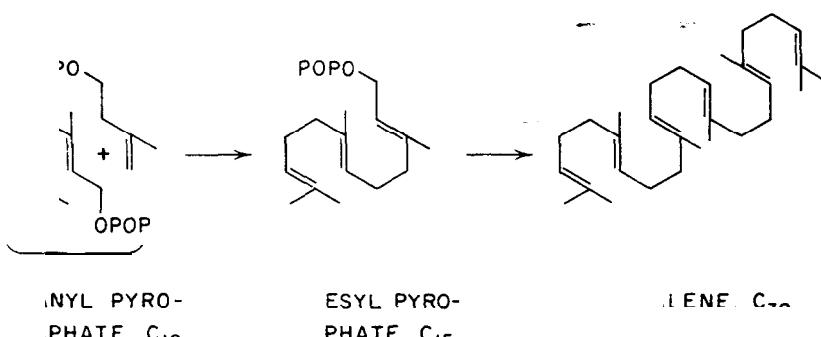


Fig. 14. Mono- and sesquiterpene precursors of squalene. (Lynen *et al.*, 1958-1959)

For those participating in the characterization of the intermediates between mevalonic acid and squalene this period was especially exciting. The chemistry was novel and essentially unexpected and the means for obtaining structural proof of microgram quantities were highly unorthodox by conventional standards. The nearly total reliance in the crucial experimentation on the tracer technique and its various extensions seems almost without parallel.

The polymerization of C₅ units by the joining of methylene groups has no precedent among enzymatic mechanisms for forming carbon-carbon bonds. One can formulate both the initial interaction of two C₅ units and subsequent C₅ additions as resulting from a nucleophilic attack of the electrons available in the exomethylene group of isopentenylpyrophosphate on an incipient cation formed by pyrophosphate elimination from the allylpyrophosphate. All these events are viewed as being concerted (Fig. 13a). This mechanism is undoubtedly attractive and useful as a working hypothesis, but it lacks the proof which only studies with purified enzymes can provide. Guided by the concepts prevalent in organic chemistry we have adopted, perhaps too readily,

the practice of formulating reactions in terpene and sterol biosynthesis as synchronous events. Yet modern enzyme chemistry points to alternative solutions. For example, covalent enzyme-substrate complexes might be formed initially between the allylpyrophosphate and the condensing enzyme with elimination of pyrophosphate (Fig. 13b). Allyl enzyme rather than the free allylpyrophosphate would then react with a second C₅ unit to form the new carbon-carbon bond.

There is every reason to believe that isopentenylpyrophosphate functions universally as the "monomeric" precursor for the great variety of linear and cyclic isoprene derivatives which occur in nature. So far only a beginning has been made in the purification of the respective polymerizing enzymes. The product specificities of these enzymes obviously extend over a wide range of molecular sizes from the monoterpenes to polyisoprenoid macromolecules. Two interesting examples of relatively high chain-length specificity are already known. Lynen's farnesyl pyrophosphate synthetase from yeast⁵³ and Popják's analogous enzyme from liver⁵⁵ afford as end products predominantly if not exclusively farnesyl pyrophosphate. Longer chains, e.g., C₂₀ pyrophosphate, are produced at less than one hundredth of the rate observed for the C₁₅ pyrophosphate. The second specific enzyme, a purified terpene synthetase isolated in our laboratory from *Micrococcus lysodeikticus* produces mainly geranylgeranyl pyrophosphate, possibly small amounts of the C₂₅ homologue but no C₁₅ pyrophosphate⁵⁶. The chain-length specificities of the two enzymes reflect in a striking way the economy of cellular processes. Termination of the chain at the C₁₅ stage is to be expected when the enzyme functions in a biosynthetic pathway that leads to squalene and to sterol, as it does in yeast and liver. In bacteria, sterol synthesis does not occur, but since *Micrococcus* produces carotenoids it is reasonable that this-bacterial terpene synthetase specializes in the formation of geranylgeranyl pyrophosphate, the C₂₀ precursor of the carotenoids.

In the formal sense, squalene formation can be viewed as resulting from the dimerization of two farnesyl units and, as Lynen has shown, these are provided by farnesyl pyrophosphate⁵⁰. The process is reductive and this accounts for the TPNH requirement which we noted in early studies of squalene synthesis from mevalonic acid⁴³. Again, we are dealing with an interaction of two reactive methylene groups as in the coupling of isopentenyl units. However, the resemblance is only superficial, since the farnesyl dimerization is reductive and involves the union of two like groups (tail to tail). In C₅ + C₅ condensations isopentenyl units are joined head to tail fashion and a reducing agent is

not needed. Several ingenious schemes have been proposed for the last step in squalene synthesis⁵⁷. From the elegant studies of Cornforth and Popják⁵⁸ we know the steric course of the coupling reaction but otherwise the mechanism has remained elusive. Some of our own contributions to this subject have unfortunately obscured rather than clarified the problem. Early experiments on squalene synthesis in D₂O or D-mevalonate in crude yeast extracts seemed to indicate that the formation of the central carbon-carbon bond in squalene entailed the loss and the reintroduction of two hydrogen atoms at the bond-forming centers⁴⁴. Cornforth and Popják therefore proposed, quite logically, that squalene synthesis occurred by way of dehydrosqualene, an intermediate with a centrally located double bond⁵⁷. However, this idea had to be abandoned when the English workers showed by mass spectrometry that only one of the hydrogen atoms linked to the reacting groups is replaced during the

C₁₅-C₁₅ coupling and not two⁵⁹. Our later results fully confirmed this finding⁶⁰. One important mechanistic detail has thus been clarified, but the mystery surrounding the C₁₅-C₁₅ condensation has yet to be solved. At the moment it would appear that the decision will lie between two mechanisms favored by Comforth and Popják and assumed to involve either an isomerization of one of the farnesyl units to a nerolidol derivative prior to condensation or alternatively a Stevens-type rearrangement of an enzyme-bound farnesyl residue⁶¹. Squalene synthetase which may or may not be a single enzyme has not yet been obtained in soluble form and this greatly restricts the approaches that can be taken to elucidate the mechanistic details of this interesting reaction. The same technical difficulties must be overcome if we are to learn more about the mechanism of squalene cyclization and about the subsequent steps in the biosynthesis of cholesterol.

The enzymatic characterization of the late stages of the cholesterol pathway has hardly begun and at present little more is known about this subject than the identity of some of the intermediates. From the nature of the chemical changes involved in the transformation of lanosterol to cholesterol, one would nevertheless estimate that they require a large number of enzymes, probably more than half of the total that comprises the entire pathway (Fig. 15). These steps necessarily include the removal of the methyl group at C₁₄ and those at C₄, the reduction of the isooctenyl side chain and the (formal) transfer of the double bond from the 8,9 to the 5,6 position in the ring system.

Lacking any clues as to the order of the transformations and expecting the intermediates to occur in traces only, we chose the "pulse" technique of radioactive labeling for detecting new metabolites and the sequence in which they

DEMETHYLATION OF LANOSTEROL

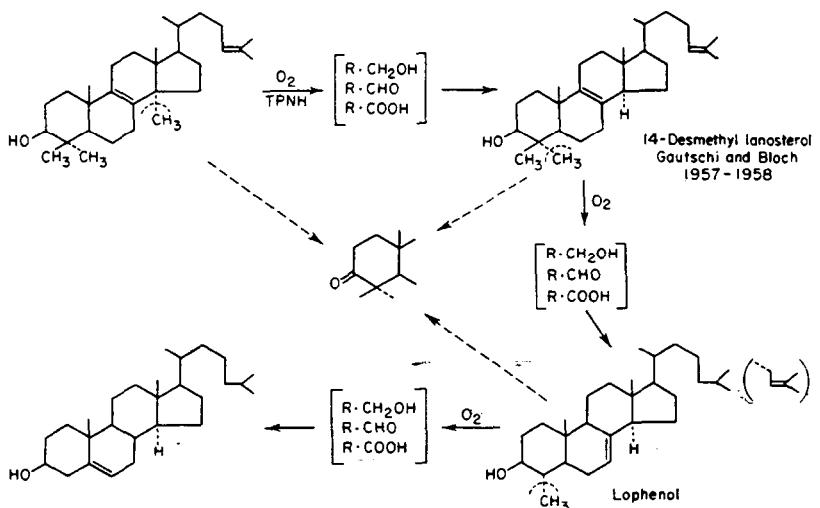


Fig. 15. The oxidative removal of methyl groups from lanosterol and the final stages in the biosynthesis of cholesterol. (Wells and Niederheiser, 1957; Djerassi *et al.*, 1958).

arise. When labeled acetate was injected into rats and the animals were sacrificed a few minutes later, several radioactive materials of unknown identity were obtained on chromatography of the nonsaponifiable tissue fraction⁶². One of the fractions, somewhat more polar than lanosterol but less polar than cholesterol, was suspected to be a partially demethylated derivative of lanosterol. This was too small in amount to be weighed or seen and structural identification of the metabolite had to be attempted without recourse to traditional chemical methodology or physical measurements. Various pieces of information could be obtained by a combination of radiochemical and enzymatic techniques similar in principle to those which led to the identification of isopentenylpyrophosphate. They all showed that the unknown differed from lanosterol only by lacking the methyl substituent at C₁₄. The structure deduced for the metabolite was, therefore, 14-desmethyllanosterol or 4,4-dimethylcholest-8,24-diene-3- β -ol^{63,64}. Conditions for accumulating this sterol in biological systems have not yet been devised and only the 24,25 dihydro derivative, not the sterol itself, has been chemically synthesized. The physical properties of 14-desmethyllanosterol are, therefore, not known. It also remains to be demonstrated that the 14-nor sterol is a metabolite of lanosterol even though no alternatives are obvious. At any rate, the identification of this short-lived intermediate and its facile conversion to cholesterol dis-

closed the order of the reactions by which lanosterol is transformed to cholesterol. The view that demethylation at C₁₄ precedes the removal of the methyl groups at C₄ was soon supported by the isolation of two mono methyl sterols, **4- α -methyl- Δ^7 -cholesten-3- β -ol** (methostenol or lophenol) from animal⁶⁵ and plant sources⁶⁶, and of **4- α -methyl- Δ^8 -cholestenol** from tumor tissue⁶⁷. Both these substances are converted to cholesterol, and are obviously metabolites of lanosterol. The sequence of demethylation reactions is not necessarily the same in all systems judging from structural evidence alone. For example, Djerassi has isolated from plant sources a 14-methylcholestane derivative, **14 α -methyl- Δ^8 -cholesten-3 β ,6 α -diol** (Mcdougallin) (ref.68), which may or may not be on the main metabolic path. In animal tissues, however, this sterol is inert (M. Slaytor and K. Bloch, unpublished).

In isolated liver the demethylation of lanosterol proceeds only in the presence of oxygen and then produces three moles of CO₂ per mole of cholesterol formed⁶⁹. Presumably each of the three methyl groups is initially hydroxylated in an oxygenase-type reaction and then oxidized to aldehydic and carboxylic derivatives. Finally, the C₁ substituent is eliminated from the sterol skeleton by decarboxylation. There is ample structural precedent for the step-wise oxidation of methyl groups in the steroids and in the cyclic terpenes (Soyasapogenol, aldosterone, abietic acid). Also, the oxygen-dependent hydroxylation of terminal methyl groups is now a well-established enzymatic reaction for aliphatic chains. One of the presumptive partially oxidized intermediates, 4-hydroxymethylene-7-cholesten-3-one has been synthesized⁷⁰ and it shows the requisite biological activity, producing a C₂₇ sterol as well as CO₂ in the usual liver system. This transformation occurs anaerobically as well as in air, as one would expect if only the initial attack on the methyl group required molecular oxygen. The corresponding carboxylic acid, **4-carboxy- Δ^7 -cholesten-3-one** is also decarboxylated by liver homogenate, but as the rate of the reaction is also quite rapid in the absence of enzyme the significance of this reaction as an enzymatic process is still in question (J. J. Britt, G. Scheuerbrandt and K. Bloch, unpublished). The enzymatic removal of the oxidized one-carbon substituents appears, in any event, to be very similar mechanistically to the corresponding nonenzymatic reactions. Decarboxylation is assisted either by a neighboring double bond or by a keto group in the β -position⁷¹.

The remaining structural changes which still have to be placed in the proper order are the elimination of the double bond in the isooctenyl side chain and the formal transfer of the nuclear double bond from position 8,9 to 5,6.

Whether the side-chain reduction occurs necessarily at one specific point and whether this involves early or late intermediates in the lanosterol-cholesterol transformation is not yet clear. The structure of desmosterol ($\Delta^{5,24}$ -cholestadienol) certainly suggests that this reductive step can be the very last one of the sequence⁷². By the same token, however, the existence in the tissues of several side-chain saturated sterols (dihydrolanosterol, lophenol, 4α -methyl- Δ^8 -cholestolen, Δ^7 -cholestolen) argues for reduction at a much earlier stage of lanosterol metabolism. Either the reducing enzyme systems are fairly non-specific⁷³ or the substrate specificity is not the same in all the tissues in which sterol synthesis occurs. For very similar reasons we do not yet know whether the double bond shift from the $\Delta^{8,9}$ to the Δ^7 position occurs early or late during the lanosterol-cholesterol transformation. It is, however, well established, mainly by the work of Frantz and his colleagues, that this isomerization precedes the entry of the 5,6 double bond^{74,75}. A $\Delta^{5,7}$ -cholestadienol is undoubtedly on the main reaction path yielding cholesterol in a final reductive step.

Perspective

The direction that further research on terpene and sterol biosynthesis will take and the aspects to be emphasized seem clearly indicated. Biosynthetic studies generally begin at the level of intact animals or whole cells and progress gradually to *in vitro* experimentation. Only with isolated and ultimately pure enzymes can one hope to understand reaction mechanisms and for many steps of cholesterol biosynthesis this is still a distant goal. We are already quite well informed about the early phase of the sequence which covers the steps from acetyl-CoA to farnesyl pyrophosphate. This is undoubtedly so because the respective enzymes are soluble and lend themselves to the traditional techniques of protein purification. It is perhaps not without significance that all the substrates for these soluble enzymes are freely water-soluble Co-enzyme A derivatives or pyrophosphate esters. By contrast, squalene and the subsequent members of the biosynthetic chain are highly hydrophobic molecules and the enzymes that act upon them are tightly bound to particulate elements of the cell. Conceivably the transformation of these lipid substrates requires more complex catalytic systems in which lipoproteins themselves may play a part. At any rate, the available methodology is clearly inadequate for manipulating many of the enzymes that act on lipophilic substrates.

A more highly developed enzymology of the cholesterol pathway is also needed for a rational approach to the problem of metabolic regulation. This is clearly a matter of broad concern transcending purely academic curiosity. A great variety of environmental, dietary and hormonal factors have been studied and shown to influence the rate of cholesterol synthesis, but the evidence is still too scanty for specifying the point or points at which physiological control is most effectively exerted. If the principle of negative feedback operates in sterol biosynthesis as it does in so many pathways, then the first specific step of the sequence should be rate limiting and it should be sensitive to cholesterol, the end product of the biosynthetic sequence. With these considerations in mind, attention has been focused on the reduction of hydroxymethylglutaryl-CoA to mevalonic acid as a likely control site. There is experimental support for this view, particularly from the work of Bucher who has shown that the profound effects of cholesterol feeding, of X-radiation and of starvation on the rates of cholesterol synthesis appear indeed to be exerted at a stage close to or immediately before the formation of mevalonic acid⁷⁶. This concept of a homeostatic control continues to receive attention", and it is to be hoped that the rapid progress that is being made in the elucidation of regulatory processes will lead also to a better understanding of the mechanisms that control cholesterol biosynthesis.

Along with the interest in chemical and enzymatic aspects of cholesterol biosynthesis there has been a growing appreciation of the role which sterols might play as fundamental cell constituents. The known metabolic transformations of cholesterol, the conversions to steroid hormones and bile acids, surely serve a specialized function since they take place only in vertebrate species. However, in many tissues and cells the function of cholesterol is clearly not metabolic. In organisms which do not metabolize sterol but nevertheless produce or require it-and this is true for all but the most primitive forms of life-sterols must play some role as structural elements of the cell. Comparative biochemistry suggests what this function might be. Sterols have not been found in any bacteria or in the blue-green algae, *i.e.* in primitively organized cells which lack the various membrane-bound intracellular organelles. The elaboration of membrane-enclosed structures devoted to specialized functions is now viewed as a landmark in evolutionary diversification⁷⁸ and it would appear that the parallel development of the biosynthetic pathway to sterols is one of the biochemical expressions of these morphological events. The sterol molecule is not distributed at random inside the differentiated cell but appears to be mainly associated with the cytoplasmic membrane and its

endoplasmic extensions. We do not yet know why and for what specific purpose the sterol molecule was selected during the evolution of organisms. One may speculate, however, that the rigidity, the planarity and the hydrophobic nature of the molecule provide a combination of features that is uniquely suitable for strengthening the otherwise fragile membrane of the more highly developed cell.

Acknowledgements

The research described has been generously supported by grants in aid from the National Institute of Health, the National Science Foundation, the Life Insurance Medical Research Fund, the Nutrition Foundation and the Eugene Higgins Trust Fund of Harvard University.

1. I. M. Heilbron, E. D. Kamm and W. M. Owens, *J. Chem Soc.*, (1926) 1630.
2. H. J. Channon, *Biochem. J.*, 20 (1926) 400.
3. D. Rittenberg, and R. Schoenheimer, *J. Biol. Chem.*, 121 (1937) 235.
4. R. Sonderhoff and H. Thomas, *Ann. Chem.*, 530 (1937) 195-213.
5. K. Bloch and D. Rittenberg, *J. Biol. Chem.*, 145 (1942) 625.
6. K. Bloch and D. Rittenberg, *J. Biol. Chem.*, 159 (1945) 45.
7. H. N. Little and K. Bloch, *J. Biol. Chem.*, 183 (1950) 33.
8. R. C. Ottke, E. L. Tatum, I. Zabin and K. Bloch, *J. Biol. Chem.*, 189 (1951) 429.
9. J. Bonner and B. Arreguin, *Arch. Biochem.*, 21 (1949) 109.
10. J. Wuersch, R. L. Huang and K. Bloch, *J. Biol. Chem.*, 195 (1952) 439.
11. R. Robinson, *J. Chem. Soc. Ind.*, 53 (1934) 1062.
12. L. Ruzicka, *Experientia*, 9 (1953) 357.
13. R. G. Langdon and K. Bloch, *J. Biol. Chem.*, 200 (1952) 129.
14. R. G. Langdon and K. Bloch, *J. Biol. Chem.*, 200 (1952) 135.
15. R. K. Maudgal, T. T. Tchen and K. Bloch, *J. Am. Chem. Soc.*, 80 (1958) 2589.
16. J. W. Cornforth, G. D. Hunter and G. Popják, *Biochem. J.*, 54 (1953) 590.
17. J. W. Cornforth, G. D. Hunter and G. Popják, *Biochem. J.*, 54 (1953) 597.
18. J. W. Cornforth, I. Y. Gore and G. Popják, *Biochem. J.*, 65 (1957) 94.
19. W. Voser, M. W. Mijovic, H. Heusser, O. Jeger and L. Ruzicka, *Helv. Chim. Acta*, 35 (1952) 2414.
20. R. B. Woodward and K. Bloch, *J. Am. Chem. Soc.*, 75 (1953) 2023.
21. W. G. Dauben, S. Abraham, S. Hotta, I. L. Chaikoff, H. L. Bradlow and A. H. Solloway, *J. Am. Chem. Soc.*, 75 (1953) 3038.

22. K. Bloch, *Helv. Chim. Acta*, 36 (1953) 1611.
23. J. W. Comforth and G. Popják, *Biochem. J.*, 58 (1954) 403.
24. T. T. Tchen and K. Bloch, *J. Am. Chem. Soc.*, 77 (1955) 6085.
25. R. B. Clayton and K. Bloch, *J. Biol. Chem.*, 218 (1956) 319.
26. N. L. R. Bucher, *J. Am. Chem. Soc.*, 75 (1953) 498.
27. A. Eschenmoser, L. Ruzicka, O. Jeger and D. Arigoni, *Helv. Chim. Acta*, 38 (1955) 1890.
28. J. W. Comforth, *Rev. Pure Appl. Chem.*, 4 (1955) 275.
29. T. T. Tchen and K. Bloch, *J. Biol. Chem.*, 226 (1957) 931.
30. J. W. Cornforth, R. H. Comforth, A. Pelter, M. G. Horning and G. Popják *Tetrahedron*, 5 (1959) 311.
31. D. W. Dicker and M. Whiting, *J. Chem. Soc.*, (1958) 1994.
32. R. Adams and B. L. Van Duuren, *J. Am. Chem. Soc.*, 75 (1953) 2377.
33. K. Bloch, *Harvey Lectures*, Ser. 48 (1952-1953) 68.
34. L. D. Wright, E. L. Cresson, H. R. Skeggs, G. D. E. MacRae, C. H. Hoffman, D. E. Wolf and K. Folkers, *J. Am. Chem. Soc.*, 78 (1956) 5273.
35. H. R. Skeggs, L. D. Wright, E. L. Cresson, G. D. E. MacRae, C. H. Hoffman, D. E. Wolf and K. Folkers, *J. Bacteriol.*, 72 (1956) 519.
36. P. A. Tavormina, M. H. Gibbs and J. W. Huff, *J. Am. Chem. Soc.*, 78 (1956) 4498.
37. H. Rudney, *J. Biol. Chem.*, 227 (1957) 363.
38. H. Rudney, in G. E. W. Wolstenholme and M. O' Connor (Eds.), *CIBA Found. Symp. on Biosynthesis of Terpenes and Sterols*, Churchill, London, 1959, p. 57.
39. F. Lynen, U. Henning, C. Bublitz, B. Sorbo and L. Kroepelin-Rueff, *Biochem. Z.*, 330 (1958) 269.
40. F. Lynen, in G.E.W. Wolstenholme and M. O'Connor (Eds.), *CIBA Found. Symp. on Biosynthesis of Terpenes and Sterols*, Churchill, London, 1959, p.59.
41. J. D. Brodie, G. Wasson and J. W. Porter, *J. Biol. Chem.*, 238 (1963) 1294.
42. D. Bloomfield and K. Bloch, *J. Biol. Chem.*, 235 (1960) 337.
43. B. H. Amdur, H. Riling and K. Bloch, *J. Am. Chem. Soc.*, 79 (1957) 2646.
44. H. Rilling, T. T. Tchen and K. Bloch, *Proc. Natl. Acad. Sci. (U.S.)*, 44 (1958) 167.
45. T. T. Tchen, *J. Am. Chem. Soc.*, 79 (1957) 6344.
46. T. T. Tchen, *J. Biol. Chem.*, 233 (1958) 1100.
47. K. Bloch, in G. E. W. Wolstenholme and M. O' Connor (Eds.), *CIBA Found. Symp. on Biosynthesis of Terpenes and Sterols*, Churchill, London, 1959, p. 4.
48. K. Bloch, *Proc. IVth Intern. Congr. Biochemistry*, Vienna, Pergamon, London, IV, 1958, p. 50.
49. S. Chaykin, J. Law, A. H. Phillips, T. T. Tchen and K. Bloch, *Proc. Natl. Acad. Sci. (U.S.)*, 44 (1958) 998-
50. F. Lynen, H. Eggerer, U. Henning and I. Kessel, *Angew. Chem.*, 70 (1958) 739.
51. M. Lindberg, C. Yuan, A. deWaard and K. Bloch, *Biochemistry*, 1 (1962) 182.
52. B. W. Agranoff, H. Eggerer, U. Henning and F. Lynen, *J. Am. Chem. Soc.*, 81 (1959) 1254.
53. F. Lynen, B. W. Agranoff, H. Eggerer, U. Henning and E. M. Moslem, *Angew. Chem.*, 71 (1959) 657.
54. G. Popják *Tetrahedron Letters*, 19 (1959) 19.

55. G. Popják, A. E. Lowe, D. Moore, L. Brown and F. A. Smith, *J. Lipid Res.*, 1 (1959) 29.
56. A. A. Kandutsch, H. Paulus, E. Levin and K. Bloch, *J. Biol. Chem.*, 239 (1964) 2507.
57. G. Popják and J. W. Cornforth, *Advan., Enzymol.*, 22 (1960) 281.
58. G. Popják, *Proc. VIth Intern. Congr. Biochemistry, New York*, VII, 1964, p. 545.
59. G. Popják, De Witt S. Goodman, J. W. Cornforth, R. H. Cornforth and R. Ryhage, *J. Biol. Chem.*, 236 (1961) 1934.
60. C. R. Childs and K. Bloch, *J. Biol. Chem.*, 237 (1962) 62.
61. G. Popják, *Proc. Roy. Soc. (London)*, Ser. B, 156 (1962) 376.
62. P. B. Schneider, R. B. Clayton and K. Bloch, *J. Biol. Chem.*, 224 (1957) 175.
63. F. Gautschi and K. Bloch, *J. Am. Chem. Soc.*, 79 (1957) 1145.
64. F. Gautschi and K. Bloch, *J. Biol. Chem.*, 233 (1958) 1343.
65. W. W. Wells and D. H. Neiderhiser, *J. Am. Chem. Soc.*, 79 (1957) 6569.
66. C. Djerassi, J. S. Mills and R. Villotti, *J. Am. Chem. Soc.*, 80 (1958) 1005.
67. A. A. Kandutsch and A. E. Russell, *J. Biol. Chem.*, 235 (1960) 2253, 2256.
68. C. Djerassi, J. C. Knight and D. I. Wilkinson, *J. Am. Chem. Soc.*, 85 (1963) 835.
69. J. A. Olsen, M. Lindberg and K. Bloch, *J. Biol. Chem.*, 226 (1957) 941.
70. J. Pudles and K. Bloch, *J. Biol. Chem.*, 235 (1960) 12.
71. M. Lindberg, F. Gautschi and K. Bloch, *J. Biol. Chem.*, 238 (1963) 1661.
72. W. M. Stokes and W. A. Fish, *J. Biol. Chem.*, 235 (1961) 2604.
73. J. Avigan, De Witt S. Goodman and D. Steinberg, *J. Biol. Chem.*, 238 (1963) 1283.
74. G. J. Schroepfer and I. D. Frantz, *J. Biol. Chem.*, 236 (1961) 3137.
75. M. E. Dempsey, J. D. Seaton, G. J. Schroepfer and R. W. Trockmann, *J. Biol. Chem.*, 239 (1964) 1381.
76. N. L. R. Bucher, in G. E. W. Wolstenholme and M. O' Connor (Eds.), *CIBA Found. Symp. on Biosynthesis of Terpenes and Sterols*, Churchill, London, 1959. p.46.
77. M. D. Siperstein and V. M. Fagan, *Advances in Enzyme Regulation*, Vol. 2, Pergamon, Oxford, 1964, p-249.
78. R. Y. Stanier and C. B. van Niel, *Archiv. Mikrobiol.*, 42 (1962) 17.

F E L I X B L O C H

The principle of nuclear induction

Nobel Lecture, December 11, 1952

It is a tribute to the inherent harmony and the organic growth of our branch of science that every advance in physics is largely due to the developments that preceded it. The discovery for which Purcell and I have received the honor of the Nobel Prize award for the year 1952 is a typical example of this situation, and before describing the principle I shall therefore present an outline of its long and distinguished background.

Both the method and the object go back ultimately to spectroscopy, a field to which modern physics owes so much in other respects. Among the various aspects of this field there are two which are of particular importance here: the Zeeman effect for introducing magnetic fields as an essential element of spectroscopy, and the hyperfine structure of spectral lines for revealing the existence of nuclear moments. The correct interpretation of hyperfine structures was first given in 1924 by Pauli¹, who proposed that atomic nuclei may possess an intrinsic angular momentum (spin) and, parallel to its orientation, a magnetic moment. The energy of interaction of this magnetic moment with the magnetic field $H_{(0)}$, produced by the atomic electrons at the position of the nucleus, depends upon the angle between them and leads thus to the observed small splitting of the energy levels. Conversely, it is possible under suitable conditions to determine from this splitting both the spin and the magnetic moment of the nucleus, and these two important quantities have indeed been determined in a great number of cases from the observation of hyperfine structures. The magnetic moments of the nuclei have been found, in all observed cases, to be of the order of the « nuclear magneton » which one obtains by substituting in the formula for the atomic Bohr magneton the mass of the proton in place of that of the electron. Nuclear moments are thus about a thousand times smaller than atomic moments, and this is plausible in view of the fact that one deals here with protons instead of electrons as elementary charged constituents. There are, however, distinct disadvantages in the optical determination of nuclear moments. In the first place the accuracy is seriously limited due to the fact that the effect consists only in such a small splitting of spectral lines that one has to

be concerned with their finite width. In the second place it is necessary for the determination of the nuclear magnetic moment from the observed hyperfine structure to have a knowledge of the field $H_{(o)}$ which is usually rather inaccurate since it involves complex electron configurations. In view of these limitations one is led to values of nuclear magnetic moments with an accuracy of a few percent at best. Finally, one is faced with the fact that hyperfine splittings tend to decrease with decreasing atomic number with the result that it is not possible, by optical means, to observe them in the case of the greatest fundamental importance, that of hydrogen.

A decisive step forward was made in 1933 by Stern², who applied his method of molecular beams to the determination of the magnetic moments of the proton and the deuteron in hydrogen molecules. Instead of the emitted light, it is here the deflection of the molecule in an inhomogeneous magnetic field which is affected by the nuclear moments. Although the observed effect was close to the limit of observability it yielded the proton moment to within ten percent with the most important result that instead of having the expected value of one nuclear magneton it is about 2.5 times larger. Of similar importance was the result that the magnetic moment of the deuteron was between 0.5 and 1 nuclear magneton, since it indicated from the simplest plausible considerations of the structure of this nucleus that one should ascribe a moment of about 2 nuclear magnetons to the neutron. I shall come back later to this point; it represents the start from which my own experimental work has followed in an almost continuous line.

Subsequent to Stern's work a number of far-reaching further developments have been achieved by Rabi in the application of atomic and molecular beams to the measurement of nuclear moments and hyperfine structures. Without attempting completeness I want to mention some aspects of method in the brilliant series of investigations which he carried out with his collaborators. One of them is based upon a paper by Breit and Rabi³, which treats the variation of the magnetic moment of an atom for the different Zeeman levels of hyperfine structure under the influence of an external magnetic field and which was applied to atomic beams where the deflection gives a direct measure of the magnetic moment. Another important aspect lies in the passage of the beam through two and later three separate field regions which can be adjusted to give zero deflection so that one deals with a null method. These innovations, besides giving many other interesting results, allowed the measurement of the hyperfine structure in the ground state of light and heavy hydrogen atoms; since the previously mentioned field $H_{(o)}$, produced

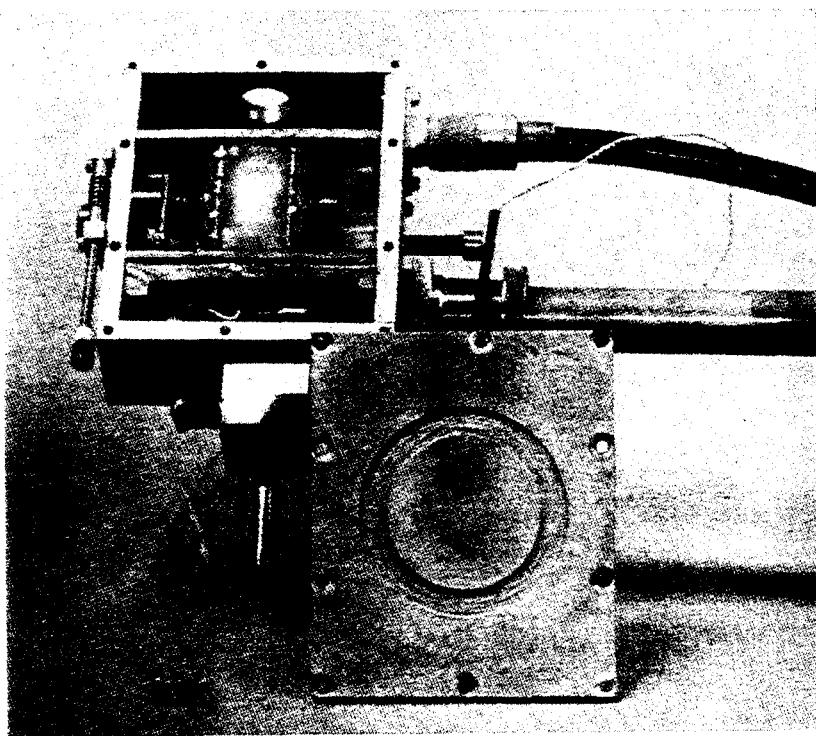


Fig. 1. The «head» of the crossed-coil arrangement with the cover plate removed. The bottom tube to the right contains the leads to the transmitter coil, which is wound in two sections visible in black in the head. The black cable leads from the receiver coil to the amplifier; the receiver coil is wound with a vertical axis inside the hollow lucite piece between the two sections of the transmitter coil. The sample test tubes are placed in its interior through the circular hole at the top of the supporting frame.

by the electron at the place of the nucleus, was given here, through a formula of Fermi⁴, from the well-known theory of the hydrogen atom, this measurement resulted in the determination of the magnetic moments of the proton and the deuteron with an accuracy of a few percent.

However, the most significant improvement in molecular and atomic beam techniques was the introduction of the magnetic resonance method. The beam here passes through a region where the magnetic field is homogeneous and constant with a weak alternating magnetic field superimposed at right angles to the strong constant field. Analogous to the resonance absorption of visible light, transitions occur here from one Zeeman level to another if the alternating field satisfies Bohr's frequency condition for the energy

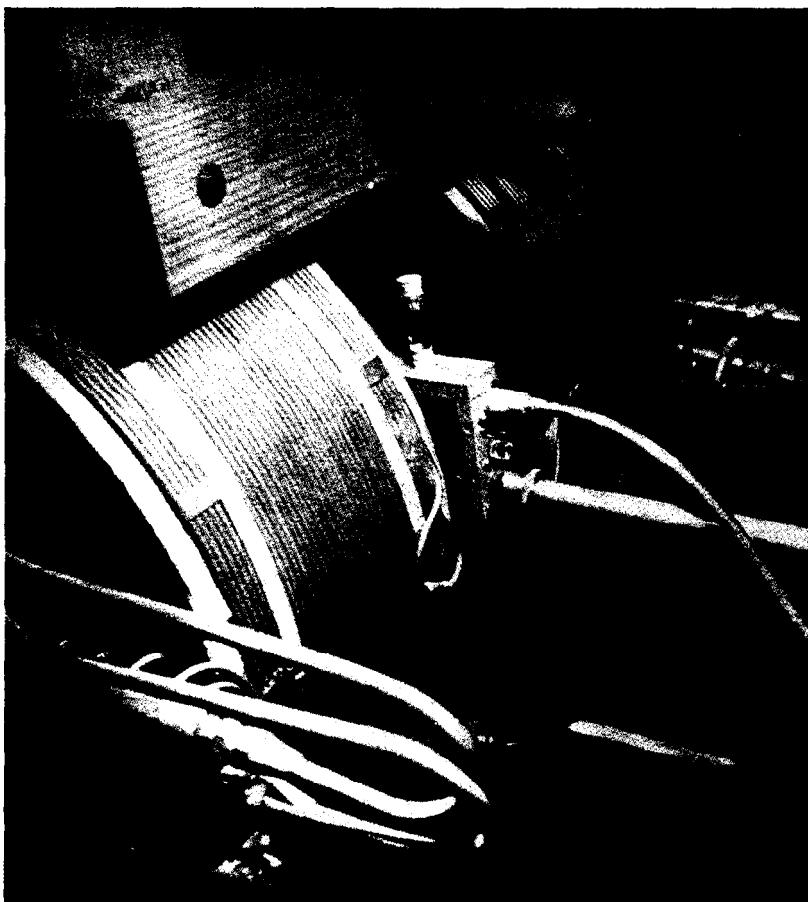


Fig. 2. The same head as in Fig. 1, about to be inserted in the gap of an electromagnet and containing a sample test tube. The two protruding lucite rods between the leads reach into the interior of the head and carry small copper disks; a fine adjustment of the coupling is achieved by rotation of these «paddles».

difference between the two levels. However, instead of optical frequencies one deals here normally with frequencies in the radio range so that this application of the magnetic resonance method, like our own, is properly labelled as belonging to the new field of radiofrequency spectroscopy. In the beam technique it has the great advantage of dispensing with a knowledge of the deflecting inhomogeneous fields, since the deflection is merely used now as an indicator for the occurrence of transitions in the homogeneous field region. A very much greater accuracy can thus be obtained; it led, for ex-



Fig. 3. A resonance line of protons in water, containing MnSO_4 as a paramagnetic catalyst and obtained from the phase component of the nuclear induction signal which corresponds to absorption. The photograph is that of the trace on a cathode-ray oscilloscope with the *vertical deflection* arising from the rectified and amplified signal, and the *horizontal deflection* corresponding to different values of the constant field.

ample, to the knowledge of the magnetic moments of the proton and the deuteron with an accuracy of about one part in a thousand and to the important discovery of a small but finite electrical quadrupole moment of the deuteron⁵ in 1939.

The first use of the magnetic resonance method was suggested in 1936 by Gorter⁶ in an attempt to detect the resonance absorption of radio quanta through the heating of a crystal. While the results of this experiment were negative, Rabi⁷ in 1937 has treated the transitions in a rotating field and has pointed out their use in atomic and molecular beams.

Coming from quite a different direction, I was led at that time to similar ideas. They originated from my preceding work which dealt with the magnetic moment of the neutron and which had been stimulated by Stern's previously mentioned measurement of the magnetic moment of the deuteron². The idea that a neutral elementary particle should possess an intrinsic magnetic moment had a particular fascination to me, since it was in such striking contrast to the then only existing theory of an intrinsic moment

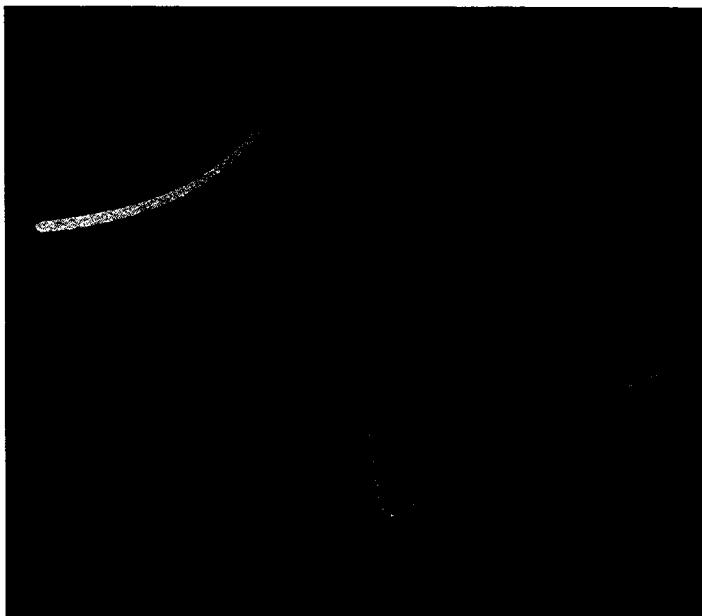


Fig. 4. The only difference between this line and that of Fig. 3 lies in the adjustment of the observed phase which is here that corresponding to dispersion.

which had been given by Dirac⁸ for the electron. Combining relativistic and quantum effects, he had shown that the magnetic moment of the electron was a direct consequence of its charge and it was clear that the magnetic moment of the neutron would have to have an entirely different origin. It seemed important to furnish a direct experimental proof for the existence of a magnetic moment of the free neutron, and I pointed out in 1936⁹ that such a proof could be obtained by observing the scattering of slow neutrons in iron. The very strongly inhomogeneous magnetic field in the neighborhood of each iron atom was shown to affect a passing neutron through its magnetic moment and thus to lead to an appreciable magnetic scattering effect; it was shown at the same time, that magnetic scattering would lead to the polarization of neutron beams. The existence of this effect was first clearly demonstrated in 1937 by a group of investigators at Columbia University¹⁰, and it opened up the possibility of further work with polarized neutron beams.

The most desirable goal to be reached here was that of accurately measuring the magnetic moment of the neutron. It occurred to me that resonance depolarization could be achieved by passing a polarized neutron beam

through a region where a weak oscillating field is superimposed on a strong constant field, provided that the frequency of the former is equal to the frequency with which the neutron moment carries out a precessional motion around the direction of the constant field. A knowledge of this field and of the corresponding resonance frequency directly determines the magnetic moment under the safe assumption that the spin of the neutron is $1/2$, and the magnetic scattering effect enters in this arrangement merely as an indicator for the occurrence of resonance depolarization. The application to polarized neutron beams was also noted by Rabi⁷ in his previously mentioned original paper on the magnetic resonance method. It was first achieved in 1939 by Alvarez and myself¹¹ with the use of the Berkeley cyclotron, and yielded a value for the magnetic moment of the neutron which was consistent with that of the deuteron if one assumed the latter to be additively composed of the moments of the proton and the neutron. The accuracy of this measurement amounted to about one percent and was partly limited by that with which the strength of the constant field could be determined. Another limit of accuracy arose from the smallness of the observed polarization effect, but a subsequent systematic investigation of neutron polarization¹², carried out with the Stanford cyclotron, showed how this effect could be greatly increased.

It was of considerable importance to improve the accuracy of the determination of the neutron moment to at least one part in a thousand in order to test small deviations from the additivity of the moments of the proton and the neutron, which could be expected in connection with the finite electric quadrupole moment of the deuteron, according to the theoretical work of Rarita and Schwinger¹³. The fact that higher accuracy hinged essentially upon that of a field calibration and the search for a suitable and convenient standard led me to new ideas when, toward the end of the last War, my thoughts turned back to the continuation of my previous work.

The essential fact of the magnetic resonance consists in the change of orientation of nuclear moments, and the methods to be employed in molecular and atomic beams as well as in neutron beams are primarily indicated by the different ways to detect this change. The acquaintance with radio techniques during the War suggested to me still another and much simpler way, that of detecting the reorientation of nuclear moments through the normal methods of radio reception. The signals to be detected would be due to the electromagnetic induction caused by nuclear reorientation and should appear as a voltage difference between the terminals of an external electric circuit.

I believe that this is the most general and distinctive feature of our discovery, and it is for this reason that I chose for it the name of « nuclear induction ». Purcell, whose independent approach was largely based on considerations of energy relations, has chosen to call it « nuclear magnetic resonance absorption », but soon after our respective initial work and despite its apparent difference, it became clear that it was based upon the same principle.

In order to understand this principle, one can start from macroscopic quantities and describe the underlying phenomenon in classical terms. Consider for this purpose, as a typical example, about one cubic centimeter of water with the protons contained in it as the nuclei under investigation. Their magnetic moments are oriented in a completely random manner in the absence of an external magnetic field; after the sample has been brought into such a field, however, there will be established a new thermal equilibrium in which the magnetic moments are distributed with a slight surplus parallel to the field. Even in relatively strong fields of the order of 10,000 gauss this surplus will at room temperature amount to no more than about one part in a million. While its direct observation would be difficult, there thus exists a « nuclear paramagnetism » in the sense that one deals with a finite macroscopic nuclear polarization which is both parallel and proportional to the applied external field. The establishment of thermal equilibrium demands the transfer of the energy released by the partial orientation of the nuclear moments into heat, and it can take place only through interaction of these moments with their molecular surroundings. The strength of this interaction determines the time interval required for the nuclear moments to adjust themselves to the equilibrium conditions; it is measured by the « relaxation time », as in the analogous case of atomic paramagnetism. The role of the relaxation time is of basic significance for our experiments, and I shall soon come back to its discussion.

For the moment, we shall return to the equilibrium state, once it is established, and to the description of the nuclear polarization under the conditions of magnetic resonance. A simple mechanical consideration of the gyroscope shows that an alternating field at right angles to the constant field has the effect of tilting the direction of the polarization with respect to the constant field and that the polarization will thereupon continue to perform a precessional rotation around this field. The angular frequency of precession is proportional to the field with a constant of proportionality which is called the « gyromagnetic ratio » of the nuclei and which is equal to the ratio of their magnetic moment and their intrinsic angular momentum. From a perfectly

macroscopic point of view, one thus deals with a situation in which the protons in our cubic centimeter of water have the effect of an invisible compass needle rotating in its interior. The « invisibility » refers actually only to observation of optical frequencies; the rotation occurs in the range of radio-frequencies, and it can very well be observed by using Faraday's law of induction. Indeed, the rotation of our compass needle is accompanied by that of a magnetic field which possesses an alternating component perpendicular to the axis of rotation, and hence by an electromotive force, induced in a suitably wound coil of wire around the sample. From here on it is merely a matter of the standard techniques of radio reception to rectify and amplify this electromotive force so that it can be recorded on a volt-meter, displayed on a cathode-ray oscilloscope, or made audible in a loudspeaker.

What amazed me most in my first calculations on this effect was the magnitude of the signals which one could expect from nuclear induction. In our example of a cubic centimeter of water in a normal field of a few thousand gauss they turned out to amount to the order of a millivolt. This magnitude is well above the noise which accompanies any radio receiver and which sets the ultimate limit of signal detection. It should be observed here that, being a phenomenon of fluctuations, the noise can always be reduced by averaging over sufficiently long times. This procedure was used later to very greatly increase the sensitivity of the method; it is characteristic of the present possibilities that my collaborators have succeeded in the last few years in detecting in natural water signals arising from deuterium and from the isotope of oxygen with atomic mass 17, despite their low abundances of 0.02 and 0.04 percent, respectively.

The existence and detection of a precessing nuclear polarization in a sample represents to my mind the basis of nuclear induction. It is, however, necessary to consider also the features which produce and counteract the tilt of the polarization with respect to the constant field. Magnetic resonance enters here as the most important means of producing the tilt, since it allows its achievement under the application of relatively weak oscillating fields. In fact, it is a common feature of every resonance phenomenon that relatively weak external forces can produce large effects if their frequency is equal to the natural frequency of the system to which they are applied. The natural frequency in question is, in our case, that with which the nuclear polarization precesses by itself around the constant field and the practical way to determine this frequency is to vary either that of the applied alternating field or the magnitude of the constant field until resonance conditions are

established and detected by a maximum of the observed nuclear induction signal. The simultaneous knowledge of resonance field and frequency then directly yields, as in the use of magnetic resonance in molecular beams, the gyromagnetic ratio and, with a knowledge of the spin, the magnetic moment of the nucleus. Actually, it is also possible to determine the spin separately by using the additional piece of information contained in the intensity of the observed signal.

To follow the analogue of mechanical resonance we must now come back to relaxation, which can be seen to act like a friction, and which counteracts the tilt produced by the alternating field. If the friction is large, i.e., if the relaxation time is short, it will either reduce the effect for a given amplitude or require a correspondingly larger amplitude of the alternating field. It will, in either case, result in a relatively broad resonance line, thus diminishing the accuracy of the measurement. While from this point of view it is undesirable to have too short a relaxation time, it is equally undesirable to have it too long, since the very circumstance of producing its tilt diminishes the magnitude of the polarization so that it requires the refreshing influence of the relaxation mechanism to bring it back to its equilibrium value.

There was not much known about the magnitude of nuclear relaxation times when Purcell and I started our first experiments on nuclear induction, and the main doubt about their success arose from the possibility of insufficient relaxation. In fact, it seems, in retrospect, that the failure of Gorter's first attempt⁶, as well as of a second one, undertaken in 1942¹⁴, was primarily due to this circumstance. While E. M. Purcell, H. C. Torrey and R. V. Pound¹⁵, toward the end of 1945, obtained their first positive results from protons in paraffin, the late W. W. Hansen, M. E. Packard, and I¹⁶ found ours a few weeks later in water without either of the groups knowing anything about the work of the other. The relaxation time of paraffin has the convenient value of about $\frac{1}{100}$ second, while pure water has a somewhat unfavorably long relaxation time of about 2 seconds. Neither of these two values had been foreseen, and I was fully prepared to find the relaxation time of pure water considerably longer and in fact too long for our method of observation. It was known, however, that the conversion of ortho- and para-hydrogen was accelerated by the presence of paramagnetic atoms and molecules; this mechanism has the common feature, with the attainment of the equilibrium polarization of protons, that it requires a random process of nuclear reorientation, and it had been understood to take place through the magnetic field of the paramagnetic catalyst acting upon the magnetic mo-

ment of the proton. An estimate showed that, depending upon the concentration of a paramagnetic salt dissolved in water, a wide range of relaxation times, going down to values of the order of 10^{-5} second, could be obtained. Before starting our observations we therefore had prepared solutions of the paramagnetic iron nitrate in water, and although the first weak signals were received from pure water we found, shortly afterward, considerably stronger signals from a solution with about one-half molar concentration of iron nitrate. The signals appeared as rather broad lines on the cathode-ray oscilloscope because of insufficient homogeneity of the constant magnetic field.

Since the width of the resonance line determines the accuracy with which magnetic moments can be determined, we shall briefly consider the conditions necessary for obtaining sharp lines. In the first place, it is necessary that the constant field have the same value in all parts of the sample; in the second place, one must not choose an excessive amplitude of the alternating field, since this too would cause an excessive broadening. The ultimate limit is given by the natural width of the line and it is closely related to the relaxation time; it can be seen, in fact, that the relative accuracy of a measurement by nuclear induction, due to the natural line width, is limited to the order of the number of cycles which the nuclear polarization carries out in its precession around the constant field during the relaxation time. As an example, we shall again consider protons in pure water and in a field of 10,000 gauss; the frequency of precession is here 42.5 megacycles per second, so that about 10^8 cycles are performed during the relaxation time of approximately 2 seconds. This means that an accuracy of about 1 part in 100 millions could be, in principle, achieved here, provided that one had a sufficiently homogeneous field available. While this limit has not yet been reached, it is noteworthy that in water, alcohol, and other liquids, resolutions of one part in 10 millions have actually been achieved. It is indeed the possibility of coherent observation over a large number of cycles which allows the use of nuclear induction as a method of high precision measurements. In fulfillment of my original plans, it was applied by H. H. Staub, D. B. Nicodemus and me to the magnetic moments of the proton, the neutron and the deuteron¹⁷, and it resulted not only in the verification of the previously mentioned deviation from additivity in the deuteron¹³, but in its measurement with an accuracy which is beyond the present scope of the theory of nuclear forces. It was particularly gratifying to me to obtain these results from experiments combining the polarization and magnetic resonance depolarization of neutrons with nuclear induction.

The description of nuclear induction which I have presented follows closely my own original thoughts on the subject but it can equally well be approached from other angles. The simplest one is probably that of Gorter⁶ in his first attempt to detect nuclear magnetic resonance. We have seen before that the alternating field tilts the nuclear polarization against the constant field. This process requires a certain amount of work which, through relaxation, will reappear in the form of heat produced in the sample. The effect in fact does not involve induction but represents pure nuclear resonance absorption; however, it would be very slight and has not yet been established. A second attempt of Gorter¹⁴, carried out later, is based upon the fact that the nuclear paramagnetic susceptibility has a maximum for radiofrequencies, corresponding to magnetic resonance conditions; it would manifest itself in the frequency of an electric oscillator of which a coil, surrounding the sample, forms the self-inductance. This scheme is actually one of the many others which can be devised for the observation of nuclear induction and, if successful, would have represented the first demonstration of the effect. Purcell's first successful experiment involved the electrodynamical aspect of absorption insofar as its occurrence under resonance conditions was manifested through the increased loss of a cavity resonator; the cavity was replaced in his succeeding arrangements by more conventional circuit elements. A particularly suitable and convenient arrangement consists of a radiofrequency bridge, which contains in one arm a coil, surrounding the sample. As a consequence of nuclear induction there occurs, under resonance conditions, a change of the impedance of this coil and thereby a readily detectable change in the balance of the bridge. It should be remarked that the change of impedance is complex, with its real part corresponding to absorption, its imaginary part to dispersion. This fact can be traced back to the phase relation between the nuclear induction signal and the applied radiofrequency field, and the phase sensitivity of the bridge allows the observation of the effect either as absorption or as dispersion or as a combination of both.

Finally, I shall give a brief description of our own original arrangement which we still use in its principal features. The essential balance which Purcell has obtained by a bridge method is here to a large extent achieved geometrically by using two radiofrequency coils with their axes oriented at right angles to each other and to the constant field. One of them, the « transmitter coil » produces the alternating field, while the other, the « receiver coil », serves for detection of the nuclear induction signal (see Figs. 1 and 2). A small amount of coupling between the two coils is admitted to produce

a voltage across the receiver coil, and its phase with respect to the superimposed voltage induced by the nuclei can be adjusted for the observation of either absorption or dispersion in similarity to the bridge method (see Figs. 3 and 4).

A considerable variety of other circuits has been introduced by different investigators. Except for the greater or lesser ease of avoiding instrumental difficulties, they lead to the same ultimate sensitivity and accuracy of the method, since they all observe the same basic phenomenon.

There is, however, one distinctive feature in the crossed-coil arrangement, which automatically yields another significant piece of information. The two coils imply a sense of rotation around the constant field; depending upon whether the nuclear polarization precesses in the same or the opposite sense of rotation there results a phase difference of 180 degrees between the voltage in the receiver coil due to coupling with the transmitter coil and the superimposed voltage due to nuclear induction. The action of the rectifier translates this phase difference into an inversion of the signal, which is directly displayed on the oscillograph or on the recording instrument. One obtains in this simple manner information about the sign of the magnetic moment of the nucleus, defined by its relative orientation to the angular momentum, since it is this sign which determines the sense of rotation of the nuclear polarization in a given field. The sign of nuclear moments represents an important clue to their interpretation in terms of nuclear structures; usually it is referred to the sign of the proton moment, which has been known for a considerable time to be positive. It has been determined in this manner for a number of nuclei where it was not previously known.

1. W. Pauli, *Naturwiss.*, 12 (1924) 741.
2. R. Frisch and O. Stern, *Z. Physik*, 85 (1933) 4.
I. Estermann and O. Stern, *Z. Physik*, 85 (1933) 17.
I. Estermann and O. Stern, *Phys. Rev.*, 46 (1934) 665.
3. G. Breit and I. I. Rabi, *Phys. Rev.*, 38 (1931) 2082.
4. E. Fermi, *Z. Physik*, 60 (1930) 320.
5. J. M. B. Kellogg, I. I. Rabi, N. F. Ramsey, and J. R. Zacharias, *Phys. Rev.*, 55 (1939) 318; 57 (1940) 677.
6. C. J. Gorter, *Physica*, 3 (1936) 995.
7. I. I. Rabi, *Phys. Rev.*, 51 (1937) 652.
8. P. A. M. Dirac, *Proc. Roy. Soc. London*, 117 (1928) 610.

9. F. Bloch, *Phys. Rev.*, 50 (1936) 259; 51 (1937) 994.
10. P. N. Powers, H. G. Beyer, and J. R. Dunning, *Phys. Rev.*, 51 (1937) 371.
11. L. W. Alvarez and F. Bloch, *Phys. Rev.*, 57 (1940) III.
12. F. Bloch, M. Hamermesh, and H. Staub, *Phys. Rev.*, 64 (1943) 47.
13. W. Rarita and J. Schwinger, *Phys. Rev.*, 59 (1941) 436.
14. C. J. Gorter and L. J. F. Broer, *Physica*, 9 (1942) 591.
15. E. M. Purcell, H. C. Torrey, and R. V. Pound, *Phys. Rev.*, 69 (1946) 37.
N. Bloembergen, E. M. Purcell, and R. V. Pound, *Phys. Rev.*, 73 (1948) 679.
16. F. Bloch, W. W. Hansen, and M. Packard, *Phys. Rev.*, 69 (1946) 127.
F. Bloch, *Phys. Rev.*, 70 (1946) 460.
17. F. Bloch, D. Nicodemus, and H. Staub, *Phys. Rev.*, 74 (1948) 1025.

NIELS BOHR

The structure of the atom

Nobel Lecture, December 11, 1922

Ladies and Gentlemen. Today, as a consequence of the great honour the Swedish Academy of Sciences has done me in awarding me this year's Nobel Prize for Physics for my work on the structure of the atom, it is my duty to give an account of the results of this work and I think that I shall be acting in accordance with the traditions of the Nobel Foundation if I give this report in the form of a survey of the development which has taken place in the last few years within the field of physics to which this work belongs.

The general picture of the atom

The present state of atomic theory is characterized by the fact that we not only believe the existence of atoms to be proved beyond a doubt, but also we even believe that we have an intimate knowledge of the constituents of the individual atoms. I cannot on this occasion give a survey of the scientific developments that have led to this result; I will only recall the discovery of the electron towards the close of the last century, which furnished the direct verification and led to a conclusive formulation of the conception of the atomic nature of electricity which had evolved since the discovery by Faraday of the fundamental laws of electrolysis and Berzelius's electrochemical theory, and had its greatest triumph in the electrolytic dissociation theory of Arrhenius. This discovery of the electron and elucidation of its properties was the result of the work of a large number of investigators, among whom Lenard and J. J. Thomson may be particularly mentioned. The latter especially has made very important contributions to our subject by his ingenious attempts to develop ideas about atomic constitution on the basis of the electron theory. The present state of our knowledge of the elements of atomic structure was reached, however, by the discovery of the atomic nucleus, which we owe to Rutherford, whose work on the radioactive substances discovered towards the close of the last century has much enriched physical and chemical science.

According to our present conceptions, an atom of an element is built up of a nucleus that has a positive electrical charge and is the seat of by far the greatest part of the atomic mass, together with a number of electrons, all having the same negative charge and mass, which move at distances from the nucleus that are very great compared to the dimensions of the nucleus or of the electrons themselves. In this picture we at once see a striking resemblance to a planetary system, such as we have in our own solar system. Just as the simplicity of the laws that govern the motions of the solar system is intimately connected with the circumstance that the dimensions of the moving bodies are small in relation to the orbits, so the corresponding relations in atomic structure provide us with an explanation of an essential feature of natural phenomena in so far as these depend on the properties of the elements. It makes clear at once that these properties can be divided into two sharply distinguished classes.

To the first class belong most of the ordinary physical and chemical properties of substances, such as their state of aggregation, colour, and chemical reactivity. These properties depend on the motion of the electron system and the way in which this motion changes under the influence of different external actions. On account of the large mass of the nucleus relative to that of the electrons and its smallness in comparison to the electron orbits, the electronic motion will depend only to a very small extent on the nuclear mass, and will be determined to a close approximation solely by the total electrical charge of the nucleus. Especially the inner structure of the nucleus and the way in which the charges and masses are distributed among its separate particles will have a vanishingly small influence on the motion of the electron system surrounding the nucleus. On the other hand, the structure of the nucleus will be responsible for the second class of properties that are shown in the radioactivity of substances. In the radioactive processes we meet with an explosion of the nucleus, whereby positive or negative particles, the so-called α - and β -particles, are expelled with very great velocities.

Our conceptions of atomic structure afford us, therefore, an immediate explanation of the complete lack of interdependence between the two classes of properties, which is most strikingly shown in the existence of substances which have to an extraordinarily close approximation the same ordinary physical and chemical properties, even though the atomic weights are not the same, and the radioactive properties are completely different. Such substances, of the existence of which the first evidence was found in the work of Soddy and other investigators on the chemical properties of the radioactive

elements, are called isotopes, with reference to the classification of the elements according to ordinary physical and chemical properties. It is not necessary for me to state here how it has been shown in recent years that isotopes are found not only among the radioactive elements, but also among ordinary stable elements; in fact, a large number of the latter that were previously supposed simple have been shown by Aston's well-known investigations to consist of a mixture of isotopes with different atomic weights.

The question of the inner structure of the nucleus is still but little understood, although a method of attack is afforded by Rutherford's experiments on the disintegration of atomic nuclei by bombardment with α -particles. Indeed, these experiments may be said to open up a new epoch in natural philosophy in that for the first time the artificial transformation of one element into another has been accomplished. In what follows, however, we shall confine ourselves to a consideration of the ordinary physical and chemical properties of the elements and the attempts which have been made to explain them on the basis of the concepts just outlined.

It is well known that the elements can be arranged as regards their ordinary physical and chemical properties in a *natural system* which displays most suggestively the peculiar relationships between the different elements. It was recognized for the first time by Mendeleev and Lothar Meyer that when the elements are arranged in an order which is practically that of their atomic weights, their chemical and physical properties show a pronounced periodicity. A diagrammatic representation of this so-called Periodic Table is given in Fig. 1, where, however, the elements are not arranged in the ordinary way but in a somewhat modified form of a table first given by Julius Thomsen, who has also made important contributions to science in this domain. In the figure the elements are denoted by their usual chemical symbols, and the different vertical columns indicate the so-called periods. The elements in successive columns which possess homologous chemical and physical properties are connected with lines. The meaning of the square brackets around certain series of elements in the later periods, the properties of which exhibit typical deviations from the simple periodicity in the first periods, will be discussed later.

In the development of the theory of atomic structure the characteristic features of the natural system have found a surprisingly simple interpretation. Thus we are led to assume that the ordinal number of an element in the Periodic Table, the so-called atomic number, is just equal to the number of electrons which move about the nucleus in the neutral atom. In an imperfect

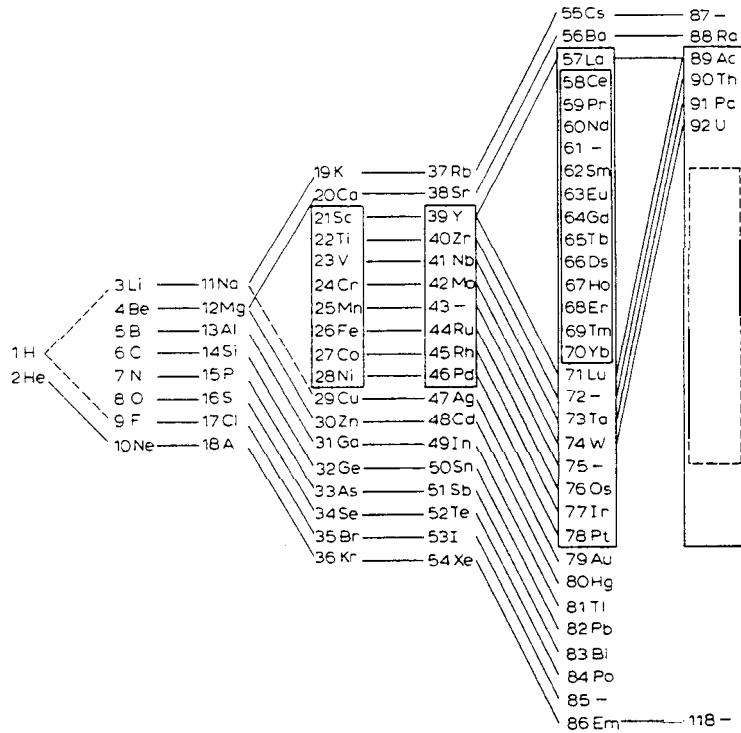


Fig. 1.

form, this law was first stated by Van den Broek; it was, however, foreshadowed by J. J. Thomson's investigations of the number of electrons in the atom, as well as by Rutherford's measurements of the charge on the atomic nucleus. As we shall see, convincing support for this law has since been obtained in various ways, especially by Moseley's famous investigations of the X-ray spectra of the elements. We may perhaps also point out, how the simple connexion between atomic number and nuclear charge offers an explanation of the laws governing the changes in chemical properties of the elements after expulsion of α - or β -particles, which found a simple formulation in the so-called radioactive displacement law.

Atomic stability ad electrodynamic theory

As soon as we try to trace a more intimate connexion between the properties of the elements and atomic structure, we encounter profound difficulties, in

that essential differences between an atom and a planetary system show themselves here in spite of the analogy we have mentioned.

The motions of the bodies in a planetary system, even though they obey the general law of gravitation, will not be completely determined by this law alone, but will depend largely on the previous history of the system. Thus the length of the year is not determined by the masses of the sun and the earth alone, but depends also on the conditions that existed during the formation of the solar system, of which we have very little knowledge. Should a sufficiently large foreign body some day traverse our solar system, we might among other effects expect that from that day the length of the year would be different from its present value.

It is quite otherwise in the case of atoms. The definite and unchangeable properties of the elements demand that the state of an atom cannot undergo permanent changes due to external actions. As soon as the atom is left to itself again, its constituent particles must arrange their motions in a manner which is completely determined by the electric charges and masses of the particles. We have the most convincing evidence of this in spectra, that is, in the properties of the radiation emitted from substances in certain circumstances, which can be studied with such great precision. It is well known that the wavelengths of the spectral lines of a substance, which can in many cases be measured with an accuracy of more than one part in a million, are, in the same external circumstances, always exactly the same within the limit of error of the measurements, and quite independent of the previous treatment of this substance. It is just to this circumstance that we owe the great importance of spectral analysis, which has been such an invaluable aid to the chemist in the search for new elements, and has also shown us that even on the most distant bodies of the universe there occur elements with exactly the same properties as on the earth.

On the basis of our picture of the constitution of the atom it is thus impossible, so long as we restrict ourselves to the ordinary mechanical laws, to account for the characteristic atomic stability which is required for an explanation of the properties of the elements.

The situation is by no means improved if we also take into consideration the well-known electrodynamic laws which Maxwell succeeded in formulating on the basis of the great discoveries of Oersted and Faraday in the first half of the last century. Maxwell's theory has not only shown itself able to account for the already known electric and magnetic phenomena in all their details, but has also celebrated its greatest triumph in the prediction of the

electromagnetic waves which were discovered by Hertz, and are now so extensively used in wireless telegraphy.

For a time it seemed as though this theory would also be able to furnish a basis for an explanation of the details of the properties of the elements, after it had been developed, chiefly by Lorentz and Larmor, into a form consistent with the atomistic conception of electricity. I need only remind you of the great interest that was aroused when Lorentz, shortly after the discovery by Zeeman of the characteristic changes that spectral lines undergo when the emitting substance is brought into a magnetic field, could give a natural and simple explanation of the main features of the phenomenon. Lorentz assumed that the radiation which we observe in a spectral line is sent out from an electron executing simple harmonic vibrations about a position of equilibrium, in precisely the same manner as the electromagnetic waves in radiotelegraphy are sent out by the electric oscillations in the antenna. He also pointed out how the alteration observed by Zeeman in the spectral lines corresponded exactly to the alteration in the motion of the vibrating electron which one would expect to be produced by the magnetic field.

It was, however, impossible on this basis to give a closer explanation of the spectra of the elements, or even of the general type of the laws holding with great exactness for the wavelengths of lines in these spectra, which had been established by Balmer, Rydberg, and Ritz. After we obtained details as to the constitution of the atom, this difficulty became still more manifest; in fact, so long as we confine ourselves to the classical electrodynamic theory we cannot even understand why we obtain spectra consisting of sharp lines at all. This theory can even be said to be incompatible with the assumption of the existence of atoms possessing the structure we have described, in that the motions of the electrons would claim a continuous radiation of energy from the atom, which would cease only when the electrons had fallen into the nucleus.

The origin of the quantum theory

It has, however, been possible to avoid the various difficulties of the electrodynamic theory by introducing concepts borrowed from the so-called quantum theory, which marks a complete departure from the ideas that have hitherto been used for the explanation of natural phenomena. This theory was originated by Planck, in the year 1900, in his investigations on the law

of heat radiation, which, because of its independence of the individual properties of substances, lent itself peculiarly well to a test of the applicability of the laws of classical physics to atomic processes.

Planck considered the equilibrium of radiation between a number of systems with the same properties as those on which Lorentz had based his theory of the Zeeman effect, but he could now show not only that classical physics could not account for the phenomena of heat radiation, but also that a complete agreement with the experimental law could be obtained if - in pronounced contradiction to classical theory - it were assumed that the energy of the vibrating electrons could not change continuously, but only in such a way that the energy of the system always remained equal to a whole number of so-called energy-quanta. The magnitude of this quantum was found to be proportional to the frequency of oscillation of the particle, which, in accordance with classical concepts, was supposed to be also the frequency of the emitted radiation. The proportionality factor had to be regarded as a new universal constant, since termed Planck's constant, similar to the velocity of light, and the charge and mass of the electron.

Planck's surprising result stood at first completely isolated in natural science, but with Einstein's significant contributions to this subject a few years after, a great variety of applications was found. In the first place, Einstein pointed out that the condition limiting the amount of vibrational energy of the particles could be tested by investigation of the specific heat of crystalline bodies, since in the case of these we have to do with similar vibrations, not of a single electron, but of whole atoms about positions of equilibrium in the crystal lattice. Einstein was able to show that the experiment confirmed Planck's theory, and through the work of later investigators this agreement has proved quite complete. Furthermore, Einstein emphasized another consequence of Planck's results, namely, that radiant energy could only be emitted or absorbed by the oscillating particle in so-called «quanta of radiation», the magnitude of each of which was equal to Planck's constant multiplied by the frequency.

In his attempts to give an interpretation of this result, Einstein was led to the formulation of the so-called « "hypothesis of light-quanta", according to which the radiant energy, in contradiction to Maxwell's electromagnetic theory of light, would not be propagated as electromagnetic waves, but rather as concrete light atoms, each with an energy equal to that of a quantum of radiation. This concept led Einstein to his well-known theory of the photoelectric effect. This phenomenon, which had been entirely unexplain-

able on the classical theory, was thereby placed in a quite different light, and the predictions of Einstein's theory have received such exact experimental confirmation in recent years, that perhaps the most exact determination of Planck's constant is afforded by measurements on the photoelectric effect. In spite of its heuristic value, however, the hypothesis of light-quanta, which is quite irreconcilable with so-called interference phenomena, is not able to throw light on the nature of radiation. I need only recall that these interference phenomena constitute our only means of investigating the properties of radiation and therefore of assigning any closer meaning to the frequency which in Einstein's theory fixes the magnitude of the light-quantum.

In the following years many efforts were made to apply the concepts of the quantum theory to the question of atomic structure, and the principal emphasis was sometimes placed on one and sometimes on the other of the consequences deduced by Einstein from Planck's result. As the best known of the attempts in this direction, from which, however, no definite results were obtained, I may mention the work of Stark, Sommerfeld, Hasenöhrl, Haas, and Nicholson.

From this period also dates an investigation by Bjerrum on infrared absorption bands, which, although it had no direct bearing on atomic structure, proved significant for the development of the quantum theory. He directed attention to the fact that the rotation of the molecules in a gas might be investigated by means of the changes in certain absorption lines with temperature. At the same time he emphasized the fact that the effect should not consist of a continuous widening of the lines such as might be expected from classical theory, which imposed no restrictions on the molecular rotations, but in accordance with the quantum theory he predicted that the lines should be split up into a number of components, corresponding to a sequence of distinct possibilities of rotation. This prediction was confirmed a few years later by Eva von Bahr, and the phenomenon may still be regarded as one of the most striking evidences of the reality of the quantum theory, even though from our present point of view the original explanation has undergone a modification in essential details.

The quantum theory of atomic constitution

The question of further development of the quantum theory was in the meantime placed in a new light by Rutherford's discovery of the atomic nu-

cleus (1911). As we have already seen, this discovery made it quite clear that by classical conceptions alone it was quite impossible to understand the most essential properties of atoms. One was therefore led to seek for a formulation of the principles of the quantum theory that could immediately account for the stability in atomic structure and the properties of the radiation sent out from atoms, of which the observed properties of substances bear witness. Such a formulation was proposed (1913) by the present lecturer in the form of two postulates, which may be stated as follows:

- (1). Among the conceivably possible states of motion in an atomic system there exist a number of so-called *stationary states* which, in spite of the fact that the motion of the particles in these states obeys the laws of classical mechanics to a considerable extent, possess a peculiar, mechanically unexplainable stability, of such a sort that every permanent change in the motion of the system must consist in a complete transition from one stationary state to another.
- (2). While in contradiction to the classical electromagnetic theory no radiation takes place from the atom in the stationary states themselves, a process of transition between two stationary states can be accompanied by the emission of electromagnetic radiation, which will have the same properties as that which would be sent out according to the classical theory from an electrified particle executing an harmonic vibration with constant frequency. This frequency ν has, however, no simple relation to the motion of the particles of the atom, but is given by the relation

$$h\nu = E' - E'',$$

where h is Planck's constant, and E' and E'' are the values of the energy of the atom in the two stationary states that form the initial and final state of the radiation process. Conversely, irradiation of the atom with electromagnetic waves of this frequency can lead to an absorption process, whereby the atom is transformed back from the latter stationary state to the former.

While the first postulate has in view the general stability of the atom, the second postulate has chiefly in view the existence of spectra with sharp lines. Furthermore, the quantum-theory condition entering in the last postulate affords a starting-point for the interpretation of the laws of series spectra.

The most general of these laws, the combination principle enunciated by Ritz, states that the frequency ν for each of the lines in the spectrum of an element can be represented by the formula

$$\nu = T'' - T',$$

where T'' and T' are two so-called « spectral terms » belonging to a manifold of such terms characteristic of the substance in question.

According to our postulates, this law finds an immediate interpretation in the assumption that the spectrum is emitted by transitions between a number of stationary states in which the numerical value of the energy of the atom is equal to the value of the spectral term multiplied by Planck's constant. This explanation of the combination principle is seen to differ fundamentally from the usual ideas of electrodynamics, as soon as we consider that there is no simple relation between the motion of the atom and the radiation sent out. The departure of our considerations from the ordinary ideas of natural philosophy becomes particularly evident, however, when we observe that the occurrence of two spectral lines, corresponding to combinations of the same spectral term with two other different terms, implies that the nature of the radiation sent out from the atom is not determined only by the motion of the atom at the-beginning of the radiation process, but also depends on the state to which the atom is transferred by the process.

At first glance one might, therefore, think that it would scarcely be possible to bring our formal explanation of the combination principle into direct relation with our views regarding the constitution of the atom, which, indeed, are based on experimental evidence interpreted on classical mechanics and electrodynamics. A closer investigation, however, should make it clear that a definite relation may be obtained between the spectra of the elements and the structure of their atoms on the basis of the postulates.

The hydrogen spectrum

The simplest spectrum we know is that of hydrogen. The frequencies of its lines may be represented with great accuracy by means of Balmer's formula:

$$\nu = K \left(\frac{1}{n''^2} - \frac{1}{n'^2} \right),$$

where K is a constant and n' and n'' are two integers. In the spectrum we accordingly meet a single series of spectral terms of the form K/n^2 , which decrease regularly with increasing term number n . In accordance with the postulates, we shall therefore assume that each of the hydrogen lines is emitted by a transition between two states belonging to a series of stationary states of the hydrogen atom in which the numerical value of the atom's energy is equal to hK/n^2 .

Following our picture of atomic structure, a hydrogen atom consists of a positive nucleus and an electron which - so far as ordinary mechanical conceptions are applicable - will with great approximation describe a periodic elliptical orbit with the nucleus at one focus. The major axis of the orbit is inversely proportional to the work necessary completely to remove the electron from the nucleus, and, in accordance with the above, this work in the stationary states is just equal to hK/n^2 . We thus arrive at a manifold of stationary states for which the major axis of the electron orbit takes on a series of discrete values proportional to the squares of the whole numbers. The accompanying Fig. 2 shows these relations diagrammatically. For the sake of simplicity the electron orbits in the stationary states are represented by circles, although in reality the theory places no restriction on the eccentricity of the orbit, but only determines the length of the major axis. The arrows represent the transition processes that correspond to the red and

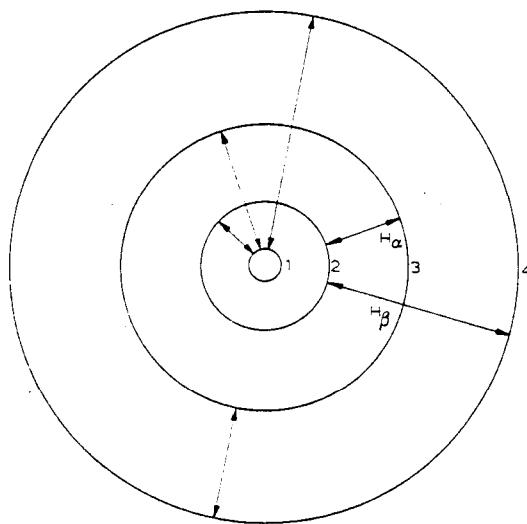


Fig. 2.

green hydrogen lines, H_{α} and H_{β} , the frequency of which is given by means of the Balmer formula when we put $n'' = 2$ and $n' = 3$ and 4 respectively. The transition processes are also represented which correspond to the first three lines of the series of ultraviolet lines found by Lyman in 1914, of which the frequencies are given by the formula when n is put equal to 1, as well as to the first line of the infrared series discovered some years previously by Paschen, which are given by the formula if n'' is put equal to 3.

This explanation of the origin of the hydrogen spectrum leads us quite naturally to interpret this spectrum as the manifestation of a process whereby the electron is bound to the nucleus. While the largest spectral term with term number 1 corresponds to the final stage in the binding process, the small spectral terms that have larger values of the term number correspond to stationary states which represent the initial states of the binding process, where the electron orbits still have large dimensions, and where the work required to remove an electron from the nucleus is still small. The final stage in the binding process we may designate as the normal state of the atom, and it is distinguished from the other stationary states by the property that, in accordance with the postulates, the state of the atom can only be changed by the addition of energy whereby the electron is transferred to an orbit of larger dimensions corresponding to an earlier stage of the binding process.

The size of the electron orbit in the normal state calculated on the basis of the above interpretation of the spectrum agrees roughly with the value for the dimensions of the atoms of the elements that have been calculated by the kinetic theory of matter from the properties of gases. Since, however, as an immediate consequence of the stability of the stationary states that is claimed by the postulates, we must suppose that the interaction between two atoms during a collision cannot be completely described with the aid of the laws of classical mechanics, such a comparison as this cannot be carried further on the basis of such considerations as those just outlined.

A more intimate connexion between the spectra and the atomic model has been revealed, however, by an investigation of the motion in those stationary states where the term number is large, and where the dimensions of the electron orbit and the frequency of revolution in it vary relatively little when we go from one stationary state to the next following. It was possible to show that the frequency of the radiation sent out during the transition between two stationary states, the difference of the term numbers of which is small in comparison to these numbers themselves, tended to coincide in frequency with one of the harmonic components into which the

electron motion could be resolved, and accordingly also with the frequency of one of the wave trams in the radiation which would be emitted according to the laws of ordinary electrodynamics.

The condition that such a coincidence should occur in this region where the stationary states differ but little from one another proves to be that the constant in the Balmer formula can be expressed by means of the relation

$$K = \frac{2\pi^2 e^4 m}{h^3},$$

where e and m are respectively the charge and mass of the electron, while h is Planck's constant. This relation has been shown to hold to within the considerable accuracy with which, especially through the beautiful investigations of Millikan, the quantities e , m , and h are known.

This result shows that there exists a connexion between the hydrogen spectrum and the model for the hydrogen atom which, on the whole, is as close as we might hope considering the departure of the postulates from the classical mechanical and electrodynamic laws. At the same time, it affords some indication of how we may perceive in the quantum theory, in spite of the fundamental character of this departure, a natural generalization of the fundamental concepts of the classical electrodynamic theory. To this most important question we shall return later, but first we will discuss how the interpretation of the hydrogen spectrum on the basis of the postulates has proved suitable in several ways, for elucidating the relation between the properties of the different elements.

Relationships between the elements

The discussion above can be applied immediately to the process whereby an electron is bound to a nucleus with any given charge. The calculations show that, in the stationary state corresponding to a given value of the number n , the size of the orbit will be inversely proportional to the nuclear charge, while the work necessary to remove an electron will be directly proportional to the square of the nuclear charge. The spectrum that is emitted during the binding of an electron by a nucleus with charge N times that of the hydrogen nucleus can therefore be represented by the formula:

$$\nu = N^2 K \left(\frac{I}{n''^2} - \frac{I}{n'^2} \right).$$

If in this formula we put $N = 2$, we get a spectrum which contains a set of lines in the visible region which was observed many years ago in the spectrum of certain stars. Rydberg assigned these lines to hydrogen because of the close analogy with the series of lines represented by the Balmer formula. It was never possible to produce these lines in pure hydrogen, but just before the theory for the hydrogen spectrum was put forward, Fowler succeeded in observing the series in question by sending a strong discharge through a mixture of hydrogen and helium. This investigator also assumed that the lines were hydrogen lines, because there existed no experimental evidence from which it might be inferred that two different substances could show properties resembling each other so much as the spectrum in question and that of hydrogen. After the theory was put forward, it became clear, however, that the observed lines must belong to a spectrum of helium, but that they were not like the ordinary helium spectrum emitted from the neutral atom. They came from an ionized helium atom which consists of a single electron moving about a nucleus with double charge. In this way there was brought to light a new feature of the relationship between the elements, which corresponds exactly with our present ideas of atomic structure, according to which the physical and chemical properties of an element depend in the first instance only on the electric charge of the atomic nucleus.

Soon after this question was settled the existence of a similar general relationship between the properties of the elements was brought to light by Moseley's well-known investigations on the characteristic X-ray spectra of the elements, which was made possible by Laue's discovery of the interference of X-rays in crystals and the investigations of W. H. and W. L. Bragg on this subject. It appeared, in fact, that the X-ray spectra of the different elements possessed a much simpler structure and a much greater mutual resemblance than their optical spectra. In particular, it appeared that the spectra changed from element to element in a manner that corresponded closely to the formula given above for the spectrum emitted during the binding of an electron to a nucleus, provided N was put equal to the atomic number of the element concerned. This formula was even capable of expressing, with an approximation that could not be without significance, the frequencies of the strongest X-ray lines, if small whole numbers were substituted for n' and n'' .

This discovery was of great importance in several respects. In the first place, the relationship between the X-ray spectra of different elements proved so simple that it became possible to fix without ambiguity the atomic number for all known substances, and in this way to predict with certainty the atomic number of all such hitherto unknown elements for which there is a place in the natural system. Fig. 3 shows how the square root of the frequency for two characteristic X-ray lines depends on the atomic number. These lines belong to the group of so-called K-lines, which are the most penetrating of the characteristic rays. With very close approximation the points lie on straight lines, and the fact that they do so is conditioned not only by our taking account of known elements, but also by our leaving an open place between molybdenum (42) and ruthenium (44), just as in Mendeleev's original scheme of the natural system of the elements.

Further, the laws of X-ray spectra provide a confirmation of the general theoretical conceptions, both with regard to the constitution of the atom and the ideas that have served as a basis for the interpretation of spectra. Thus

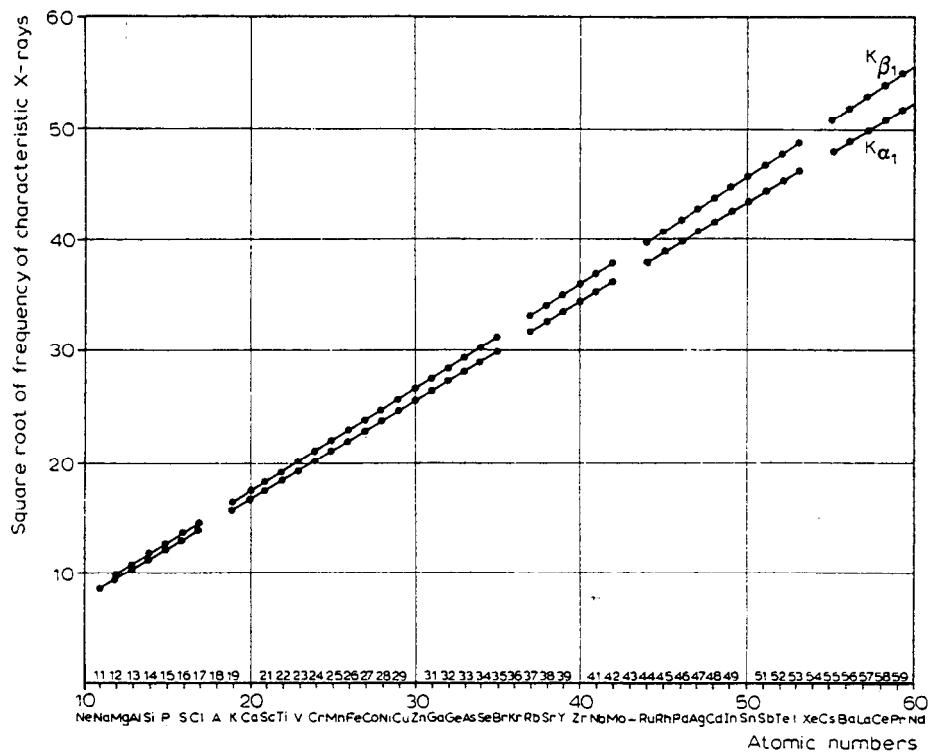


Fig. 3.

the similarity between X-ray spectra and the spectra emitted during the binding of a single electron to a nucleus may be simply interpreted from the fact that the transitions between stationary states with which we are concerned in X-ray spectra are accompanied by changes in the motion of an electron in the inner part of the atom, where the influence of the attraction of the nucleus is very great compared with the repulsive forces of the other electrons.

The relations between other properties of the elements are of a much more complicated character, which originates in the fact that we have to do with processes concerning the motion of the electrons in the outer part of the atom, where the forces that the electrons exert on one another are of the same order of magnitude as the attraction towards the nucleus, and where, therefore, the details of the interaction of the electrons play an important part. A characteristic example of such a case is afforded by the spatial extension of the atoms of the elements. Lothar Meyer himself directed attention to the characteristic periodic change exhibited by the ratio of the atomic weight to the density, the so-called atomic volume, of the elements in the natural system. An idea of these facts is given by Fig. 4, in which the atomic volume is represented as a function of the atomic number. A greater difference be-

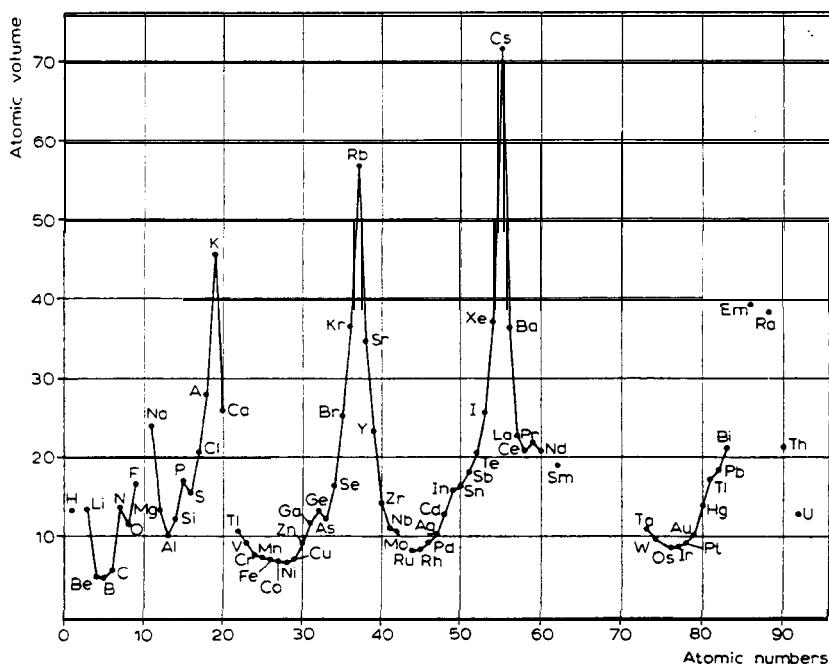


Fig. 4.

tween this and the previous figure could scarcely be imagined. While the X-ray spectra vary uniformly with the atomic number, the atomic volumes show a characteristic periodic change which corresponds exactly to the change in the chemical properties of the elements.

Ordinary optical spectra behave in an analogous way. In spite of the dissimilarity between these spectra, Rydberg succeeded in tracing a certain general relationship between the hydrogen spectrum and other spectra. Even though the spectral lines of the elements with higher atomic number appear as combinations of a more complicated manifold of spectral terms which is not so simply co-ordinated with a series of whole numbers, still the spectral terms can be arranged in series each of which shows a strong similarity to the series of terms in the hydrogen spectrum. This similarity appears in the fact that the terms in each series can, as Rydberg pointed out, be very accurately represented by the formula $K/(n + \alpha)^2$, where K is the same constant that occurs in the hydrogen spectrum, often called the Rydberg constant, while n is the term number, and α a constant which is different for the different series.

This relationship with the hydrogen spectrum leads us immediately to regard these spectra as the *last step of a process whereby the neutral atom is built up by the capture and binding of electrons to the nucleus*, one by one. In fact, it is clear that the last electron captured, so long as it is in that stage of the binding process in which its orbit is still large compared to the orbits of the previously bound electrons, will be subjected to a force from the nucleus and these electrons, that differs but little from the force with which the electron in the hydrogen atom is attracted towards the nucleus while it is moving in an orbit of corresponding dimensions.

The spectra so far considered, for which Rydberg's laws hold, are excited by means of electric discharge under ordinary conditions and are often called arc spectra. The elements emit also another type of spectrum, the so-called spark spectra, when they are subjected to an extremely powerful discharge. Hitherto it was impossible to disentangle the spark spectra in the same way as the arc spectra. Shortly after the above view on the origin of arc spectra was brought forward, however, Fowler found (1914) that an empirical expression for the spark spectrum lines could be established which corresponds exactly to Rydberg's laws with the single difference that the constant K is replaced by a constant four times as large. Since, as we have seen, the constant that appears in the spectrum sent out during the binding of an electron to a helium nucleus is exactly equal to $4 K$, it becomes evident that spark

spectra are due to the ionized atom, and that their emission corresponds to *the last step but one in the formation of the neutral atom* by the successive capture and binding of electrons.

Absorption and excitation of spectral lines

The interpretation of the origin of the spectra was also able to explain the characteristic laws that govern absorption spectra. As Kirchhoff and Bunsen had already shown, there is a close relation between the selective absorption of substances for radiation and their emission spectra, and it is on this that the application of spectrum analysis to the heavenly bodies essentially rests. Yet on the basis of the classical electromagnetic theory, it is impossible to understand why substances in the form of vapour show absorption for certain lines in their emission spectrum and not for others.

On the basis of the postulates given above we are, however, led to assume that the absorption of radiation corresponding to a spectral line emitted by a transition from one stationary state of the atom to a state of less energy is brought about by the return of the atom from the last-named state to the first. We thus understand immediately that in ordinary circumstances a gas or vapour can only show selective absorption for spectral lines that are produced by a transition from a state corresponding to an earlier stage in the binding process to the normal state. Only at higher temperatures or under the influence of electric discharges whereby an appreciable number of atoms are being constantly disrupted from the normal state, can we expect absorption for other lines in the emission spectrum in agreement with the experiments.

A most direct confirmation for the general interpretation of spectra on the basis of the postulates has also been obtained by investigations on the excitation of spectral lines and ionization of atoms by means of impact of free electrons with given velocities. A decided advance in this direction was marked by the well-known investigations of Franck and Hertz (1914). It appeared from their results that by means of electron impacts it was impossible to impart to an atom an arbitrary amount of energy, but only such amounts as corresponded to a transfer of the atom from its normal state to another stationary state of the existence of which the spectra assure us, and the energy of which can be inferred from the magnitude of the spectral term.

Further, striking evidence was afforded of the independence that, accord-

ing to the postulates, must be attributed to the processes which give rise to the emission of the different spectral lines of an element. Thus it could be shown directly that atoms that were transferred in this manner to a stationary state of greater energy were able to return to the normal state with emission of radiation corresponding to a single spectral line.

Continued investigations on electron impacts, in which a large number of physicists have shared, have also produced a detailed confirmation of the theory concerning the excitation of series spectra. Especially it has been possible to show that for the *ionization* of an atom by electron impact an amount of energy is necessary that is exactly equal to the work required, according to the theory, to remove the last electron captured from the atom. This work can be determined directly as the product of Planck's constant and the spectral term corresponding to the normal state, which, as mentioned above, is equal to the limiting value of the frequencies of the spectral series connected with selective absorption.

The quantum theory of multiply-periodic systems

While it was thus possible by means of the fundamental postulates of the quantum theory to account directly for certain general features of the properties of the elements, a closer development of the ideas of the quantum theory was necessary in order to account for these properties in further detail. In the course of the last few years a more general theoretical basis has been attained through the development of formal methods that permit the fixation of the stationary states for electron motions of a more general type than those we have hitherto considered. For a simply periodic motion such as we meet in the pure harmonic oscillator, and at least to a first approximation, in the motion of an electron about a positive nucleus, the manifold of stationary states can be simply co-ordinated to a series of whole numbers. For motions of the more general class mentioned above, the so-called *multiply-periodic* motions, however, the stationary states compose a more complex manifold, in which, according to these formal methods, each state is characterized by several whole numbers, the so-called "quantum numbers".

In the development of the theory a large number of physicists have taken part, and the introduction of several quantum numbers can be traced back to the work of Planck himself. But the definite step which gave the impetus to further work was made by Sommerfeld (1915) in his explanation of the

fine structure shown by the hydrogen lines when the spectrum is observed with a spectroscope of high resolving power. The occurrence of this fine structure must be ascribed to the circumstance that we have to deal, even in hydrogen, with a motion which is not exactly simply periodic. In fact, as a consequence of the change in the electron's mass with velocity that is claimed by the theory of relativity, the electron orbit will undergo a very slow precession in the orbital plane. The motion will therefore be doubly periodic, and besides a number characterizing the term in the Balmer formula, which we shall call the *principal quantum number* because it determines in the main the energy of the atom, the fixation of the stationary states demands another quantum number which we shall call the *subordinate quantum number*.

A survey of the motion in the stationary states thus fixed is given in the diagram (Fig. 5), which reproduces the relative size and form of the electron orbits. Each orbit is designated by a symbol n_k , where n is the principal quantum number and k the subordinate quantum number. All orbits with the same principal quantum number have, to a first approximation, the same major axis, while orbits with the same value of k have the same parameter, i.e. the same value for the shortest chord through the focus. Since the energy values for different states with the same value of n but different values of k differ a little from each other, we get for each hydrogen line corresponding to definite values of n' and n'' in the Balmer formula a number of different transition processes, for which the frequencies of the emitted radia-

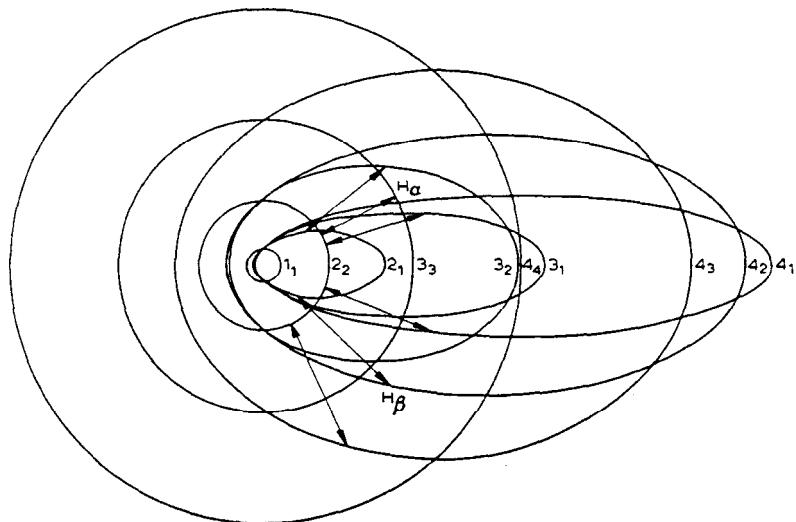


Fig. 5.

tion as calculated by the second postulate are not exactly the same. As Sommerfeld was able to show, the components this gives for each hydrogen line agree with the observations on the fine structure of hydrogen lines to within the limits of experimental error. In the figure the arrows designate the processes that give rise to the components of the red and green lines in the hydrogen spectrum, the frequencies of which are obtained by putting $n'' = 2$ and $n' = 3$ or 4 respectively in the Balmer formula.

In considering the figure it must not be forgotten that the description of the orbit is there incomplete, in so much as with the scale used the slow precession does not show at all. In fact, this precession is so slow that even for the orbits that rotate most rapidly the electron performs about 40,000 revolutions before the perihelion has gone round once. Nevertheless, it is this precession alone that is responsible for the multiplicity of the stationary states characterized by the subordinate quantum number. If, for example, the hydrogen atom is subjected to a small disturbing force which perturbs the regular precession, the electron orbit in the stationary states will have a form altogether different from that given in the figure. This implies that the fine structure will change its character completely, but the hydrogen spectrum will continue to consist of lines that are given to a close approximation by the Balmer formula, due to the fact that the approximately periodic character of the motion will be retained. Only when the disturbing forces become so large that even during a single revolution of the electron the orbit is appreciably disturbed, will the spectrum undergo essential changes. The statement often advanced that the introduction of two quantum numbers should be a necessary condition for the explanation of the Balmer formula must therefore be considered as a misconception of the theory.

Sommerfeld's theory has proved itself able to account not only for the fine structure of the hydrogen lines, but also for that of the lines in the helium spark spectrum. Owing to the greater velocity of the electron, the intervals between the components into which a line is split up are here much greater and can be measured with much greater accuracy. The theory was also able to account for certain features in the fine structure of X-ray spectra, where we meet frequency differences that may even reach a value more than a million times as great as those of the frequency differences for the components of the hydrogen lines.

Shortly after this result had been attained, Schwarzschild and Epstein (1916) simultaneously succeeded, by means of similar considerations, in accounting for the characteristic changes that the hydrogen lines undergo in

an electric field, which had been discovered by Stark in the year 1914. Next, an explanation of the essential features of the Zeeman effect for the hydrogen lines was worked out at the same time by Sommerfeld and Debye (1917). In this instance the application of the postulates involved the consequence that only certain orientations of the atom relative to the magnetic field were allowable, and this characteristic consequence of the quantum theory has quite recently received a most direct confirmation in the beautiful researches of Stern and Gerlach on the deflexion of swiftly moving silver atoms in a nonhomogenous magnetic field.

The correspondence principle

While this development of the theory of spectra was based on the working out of formal methods for the fixation of stationary states, the present lecturer succeeded shortly afterwards in throwing light on the theory from a new viewpoint, by pursuing further the characteristic connexion between the quantum theory and classical electrodynamics already traced out in the hydrogen spectrum. In connexion with the important work of Ehrenfest and Einstein these efforts led to the formulation of the so-called *correspondence principle*, according to which the occurrence of transitions between the stationary states accompanied by emission of radiation is traced back to the harmonic components into which the motion of the atom may be resolved and which, according to the classical theory, determine the properties of the radiation to which the motion of the particles gives rise.

According to the correspondence principle, it is assumed that every transition process between two stationary states can be co-ordinated with a corresponding harmonic vibration component in such a way that the probability of the occurrence of the transition is dependent on the amplitude of the vibration. The state of polarization of the radiation emitted during the transition depends on the further characteristics of the vibration, in a manner analogous to that in which on the classical theory the intensity and state of polarization in the wave system emitted by the atom as a consequence of the presence of this vibration component would be determined respectively by the amplitude and further characteristics of the vibration.

With the aid of the correspondence principle it has been possible to confirm and to extend the above-mentioned results. Thus it was possible to develop a complete quantum theory explanation of the Zeeman effect for the

hydrogen lines, which, in spite of the essentially different character of the assumptions that underlie the two theories, is very similar throughout to Lorentz's original explanation based on the classical theory. In the case of the Stark effect, where, on the other hand, the classical theory was completely at a loss, the quantum theory explanation could be so extended with the help of the correspondence principle as to account for the polarization of the different components into which the lines are split, and also for the characteristic intensity distribution exhibited by the components. This last question has been more closely investigated by Kramers, and the accompanying figure will give some impression of how completely it is possible to account for the phenomenon under consideration.

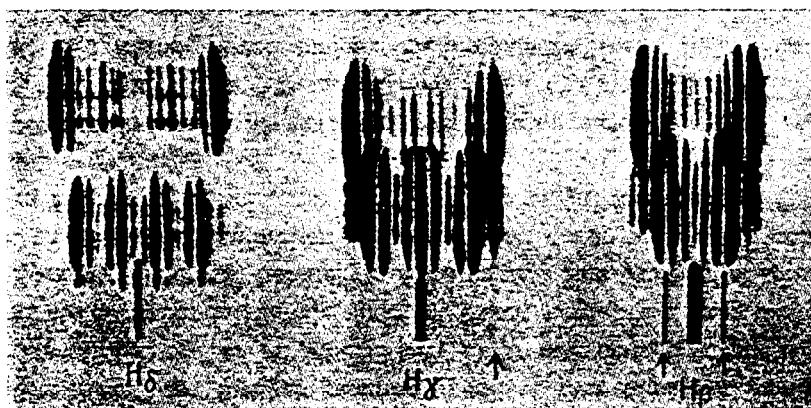


Fig. 6.

Fig. 6 reproduces one of Stark's well-known photographs of the splitting up of the hydrogen lines. The picture displays very well the varied nature of the phenomenon, and shows in how peculiar a fashion the intensity varies from component to component. The components below are polarized perpendicular to the field, while those above are polarized parallel to the field.

Fig. 7 gives a diagrammatic representation of the experimental and theoretical results for the line $H\gamma$, the frequency of which is given by the Balmer formula with $n'' = 2$ and $n' = 5$. The vertical lines denote the components into which the line is split up, of which the picture on the right gives the components which are polarized parallel to the field and that on the left those that are polarized perpendicular to it. The experimental results are represented in the upper half of the diagram, the distances from the dotted line representing the measured displacements of the components, and the lengths

of the lines being proportional to the relative intensity as estimated by Stark from the blackening of the photographic plate. In the lower half is given for comparison a representation of the theoretical results from a drawing in Kramers' paper.

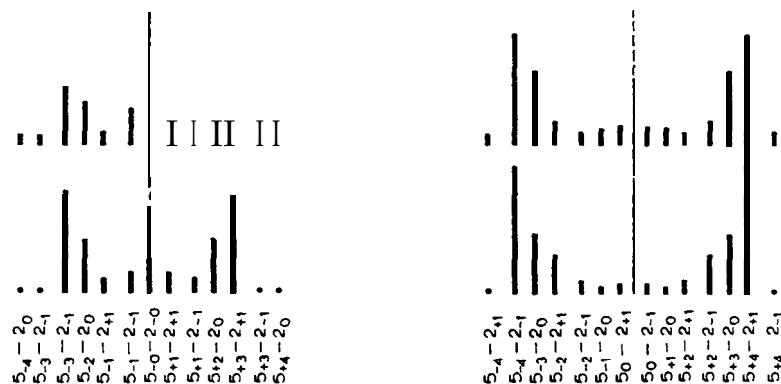


Fig. 7.

The symbol $(n'_s - n''_s)$ attached to the lines gives the transitions between the stationary states of the atom in the electric field by which the components are emitted. Besides the principal quantum integer n , the stationary states are further characterized by a subordinate quantum integer s , which can be negative as well as positive and has a meaning quite different from that of the quantum number k occurring in the relativity theory of the fine structure of the hydrogen lines, which fixed the form of the electron orbit in the undisturbed atom. Under the influence of the electric field both the form of the orbit and its position undergo large changes, but certain properties of the orbit remain unchanged, and the subordinate quantum number s is connected with these. In Fig. 7 the position of the components corresponds to the frequencies calculated for the different transitions, and the lengths of the lines are proportional to the probabilities as calculated on the basis of the correspondence principle, by which also the polarization of the radiation is determined. It is seen that the theory reproduces completely the main feature of the experimental results, and in the light of the correspondence principle we can say that the Stark effect reflects down to the smallest details the action of the electric field on the orbit of the electron in the hydrogen atom, even though in this case the reflection is so distorted that, in contrast with the case of the Zeeman effect, it would scarcely be possible directly to

recognize the motion on the basis of the classical ideas of the origin of electromagnetic radiation.

Results of interest were also obtained for the spectra of elements of higher atomic number, the explanation of which in the meantime had made important progress through the work of Sommerfeld, who introduced several quantum numbers for the description of the electron orbits. Indeed, it was possible, with the aid of the correspondence principle, to account completely for the characteristic rules which govern the seemingly capricious occurrence of combination lines, and it is not too much to say that the quantum theory has not only provided a simple interpretation of the combination principle, but has further contributed materially to the clearing up of the mystery that has long rested over the application of this principle.

The same viewpoints have also proved fruitful in the investigation of the so-called band spectra. These do not originate, as do series spectra, from individual atoms, but from molecules; and the fact that these spectra are so rich in lines is due to the complexity of the motion entailed by the vibrations of the atomic nuclei relative to each other and the rotations of the molecule as a whole. The first to apply the postulates to this problem was Schwarzschild, but the important work of Heurhnger especially has thrown much light on the origin and structure of band spectra. The considerations employed here can be traced back directly to those discussed at the beginning of this lecture in connexion with Bjerrum's theory of the influence of molecular rotation on the infrared absorption lines of gases. It is true we no longer think that the rotation is reflected in the spectra in the way claimed by classical electrodynamics, but rather that the line components are due to transitions between stationary states which differ as regards rotational motion. That the phenomenon retains its essential feature, however, is a typical consequence of the correspondence principle.

The natural system of the elements

The ideas of the origin of spectra outlined in the preceding have furnished the basis for a theory of the structure of the atoms of the elements which has shown itself suitable for a general interpretation of the main features of the properties of the elements, as exhibited in the natural system. This theory is based primarily on considerations of the manner in which the atom can be imagined to be built up by the capture and binding of electrons to the nu-

cleus, one by one. As we have seen, the optical spectra of elements provide us with evidence on the progress of the last steps in this building-up process.

An insight into the kind of information that the closer investigation of the spectra has provided in this respect may be obtained from Fig. 8, which gives a diagrammatic representation of the orbital motion in the stationary states corresponding to the emission of the arc-spectrum of potassium. The curves show the form of the orbits described in the stationary states by the last electron captured in the potassium atom, and they can be considered as stages in the process whereby the 19th electron is bound after the 18 previous electrons have already been bound in their normal orbits. In order not to complicate the figure, no attempt has been made to draw any of the orbits of these inner electrons, but the region in which they move is enclosed by a dotted circle. In an atom with several electrons the orbits will, in general, have a complicated character. Because of the symmetrical nature of the field of force about the nucleus, however, the motion of each single electron can be approximately described as a plane periodic motion on which is superimposed a uniform rotation in the plane of the orbit. The orbit of each electron will therefore be to a first approximation doubly periodic, and will be fixed by two quantum numbers, as are the stationary states in a hydrogen atom when the relativity precession is taken into account.

In Fig. 8, as in Fig. 5, the electron orbits are marked with the symbol n_k

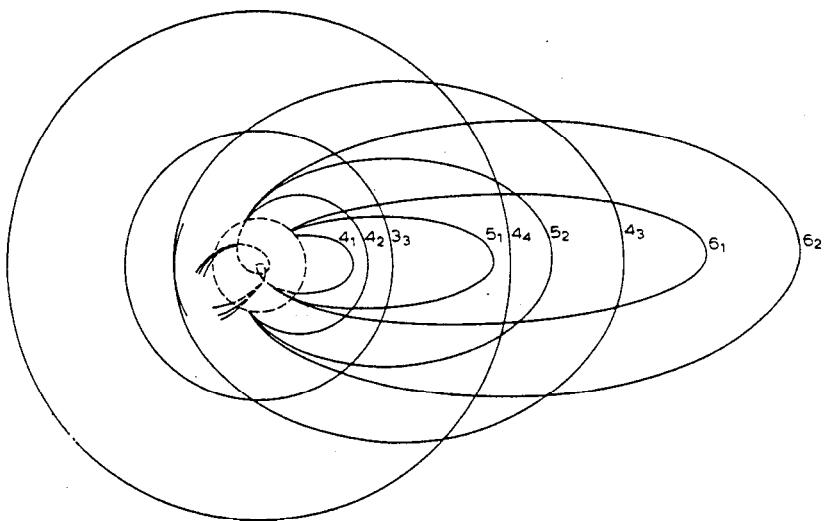


Fig. 8.

where n is the principal quantum number and k the subordinate quantum number. While for the initial states of the binding process, where the quantum numbers are large, the orbit of the last electron captured lies completely outside of those of the previously bound electrons, this is not the case for the last stages. Thus, in the potassium atom, the electron orbits with subordinate quantum numbers 2 and 1 will, as indicated in the figure, penetrate partly into the inner region. Because of this circumstance, the orbits will deviate very greatly from a simple Kepler motion, since they will consist of a series of successive outer loops that have the same size and form, but each of which is turned through an appreciable angle relative to the preceding one. Of these outer loops only one is shown in the figure. Each of them coincides very nearly with a piece of a Kepler ellipse, and they are connected, as indicated, by a series of inner loops of a complicated character in which the electron approaches the nucleus closely. This holds especially for the orbit with subordinate quantum number 1, which, as a closer investigation shows, will approach nearer to the nucleus than any of the previously bound electrons.

On account of this penetration into the inner region, the strength with which an electron in such an orbit is bound to the atom will - in spite of the fact that for the most part it moves in a field of force of the same character as that surrounding the hydrogen nucleus - be much greater than for an electron in a hydrogen atom that moves in an orbit with the same principal quantum number, the maximum distance of the electron from the nucleus at the same time being considerably less than in such a hydrogen orbit. As we shall see, this feature of the binding process in atoms with many electrons is of essential importance in order to understand the characteristic periodic way in which many properties of the elements as displayed in the natural system vary with the atomic number.

In the accompanying table (Fig. 9) is given a summary of the results concerning the structure of the atoms of the elements to which the author has been led by a consideration of successive capture and binding of electrons to the atomic nucleus. The figures before the different elements are the atomic numbers, which give the total number of electrons in the neutral atom. The figures in the different columns give the number of electrons in orbits corresponding to the values of the principal and subordinate quantum numbers standing at the top. In accordance with ordinary usage we will, for the sake of brevity, designate an orbit with principal quantum number n as an n -quantum orbit. The first electron bound in each atom moves in an orbit

	1 ₁	2 ₁ 2 ₂	3 ₁ 3 ₂ 3 ₃	4 ₁ 4 ₂ 4 ₃ 4 ₄	5 ₁ 5 ₂ 5 ₃ 5 ₄ 5 ₅	6 ₁ 6 ₂ 6 ₃ 6 ₄ 6 ₅ 6 ₆	7 ₁ 7 ₂
1 H	1						
2 He	2						
3 Li	2	1					
4 Be	2	2					
5 B	2	2 (1)					
--	--	--	--				
10 Ne	2	4 4					
11 Na	2	4 4	1				
12 Mg	2	4 4	2				
13 Al	2	4 4	2 1				
--	--	--	--				
18 Ar	2	4 4	4 4				
19 K	2	4 4	4 4	1			
20 Ca	2	4 4	4 4	2			
21 Sc	2	4 4	4 4 1	(2)			
22 Ti	2	4 4	4 4 2	(2)			
--	--	--	--	--			
29 Cu	2	4 4	6 6 6	1			
30 Zn	2	4 4	6 6 6	2			
31 Ga	2	4 4	6 6 6	2 1			
--	--	--	--	--			
36 Kr	2	4 4	6 6 6	4 4			
37 Rb	2	4 4	6 6 6	4 4	1		
38 Sr	2	4 4	6 6 6	4 4	2		
39 Y	2	4 4	6 6 6	4 4 1	(2)		
40 Zr	2	4 4	6 6 6	4 4 2	(2)		
--	--	--	--	--	--		
47 Ag	2	4 4	6 6 6	6 6 6	1		
48 Cd	2	4 4	6 6 6	6 6 6	2		
49 In	2	4 4	6 6 6	6 6 6	2 1		
--	--	--	--	--	--		
54 X	2	4 4	6 6 6	6 6 6	4 4		
55 Cs	2	4 4	6 6 6	6 6 6	4 4	I	
56 Ba	2	4 4	6 6 6	6 6 6	4 4	2	
57 La	2	4 4	6 6 6	6 6 6	4 4 1	(2)	
58 Ce	2	4 4	6 6 6	6 6 6 1	4 4 1	(2)	
59 Pr	2	4 4	6 6 6	6 6 6 2	4 4 1	(2)	
--	--	--	--	--	--	--	
71 Cp	2	4 4	6 6 6	8 8 8 8	4 4 1	(2)	
72 -	2	4 4	6 6 6	8 8 8 8	4 4 2	(2)	
--	--	--	--	--	--	--	
79 Au	2	4 4	6 6 6	8 8 8 8	6 6 6		
80 Hg	2	4 4	6 6 6	8 8 8 8	6 6 6	2	
81 Tl	2	4 4	6 6 6	8 8 8 8	6 6 6	2 1	
--	--	--	--	--	--	--	
86 Em	2	4 4	6 6 6	8 8 8 8	6 6 6	4 4	
87 -	2	4 4	6 6 6	8 8 8 8	6 6 6	4 4	I
88 Ra	2	4 4	6 6 6	8 8 8 8	6 6 6	4 4	2
89 Ac	2	4 4	6 6 6	8 8 8 8	6 6 6	4 4 1	(2)
90 Th	2	4 4	6 6 6	8 8 8 8	6 6 6	4 4 2	(2)
--	--	--	--	--	--	--	
118 ?	2	4 4	6 6 6	8 8 8 8	8 8 8 8	6 6 6	4 4

Fig. 9.

that corresponds to the normal state of the hydrogen atom with quantum symbol 1.₁ In the hydrogen atom there is of course only one electron; but we must assume that in the atoms of other elements the next electron also will be bound in such a r-quantum orbit of type 1₁. As the table shows, the following electrons are bound in 2-quantum orbits. To begin with, the binding will result in a 2₁ orbit, but later electrons will be bound in 2₂ orbits, until, after binding the first 10 electrons in the atom, we reach a closed configuration of the a-quantum orbits in which we assume there are four orbits of each type. This configuration is met for the first time in the neutral neon atom, which forms the conclusion of the second period in the system of the elements. When we proceed in this system, the following electrons are bound in 3-quantum orbits, until, after the conclusion of the third period of the system, we encounter for the first time, in elements of the fourth period, electrons in 4-quantum orbits, and so on.

This picture of atomic structure contains many features that were brought forward by the work of earlier investigators. Thus the attempt to interpret the relations between the elements in the natural system by the assumption of a division of the electrons into groups goes as far back as the work of J. J. Thomson in 1904. Later, this viewpoint was developed chiefly by Kossel (1916), who, moreover, has connected such a grouping with the laws that investigations of X-ray spectra have brought to light.

Also G. R. Lewis and I. Langmuir have sought to account for the relations between the properties of the elements on the basis of a grouping inside the atom. These investigators, however, assumed that the electrons do not move about the nucleus, but occupy positions of equilibrium. In this way, though, no closer relation can be reached between the properties of the elements and the experimental results concerning the constituents of the atoms. Statical positions of equilibrium for the electrons are in fact not possible in cases in which the forces between the electrons and the nucleus even approximately obey the laws that hold for the attractions and repulsions between electrical charges.

The possibility of an interpretation of the properties of the elements on the basis of these latter laws is quite characteristic for the picture of atomic structure developed by means of the quantum theory. As regards this picture, the idea of connecting the grouping with a classification of electron orbits according to increasing quantum numbers was suggested by Moseley's discovery of the laws of X-ray spectra, and by Sommerfeld's work on the fine structure of these spectra. This has been principally emphasized by Vegard,

who some years ago in connexion with investigations of X-ray spectra proposed a grouping of electrons in the atoms of the elements, which in many ways shows a likeness to that which is given in the above table.

A satisfactory basis for the further development of this picture of atomic structure has, however, only recently been created by the study of the binding processes of the electrons in the atom, of which we have experimental evidence in optical spectra, and the characteristic features of which have been elucidated principally by the correspondence principle. It is here an essential circumstance that the restriction on the course of the binding process, which is expressed by the presence of electron orbits with higher quantum numbers in the normal state of the atom, can be naturally connected with the general condition for the occurrence of transitions between stationary states, formulated in that principle.

Another essential feature of the theory is the influence, on the strength of binding and the dimensions of the orbits, of the penetration of the later bound electrons into the region of the earlier bound ones, of which we have seen an example in the discussion of the origin of the potassium spectrum. Indeed, this circumstance may be regarded as the essential cause of the pronounced periodicity in the properties of the elements, in that it implies that the atomic dimensions and chemical properties of homologous substances in the different periods, as, for example, the alkali-metals, show a much greater similarity than that which might be expected from a direct comparison of the orbit of the last electron bound with an orbit of the same quantum number in the hydrogen atom.

The increase of the principal quantum number which we meet when we proceed in the series of the elements, affords also an immediate explanation of the characteristic deviations from simple periodicity which are exhibited by the natural system and are expressed in Fig. 1 by the bracketing of certain series of elements in the later periods. The first time such a deviation is met with is in the 4th period, and the reason for it can be simply illustrated by means of our figure of the orbits of the last electron bound in the atom of potassium, which is the first element in this period. Indeed, in potassium we encounter for the first time in the sequence of the elements a case in which the principal quantum number of the orbit of the last electron bound is, in the normal state of the atom, larger than in one of the earlier stages of the binding process. The normal state corresponds here to a 4₁orbit, which, because of the penetration into the inner region, corresponds to a much stronger binding of the electron than a 4-quantum orbit in the hydrogen

atom. The binding in question is indeed even stronger than for a 2-quantum orbit in the hydrogen atom, and is therefore more than twice as strong as in the circular 3_3 orbit which is situated completely outside the inner region, and for which the strength of the binding differs but little from that for a 3-quantum orbit in hydrogen.

This will not continue to be true, however, when we consider the binding of the 19th electron in substances of higher atomic number, because of the much smaller relative difference between the field of force outside and inside the region of the first eighteen electrons bound. As is shown by the investigation of the spark spectrum of calcium, the binding of the 19th electron in the 4_1 orbit is here but little stronger than in 3_3 orbits, and as soon as we reach scandium, we must assume that the 3_3 orbit will represent the orbit of the 19th electron in the normal state, since this type of orbit will correspond to a stronger binding than a 4_1 orbit. While the group of electrons in 2-quantum orbits has been entirely completed at the end of the 2nd period, the development that the group of 3-quantum orbits undergoes in the course of the 3rd period can therefore only be described as a provisional completion, and, as shown in the table, this electron group will, in the bracketed elements of the 4th period, undergo a stage of further development in which electrons are added to it in 3-quantum orbits.

This development brings in new features, in that the development of the electron group with 4-quantum orbits comes to a standstill, so to speak, until the 3-quantum group has reached its final closed form. Although we are not yet in a position to account in all details for the steps in the gradual development of the 3-quantum electron group, still we can say that with the help of the quantum theory we see at once why it is in the 4th period of the system of the elements that there occur for the first time successive elements with properties that resemble each other as much as the properties of the iron *group*; indeed, we can even understand why these elements show their well-known paramagnetic properties. Without further reference to the quantum theory, Eadenburg had on a previous occasion already suggested the idea of relating the chemical and magnetic properties of these elements with the development of an inner electron group in the atom.

I will not enter into many more details, but only mention that the peculiarities we meet with in the 5th period are explained in much the same way as those in the 4th period. Thus the properties of the bracketed elements in the 5th period as it appears in the table, depend on a stage in the development of the 4-quantum electron group that is initiated by the entrance in the

normal state of electrons in 4_3 orbits. In the 6th period, however, we meet new features. In this period we encounter not only a stage of the development of the electron groups with 5- and 6-quantum orbits, but also the final completion of the development of the 4-quantum electron group, which is initiated by the entrance for the first time of electron orbits of the 4_4 type in the normal state of the atom. This development finds its characteristic expression in the occurrence of the peculiar family of elements in the 6th period, known as the *rare-earths*. These show, as we know, a still greater mutual similarity in their chemical properties than the elements of the iron family. This must be ascribed to the fact that we have here to do with the development of an electron group that lies deeper in the atom. It is of interest to note that the theory can also naturally account for the fact that these elements, which resemble each other in so many ways, still show great differences in their magnetic properties.

The idea that the occurrence of the rare-earths depends on the development of an inner electron group has been put forward from different sides. Thus it is found in the work of Vegard, and at the same time as my own work, it was proposed by Bury in connexion with considerations of the systematic relation between the chemical properties and the grouping of the electrons inside the atom from the point of view of Langmuir's static atomic model. While until now it has not been possible, however, to give any theoretical basis for such a development of an inner group, we see that our extension of the quantum theory provides us with an unforced explanation. Indeed, it is scarcely an exaggeration to say that if the existence of the rare-earths had not been established by direct chemical investigation, the occurrence of a family of elements of this character within the 6th period of the natural system of the elements might have been theoretically predicted.

When we proceed to the 7th period of the system, we meet for the first time with 7-quantum orbits, and we shall expect to find within this period features that are essentially similar to those in the 6th period, in that besides the first stage in the development of the 7-quantum orbits, we must expect to encounter further stages in the development of the group with 6- or 5-quantum orbits. However, it has not been possible directly to confirm this expectation, because only a few elements are known in the beginning of the 7th period. The latter circumstance may be supposed to be intimately connected with the instability of atomic nuclei with large charges, which is expressed in the prevalent radioactivity among elements with high atomic number.

X-ray spectra and atomic constitution

In the discussion of the conceptions of atomic structure we have hitherto placed the emphasis on the formation of the atom by successive capture of electrons. Our picture would, however, be incomplete without some reference to the confirmation of the theory afforded by the study of X-ray spectra. Since the interruption of Moseley's fundamental researches by his untimely death, the study of these spectra has been continued in a most admirable way by Prof. Siegbahn in Lund. On the basis of the large amount of experimental evidence adduced by him and his collaborators, it has been possible recently to give a classification of X-ray spectra that allows an immediate interpretation on the quantum theory. In the first place it has been possible, just as in the case of the optical spectra, to represent the frequency of each of the X-ray lines as the difference between two out of a manifold of spectral terms characteristic of the element in question. Next, a direct connexion with the atomic theory is obtained by the assumption that each of these spectral terms multiplied by Planck's constant is equal to the work which must be done on the atom to remove one of its inner electrons. In fact, the removal of one of the inner electrons from the completed atom may, in accordance with the above considerations on the formation of atoms by capture of electrons, give rise to transition processes by which the place of the electron removed is taken by an electron belonging to one of the more loosely bound electron groups of the atom, with the result that after the transition an electron will be lacking in this latter group.

The X-ray lines may thus be considered as giving evidence of stages in a process by which the atom undergoes a *reorganization* after a disturbance in its interior. According to our views on the stability of the electronic configuration such a disturbance must consist in the removal of electrons from the atom, or at any rate in their transference from normal orbits to orbits of higher quantum numbers than those belonging to completed groups; a circumstance which is clearly illustrated in the characteristic difference between selective absorption in the X-ray region, and that exhibited in the optical region.

The classification of the X-ray spectra, to the achievement of which the above-mentioned work of Sommerfeld and Kossel has contributed materially, has recently made it possible, by means of a closer examination of the manner in which the terms occurring in the X-ray spectra vary with the atomic number, to obtain a very direct test of a number of the theoretical

conclusions as regards the structure of the atom. In Fig. 9 the abscissæ are the atomic numbers and the ordinates are proportional to the square roots of the spectral terms, while the symbols K, L, M, N, O, for the individual

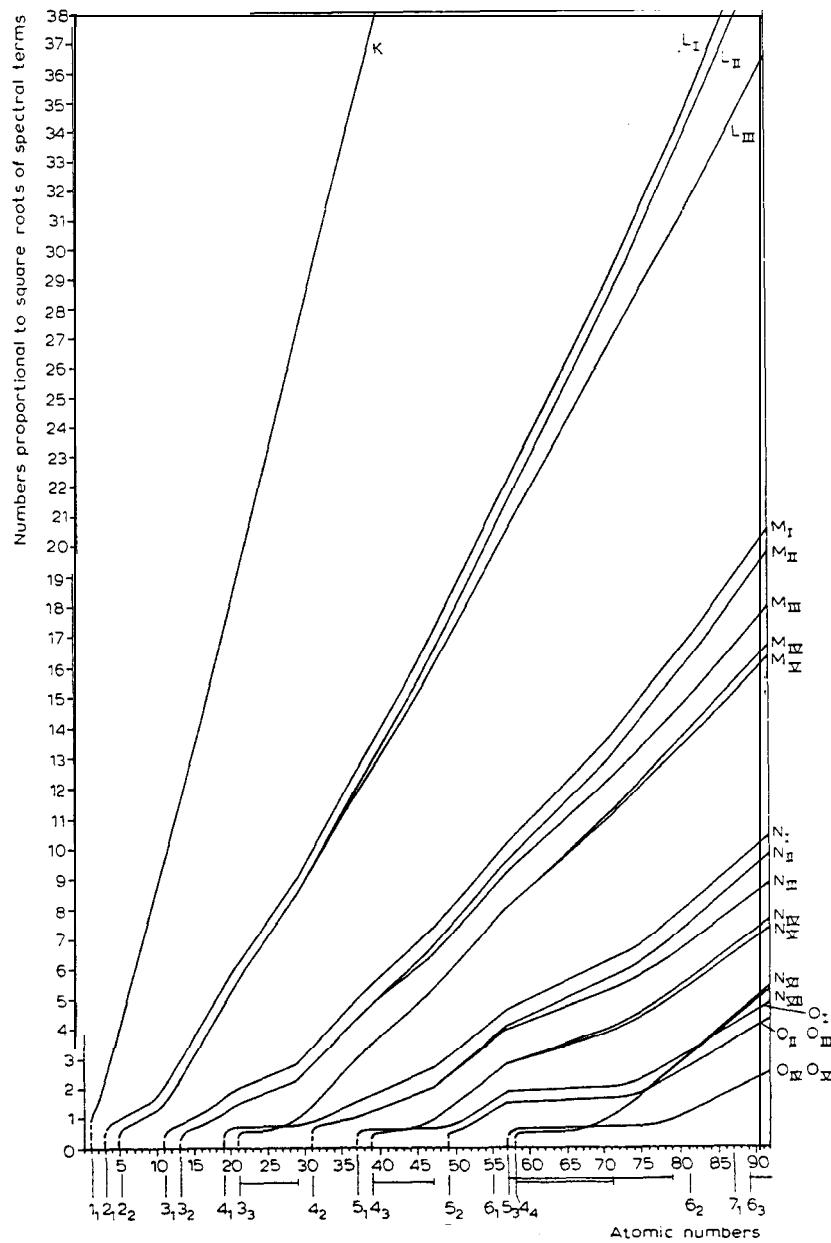


Fig. 10.

terms refer to the characteristic discontinuities in the selective absorption of the elements for X-rays; these were originally found by Barkla before the discovery of the interference of X-rays in crystals had provided a means for the closer investigation of X-ray spectra. Although the curves generally run very uniformly, they exhibit a number of deviations from uniformity which have been especially brought to light by the recent investigation of Coster, who has for some years worked in Siegbahn's laboratory.

These deviations, the existence of which was not discovered until after the publication of the theory of atomic structure discussed above, correspond exactly to what one might expect from this theory. At the foot of the figure the vertical lines indicate where, according to the theory, we should first expect, in the normal state of the atom, the occurrence of n_k orbits of the type designated. We see how it has been possible to connect the occurrence of every spectral term with the presence of an electron moving in an orbit of a definite type, to the removal of which this term is supposed to correspond. That in general there corresponds more than one curve to each type of orbit n_k is due to a complication in the spectra which would lead us too far afield to enter into here, and may be attributed to the deviation from the previously described simple type of motion of the electron arising from the interaction of the different electrons within the same group.

The intervals in the system of the elements, in which a further development of an inner electron group takes place because of the entrance into the normal atom of electron orbits of a certain type, are designated in the figure by the horizontal lines, which are drawn between the vertical lines to which the quantum symbols are affixed. It is clear that such a development of an inner group is everywhere reflected in the curves. Particularly the course of the N- and O-curves may be regarded as a direct indication of that stage in the development of the electron groups with 4-quantum orbits of which the occurrence of the rareearths bears witness. Although the apparent complete absence of a reflection in the X-ray spectra of the complicated relationships exhibited by most other properties of the elements was the typical and important feature of Moseley's discovery, we can recognize, nevertheless, in the light of the progress of the last years, an intimate connexion between the X-ray spectra and the general relationships between the elements within the natural system.

Before concluding this lecture I should like to mention one further point in which X-ray investigations have been of importance for the test of the theory. This concerns the properties of the hitherto unknown element with

atomic number 72. On this question opinion has been divided in respect to the conclusions that could be drawn from the relationships within the Periodic Table, and in many representations of the table a place is left open for this element in the rare-earth family. In Julius Thomsen's representation of the natural system, however, this hypothetical element was given a position homologous to titanium and zirconium in much the same way as in our representation in Fig. 1. Such a relationship must be considered as a necessary consequence of the theory of atomic structure developed above, and is expressed in the table (Fig. 9) by the fact that the electron configurations for titanium and zirconium show the same sort of resemblances and differences as the electron configurations for zirconium and the element with atomic number 72. A corresponding view was proposed by Bury on the basis of his above-mentioned systematic considerations of the connexion between the grouping of the electrons in the atom and the properties of the elements.

Recently, however, a communication was published by Dauvillier announcing the observation of some weak lines in the X-ray spectrum of a preparation containing rareearths. These were ascribed to an element with atomic number 72 assumed to be identical with an element of the rare-earth family, the existence of which in the preparation used had been presumed by Urbain many years ago. This conclusion would, however, if it could be maintained, place extraordinarily great, if not unsurmountable, difficulties in the way of the theory, since it would claim a change in the strength of the binding of the electrons with the atomic number which seems incompatible with the conditions of the quantum theory. In these circumstances Dr. Coster and Prof. Hevesy, who are both for the time working in Copenhagen, took up a short time ago the problem of testing a preparation of zircon-bearing minerals by X-ray spectroscopic analysis. These investigators have been able to establish the existence in the minerals investigated of appreciable quantities of an element with atomic number 72, the chemical properties of which show a great similarity to those of zirconium and a decided difference from those of the rareearths.*

I hope that I have succeeded in giving a summary of some of the most important results that have been attained in recent years in the field of atomic theory, and I should like, in concluding, to add a few general remarks concerning the viewpoint from which these results may be judged, and par-

* For the result of the continued work of Coster and Hevesy with the new element, for which they have proposed the name hafnium, the reader may be referred to their letters in *Nature* of January 20, February 10 and 24, and April 7.

ticularly concerning the question of how far, with these results, it is possible to speak of an explanation, in the ordinary sense of the word. By a theoretical explanation of natural phenomena we understand in general a classification of the observations of a certain domain with the help of analogies pertaining to other domains of observation, where one presumably has to do with simpler phenomena. The most that one can demand of a theory is that this classification can be pushed so far that it can contribute to the development of the field of observation by the prediction of new phenomena.

When we consider the atomic theory, we are, however, in the peculiar position that there can be no question of an explanation in this last sense, since here we have to do with phenomena which from the very nature of the case are simpler than in any other field of observation, where the phenomena are always conditioned by the combined action of a large number of atoms. We are therefore obliged to be modest in our demands and content ourselves with concepts which are formal in the sense that they do not provide a visual picture of the sort one is accustomed to require of the explanations with which natural philosophy deals. Bearing this in mind I have sought to convey the impression that the results, on the other hand, fulfill, at least in some degree, the expectations that are entertained of any theory; in fact, I have attempted to show how the development of atomic theory has contributed to the classification of extensive fields of observation, and by its predictions has pointed out the way to the completion of this classification. It is scarcely necessary, however, to emphasize that the theory is yet in a very preliminary stage, and many fundamental questions still await solution.

The statistical interpretation of quantum mechanics

Nobel Lecture, December 11, 1954

The work, for which I have had the honour to be awarded the Nobel Prize for 1954, contains no discovery of a fresh natural phenomenon, but rather the basis for a new mode of thought in regard to natural phenomena. This way of thinking has permeated both experimental and theoretical physics to such a degree that it hardly seems possible to say anything more about it that has not been already so often said. However, there are some particular aspects which I should like to discuss on what is, for me, such a festive occasion. The first point is this: the work at the Göttingen school, which I directed at that time (1926-1927), contributed to the solution of an intellectual crisis into which our science had fallen as a result of Planck's discovery of the quantum of action in 1900. Today, physics finds itself in a similar crisis - I do not mean here its entanglement in politics and economics as a result of the mastery of a new and frightful force of Nature, but I am considering more the logical and epistemological problems posed by nuclear physics. Perhaps it is well at such a time to recall what took place earlier in a similar situation, especially as these events are not without a definite dramatic flavour.

The second point I wish to make is that when I say that the physicists had accepted the concepts and mode of thought developed by us at the time, I am not quite correct. There are some very noteworthy exceptions, particularly among the very workers who have contributed most to building up the quantum theory. Planck, himself, belonged to the sceptics until he died. Einstein, De Broglie, and Schrödinger have unceasingly stressed the unsatisfactory features of quantum mechanics and called for a return to the concepts of classical, Newtonian physics while proposing ways in which this could be done without contradicting experimental facts. Such weighty views cannot be ignored. Niels Bohr has gone to a great deal of trouble to refute the objections. I, too, have ruminated upon them and believe I can make some contribution to the clarification of the position. The matter concerns the borderland between physics and philosophy, and so my physics lecture

will partake of both history and philosophy, for which I must crave your indulgence.

First of all, I will explain how quantum mechanics and its statistical interpretation arose. At the beginning of the twenties, every physicist, I think, was convinced that Planck's quantum hypothesis was correct. According to this theory *energy* appears in finite quanta of magnitude $h\nu$ in oscillatory processes having a specific frequency ν (e.g. in light waves). Countless experiments could be explained in this way and always gave the same value of Planck's constant h . Again, Einstein's assertion that light quanta have *momentum* $h\nu/c$ (where c is the speed of light) was well supported by experiment (e.g. through the Compton effect). This implied a revival of the corpuscular theory of light for a certain complex of phenomena. The wave theory still held good for other processes. Physicists grew accustomed to this *duality* and learned how to cope with it to a certain extent.

In 1913 Niels Bohr had solved the riddle of *line spectra* by means of the quantum theory and had thereby explained broadly the amazing stability of the atoms, the structure of their electronic shells, and the Periodic System of the elements. For what was to come later, the most important assumption of his teaching was this: an atomic system cannot exist in all mechanically possible states, forming a continuum, but in a series of discrete « stationary » states. In a transition from one to another, the difference in energy $E_m - E_n$ is emitted or absorbed as a light quantum $h\nu_{mn}$ (according to whether E_m is greater or less than E_n). This is an interpretation in terms of energy of the fundamental law of spectroscopy discovered some years before by W. Ritz. The situation can be taken in at a glance by writing the energy levels of the stationary states twice over, horizontally and vertically. This produces a square array

	E_1 ,	E_2 ,	E_3 . . .	
E_1	11	12	13	-
E_2	21	22	23	-
E_3	31	32	33	-
-	-	-	-	-

in which positions on a diagonal correspond to states, and non-diagonal positions correspond to transitions.

It was completely clear to Bohr that the law thus formulated is in conflict with mechanics, and that therefore the use of the energy concept in this

connection is problematical. He based this daring fusion of old and new on his *principle of correspondence*. This consists in the obvious requirement that ordinary classical mechanics must hold to a high degree of approximation in the limiting case where the numbers of the stationary states, the so-called quantum numbers, are very large (that is to say, far to the right and to the lower part in the above array) and the energy changes relatively little from place to place, in fact practically continuously.

Theoretical physics maintained itself on this concept for the next ten years. The problem was this: an harmonic oscillation not only has a frequency, but also an intensity. For each transition in the array there must be a corresponding intensity. The question is how to find this through the considerations of correspondence? It meant guessing the unknown from the available information on a known limiting case. Considerable success was attained by Bohr himself, by Kramers, Sommerfeld, Epstein, and many others. But the decisive step was again taken by Einstein who, by a fresh derivation of Planck's radiation formula, made it transparently clear that the classical concept of intensity of radiation must be replaced by the statistical concept of *transition probability*. To each place in our pattern or array there belongs (together with the frequency $v_{mn} = (E_n - E_m)/h$) a definite probability for the transition coupled with emission or absorption.

In Göttingen we also took part in efforts to distil the unknown mechanics of the atom from the experimental results. The logical difficulty became ever sharper. Investigations into the scattering and dispersion of light showed that Einstein's conception of transition probability as a measure of the strength of an oscillation did not meet the case, and the idea of an *amplitude* of oscillation associated with each transition was indispensable. In this connection, work by Ladenburg¹, Kramer², Heisenberg³, Jordan and me⁴ should be mentioned. The art of guessing correct formulae, which deviate from the classical formulae, yet contain them as a limiting case according to the correspondence principle, was brought to a high degree of perfection. A paper of mine, which introduced, for the first time I think, the expression *quantum mechanics* in its title, contains a rather involved formula (still valid today) for the reciprocal disturbance of atomic systems.

Heisenberg, who at that time was my assistant, brought this period to a sudden end⁵. He cut the Gordian knot by means of a philosophical principle and replaced guess-work by a mathematical rule. The principle states that concepts and representations that do not correspond to physically observable facts are not to be used in theoretical description. Einstein used the

same principle when, in setting up his theory of relativity, he eliminated the concepts of absolute velocity of a body and of absolute simultaneity of two events at different places. Heisenberg banished the picture of electron orbits with definite radii and periods of rotation because these quantities are not observable, and insisted that the theory be built up by means of the square arrays mentioned above. Instead of describing the motion by giving a coordinate as a function of time, $x(t)$, an array of transition amplitudes x_{mn} should be determined. To me the decisive part of his work is the demand to determine a rule by which from a given

$$\text{array } \begin{bmatrix} x_{11} & x_{12} & \dots \\ x_{21} & x_{22} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \quad \text{the array for the square } \begin{bmatrix} (x^2)_{11} & (x^2)_{12} & \dots \\ (x^2)_{21} & (x^2)_{22} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

can be found (or, more general, the *multiplication rule* for such arrays).

By observation of known examples solved by guess-work he found this rule and applied it successfully to simple examples such as the harmonic and anharmonic oscillator.

This was in the summer of 1925. Heisenberg, plagued by hay fever took leave for a course of treatment by the sea and gave me his paper for publication if I thought I could do something with it.

The significance of the idea was at once clear to me and I sent the manuscript to the *Zeitschrift für Physik*. I could not take my mind off Heisenberg's multiplication rule, and after a week of intensive thought and trial I suddenly remembered an algebraic theory which I had learned from my teacher, Professor Rosanes, in Breslau. Such square arrays are well known to mathematicians and, in conjunction with a specific rule for multiplication, are called matrices. I applied this rule to Heisenberg's quantum condition and found that this agreed in the diagonal terms. It was easy to guess what the remaining quantities must be, namely, zero; and at once there stood before me the peculiar formula

$$pq - qp = h/2\pi i$$

This meant that coordinates q and momenta p cannot be represented by figure values but by symbols, the product of which depends upon the order of multiplication - they are said to be « non-commuting ».

I was as excited by this result as a sailor would be who, after a long voyage, sees from afar, the longed-for land, and I felt regret that Heisenberg was not

there. I was convinced from the start that we had stumbled on the right path. Even so, a great part was only guess-work, in particular, the disappearance of the non-diagonal elements in the above-mentioned expression. For help in this problem I obtained the assistance and collaboration of my pupil Pascual Jordan, and in a few days we were able to demonstrate that I had guessed correctly. The joint paper by Jordan and myself⁶ contains the most important principles of quantum mechanics including its extension to electrodynamics. There followed a hectic period of collaboration among the three of us, complicated by Heisenberg's absence. There was a lively exchange of letters; my contribution to these, unfortunately, have been lost in the political disorders. The result was a three-author paper⁷ which brought the formal side of the investigation to a definite conclusion. Before this paper appeared, came the *first dramatic surprise*: Paul Dirac's paper on the same subject⁸. The inspiration afforded by a lecture of Heisenberg's in Cambridge had led him to similar results as we had obtained in Göttingen except that he did not resort to the known matrix theory of the mathematicians, but discovered the tool for himself and worked out the theory of such non-commutating symbols.

The first non-trivial and physically important application of quantum mechanics was made shortly afterwards by W. Pauli⁹ who calculated the stationary energy values of the *hydrogen atom* by means of the matrix method and found complete agreement with Bohr's formulae. From this moment onwards there could no longer be any doubt about the correctness of the theory.

What this formalism really signified was, however, by no means clear. Mathematics, as often happens, was cleverer than interpretative thought. While we were still discussing this point there came the *second dramatic surprise*, the appearance of Schrödinger's famous papers¹⁰. He took up quite a different line of thought which had originated from Louis de Broglie¹¹.

A few years previously, the latter had made the bold assertion, supported by brilliant theoretical considerations, that wave-corpuscle duality, familiar to physicists in the case of light, must also be valid for electrons. To each electron moving free of force belongs a plane wave of a definite wavelength which is determined by Planck's constant and the mass. This exciting dissertation by De Broglie was well known to us in Göttingen. One day in 1925 I received a letter from C. J. Davisson giving some peculiar results on the reflection of electrons from metallic surfaces. I, and my colleague on the experimental side, James Franck, at once suspected that these curves of

Davisson's were crystal-lattice spectra of De Broglie's electron waves, and we made one of our pupils, Elsasser¹², to investigate the matter. His result provided the first preliminary confirmation of the idea of De Broglie's, and this was later proved independently by Davisson and Germer¹³ and G. P. Thomson¹⁴ by systematic experiments.

But this acquaintance with De Broglie's way of thinking did not lead us to an attempt to apply it to the electronic structure in atoms. This was left to Schrödinger. He extended De Broglie's wave equation which referred to force-free motion, to the case where the effect of force is taken into account, and gave an exact formulation of the *subsidiary conditions*, already suggested by De Broglie, to which the wave function ψ must be subjected, namely that it should be single-valued and finite in space and time. And he was successful in deriving the stationary states of the hydrogen atom in the form of those monochromatic solutions of his wave equation which do not extend to infinity.

For a brief period at the beginning of 1926, it looked as though there were, suddenly, two self-contained but quite distinct systems of explanation extant: matrix mechanics and wave mechanics. But Schrödinger himself soon demonstrated their complete equivalence.

Wave mechanics enjoyed a very great deal more popularity than the Göttingen or Cambridge version of quantum mechanics. It operates with a wave function ψ , which in the case of *one* particle at least, can be pictured in space, and it uses the mathematical methods of partial differential equations which are in current use by physicists. Schrödinger thought that his wave theory made it possible to return to deterministic classical physics. He proposed (and he has recently emphasized his proposal anew's), to dispense with the particle representation entirely, and instead of speaking of electrons as particles, to consider them as a continuous density distribution $|\psi|^2$ (or electric density $e|\psi|^2$).

To us in Göttingen this interpretation seemed unacceptable in face of well established experimental facts. At that time it was already possible to count particles by means of scintillations or with a Geiger counter, and to photograph their tracks with the aid of a Wilson cloud chamber.

It appeared to me that it was not possible to obtain a clear interpretation of the ψ -function, by considering bound electrons. I had therefore, as early as the end of 1925, made an attempt to extend the matrix method, which obviously only covered oscillatory processes, in such a way as to be applicable to aperiodic processes. I was at that time a guest of the Mas-

sachusetts Institute of Technology in the USA, and I found there in Norbert Wiener an excellent collaborator. In our joint paper¹⁶ we replaced the matrix by the general concept of an operator, and thus made it possible to describe aperiodic processes. Nevertheless we missed the correct approach. This was left to Schrödinger, and I immediately took up his method since it held promise of leading to an interpretation of the ψ -function. Again an idea of Einstein's gave me the lead. He had tried to make the duality of particles - light quanta or photons - and waves comprehensible by interpreting the square of the optical wave amplitudes as probability density for the occurrence of photons. This concept could at once be carried over to the ψ -function: $|\psi|^2$ ought to represent the probability density for electrons (or other particles). It was easy to assert this, but how could it be proved?

The atomic collision processes suggested themselves at this point. A swarm of electrons coming from infinity, represented by an incident wave of known intensity (i.e., $|\psi|^2$), impinges upon an obstacle, say a heavy atom. In the same way that a water wave produced by a steamer causes secondary circular waves in striking a pile, the incident electron wave is partially transformed into a secondary spherical wave whose amplitude of oscillation ψ differs for different directions. The square of the amplitude of this wave at a great distance from the scattering centre determines the relative probability of scattering as a function of direction. Moreover, if the scattering atom itself is capable of existing in different stationary states, then Schrödinger's wave equation gives automatically the probability of excitation of these states, the electron being scattered with loss of energy, that is to say, inelastically, as it is called. In this way it was possible to get a theoretical basis¹⁷ for the assumptions of Bohr's theory which had been experimentally confirmed by Franck and Hertz. Soon Wentzel¹⁸ succeeded in deriving Rutherford's famous formula for the scattering of α -particles from my theory.

However, a paper by Heisenberg¹⁹, containing his celebrated uncertainty relationship, contributed more than the above-mentioned successes to the swift acceptance of the statistical interpretation of the ψ -function. It was through this paper that the revolutionary character of the new conception became clear. It showed that not only the determinism of classical physics must be abandoned, but also the naive concept of reality which looked upon the particles of atomic physics as if they were very small grains of sand. At every instant a grain of sand has a definite position and velocity. This is not the case with an electron. If its position is determined with increasing accuracy, the possibility of ascertaining the velocity becomes less and *vice*

versa. I shall return shortly to these problems in a more general connection, but would first like to say a few words about the theory of collisions.

The mathematical approximation methods which I used were quite primitive and soon improved upon. From the literature, which has grown to a point where I cannot cope with, I would like to mention only a few of the first authors to whom the theory owes great progress: Faxén in Sweden, Holtsmark in Norway²⁰, Bethe in Germany²¹, Mott and Massey in England²².

Today, collision theory is a special science with its own big, solid textbooks which have grown completely over my head. Of course in the last resort all the modern branches of physics, quantum electrodynamics, the theory of mesons, nuclei, cosmic rays, elementary particles and their transformations, all come within range of these ideas and no bounds could be set to a discussion on them.

I should also like to mention that in 1926 and 1927 I tried another way of supporting the statistical concept of quantum mechanics, partly in collaboration with the Russian physicist Fock²³. In the above-mentioned three-author paper there is a chapter which anticipates the Schrödinger function, except that it is not thought of as a function $\psi(x)$ in space, but as a function ψ_n of the discrete index $n = 1, 2, \dots$ which enumerates the stationary states. If the system under consideration is subject to a force which is variable with time, ψ_n becomes also time-dependent, and $|\psi_n(t)|^2$ signifies the probability for the existence of the state n at time t . Starting from an initial distribution where there is only one state, transition probabilities are obtained, and their properties can be examined. What interested me in particular at the time, was what occurs in the adiabatic limiting case, that is, for very slowly changing action. It was possible to show that, as could have been expected, the probability of transitions becomes ever smaller. The theory of transition probabilities was developed independently by Dirac with great success. It can be said that the whole of atomic and nuclear physics works with this system of concepts, particularly in the very elegant form given to them by Dirac²⁴. Almost all experiments lead to statements about relative frequencies of events, even when they occur concealed under such names as effective cross section or the like.

How does it come about then, that great scientists such as Einstein, Schrödinger, and De Broglie are nevertheless dissatisfied with the situation? Of course, all these objections are levelled not against the correctness of the formulae, but against their interpretation. Two closely knitted points of view

are to be distinguished: the question of determinism and the question of reality.

Newtonian mechanics is deterministic in the following sense:

If the initial state (positions and velocities of all particles) of a system is accurately given, then the state at any other time (earlier or later) can be calculated from the laws of mechanics. All the other branches of classical physics have been built up according to this model. Mechanical determinism gradually became a kind of article of faith: the world as a machine, an automaton. As far as I can see, this idea has no forerunners in ancient and medieval philosophy. The idea is a product of the immense success of Newtonian mechanics, particularly in astronomy. In the 19th century it became a basic philosophical principle for the whole of exact science. I asked myself whether this was really justified. Can absolute predictions really be made for all time on the basis of the classical equations of motion? It can easily be seen, by simple examples, that this is only the case when the possibility of absolutely exact measurement (of position, velocity, or other quantities) is assumed. Let us think of a particle moving without friction on a straight line between two end-points (walls), at which it experiences completely elastic recoil. It moves with constant speed equal to its initial speed v_0 backwards and forwards, and it can be stated exactly where it will be at a given time provided that v_0 is accurately known. But if a small inaccuracy Δv_0 is allowed, then the inaccuracy of prediction of the position at time t is $t\Delta v_0$ which increases with t . If one waits long enough until time $t_c = l/\Delta v_0$ where l is the distance between the elastic walls, the inaccuracy Δx will have become equal to the whole space l . Thus it is impossible to forecast anything about the position at a time which is later than t_c . Thus determinism lapses completely into indeterminism as soon as the slightest inaccuracy in the data on velocity is permitted. Is there any sense - and I mean any physical sense, not metaphysical sense - in which one can speak of absolute data? Is one justified in saying that the coordinate $x = \pi$ cm where $\pi = 3.1415\ldots$ is the familiar transcendental number that determines the ratio of the circumference of a circle to its diameter? As a mathematical tool the concept of a real number represented by a nonterminating decimal fraction is exceptionally important and fruitful. As the measure of a physical quantity it is nonsense. If π is taken to the 20th or the 25th place of decimals, two numbers are obtained which are indistinguishable from each other and the true value of π by any measurement. According to the heuristic principle used by Einstein in the theory of relativity, and by Heisenberg in the quantum theory, concepts which correspond to no conceivable observation should be eliminated

from physics. This is possible without difficulty in the present case also. It is only necessary to replace statements like $x = \pi$ cm by: the probability of distribution of values of x has a sharp maximum at $x = \pi$ cm; and (if it is desired to be more accurate) to add: of such and such a breadth. In short, ordinary mechanics must also be statistically formulated. I have occupied myself with this problem a little recently, and have realized that it is possible without difficulty. This is not the place to go into the matter more deeply. I should like only to say this: the determinism of classical physics turns out to be an illusion, created by overrating mathematico-logical concepts. It is an idol, not an ideal in scientific research and cannot, therefore, be used as an objection to the essentially indeterministic statistical interpretation of quantum mechanics.

Much more difficult is the objection based on reality. The concept of a particle, e.g. a grain of sand, implicitly contains the idea that it is in a definite position and has definite motion. But according to quantum mechanics it is impossible to determine simultaneously with any desired accuracy both position and velocity (more precisely : momentum, i.e. mass times velocity). Thus two questions arise: what prevents us, in spite of the theoretical assertion, to measure both quantities to any desired degree of accuracy by refined experiments? Secondly, if it really transpires that this is not feasible, are we still justified in applying to the electron the concept of particle and therefore the ideas associated with it?

Referring to the first question, it is clear that if the theory is correct - and we have ample grounds for believing this - the obstacle to simultaneous measurement of position and motion (and of other such pairs of so-called conjugate quantities) must lie in the laws of quantum mechanics themselves. In fact, this is so. But it is not a simple matter to clarify the situation. Niels Bohr himself has gone to great trouble and ingenuity²⁵ to develop a theory of measurements to clear the matter up and to meet the most refined and ingenious attacks of Einstein, who repeatedly tried to think out methods of measurement by means of which position and motion could be measured simultaneously and accurately. The following emerges: to measure space coordinates and instants of time, rigid measuring rods and clocks are required. On the other hand, to measure momenta and energies, devices are necessary with movable parts to absorb the impact of the test object and to indicate the size of its momentum. Paying regard to the fact that quantum mechanics is competent for dealing with the interaction of object and apparatus, it is seen that no arrangement is possible that will fulfil both require-

ments simultaneously. There exist, therefore, mutually exclusive though complementary experiments which only as a whole embrace everything which can be experienced with regard to an object.

This idea of *complementarity* is now regarded by most physicists as the key to the clear understanding of quantum processes. Bohr has generalized the idea to quite different fields of knowledge, e.g. the connection between consciousness and the brain, to the problem of free will, and other basic problems of philosophy. To come now to the last point: can we call something with which the concepts of position and motion cannot be associated in the usual way, a thing, or a particle? And if not, what is the reality which our theory has been invented to describe?

The answer to this is no longer physics, but philosophy, and to deal with it thoroughly would mean going far beyond the bounds of this lecture. I have given my views on it elsewhere²⁶. Here I will only say that I am emphatically in favour of the retention of the particle idea. Naturally, it is necessary to redefine what is meant. For this, well-developed concepts are available which appear in mathematics under the name of invariants in transformations. Every object that we perceive appears in innumerable aspects. The concept of the object is the invariant of all these aspects. From this point of view, the present universally used system of concepts in which particles and waves appear simultaneously, can be completely justified.

The latest research on nuclei and elementary particles has led us, however, to limits beyond which this system of concepts itself does not appear to suffice. The lesson to be learned from what I have told of the origin of quantum mechanics is that probable refinements of mathematical methods will not suffice to produce a satisfactory theory, but that somewhere in our doctrine is hidden a concept, unjustified by experience, which we must eliminate to open up the road.

1. R. Ladenburg, *Z. Physik*, 4 (1921) 451 ; R. Ladenburg and F. Reiche, *Naturwiss.*, 11 (1923) 584.
2. H. A. Kramers, *Nature*, 113 (1924) 673.
3. H. A. Kramers and W. Heisenberg, *Z. Physik*, 31 (1925) 681.
4. M. Born, *Z. Physik*, 26 (1924) 379; M. Born and P. Jordan, *Z. Physik*, 33 (1925) 479.
5. W. Heisenberg, *Z. Physik*, 33 (1925) 879.
6. M. Born and P. Jordan, *Z. Physik*, 34 (1925) 858.

7. M. Born, W. Heisenberg, and P. Jordan, *Z. Physik*, 35 (1926) 557.
8. P. A. M. Dirac, *Proc. Roy. Soc. (London)*, A 109 (1925) 642.
9. W. Pauli, *Z. Physik*, 36 (1926) 336.
10. E. Schrödinger, *Ann. Physik*, [4] 79 (1926) 361,489,734; 80 (1926) 437; 81(1926) 109.
11. L. de Broglie, *Thesis Paris*, 1924; *Ann. Phys. (Paris)*, [10] 3 (1925) 22.
12. W. Elasser, *Naturwiss.*, 13 (1925) 711.
13. C. J. Davisson and L. H. Germer, *Phys. Rev.*, 30 (1927) 707.
14. G. P. Thomson and A. Reid, *Nature*, 119 (1927) 890; G. P. Thomson, *Proc. Roy. Soc. (London)*, A 117 (1928) 600.
15. E. Schrödinger, *Brit. J. Phil. Sci.*, 3 (1952) 109, 233.
16. M. Born and N. Wiener, *Z. Physik*, 36 (1926) 174.
17. M. Born, *Z. Physik*, 37 (1926) 863 ; 38 (1926) 803 ; *Göttinger Nachr. Math. Phys. Kl.*, (1926) 146.
18. G. Wentzel, *Z. Physik*, 40 (1926) 590.
19. W. Heisenberg, *Z. Physik*, 43 (1927) 172.
20. H. Faxén and J. Holtsmark, *Z. Physik*, 45 (1927) 307.
21. H. Bethe, *Ann. Physik*, 5 (1930) 325.
22. N. F. Mott, *Proc. Roy. Soc. (London)*, A 124 (1929) 422, 425; *Proc. Cambridge Phil. Soc.*, 25 (1929) 304.
23. M. Born, *Z. Physik*, 40 (1926) 167; M. Born and V. Fock, *Z. Physik*, 51 (1928) 165.
24. P. A. M. Dirac, *Proc. Roy. Soc. (London)*, A 109 (1925) 642; 110 (1926) 561; 111 (1926) 281; 112 (26) 674.
25. N. Bohr, *Naturwiss.*, 16 (1928) 245; 17 (1929) 483; 21 (1933) 13 . «Kausalität und Komplementarität» (Causality and Complementarity), *Die Erkenntnis*, 6 (1936) 293.
26. M. Born, *Phil. Quart.*, 3 (1953) 134; *Physik. Bl.*, 10 (1954) 49.

ERNST B. CHAIN

The chemical structure of the penicillins

Nobel Lecture, March 20, 1946

Before beginning with the subject proper of this lecture let me give you a few details of the historical development of the chemical work on penicillin and its organization. Work on the purification and the structure of penicillin was started at Oxford immediately after the extraordinary chemotherapeutic value of the compound had been established conclusively by our group. The initial chemical work was done by my colleague Dr. E. P. Abraham and myself in the Department of Pathology. Towards the end of 1942 we joined forces with Dr. W. Baker (now Professor of Organic Chemistry at Bristol) and Sir Robert Robinson. This group of chemists - Dr. Abraham, Dr. Baker, Sir R. Robinson and myself - have formed the nucleus of research workers whose efforts have led to the elucidation of the chemical structure of the penicillins and the synthesis of some of their degradation products. The success of this work has been due to the combined efforts of all the members of our group, and I should like you to regard me tonight merely as its representative.

Shortly after the chemical work had been started at Oxford, a number of other British research centres, both academic and industrial, began similar studies. Of these I should like to mention in particular the Imperial College of Science whose group was under the leadership of Dr. A. H. Cook and Professor Sir Ian Heilbron, the chemical laboratories of Burroughs Wellcome Ltd. in which the work was directed by Dr. S. Smith, the laboratories of Imperial Chemical Industries Ltd., and the laboratories of the firm of Glaxo, under the direction of Dr. F. A. Robinson.

Simultaneously with the work in England, American chemists began an intensive study of the structure of penicillin with the aim of quickly achieving a synthesis. This work was carried out on a very large scale, with something like 200 academic and industrial research chemists taking part in the project. Until May 1944 this work was entirely independent of the British effort, and we in Britain had no information about the state of the American investigations, except for a few fragmentary rumours.

In 1943 the British and U.S. Governments imposed a ban on the publica-

tion of all chemical work on penicillin and simultaneously negotiations were begun between the two governments for the purpose of finding a suitable method for a complete exchange of information between the various groups of workers on both sides of the Atlantic. These negotiations were protracted, and while they were in progress we at Oxford got on with our studies and were able to propose the first complete structural formulae for penicillin in October, 1943. In February, 1944, agreement for exchange of information between the British and American workers was reached; in Britain the *Medical Research Council* (M.R.C.) formed the "Penicillin Synthesis Committee" to which were sent papers by British authors; in America the *Office of Scientific Research and Development* (O.S.R.D.) delegated Dr. Hans T. Clarke of Columbia University to co-ordinate the chemical research work on penicillin in the U.S.A. and to receive monthly reports from its contractors. These two bodies, the M.R.C. and O.S.R.D., agreed to exchange their reports at monthly intervals, and in April 1944 we received the first American reports on penicillin. As I have already mentioned, the Americans have put a tremendous effort into the investigations on the chemistry of penicillin, and the following groups of chemists in the U.S.A. have participated in the project: *Academic* - Dr. Du Vigneaud and his collaborators, of Cornell University, New York, Dr. W. Bachmann of Michigan University; Dr. Woodward of Harvard University. *Industrial* - the Merck group, who have made the most extensive and valuable contributions in the degradation work as well as in the synthetic studies; the Squibb group; the Pfizer group; the N.R.R.L. group of the U.S. Department of Agriculture at Peoria, Illinois; the Abbott group; the Eli Lilly group; the Upjohn group; the Shell group; and others.

We at Oxford have been greatly handicapped in our work by lack of material. Altogether we had about 2 g of penicillin at our disposal; of this 1.5 g were about 50% pure and only about 500 mg were about 90% pure. The American workers were in a more fortunate position; the Merck group alone has used up many hundred grams of pure crystalline penicillin.

The Anglo-American collaboration continued until October 1945, and altogether about 700 reports were sent to the coordinating government committees. These reports contain partly work directly concerned with the degradation of the penicillins, and partly synthetic work, concerned with the synthesis of degradation products, intermediates and model compounds. It is obviously impossible to give you a complete account of all the work embodied in the 700 reports, in which a good many new compounds have

been added to Beilstein. I shall limit myself to work bearing directly on the purification and structure of the penicillins and shall quote only as much of the synthetic work as is relevant to the arguments about the structure. For the sake of presenting a coherent and clear picture it will not be always convenient to follow strictly the historical course of events, but I shall try to do so whenever possible. A comprehensive account of all the chemical work on penicillin is being published in form of an Anglo-American monograph under the auspices of the National Academy of Sciences, Washington, U.S.A.

During the purification studies it became clear that there existed several penicillins which had very similar biological and chemical properties, but which differed in their chemical composition. Later work showed that all penicillins contain a common nucleus, but differ in the structure of their side chains. So far four different penicillins have been obtained in the form of their crystalline sodium salts. They are designated in England as penicillins I-IV, according to the sequence of their historical discovery; in America they are termed, F, G, X and K.

Let me briefly bring back to your memory the most important physical and chemical properties of penicillin. The penicillins are organic acids, readily soluble in different organic solvents, such as esters, chloroform or ether, but insoluble or only sparingly soluble in hydrocarbons. They are stable in water only in the form of their salts, in a pH ranging between 5 and 8, and rapidly lose their biological activity in aqueous solutions of higher acidity or alkalinity. In addition to acid and alkali, the penicillins are also inactivated by many other reagents, for example by most heavy-metal ions, including those of Zn and Cd, by primary alcohols and amines, thiols, aldehydic or ketonic reagents, oxidizing reagents and a specific enzyme, penicillinase, which occurs in some penicillin-resistant strains of bacteria.

There is not time to describe in detail the methods of purifying the penicillins and a few general remarks about them must suffice. In view of the high sensitivity of the penicillins to many reagents commonly used in purification processes we were limited almost exclusively to distribution of penicillin between different solvents and to various forms of chromatography. In particular, extensive use has been made of modifications of the method of partition chromatography, a method invented in England by Martin and Synge, which is capable of wide applicability.

The success of the purification process depends entirely on the nature of the starting material, in other words on the composition of the culture me-

dium, on the conditions of fermentation, and on the strain of *Penicillium notatum* used. With the starting material as it is now available, the purification of penicillins II and III presents no difficulty, and crystalline sodium penicillin II has become a readily available substance.

The first penicillin to be obtained pure was penicillin II, which was crystallized in the form of its sodium salt. This was achieved about July 1943 by Wintersteiner and MacPhillamy, working at the Squibb Institute in New Jersey. About one week later, we at Oxford obtained the sodium salt of penicillin I in the crystalline state. Only the alkali salts of the penicillins and their salts with a few simple organic cations have so far been obtained crystalline. Despite many attempts it has not yet been possible to obtain crystalline their salts with any divalent metals. The sodium salts of penicillins I, II, and III can be crystallized from a mixture of water and butanol (1:20). Crystalline sodium salt of penicillin II is now produced on an industrial scale.

The crystalline sodium salts of the penicillins are colourless needles. The pure substances are strongly dextro-rotatory, $[\alpha]_D$ of penicillin I and II being +305°. Elementary analysis of the crystalline sodium salts has shown that the penicillins have the following composition:

Penicillin I	$C_{14}H_{20}O_4N_2S$
Penicillin II	$C_{16}H_{18}O_4N_2S$
Penicillin III	$C_{16}H_{18}O_5N_2S$
Penicillin IV	$C_{16}H_{26}O_4N_2S$

On catalytic hydrogenation with Pt or Pd, penicillin-I takes up one mol of H₂. The other penicillins do not react with catalytically activated hydrogen.

Analysis of the salts and electrometric titration curves have shown that the penicillins are strong monobasic acids having pK's about 2.9 (Fig.1). There is no indication of the presence of any basic group in the electrotitration curve. This fact has played an important role in structural considerations.

The acid group in the penicillins is a carboxyl group that can be esterified by the action of CH₂N₂. The methyl ester has been obtained in the crystalline state. Its activity *in vitro*, about 70 u./mg, is much less than that of penicillin salts, but *in vivo* it possesses about the same activity as the salts. This is due to the fact that it is hydrolyzed easily by enzymes occurring in the body tissues. The methyl ester of penicillin cannot be hydrolyzed chemically even under mild conditions (pyridine and one equivalent of alkali at 0°C) without appreciable loss of antibacterial activity.

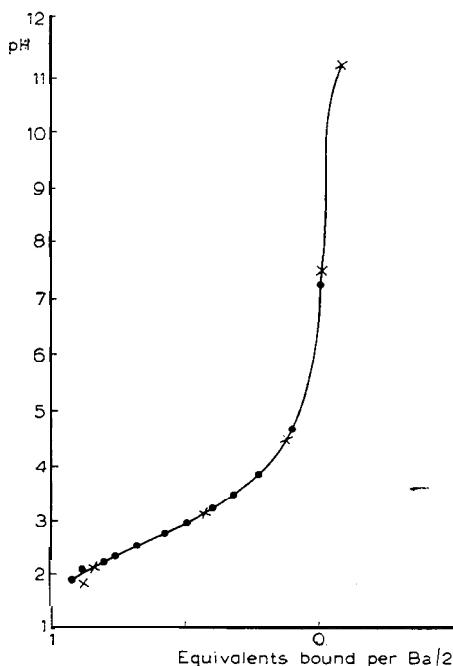


Fig. 1. Electrometric titration curve of 2-pentenylpenicillin (0°C).

Molecular weight determinations of the penicillins by several methods have shown that their molecular weights correspond to the simple formulae, shown above. Penicillin I and IV have no characteristic u.v. absorption, but penicillin II and III show clearly the fine structure of a benzene ring. When penicillin is inactivated by keeping at acid pH (Fig. 2), the electrometric titration shows that a new very strong acidic group, about $\text{pK } 1.5$, and a basic group $\text{pK } 7.6$, is formed. The reaction product is insoluble in organic solvents, in accordance with its zwitterionic structure.

When penicillin is inactivated by alkali at pH 10, it is also converted into a zwitterion with the formation of new acidic and basic groups, but this compound differs from the product of acid inactivation, the newly formed acidic group having a pK of 1.8, the new basic group a pK of 5. Both products, that of acid as well as that of alkaline inactivation, have been obtained in the crystalline state. The product of acid inactivation is isomeric with penicillin and is termed penillic acid. The product of alkaline inactivation contains an additional molecule of H_2O ; it is thus a hydrolysis product and is termed penicilloic acid. We shall discuss the structure of these important degradation products later on.

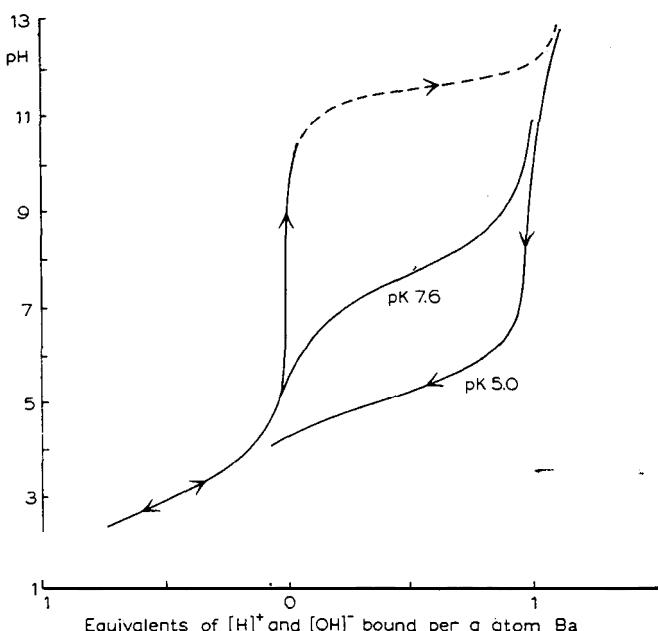
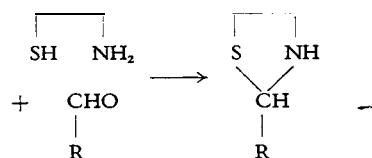


Fig. 2. Electrometric titration of 2-pentenylpenicillin (0°C) before and after inactivation with acid and alkali.

As a starting-point in the elucidation of the structure of the penicillin molecule we decided to investigate the nature of the two nitrogen atoms shown to be present by analysis. Some indication of the nature of these nitrogen atoms was obtained by hydrolysis of penicillin at 100° with normal acid. After short hydrolysis (30 min-1h) one of the two nitrogen atoms appeared in the form of NH_2 -nitrogen, estimable by the Van Slyke procedure for the determination of α -amino groups; during prolonged hydrolysis (24 h) the other N was gradually liberated as ammonia. The acid hydrolysate of penicillin gave a strong ninhydrin reaction, confirming the presence of the α -amino acid suggested by the Van Slyke determination. This amino acid was the first degradation product of penicillin to be isolated in crystalline form. It is precipitated by HgCl_2 and obtained crystalline after decomposition of the mercury complex with H_2S . Elementary analysis shows that it is a hydrochloride, having the formula $\text{C}_5\text{H}_{11}\text{O}_2\text{NSHCl}$, and it is thus the moiety of the penicillin molecule which contains the sulphur atom. The S-containing amino acid was termed penicillamine; it gives strong nitroprusside and ferric chloride reactions for SH. On oxidation with bromine it yields a crystalline compound which was termed penicillaminic acid. This

substance afforded better analyses than penicillamine hydrochloride. It contained three additional oxygen atoms indicating that it was the sulphonic acid corresponding to penicillamine. The titration curve showed that it contained two acid groups (the sulphonic acid and carboxylic groups) and one basic group (the α -NH₂ group), but no SH group. Like penicillamine, it gives a strong ninhydrin reaction, and all its nitrogen appears as α -amino acid nitrogen in the Van Slyke determination. That the amino and thiol groups in penicillamine are in juxtaposition, is shown by the easy formation of thiazolidines when the substance is warmed with ketones and aldehydes :



The titration of penicillamine shows clearly three proton binding centres that correspond to the carboxyl group (pK 1.8), the α -amino group (pK 7.9) and the SH group (pK 10.5) (Fig. 3.)

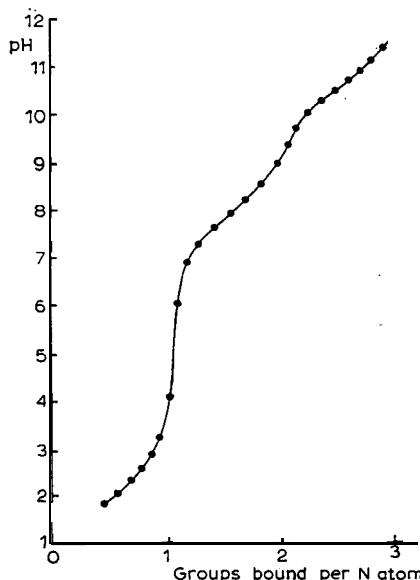
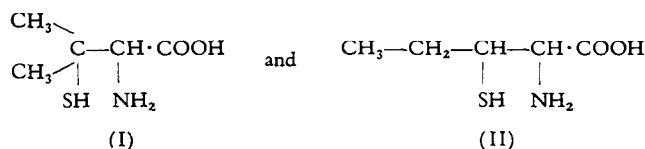
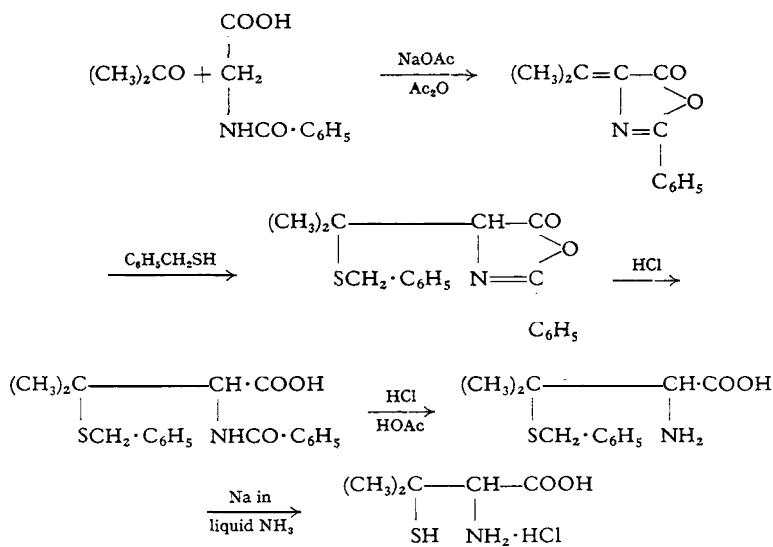


Fig. 3. Electrometric titration curve of penicillamine hydrochloride (25°C).

Of the two possible isomeric structures for penicillamine, (II) appeared



improbable because $\text{C}-\text{CH}_3$ determinations gave very low values; on the other hand this finding was in accordance with structure (I), as it is known from the work of Kuhn and Roth that *gem*-dimethyl groups such as are present in structure (I) are not oxidized to acetic acid by chromic acid under the conditions of the $\text{C}-\text{CH}_3$ determination method. We concluded therefore that penicillamine had structure (I). This was conclusively proved by synthesis. This synthesis is based on a method evolved by Carter, Stevens and Ney (*J. Biol. Chem.*, 139 (1941) 247) for the synthesis of methyl cysteine; it involves the addition of benzyl mercaptan to the double bond of the azlactone obtained by condensation of hippuric acid with acetone. The steps in this synthesis are indicated in the following scheme:



Several other methods of synthesising penicillamine have now been developed.

The resolution of this amino acid is best achieved through fractionation of the brucine salt of the N-formyl compound. Natural penicillamine belongs

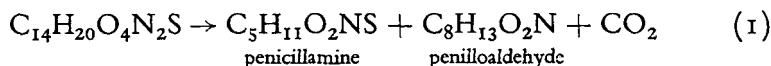
to the "unnatural" *d*-configuration. This was anticipated from the optical behaviour of penicillamine and its acetone derivative which was analogous to that of *d*-cysteine. The *d*-configuration of penicillamine was finally proved by treatment of the phenylureido derivative with Raney nickel, which led to the phenylureido derivative of *d*-valine. Penicillamine is a new amino acid which so far has not been found in any other biological material and it is yet another example of an amino acid of "unnatural" configuration produced by micro-organisms. These occur, for example, in the antibiotics gramicidin and tyrocidine and in the antigen from *Bacillus mesentericus*.

Penicillamine is similar to cysteine in many respects but it is much more soluble in water. The same applies to the disulphide. The disulphide, however, differs from cysteine in its far greater stability towards reducing agents; thus, unlike cysteine, it cannot be reduced by KCN. Neither *d*- nor *l*-penicillamine, nor their disulphides, are attacked by enzymes occurring in animal tissues.

Penicillamine, which is a constituent common to all penicillins, accounts for five of the 14 C atoms present in penicillin I. Another carbon atom is accounted for in the form of CO₂, one molecule of which is liberated when the free penicillin is heated to about 60°. The remaining eight carbon atoms are found in an aldehyde C₈H₁₃O₂N which is isolated in small amounts from the acid hydrolysates of penicillin I, after removal of the penicillamine by HgCl₂. This aldehyde was obtained in the form of the 2,4-dinitrophenylhydrazone and as the dimedone derivative. It is obtained in larger amounts after treatment of penicillin with alkali ("alkali inactivation") and subsequent treatment of the solution with HgCl₂. This precipitates penicillamine in the form of its mercuric chloride complex and simultaneously one molecule of CO₂ is liberated; the supernatant solution now gives in good yield a precipitate with 2,4-dinitrophenylhydrazine of the hydrazone of the aldehyde C₈H₁₃O₂N. This aldehyde was termed penillo-aldehyde. Thus, all the 14 carbon atoms in penicillin I had been accounted for and the equation C₁₄H₂₀O₄N₂S + 2H₂O = C₅H₁₁O₂NS + CO₂ + C₈H₁₃O₂N could be written. The constitution of the aldehyde C₈H₁₃O₂N was elucidated as follows:

Oxidation with Ag₂O of penillo-aldehyde gave a crystalline acid C₈H₁₃O₃N. Information about the nature of the nitrogen in this acid was obtained by hydrolysis at 100° with N HCl: the hydrolysate gave a strong ninhydrin reaction for **α-amino** acids and about 70% of the nitrogen present in it was detectable as NH₂-nitrogen by the Van Slyke procedure. Hence it was con-

cluded that the acid $C_8H_{13}O_3N$ contained a peptide linkage. Its exact constitution was deduced from information about the composition and behaviour on degradation of the American penicillin. The empirical formula of this penicillin was telegraphed to the M.R.C. in July 1943; it was $C_{16}H_{18}O_4N_2S$. Now we know that on acid hydrolysis the English penicillin decomposed according to the equation



As we were informed that the American penicillin afforded the same amino acid penicillamine on acid hydrolysis, we assumed that its hydrolysis proceeded according to the equation



We had heard that the American workers had isolated phenylacetic acid from acid hydrolysis of their penicillin. This is an easily recognizable substance which we never encountered among our own degradation products. When we were informed that the American workers had isolated phenylacetic acid from a crystalline degradation product of penicillin, we then knew for certain that the American penicillin differed in chemical composition from our own penicillin and that the difference could only be in the penilloaldehyde moiety. Assuming that the American penilloaldehyde contained a phenylacetyl group and, like ours, a peptide linkage, its structure could only be $C_6H_5CH_2CONHCH_2\cdot CHO$, that is phenylacetylaminooacetaldehyde.

If this assumption were correct our C_8 aldehyde should have the structure $C_5H_9CO\cdot NH\cdot CH_2\cdot CHO$, of a hexenoylaminooacetaldehyde, and our C_8 acid, derived from the aldehyde by oxidation should have the structure $C_5H_9CONH\cdot CH_2\cdot COOH$, of a hexenoylglycine. The presence of glycine in this acid was proved by its isolation in the form of the naphthalene-sulphonyl derivative. The structure of the unsaturated fatty acid C_5H_9COOH was established by oxidation with cold permanganate which gave propional, locating the double bond in the β,γ -position, and proving the structure as $CH_3H_2CH=CH\cdot CH_2COOH$. This structure was also confirmed by work at the Imperial College, where caproic acid, in the form of its *p*-bromophenyl-phenacyl derivative was isolated after hydrogenation of the

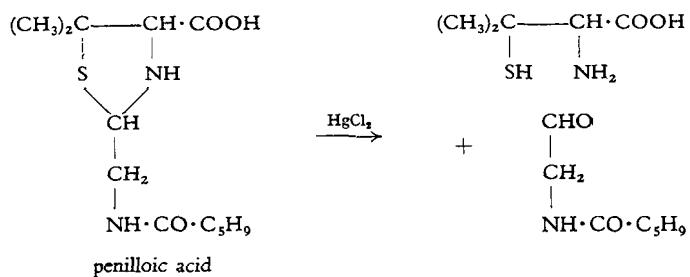
acid and subsequent acid hydrolysis. The structure of penillo-aldehyde-I was thus proved to be da-hexenoylamino-acetaldehyde, $\text{CH}_3\text{CH}_2\text{CH}=\text{CHCH}_2\text{-CONHCH}_2\cdot\text{CHO}$. The acetal of this aldehyde was synthesized from the acetal of aminoacetaldehyde and Δ^2 -hexenoylchloride. Treatment of the products with 2,4-dinitrophenylhydrazine in 2*N* H_2SO_4 gave a 2,4-dinitrophenylhydrazone identical with that obtained from natural penillo-aldehyde-I.

The hexenoyl group in the penilloaldehyde moiety of the penicillin I molecule is responsible for the uptake of one molecule of H_2 when penicillin I is treated with H_2 and Pt or Pd. Later we obtained the information that the American penicillin II did in fact yield a penilloaldehyde of the constitution we had postulated, namely phenylacetylaminocetaldehyde. It had thus been established that the penicillin molecule is built up from three parts: (1) the thiolamino acid penicillamine, (2) a labile carboxyl group that readily yields CO_2 on heating free penicillin to 60°, or on treating alkali-inactivated penicillin with HgCl_2 , and (3) an acylated aminoacetaldehyde termed penilloaldehyde. The first two components, penicillamine and the labile carboxyl group, are common to all penicillins. The penilloaldehyde moiety varies in the different penicillins. In penicillin I it is β,γ -hexenoylamino-acetaldehyde, in penicillin II phenylacetylaldehyde.

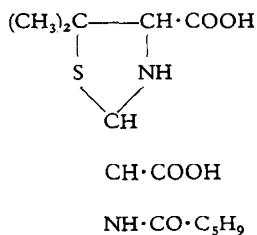
The two other penicillins that have been obtained in crystalline state yielded, on degradation, penilloaldehydes which were recognized as *p*-hydroxyphenylacetylaminocetaldehyde in penicillin III, and *n*-heptoaminocetaldehyde in penicillin IV. The question now remaining to be answered was the manner in which the three components were linked together in the penicillin molecule. It was hoped to obtain information on this point by obtaining larger breakdown products of penicillin. The study of the reactions occurring during the inactivation of penicillin by various reagents led to the isolation of such products.

Let us consider first the product that is obtained on inactivation of penicillin with alkali. We have seen that after alkali inactivation and subsequent addition of HgCl_2 the mercury complex of penicillamine is precipitated and the supernatant solution contains the free penilloaldehyde. Penilloaldehyde appears only after the addition of HgCl_2 ; before this, no precipitate is obtained with 2,4-nitrophenylhydrazine. Similarly, no reaction for SH or $\text{NH}_2\text{-N}$ is given by alkali-inactivated penicillin, thus showing that no free penicillamine is present in solution; the latter is formed only through the action of HgCl_2 on the alkali inactivation product of penicillin. From these

facts we concluded that penicillamine and penilloaldehyde were bound in solution in a thiazolidine ring which was broken by $HgCl_2$, in a manner characteristic of thiazolidines:



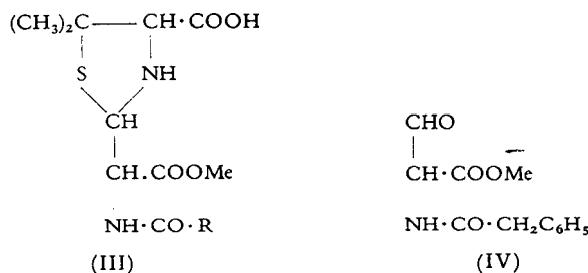
Very soon after the precipitation of penicillamine with $HgCl_2$ from alkali-inactivated penicillin, CO_2 development sets in and finally one molecule of CO_2 is liberated. The ease of liberation of CO_2 could best be explained by the assumption that it derived from a carboxyl group in the β -position to the potential aldehyde carboxyl group of penilloaldehyde. The most probable structure of the alkali-inactivation product of penicillin was therefore a thiazolidine with the formula:



This compound has been isolated in the form of its crystalline sodium salt and various crystalline derivatives, such as different esters, amides, and N-acylated derivatives of these (Merck group). It is one of the most important degradation products of penicillin and has been given the name penilloic acid. Its structure was proved by degradation and synthesis. The information leading to the certain elucidation of its structure was obtained from the study of the reaction products of penicillin with methyl alcohol and with benzylamine, reagents which, as I mentioned before, readily inactivate penicillin. When the sodium salt of penicillin I is dissolved in methyl alcohol its antibacterial activity is lost in a few hours. The product of the reaction, a monobasic acid like the original penicillin with roughly the same solubility

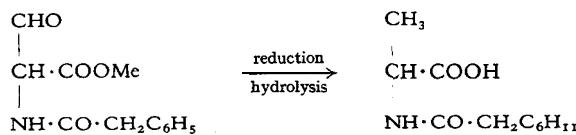
properties, contains one CH_3O group. This group is easily split off by mild alkaline hydrolysis, e.g. at pH 10 at room temperature, with the appearance of a new acid group. The resulting dicarboxylic acid behaves in every respect like alkali-inactivated penicillin (penicilloic acid), giving an identical electrometric titration curve and, on decomposition with HgCl_2 , yielding penicillamine, penilloaldehyde and CO_2 . This suggested that the product of methanol inactivation of penicillin was a mono-methyl ester of penicilloic acid of the structure (III).

This structure was proved (Merck group) by degradation of penicillin II

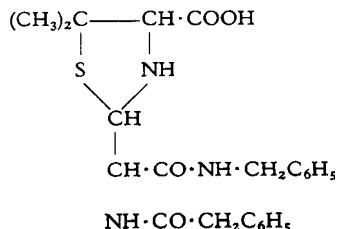


with HgCl_2 which produced penicillamine and the methyl ester of a β -aldehydic acid, termed penaldic acid, which was obtained as the crystalline 2,4-dinitrophenylhydrazone and as the amide (IV).

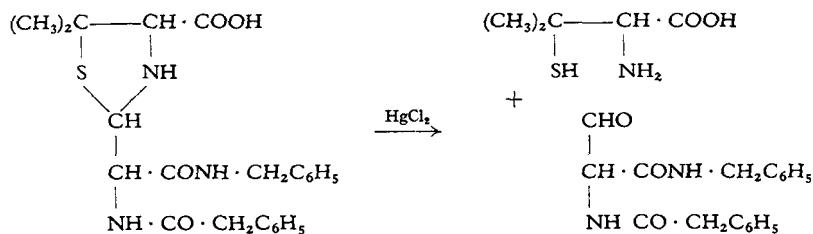
The structure of this aldehydic acid was proved by catalytic reduction to hexahydrophenylacetylalanine.



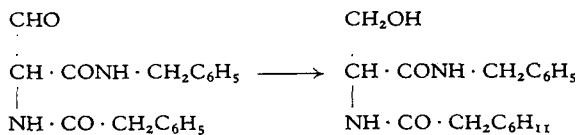
In a similar manner benzylamine reacts with free penicillin II in ether to give a crystalline compound which was shown to be the benzylamine salt of the benzylamide of penicilloic-II-acid (Merck group):



This compound is decomposed by $HgCl_2$ into penicillamine and the benzylamide of II-penaldic acid.



The structure of the benzylamide of penaldic acid was proved by reduction to the benzylamide of hexahydrophenylacetylserine.



The latter compound was synthesized by phenylacetylation of serine, esterifying the product with CH_2N_2 , treatment of the methyl ester of phenylacetylserine with benzylamine, and catalytic reduction. The isolation of penicillamine after $HgCl_2$ degradation of the benzylamine inactivation product of penicillin (benzylamide of penicilloic acid) has also proved conclusively that the free carboxyl group in penicillin belongs to the penicillamine moiety.

It may be helpful to say a few words about the nomenclature of the penicilloic acids. The two carboxyl groups are termed α and β . The two ester groups are hydrolyzed by alkali with different velocities, the α -group coming off very easily at pH 10, the β -group remaining untouched under these conditions. It is thus possible to carry out a stepwise hydrolysis of penicilloic acid di-esters. Four stereo-isomers are theoretically possible and three of them have been synthesized in the form of the N-benzoyl derivatives of the α -methyleneesters. Their melting points and specific rotations differ considerably as shown in Table 1. The fourth has been obtained in the form of its crystalline copper salt by the action of copper sulphate on sodium penicillin II or the α -isomer of penicilloic-II-acid. The isomers were prepared by mutarotating the synthetic material (which is predominantly the γ -form) in methanol, benzoylating the crude mixture and fractionating the N-benzoyl

derivative by crystallization from ether. Frequent use was also made of the benzylamine salts and the perchlorates.

Table 1.

	<i>Melting point of N-benzoyl derivatives</i>	$[\alpha]_D^{22}$ (alcohol)
α -Methyl <i>d</i> - α -II-penicilloate	171-173°	+ 60.0
α -Methyl <i>d</i> - β -II-penicilloate	236-237°	- 13.2
α -Methyl <i>d</i> - γ -II-penicilloate	190-192°	+ 123.0

Before we turn to discussing possible formulae for penicillin it is necessary to consider the structure of another degradation product obtained in good yield on acid inactivation. When free penicillin acid, obtained-for example by treating the barium salt with one equivalent of H_2SO_4 , is left in aqueous solution at room temperature for about 30 minutes, about 80% of the material becomes insoluble in organic solvents, and on evaporation of the aqueous phase a nicely crystalline compound is obtained in good yield. This compound has been termed penillic acid and was one of the first degradation products of penicillin obtained in the crystalline state (Duffin and Smith, *Nature* 151 (1943) 251). Its composition is the same as that of penicillin, but its chemical and physical properties are totally different. It is thus a product of some intramolecular re-arrangement of the penicillin molecule. Penillic acid contains two acid groups and one basic group (Fig. 4). It is more strongly dextro-rotatory than penicillin, having $[\alpha]_D + 529^\circ$, and it has a characteristic u.v. absorption spectrum, with a maximum at 2350 Å. On heating with acid it yields the same products as are obtained on acid hydrolysis of penicillin, i.e. penillamine, penilloaldehyde and CO_2 .

The structure of penillic-I-acid was deduced from a crystalline degradation product obtained by the action on it of $HgCl_2$. When $HgCl_2$ is added to a solution of penillic acid, one molecule of CO_2 is liberated and the mercury complex of a base $C_{13}H_{20}O_2N_2S$ precipitates. The hydrochloride of this base is obtained on decomposition of the mercury complex with H_2S . The base was termed penillamine. It gives a strong SH test with ferric chloride and sodium nitroprusside. Electrotitration (Fig. 5) shows up the SH group indicated by the colour tests, and reveals in addition the presence of one carboxyl group and one basic group. On heating it with 2,4-dinitrophenylhydrazine, there is obtained the dinitrophenylosazone of glyoxal. On oxida-

tion with bromine it yields penicillaminic acid. The only possible formula which could be constructed on the basis of these facts was (V). This formula was later proved by synthesis (Abraham, Baker, Chain, and Robinson).

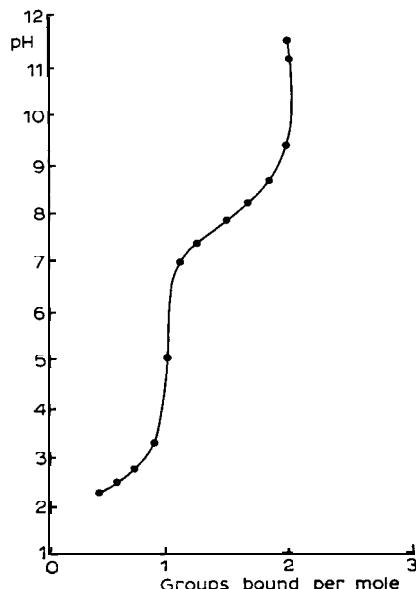


Fig. 4. Electrometric titration curve of 2-pentenylpenillic acid (25°C).

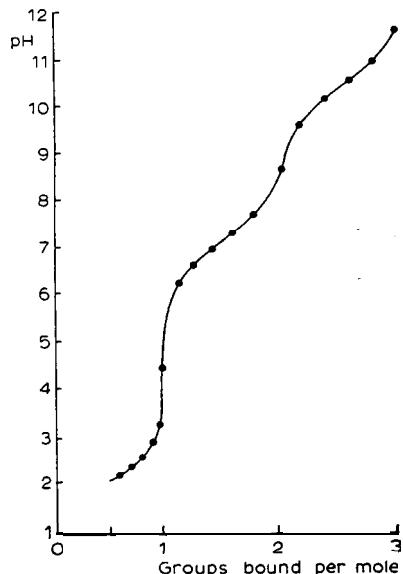
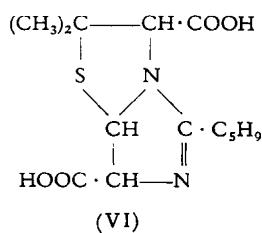
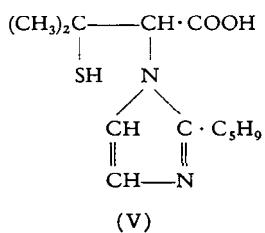
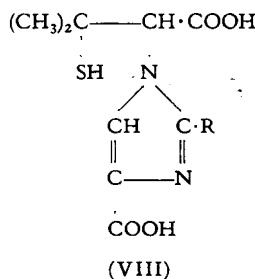
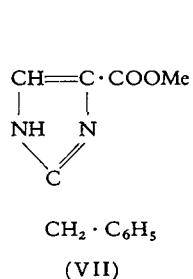


Fig. 5. Electrometric titration curve of 2-pentenylpenillamine hydrochloride (25°C).

The construction of a formula for penillic-I-acid on the basis of the formula for I-penillamine was a matter of placing into the right position the labile carboxylic group that appears in the form of CO_2 when penillic acid is treated with HgCl_2 or is heated with aqueous HCl. The most reasonable as-



sumption was that this carboxylic group was in B-position to a potential aldehydic carbonyl group, and on this assumption penillic acid could only be formulated as (VI). This formula for penillic-I-acid was in accord with all its known properties. It accounted for its two carboxylic groups and the basic group revealed in the electrotitrations, for its solubility properties and its easy transformation into penillamine by HgCl_2 . Treatment with HgCl_2 involves opening of the thiazolidine ring by hydrolysis, loss of CO_2 from the aldehyde-ammonia compound formed and subsequent or simultaneous elimination of H_2O through the tendency of this compound to go over into the very stable imidazole ring system. Further evidence for this formula was obtained much later from the mild thermal decomposition of dimethyl penillate- at 115° *in vacuo*, which gave (VII), a base which was considered to be 2-benzyl-4-carbomethoxy-imidazole. This assumption was proved correct by synthesis.



The formula of penillic acid was finally confirmed by total synthesis (Merck group).

When penillic acid I or II is treated with alkali or simply heated in meth-

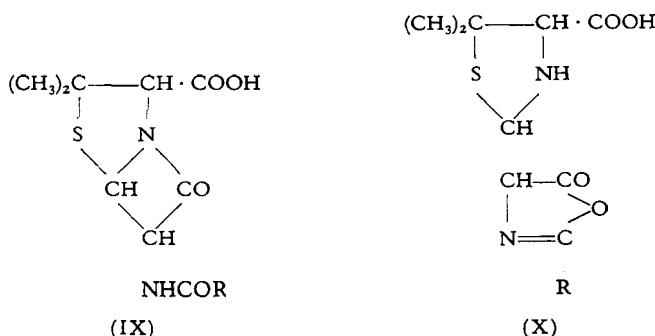
anol, the thiazolidine ring is opened and a new isomeric crystalline compound, termed isopenillic acid, is obtained. This contains a free SH group, as indicated by the colour reaction with nitroprusside and FeCl_3 , and has the structure (VIII) (Oxford workers). This structure has also been obtained by synthesis (Merck group).

Knowing the structure of penicilloic and penillic acids it is possible to construct formulae for the penicillin molecule. In the attempt to do this the main considerations to be borne in mind are the following:

- (1) Penicillin is a monobasic acid, but the degradation products are dibasic acids. Penicillin must therefore have in bound form one carboxyl group that is easily liberated by alkali, methanol and primary amines at room temperature.
- (2) The free carboxylic acid in penicillin is the penicillamine-carboxylic group.
- (3) Penicillin has no basic group, not even of the weakest type.
- (4) The penicillin molecule is capable of undergoing a facile rearrangement to an imidazoline derivative.

On the basis of these considerations our group at Oxford proposed two formulae for penicillin later known as the " β -lactam" structure (IX) and the "thiazolidine-oxazolone" structure (X).

The main guiding principle leading to the construction of Formula (IX) was the non-basicity of penicillin, and the only feasible manner in which the penicilloic NH could be rendered non-basic was to connect it with the labile carboxyl group, producing a peptide linkage. This linkage produced an admittedly very unusual four-membered ring which has not previously been observed in any natural product, but we were prepared to accept its existence because we could not find any other reasonable way of producing two non-basic nitrogens in the penicillin molecule. Furthermore we thought that



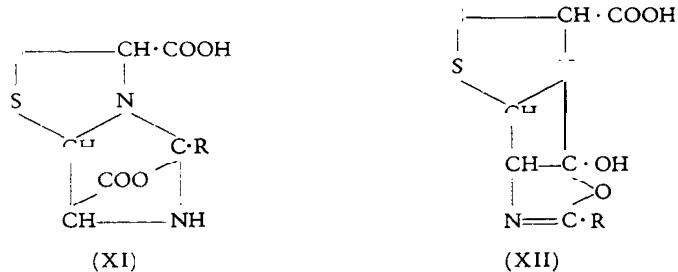
the strain inherent in four-membered rings might account for its reactions with methanol, primary amines, etc. Two arguments of purely chemical character were advanced against Formula (IX). Firstly the possibility of the existence of the four-membered ring system was considered unlikely; secondly, no reaction mechanism could be conceived which could explain in a satisfactory manner the penillic acid rearrangement since the NH-CO-R linkage as assumed in the β -lactam structure would be expected to be relatively non-reactive and would certainly not be expected to pass over into an imidazoline derivative on treatment with very dilute acid at room temperature.

The oxazolone-thiazolidine formula contained the well-known five-membered azlactone ring and appeared to explain very well the reactivity of penicillin towards CH_3OH , etc. Furthermore, a plausible reaction mechanism for the penillic acid re-arrangement on the basis of the electronic theory of Sir R. Robinson was suggested.

At the same time when these formulae were proposed, very little was known about thiazolidines or azlactones, but even then it was difficult for some of us to see any reason for the non-basicity of the NH nitrogen in the thiazolidine ring of structure (X). A nitrogen atom can only be made non-basic by being bound to a strong electron-accepting centre, such as the C=O group, and such centre was not present in Formula (X). It was argued that the basic strength of the NH could be depressed by intraspatial interaction of the carbonyl group in the azlactone group or by other factors of hitherto unknown nature, but this argument appeared unacceptable for quantitative reasons. The pK of penicilloic acid was known to be about 5 and to explain the non-basicity of penicillin a shift of about 4 pK units would have to be postulated to occur through the influence of unknown factors. However, despite these considerations the oxazolone-thiazolidine structure was for a long time most favoured by the majority of workers. In general, the attitude of the investigators to the two formulae varied according to whether they attached more weight to physico-chemical considerations, or to purely chemical arguments, based on the likelihood of reaction mechanisms that could be derived from the two formulae, to explain the various rearrangements.

Apart from Formulae (IX) and (X) many other formulae for penicillin can be constructed on paper from the elements of penaldic acid and penicillamine by the elimination of two molecules of water. Most of these formulae can be excluded *a priori* on account of their obvious disagreement with the

properties of penicillin. Two of these (XI and XII) have, however, received more serious attention in various quarters, in particular by those workers who saw the difficulties inherent in the thiazolidine-oxazolone structure, but were not prepared to accept the β -lactam structure.



Formula (XI) (Imperial College workers) contains the penillic acid skeleton performed, and explains thus in an easy way the facile formation of penillic acid on acid inactivation of penicillin. It does not, however, explain the reaction with methanol or primary amines, which in fact gives penilloic acid derivatives whereas Formula (XI) would be expected to lead to penillic acid derivatives. Furthermore, Formula (XI) gave no satisfactory explanation of the non-basicity of the thiazolidine nitrogen, which appears in penillic acid as a fairly strong basic group. One would, in fact, expect a compound of structure (XI) to possess two basic centres; Formula (XI) was finally eliminated by X-ray diffraction analysis, and the same applies to Formula (XII), an intermediate between the azlactone and β -lactam formulae. On chemical grounds Formula (XII) (Stodola, Northern Regional Research Laboratory) seemed unacceptable because it contained a carbinolamine group; these groups are known to be strong bases and thus could not explain the non-basicity of penicillin. Furthermore, carbinol-amines react with primary alcohols readily to give alkyl ethers, and Formula (XII) therefore provided no satisfactory explanation for the formation of penilloic acid derivatives under the influence of primary alcohols.

Formulae (IX) and (X) remained strong rivals for a long time, and were the object of many spirited discussions. As the work progressed, more and more evidence came in which was quite incompatible with Formula (X), but in good agreement with Formula (IX). This evidence was derived partly from synthetic model compounds, and partly from degradation studies. Let us first consider the evidence derived from synthetic work. As I have mentioned before, at the outset of this work very little was known about either

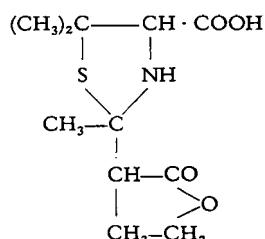
azlactones or thiazolidines. What was known was not compatible with structure (X). The few known thiazolidines derived from cysteine and a few simple aldehydes and ketones all showed a definitely basic group, having a pK of about 7. During the course of the work, a great number of thiazolidines deriving from penicillamine and from cysteine by condensation with a large variety of aldehydes and ketones of widely differing chemical constitutions were synthesized, as well as N-acyl derivatives of these. The thiazolidines are formed very easily by simple fusion of the thiolamino acid and the corresponding aldehyde and ketone at 80-100°C, with or without solvents, or by heating the acetals of the aldehydes and ketones with the hydrochlorides of the thiolamino acids at temperatures from 80° to 110°. In most cases they crystallize easily, or crystalline derivatives are formed readily.

The investigation of the properties of the newly synthesized thiazolidines has given the following general result:

All thiazolidines with non-acylated amino groups, without a single exception, have properties which are widely different from those of penicillin.

The N-acylated thiazolidines, on the other hand, resemble penicillin in many of their properties. In particular the following facts are of interest in this respect:

(1) All thiazolidines containing a non-acylated NH group possess a basic group which manifests itself clearly through facile formation of salts and in electrotitrations. The pK of the thiazolidines deriving from penicillamine is somewhere near 5. The imido group of the thiazolidines can readily be acylated, and the N-acyl thiazolidines, as is to be expected, possess no basic group. Penicillin cannot be acylated, even by the most active acylating reagents, such as ketene or acid chlorides and pyridine; its biological activity is unimpaired by these reagents. In order to investigate whether a carbonyl group in the same position as the carbonyl group of the oxazolone postulated in the thiazolidine-oxazolone structure could depress significantly the basicity of the imido group of thiazolidines the following compound was synthesized from α -acetyl-butyrolactone and penicillamine (Abbott group) :



The pK of the imino group of this substance was very similar to that of other thiazolidines, being about 4 (Eli Lilly group). Thus the carbonyl group of the lactone ring had no significant influence on the basicity of the thiazolidine nitrogen atom and the hypothesis of an intraspatial influence of the carbonyl group on the basic strength of the thiazolidine N had been rendered untenable.

During one stage of the discussion on the formula of penicillin the formation of hydrogen bonds between NH and CO of structure (X) was considered by some workers as another possible explanation for the non-basicity of the thiazolidine nitrogen. To test this inherently unlikely hypothesis, thiazolidines derived from penicillamine and *p*, *m*- or *o*-sahcylaldehyde were made. All these compounds readily formed hydrochlorides and, when titrated electrometrically, showed no significant differences in the pK values of their NH groups thus excluding possibility that there was any influence by the formation of hydrogen bonds on the pK's of the thiazolidine NH.

(2) All thiazolidines containing non-acylated NH groups react easily with I₂ to form the corresponding S-S compounds. This is because in solution there exists an equilibrium between the thiazolidine and the free thiolamino acid and the carbonyl compound; this equilibrium is displaced by oxidation of the SH compound to S-S. N-acylated thiazolidines are much more stable and do not react with I₂. Penicillin behaves like an N-acylated thiazolidine in that it does not react easily with I₂. Its degradation products, penillic or penicilloic acid react readily, however, with iodine, as is to be expected from their structures.

(3) All thiazolidines with free imino groups invariably poison Pt or Pd hydrogenation catalysts. Not only is it impossible to reduce catalytically unsaturated groups in N-non-acylated thiazolidines, but the presence of even small amounts of such thiazolidines prevents completely the catalytic hydrogenation of easily reducible substances, such as cinnamic acid. The reason for this is the formation of free SH groups, well-known catalyst poisons. N-acylated thiazolidines, on the other hand, are inert towards hydrogenation catalysts because of the greater stability of the thiazolidine ring (Eli Lilly group). Penicillin I, which contains an unsaturated hexenyl side chain, is easily reduced by Pt or Pd catalysts and behaves thus also in this respect like an N-acylated thiazolidine. Again, the breakdown products penillic and penicilloic acids behave like all other ordinary thiazolidines towards catalytic hydrogenation; i.e. they poison the catalysts irreversibly.

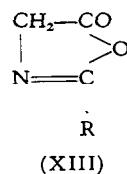
(4) On oxidation with KMnO₄, thiazolidines with free imino groups are

oxidized to the corresponding sulphonic acids while N-acylated thiazolidines yield the corresponding sulphones. Penicillin methyl ester on oxidation with KMnO_4 gives a sulphone (Merck group), a fact which is incompatible with structure (X) but in agreement with structure (IX).

To sum up, the following can be said: The examination of the properties of many model thiazolidines has shown unequivocally that penicillin behaves in no way like a normal thiazolidine with a non-acylated free imino group; its behaviour is therefore not compatible with the thiazolidine oxazolone structure (X). It resembles much more a N-acylated thiazolidine, which is in accordance with the β -lactam structure (IX). No evidence could be found for the hypothesis that the basicity of the thiazolidine NH can be depressed through the intraspatial influence of neighbouring groups.

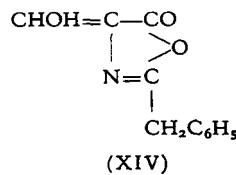
So much about thiazolidines. Apart from these, a great deal of effort has been expended in the preparation of oxazolones and the study of their properties. The outcome of this work, which time does not permit me to discuss in detail, has given the following main results:

In accordance with earlier works in the literature no oxazolone of the type (XIII) is stable in water at any pH; all are hydrolyzed more or less rapidly to

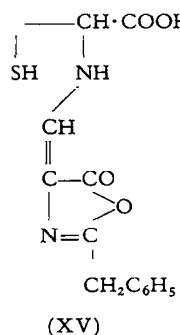


the corresponding acylated amino acids. Penicillin salts are, in contradistinction, stable in water for an indefinite time.

4-Hydroxymethylene oxazolones are stable in water in the form of their alkali salts, but are rapidly decomposed in acid medium. All oxazolones, even the most stable ones, react with liquid ammonia to give the-corresponding amide with ring opening. Penicillin, on the other hand, is quite inert towards liquid ammonia. Particular attention has been given to the preparation and study of 2-benzyl-4-hydroxymethyleneoxazolone (XIV), and various methods of preparation of this compound have been worked out.



A few words about the properties of 4-hydroxymethylene oxazolones. Their discussion is necessary for the understanding of certain degradation reactions of penicillin which will be considered shortly. The hydroxymethylene oxazolones give a strong blue colouration with FeCl_3 and have no pronounced aldehydic character. Their reactivity is more like that of acid chlorides than of aldehydes. Thus they react with diazomethane to give methoxymethylene compounds and combine readily with amines to give the corresponding aminomethylene compounds. With amino acids a similar reaction occurs. Thus crystalline aminomethylene derivatives have been obtained by combining 2-benzyl-4-ethoxymethyleneoxazolone with the amino acids glycine, alanine, and valine. With thiolamino acids hydroxymethylene oxazolones do not form thiazolidines, like normal aldehydes, but both NH_2 and SH groups react separately and independently. When molecular proportions of thiolamino acids and hydroxy- or alkoxy-oxazolones are combined, the amino group reacts preferentially; thus penicillamine and 2-benzyl-4-hydroxymethyleneoxazolone give the compound (XV)



Compounds of this type are of interest because, as will be shown later, they are degradation products of penicillin; they have been given the name penicillenic acids. 4-Aminomethylene oxazolones are recognized easily by their characteristic absorption spectra; they have two absorption maxima, at 3,200 Å (E_M 25,000) and (a weaker one) at 2,700 Å (E_M 5,000).

The aminomethylene oxazolones tend to be more stable in acid solution than the alkoxy- or hydroxy-methylene oxazolones; in alkali they behave like the alkoxy-methylene compounds, i.e. they are hydrolyzed to the sodium salt of the hydroxymethylene derivatives. Thus sodium penicillenate is hydrolyzed by alkali to the sodium salt of 2-benzyl-4-hydroxymethyleneoxazolone.

Let us now return to evidence based on the degradation studies. The Merck

workers found that when penicillin methylester in ether solution is treated with $HgCl_2$ and the resulting precipitate is decomposed with H_2S , an amorphous substance is obtained which exhibits the characteristic absorption spectrum of penicillenic acid with two maxima at 3,150 Å and 2,700 Å.

This finding attracted a good deal of interest. Ever since the thiazolidine oxazolone structure for penicillin was proposed, continuous attempts were made, particularly by the Merck group, to isolate the 2-benzyl-4-hydroxy-methyleneoxazolone which formed one component of this structure. The most obvious way to obtain this oxazolone, which was known to be quite stable in alkali, was to try to split sodium penicillin by the action of $HgCl_2$. All normal N-non-acylated thiazolidines are instantly decomposed by $HgCl_2$ into the mercury complex of the thiolamino acid and the carbonyl component. However, when the effect of $HgCl_2$ on sodium penicillin was tried, it was found that - unlike the normal N-non-acylated thiazolidine - it did not react instantaneously, but only very slowly after an interval of many hours; and examination of the degradation products showed that it had been split into penicillamine and penaldic acid, but no trace of 2-benzyl-4-hydroxymethyleneoxazolone, easily detectable by its characteristic u.v. absorption spectrum, was ever observed. The behaviour of sodium penicillin towards $HgCl_2$ was in fact additional evidence against the thiazolidine-oxazolone structure and in favour of the β -lactam structure. Now, if the product obtained after reaction of $HgCl_2$ on methyl penicillin was really penicillenic acid, then it would have been definitely proved that 2-benzyl-4-hydroxymethyleneoxazolone can be obtained from penicillin by a mild degradation process and this finding would have to be taken into account in the considerations of the structure of the penicillin molecule. The $HgCl_2$ degradation product of methyl penicillin was therefore examined very carefully, and the result of the examination left no doubt that it was composed predominantly of penicillenic acid from *d*-penicillamine and 4-hydroxymethylene-oxazolone; it was therefore an easily available compound whose properties could be studied without difficulty. Penicillenic acid has two characteristic reactions:

- (1) On addition of benzylamine to penicillenic acid the characteristic u.v. absorption disappears and the α -benzylamide of penicilloic acid is formed. The $HgCl_2$ degradation product of methyl penicillin behaves in the same manner; after addition of benzylamine, the α -benzylamide of penicilloic acid was isolated in the crystalline state and was shown to be identical with the synthetic material.

(2) Synthetic penicillenic acid is hydrolyzed by alkali to penicillamine and 2-benzyl-4-hydroxymethyleneoxazolone which can be isolated as the crystalline Na salt and in form of crystalline derivatives. The $HgCl_2$ -degradation product of methylpenicillin behaves in an identical manner, and the Merck workers succeeded in isolating from its alkaline hydrolysates the sodium salt of 2-benzyl-4-hydroxymethyleneoxazolone in the crystalline state and characterized it by various crystalline derivatives which were identified with synthetic specimens.

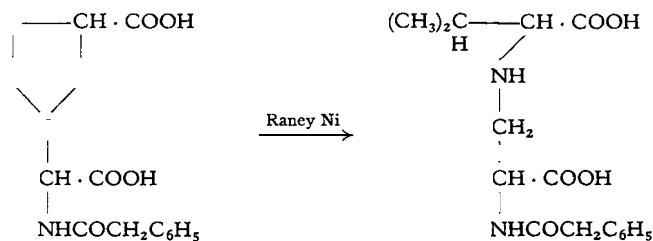
If we now consider the implications of the isolation of 2-benzyl-4-hydroxymethyleneoxazolone as a degradation product of penicillin for arguments concerning the structural formula of penicillin, it must be admitted that, in the absence of all other evidence, the isolation of one component of the postulated thiazolidine-oxazolone structure would naturally be considered as strong evidence in favour of this structure; in fact it is the strongest evidence that could be obtained from straightforward chemical degradation reactions. However, at the time when the oxazolone was isolated by the Merck workers, a great deal of very strong evidence against the thiazolidine-oxazolone structure had already been accumulated and this evidence could not simply be disregarded in front of the new finding. Consequently, the less simple explanation for the formation of the oxazolone during the degradation of penicillin had to be taken into consideration, namely that it was formed as the result of a novel type of intramolecular rearrangement of the four-membered ring of the β -lactam structure present in the original penicillin molecule, induced by the reaction of $HgCl_2$ with methyl penicillin. A distinct aversion to this assumption was noticeable among many chemists because no analogous reaction was known in the literature and no plausible reaction mechanisms for the rearrangement could be suggested.

While the discussion on the significance of the isolation of the oxazolone for the structure of the penicillins was still in full swing, the Merck group isolated several new crystalline degradation products of penicillin II, and the elucidation of their structure neutralized completely all the arguments in favour of the oxazolone-thiazolidine structure that could be advanced by the defenders of this structure from the isolation of the oxazolone. They succeeded, in fact, in isolating in good yield and by a very mild degradation procedure, a product which was shown to possess the β -lactam structure.

It will be remembered that Mozingo, of the Merck Institute, had developed a new method for desulphurization, by hydrogenolysis with Raney nickel, which led to an important advance in the elucidation in the structure

of biotin. This method consists in heating the sulphur-containing compound for a short time with a suspension of finely divided Raney nickel through which hydrogen had been passed. The sulphur is thereby removed as nickel sulphide and is replaced by two atoms of hydrogen in a very smooth reaction. This method has a wide applicability and has proved very useful in the studies aimed at elucidating the chemical structure of the penicillins.

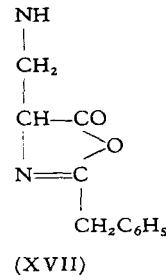
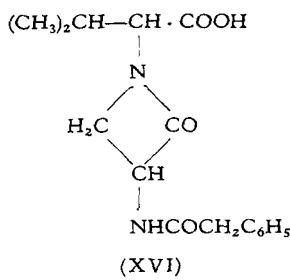
When sodium penicillin II is treated with Raney nickel at 90°C for 1 min, the sulphur is eliminated and two crystalline compounds $C_{16}H_{20}O_4N_2$ and $C_{16}H_{22}O_4N_2$ are obtained, the first in very good yield, the latter in smaller amounts. The first compound has the elementary composition of penicillin II except that the sulphur is removed and replaced by two hydrogen atoms, and it has accordingly been termed desthiopenicillin. The acid-base properties of desthiopenicillin are the same as those of penicillin, i.e. it is a mono-basic acid with no detectable basic group. Chemically it is much more stable than penicillin; it does not react with acid, alkali, primary amines, or alcohols at room temperature. However, on heating for a short time in acid or alkali, desthiopenicilloic acid is obtained, a compound which was found to be identical with desthiopenicilloic acid prepared from natural penicilloic acid by Raney treatment.



Electrometric titration of desthiopenicilloic acid shows that it possesses two acid groups and one basic group, pK 8.2.

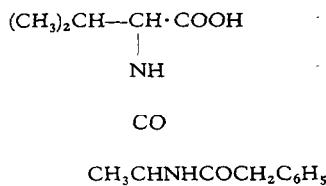
When desthiopenicillin is heated with benzylamine for three hours in refluxing dioxan, the benzylamide of desthiopenicilloic acid is obtained. This is identical with the compound obtained by hydrogenolysis of the penicilloic acid benzylamide derived from natural penicillin.

Only one formula (XVI) could be constructed for desthiopenicillin II that was in agreement with its properties and chemical reactions. This formula contained the four-membered ring postulated in the β -lactam structure for penicillin. An alternative formula (XVII) containing the oxazolone ring could be disregarded, firstly because of the stability of desthiopenicillin and



secondly because of the absence of a basic group which should certainly have been present if the oxazolone structure containing the NH group was correct.

Thus the rather curious situation had arisen that the degradation work had furnished apparent support both for the thiazohdine-oxazolone and the β -lactam structure for penicillin; constituents of both formulae, an oxazolone and a substance containing the four-membered β -lactam ring had been isolated from penicillin by very mild degradation procedures. It then remained to decide which of the two ring systems was originally present in penicillin, and which the result of a rearrangement process. Which of the two rearrangements could be considered the more plausible was largely a matter of personal opinion, and as might be expected the views on this question differed widely among the various investigators. The elucidation of the structure of the second degradation product which had been obtained by treatment of penicillin with Raney nickel did not add anything decisive. It appeared that this substance, $\text{C}_{16}\text{H}_{22}\text{O}_4\text{N}_2\text{S}$ (m-p. 206-207°), was phenyl acetyl C (+) alanyl d (-) valine;

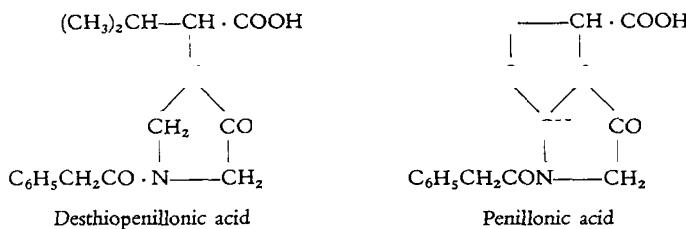


this was proved by synthesis (from the azide of phenylacetyl l (+) alanine and d-valine). The importance of the isolation of phenylacetyl l (+) alanyl d (-) valine in respect to the structure of penicillin lies in the fact that it has settled the optical configuration of another of the three asymmetric carbon atoms in the penicillin molecule. The penicillamine radical had, as was point-

ed out before, the unnatural *d*-configuration; the alanine radical was now shown to possess the natural *l*-configuration. The simplest explanation for the appearance of this substance was that it derived by hydrogenolysis from the β -lactam which contained the two N-CO linkages preformed. It could, however, also have arisen from the oxazolone by an internal acylation followed by a rearrangement.

To make the situation somewhat more confused a new crystalline isomerization product of penicillin was found which differed from the two other isomers penillic and penicillenic acids. This new product, termed penillonic acid, is formed when methyl penicillin is heated in toluene in the presence of a small amount of I₂ (Merck group). As in the case of all the other crystalline degradation products, the structure of the new isomer penillonic acid was eagerly studied in the hope of finding a new line of approach to the problem of the structure of penicillin. With the formulae for penicillin, penillic acid and penicillenic acid already disposed of, the possibilities for new structural arrangements were becoming rather limited and it became quite difficult to think of yet another structural isomer of penicillin.

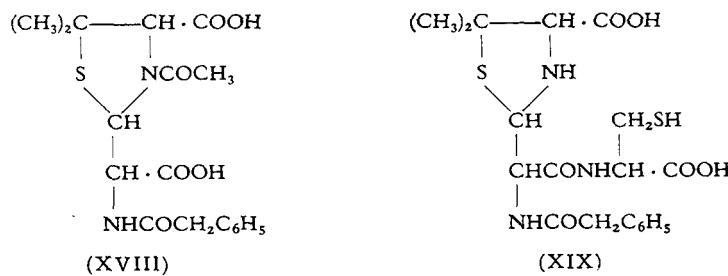
It was found that methyl penillonate was also obtained from methyl penicillenate, both natural and synthetic, by heating in toluene in the presence of a small amount of I₂; but in addition it was formed by simple sublimation *in vacuo* of methyl penicillin, whereas methyl penicillenate gave no penillonic acid under these conditions. Penillonic acid is therefore not merely a rearrangement product of methyl penicillenate, which would be of secondary interest to structural considerations concerning penicillin, but is also formed directly from penicillin, as the result of yet another rearrangement, in addition to the penillic and penicillenic acid rearrangements. The structure of penillonic acid, a monobasic acid with no basic centre, was finally elucidated by degradation of desthiopenillonic acid, C₁₆H₂₀O₄N, obtained by treatment of methyl penillonate with Raney nickel and subsequent saponification. This compound is isomeric with desthiopenicillin, but quite different in its physical and chemical properties. It does not react with benzylamine, mercuric chloride or methanol, even under drastic conditions. It does not react with acid or alkali at room temperature or when heated to 110° for a short time. Prolonged hydrolysis with NaOH or conc. HCl at 100° yields phenaceturic acid, valine and one molecule of formaldehyde. These findings were best reconciled with the following structures for desthiopenillonic II acid and penillonic-II acid:



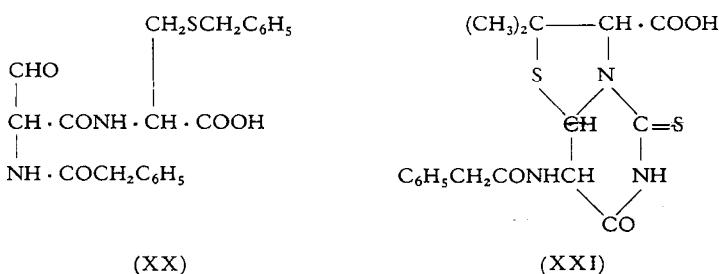
The elucidation of the structure of penillonic acid, like that of the other degradation products of penicillin, was thus of little use for obtaining unequivocal evidence about the structure of penicillin and was disappointing in this respect. All it showed was the occurrence of yet another extraordinary rearrangement involving a reaction mechanism which was difficult to explain on the basis of any formula. Ring expansion from the four-membered to a five-membered ring would have to occur on the basis of the β -lactam structure, a rearrangement of the oxazolone involving a migration of nitrogen on the basis of the oxazolone structure. The latter rearrangement does in fact occur, as penillonic acid is formed from synthetic penicillenic acid which has a known structure containing the oxazolone ring; but no satisfactory explanation for the reaction mechanism of this rearrangement has yet been put forward.

A great deal of most interesting degradation work on penicillin has been carried out in addition to that mentioned above. Time unfortunately does not permit me to give anything like an adequate account of this work, but I want to mention the broad results. These were quite as ambiguous as those obtained from the other degradation studies, and did not allow of a definite decision in favour of one or the other of the two formulae under discussion.

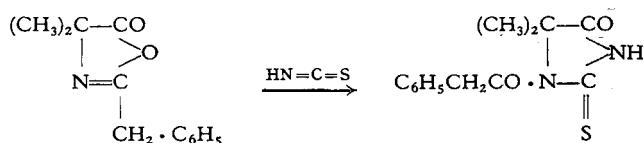
Thus, it was found, that free penicillin-II acid is easily inactivated by acetic acid in an organic solvent (Squibb group); N-acetyl-II-penicilloic acid (XVIII) is formed; the constitution of this compound was proved by synthesis. Sodium penicillin II is inactivated by cysteine at pH 7 (Squibb group).



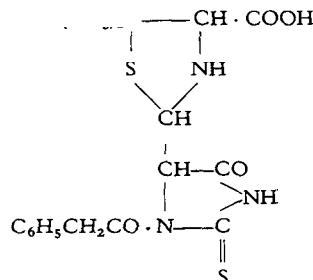
The compound formed is a peptide of penicilloic acid and cysteine, with the SH group free. The constitution of this compound (XIX) was proved by benzylation with benzylchloride and splitting with $HgCl_2$, which yielded the aldehyde (XX). The two reactions just mentioned could be explained on the basis of either structure (IX) or (X). With $HN=C=S$, penicillin-II methyl ester reacts to give in good yield a crystalline product on which a considerable amount of very ingenious degradation work was carried out by the Cornell and Squibb groups of workers. This cannot be reported in any detail but has led to the elucidation of the structure of this product (XXI). This



structure could be derived easily from structure (IX), but was not in good agreement with structure (X). It was known from the literature that N-acylated amino acids on heating with thiocyanic acid and acetic anhydride gave I-acyl-2-thiohydantoins and it was assumed that this reaction proceeds via the azlactones. A great deal of work on model oxazolones has shown that I-acyl-2-thiohydantoins are indeed easily formed when they are treated with thiocyanic acid at room temperature. Thus 2-benzyl-4,4-dimethyloxazolone gives I-phenylacetyl-2-thio-5,5-dimethyl-thiohydantoin.



If penicillin had an oxazolone ring as one of the constituents of its molecule, a thiohydantoin of the structure (XXII) would be expected to be formed on treatment of penicillin-II methyl ester with thiocyanic acid. The actual reaction product has a different constitution, as pointed out above, containing a thiouracil nucleus. Thus, the structure of the thiocyanic degradation product of penicillin is really against structure (X), though an intramolecular re-



(XXII)

arrangement from an originally present oxazolone to a β -lactam under the conditions of the reaction cannot, of course, be excluded.

I have now come to the end of my account of the degradation studies on penicillin. These certainly do not lack in variety and surprises and have led to the discovery of several entirely novel rearrangements.

Summarizing the evidence for the structure of penicillin obtained from the degradation work it can be stated that no absolutely unequivocal conclusion could be derived from it although the balance of the work was more in favour of the plactam structure than of the thiazolidine-oxazolone structure. In particular the formation of the sulphone and of the reaction product with thiocyanic acid was extremely difficult to explain except on the basis of the β -lactam formula.

The final solution of the problem of the structure of penicillin came from crystallographic X-ray studies. This work, in which Mrs. D. Crowfoot and her colleague Mrs. Barbara Rogers-Low have played a predominant rôle, has led to the definite exclusion of the thiazolidine-oxazolone structure and to the conclusive proof of the β -lactam structure. Through a series of Fourier analyses of electron diffraction densities, obtained by X-ray pictures of single crystals of the rubidium and potassium salts of penicillin II, Mrs. Crowfoot and Mrs. Rogers-Low succeeded in measuring all bond distances between the atoms in the penicillin molecule with an accuracy of 0.2 Å and in thus mapping out clearly the whole penicillin molecule. The alkali metal and the sulphur atom served as landmarks in the Fourier analyses. The measurements of the atomic distances show clearly and unequivocally that there exists a normal bond between the thiazolidine nitrogen and the carbonyl group of the labile carbonyl group, but no bond exists between this carbonyl group and the oxygen atom of the peptide side-chain. The four-membered β -lactam ring is clearly visible. These calculations were completely con-

firmed by an independent group of X-ray workers, Dr. C. W. Bunn and his colleagues of Imperial Chemical Industries Ltd. Thus the structure of penicillin was definitely proved to be the β -lactam structure (IX). The work of Mrs. Crowfoot and Mrs. Rogers-Low is a considerable achievement; for the first time the structure of a whole molecule has been calculated from X-ray data, and it is the more remarkable that this should have been possible in the case of a substance having the complexity of the penicillin molecule. The ever-increasing importance of crystallographic X-ray work for the elucidation of chemical structures in collaboration with the organic chemist has been demonstrated in an impressive manner by these investigations.

The elucidation of the chemical structure of the penicillins has been a most interesting and fascinating task in every respect from the very beginning. It became apparent very soon that the chemical behaviour of penicillin was in no way less interesting and original than its biological properties. Penicillin is a simple dipeptide, composed of two simple amino acids: β -thiol-valine and an acylated serine in which the alcohol group has been oxidized to the aldehyde group. Through the incorporation of a peptide linkage in a peculiar ring system so far not observed in any other natural product, this peptide linkage has acquired a high reactivity, and it is the fusion of thiazolidine and β -lactam rings which confers to the penicillin molecule its unique biological and biochemical properties. This is a fact which deserves attention beyond the limited field of penicillin chemistry. The question of the nature of the linkages by which naturally occurring peptides acquire their specific and often very pronounced biological properties has been, and is, one of the major problems in biochemistry. Now it has been shown in the case of the penicillin molecule that a simple dipeptide can acquire most characteristic and specific chemical and biological properties through a novel, but very simple type of linkage of the two amino acids. Naturally, one is tempted to ask immediately whether the occurrence of the β -lactam type of peptide linkage is limited to the penicillin molecule, and whether it does not occur also in other natural products, such as proteins. Perhaps the β -lactam structure is an important feature of many natural products and has escaped discovery up to the present time only because it does not manifest itself by some characteristic property serving as an indicator for its existence, as, in the case of penicillin, the striking bacterial action of this compound. This applies also to the various rearrangements which the penicillin molecule has been shown to undergo readily. These rearrangements, in particular the formation of

imidazole derivatives and oxazolones from amides under very mild conditions, are all of great general biochemical interest. The search for suitable reaction mechanisms for these entirely novel rearrangements will provide an attractive field of research for the theoretical organic chemist.

To end this lecture, just a few words may be said about the present state of the work on the synthesis of penicillin. Despite the apparent simplicity of the penicillin molecule and despite a tremendous effort on the part of many competent chemists, no workable method of synthesis has as yet been evolved. All feasible routes that could possibly lead to such synthesis have been explored, but have not given positive results. In attempts to synthesize the oxazolone-thiazolidine structures, traces of biologically active material have been obtained both by the Merck group and the Oxford workers through condensation of 2-benzyl-4-methoxymethyleneoxazolone, and similar compounds with *d*-penicillamine. *l*-Penicillamine, or *d*- and *l*-cysteine led to no activity. There can be no doubt that the active material synthesized in this manner is in fact penicillin. It acts against the same bacteria as does natural penicillin, and is inactivated by the same specific reagents that inactivate natural penicillin: acid and alkali, methanol and the specific enzyme penicillinase. Furthermore, penicillamine containing radioactive sulphur has been used for the condensation with 2-benzyl-4-methoxymethyleneoxazolone (Cornell group); when a large amount of natural crystalline penicillin II was added to the active product thus obtained and the mixture crystallized, it was found that on recrystallization, the radioactivity always followed the crystalline penicillin fraction, even after 14 recrystallizations of the sodium salt and a further 14 recrystallizations of the acid inactivation product penillic acid. This showed that the solubility properties of the synthetic active material were extremely similar to those of natural penicillin II, so that the identity of the synthetic with the natural material was made even more probable. All attempts to improve the very small yield of synthetic penicillin (about 0.1%) have failed, and it appears improbable that a synthetic process will be evolved that could compete successfully with the cheap biological production of penicillin.

ALBERT EINSTEIN

Fundamental ideas and problems of the theory of relativity

*Lecture delivered to the Nordic Assembly of Naturalists at Gothenburg**

July 11, 1923

If we consider that part of the theory of relativity which may nowadays in a sense be regarded as bona fide scientific knowledge, we note two aspects which have a major bearing on this theory. The whole development of the theory turns on the question of whether there are physically preferred states of motion in Nature (physical relativity problem). Also, concepts and distinctions are only admissible to the extent that observable facts can be assigned to them without ambiguity (stipulation that concepts and distinctions should have meaning). This postulate, pertaining to epistemology, proves to be of fundamental importance.

These two aspects become clear when applied to a special case, e.g. to classical mechanics. Firstly we see that at any point filled with matter there exists a preferred state of motion, namely that of the substance at the point considered. Our problem starts however with the question whether physically preferred states of motion exist in reference to *extensive* regions. From the viewpoint of classical mechanics the answer is in the affirmative; the physically preferred states of motion from the viewpoint of mechanics are those of the inertial frames.

This assertion, in common with the basis of the whole of mechanics as it generally used to be described before the relativity theory, far from meets the above "stipulation of meaning". Motion can only be conceived as the relative motion of bodies. In mechanics, motion relative to the system of coordinates is implied when merely motion is referred to. Nevertheless this interpretation does not comply with the "stipulation of meaning" if the coordinate system is considered as something purely imaginary. If we turn our attention to experimental physics we see that there the coordinate system is invariably represented by a "practically rigid" body. Furthermore it is assumed that such rigid bodies can be positioned in rest relative to one another

* The Lecture was not delivered on the occasion of the Nobel Prize award, and did not, therefore, concern the discovery of the photoelectric effect.

in common with the bodies of Euclidian geometry. Insofar as we may think of the rigid measuring body as existing as an object which can be experienced, the "system of coordinates" concept as well as the concept of the motion of matter relative thereto can be accepted in the sense of the "stipulation of meaning". At the same time Euclidian geometry, by this conception, has been adapted to the requirements of the physics of the "stipulation of meaning". The question whether Euclidian geometry is valid becomes physically significant; its validity is assumed in classical physics and also later in the special theory of relativity.

In classical mechanics the inertial frame and time are best defined together by a suitable formulation of the law of inertia: It is possible to fix the time and assign a state of motion to the system of coordinates (inertial frame) such that, with reference to the latter, force-free material points undergo no acceleration; furthermore it is assumed that this time can be measured without disagreement by identical clocks (systems which run down periodically) in any arbitrary state of motion. There are then an infinite number of inertial frames which are in uniform translational motion relative to each other, and hence there is also an infinite number of mutually equivalent, physically preferred states of motion. Time is absolute, i.e. independent of the choice of the particular inertial frame; it is defined by more characteristics than logically necessary, although - as implied by mechanics - this should not lead to contradictions with experience. Note in passing that the logical weakness of this exposition from the point of view of the stipulation of meaning is the lack of an experimental criterion for whether a material point is force-free or not; therefore the concept of the inertial frame remains rather problematical. This deficiency leads to the general theory of relativity. We shall not consider it for the moment.

The concept of the rigid body (and that of the clock) has a key bearing on the foregoing consideration of the fundamentals of mechanics, a bearing which there is some justification for challenging. The rigid body is only approximately achieved in Nature, not even with desired approximation; this concept does not therefore strictly satisfy the "stipulation of meaning". It is also logically unjustifiable to base all physical consideration on the rigid or solid body and then finally reconstruct that body atomically by means of elementary physical laws which in turn have been determined by means of the rigid measuring body. I am mentioning these deficiencies of method because in the same sense they are also a feature of the relativity theory in the schematic exposition which I am advocating here. Certainly it would be

logically more correct to begin with the whole of the laws and to apply the "stipulation of meaning" to this whole first, i.e. to put the unambiguous relation to the world of experience last instead of already fulfilling it in an imperfect form for an artificially isolated part, namely the space-time metric. We are not, however, sufficiently advanced in our knowledge of Nature's elementary laws to adopt this more perfect method without going out of our depth. At the close of our considerations we shall see that in the most recent studies there is an attempt, based on ideas by Levi-Civita, Weyl, and Eddington, to implement that logically purer method.

It more clearly follows from the above what is implied by "preferred states of motion". They are preferred as regards the laws of Nature. States of motion are preferred when, relative to the formulation of the laws of Nature, coordinate systems within them are distinguished in that with respect to them those laws assume a form preferred by simplicity. According to classical mechanics the states of motion of the inertial frames in this sense are physically preferred. Classical mechanics permits a distinction to be made between (absolutely) unaccelerated and accelerated motions; it also claims that velocities have only a relative existence (dependent on the selection of the inertial frame), while accelerations and rotations have an absolute existence (independent of the selection of the inertial frame). This state of affairs can be expressed thus: According to classical mechanics "velocity relativity" exists, but not "acceleration relativity". After these preliminary considerations we can pass to the actual topic of our contemplations, the relativity theory, by characterizing its development so far in terms of principles.

The special theory of relativity is an adaptation of physical principles to Maxwell-Lorentz electrodynamics. From earlier physics it takes the assumption that Euclidian geometry is valid for the laws governing the position of rigid bodies, the inertial frame and the law of inertia. The postulate of equivalence of inertial frames for the formulation of the laws of Nature is assumed to be valid for the whole of physics (special relativity principle). From Maxwell-Lorentz electrodynamics it takes the postulate of invariance of the velocity of light in a vacuum (light principle).

To harmonize the relativity principle with the light principle, the assumption that an absolute time (agreeing for all inertial frames) exists, had to be abandoned. Thus the hypothesis is abandoned that arbitrarily moved and suitably set identical clocks function in such a way that the times shown by two of them, which meet, agree. A specific time is assigned to each inertial frame; the state of motion and the time of the inertial frame are defined, in

accordance with the stipulation of meaning, by the requirement that the light principle should apply to it. The existence of the inertial frame thus defined and the validity of the law of inertia with respect to it are assumed. The time for each inertial frame is measured by identical clocks that are stationary relative to the frame.

The laws of transformation for space coordinates and time for the transition from one inertial frame to another, the Lorentz transformations as they are termed, are unequivocally established by these definitions and the hypotheses concealed in the assumption that they are free from contradiction. Their immediate physical significance lies in the effect of the motion relative to the used inertial frame on the form of rigid bodies (Lorentz contraction) and on the rate of the clocks. According to the special relativity principle the laws of Nature must be covariant relative to Lorentz transformations; the theory thus provides a criterion for general laws of Nature. It leads in particular to a modification of the Newtonian point motion law in which the velocity of light in a vacuum is considered the limiting velocity, and it also leads to the realization that energy and inertial mass are of like nature.

The special relativity theory resulted in appreciable advances. It reconciled mechanics and electrodynamics. It reduced the number of logically independent hypotheses regarding the latter. It enforced the need for a clarification of the fundamental concepts in epistemological terms. It united the momentum and energy principle, and demonstrated the like nature of mass and energy. Yet it was not entirely satisfactory - quite apart from the quantum problems, which all theory so far has been incapable of really solving. In common with classical mechanics the special relativity theory favours certain states of motion - namely those of the inertial frames - to all other states of motion. This was actually more difficult to tolerate than the preference for a single state of motion as in the case of the theory of light with a stationary ether, for this imagined a real reason for the preference, i.e. the light ether. A theory which from the outset prefers no state of motion should appear more satisfactory. Moreover the previously mentioned vagueness in the definition of the inertial frame or in the formulation of the law of inertia raises doubts which obtain their decisive importance, owing to the empirical principle for the equality of the inertial and heavy mass, in the light of the following consideration.

Let K be an inertial frame without a gravitational field, K' a system of co-ordinates accelerated uniformly relative to K . The behaviour of material points relative to K' is the same as if K' were an inertial frame in respect

of which a homogeneous gravitational field exists. On the basis of the empirically known properties of the gravitational field, the definition of the inertial frame thus proves to be weak. The conclusion is obvious that any arbitrarily moved frame of reference is equivalent to any other for the formulation of the laws of Nature, that there are thus no physically preferred states of motion at all in respect of regions of finite extension (general relativity principle).

The implementation of this concept necessitates an even more profound modification of the geometric-kinematical principles than the special relativity theory. The Lorentz contraction, which is derived from the latter, leads to the conclusion that with regard to a system K' arbitrarily moved relative to a (gravity field free) inertial frame K , the laws of Euclidian geometry governing the position of rigid (at rest relative to K') bodies do not apply. Consequently the Cartesian system of coordinates also loses its significance in terms of the stipulation of meaning. Analogous reasoning applies to time; with reference to K' the time can no longer meaningfully be defined by the indication on identical clocks at rest relative to K' , nor by the law governing the propagation of light. Generalizing, we arrive at the conclusion that gravitational field and metric are only different manifestations of the same physical field.

We arrive at the formal description of this field by the following consideration. For each infinitesimal point-environment in an arbitrary gravitational field a local frame of coordinates can be given for such a state of motion that relative to this local frame no gravitational field exists (local inertial frame). In terms of this inertial frame we may regard the results of the special relativity theory as correct to a first approximation for this infinitesimally small region. There are an infinite number of such local inertial frames at any space-time point; they are associated by Lorentz transformations. These latter are characterised in that they leave invariant the "distance" ds of two infinitely adjacent point events - defined by the equation:

$$ds^2 = c^2 dt^2 - dx^2 - dy^2 - dz^2$$

which distance can be measured by means of scales and clocks. For, x, y, z, t represent coordinates and time measured with reference to a local inertial frame.

To describe space-time regions of finite extent arbitrary point coordinates in four dimensions are required which serve no other purpose than to pro-

vide an unambiguous designation of the space-time points by four numbers each, x_1, x_2, x_3 and x_4 , which takes account of the continuity of this four-dimensional manifold (Gaussian coordinates). The mathematical expression of the general relativity principle is then, that the systems of equations expressing the general laws of Nature are equal for all such systems of coordinates.

Since the coordinate differentials of the local inertial frame are expressed linearly by the differentials dx_v of a Gaussian system of coordinates, when the latter is used, for the distance ds of two events an expression of the form

$$ds^2 = \Sigma g_{\mu\nu} dx_\mu dx_\nu \quad (g_{\mu\nu} = g_{\nu\mu})$$

is obtained. The $g_{\mu\nu}$ which are continuous functions of x_ν , determine the metric in the four-dimensional manifold where ds is defined as an (absolute) parameter measurable by means of rigid scales and clocks. These same parameters $g_{\mu\nu}$ however also describe with reference to the Gaussian system of coordinates the gravitational field which we have previously found to be identical with the physical cause of the metric. The case as to the validity of the special relativity theory for finite regions is characterised in that when the system of coordinates is suitably chosen, the values of $g_{\mu\nu}$ for finite regions are independent of x_ν .

In accordance with the general theory of relativity the law of point motion in the pure gravitational field is expressed by the equation for the geodetic line. Actually the geodetic line is the simplest mathematically which in the special case of constant $g_{\mu\nu}$ becomes rectilinear. Here therefore we are confronted with the transfer of Galileo's law of inertia to the general theory of relativity.

In mathematical terms the search for the field equations amounts to ascertaining the simplest generally covariant differential equations to which the gravitational potentials $g_{\mu\nu}$ can be subjected. By definition these equations should not contain higher derivatives of $g_{\mu\nu}$ with respect to x_ν than the second, and these only linearly, which condition reveals these equations to be a logical transfer of the Poisson field equation of the Newtonian theory of gravity to the general theory of relativity.

The considerations mentioned led to the theory of gravity which yields the Newtonian theory as a first approximation and furthermore it yields the motion of the perihelion of Mercury, the deflection of light by the sun, and the red shift of spectral lines in agreement with experience.*

* As regards the red shift, the agreement with experience is not yet completely assured, however.

To complete the basis of the general theory of relativity, the electromagnetic field must still be introduced into it which, according to our present conviction, is also the material from which we must build up the elementary structures of matter. The Maxwellian field equations can readily be adopted into the general theory of relativity. This is a completely unambiguous adoption provided it is assumed that the equations contain no differential quotients of g_{ν} higher than the first, and that in the customary Maxwellian form they apply in the local inertial frame. It is also easily possible to supplement the gravitational field equations by electromagnetic terms in a manner specified by the Maxwellian equations so that they contain the gravitational effect of the electromagnetic field.

These field equations have not provided a theory of matter. To incorporate the field generating effect of ponderable masses in the theory, matter had therefore (as in classical physics) to be introduced into the theory in an approximate, phenomenological representation.

And that exhausts the direct consequences of the relativity principle. I shall turn to those problems which are related to the development which I have traced. Already Newton recognized that the law of inertia is unsatisfactory in a context so far unmentioned in this exposition, namely that it gives no real cause for the special physical position of the states of motion of the inertial frames relative to all other states of motion. It makes the observable material bodies responsible for the gravitational behaviour of a material point, yet indicates no material cause for the inertial behaviour of the material point but devises the cause for it (absolute space or inertial ether). This is not logically inadmissible although it is unsatisfactory. For this reason E. Mach demanded a modification of the law of inertia in the sense that the inertia should be interpreted as an acceleration resistance of the bodies against *one another* and not against "space". This interpretation governs the expectation that accelerated bodies have concordant accelerating action in the same sense on other bodies (acceleration induction).

This interpretation is even more plausible according to general relativity which eliminates the distinction between inertial and gravitational effects. It amounts to stipulating that, apart from the arbitrariness governed by the free choice of coordinates, the g_{ν} -field shall be completely determined by the matter. Mach's stipulation is favoured in general relativity by the circumstance that acceleration induction in accordance with the gravitational field equations really exists, although of such slight intensity that direct detection by mechanical experiments is out of the question.

Mach's stipulation can be accounted for in the general theory of relativity by regarding the world in spatial terms as finite and self-contained. This hypothesis also makes it possible to assume the mean density of matter in the world as finite, whereas in a spatially infinite (quasi-Euclidian) world it should disappear. It cannot, however, be concealed that to satisfy Mach's postulate in the manner referred to a term with no experimental basis whatsoever must be introduced into the field equations, which term logically is in no way determined by the other terms in the equations. For this reason this solution of the "cosmological problem" will not be completely satisfactory for the time being.

A second problem which at present is the subject of lively interest is the identity between the gravitational field and the electromagnetic field. The mind striving after unification of the theory cannot be satisfied that two fields should exist which, by their nature, are quite independent. A mathematically unified field theory is sought in which the gravitational field and the electromagnetic field are interpreted only as different components or manifestations of the same uniform field, the field equations where possible no longer consisting of logically mutually independent summands.

The gravitational theory, considered in terms of mathematical formalism, i.e. Riemannian geometry, should be generalized so that it includes the laws of the electromagnetic field. Unfortunately we are unable here to base ourselves on empirical facts as when deriving the gravitational theory (equality of the inertial and heavy mass), but we are restricted to the criterion of mathematical simplicity which is not free from arbitrariness. The attempt which at present appears the most successful is that, based on the ideas of Levi-Civita, Weyl and Eddington, to replace Riemannian metric geometry by the more general theory of affine correlation.

The characteristic assumption of Riemannian geometry is the attribution to two infinitely adjacent points of a "distance" ds , the square of which is a homogeneous second order function of the coordinate differentials. It follows from this that (apart from certain conditions of reality) Euclidian geometry is valid in any infinitely small region. Hence to every line element (or vector) at a point P is assigned a parallel and equal line element (or vector) through any given infinitesimally adjacent point P' (affine correlation). Riemannian metric determines an affine correlation. Conversely, however, when an affine correlation (law of infinitesimal parallel displacement) is mathematically given, generally no Riemannian metric determination exists from which it can be derived.

The most important concept of Riemannian geometry, "space curvature", on which the gravitational equations are also based, is based exclusively on the "affine correlation". If one is given in a continuum, without first proceeding from a metric, it constitutes a generalization of Riemannian geometry but which still retains the most important derived parameters. By seeking the simplest differential equations which can be obeyed by an affine correlation there is reason to hope that a generalization of the gravitation equations will be found which includes the laws of the electromagnetic field. This hope has in fact been fulfilled although I do not know whether the formal connection so derived can really be regarded as an enrichment of physics as long as it does not yield any new physical connections. In particular a field theory can, to my mind, only be satisfactory when it permits the elementary electrical bodies to be represented as solutions free from singularities.

Moreover it should not be forgotten that a theory relating to the elementary electrical structures is inseparable from the quantum theory problems. So far also relativity theory has proved ineffectual in relation to this most profound physical problem of the present time. Should the form of the general equations some day, by the solution of the quantum problem, undergo a change however profound, even if there is a complete change in the parameters by means of which we represent the elementary process, the relativity principle will not be relinquished and the laws previously derived therefrom will at least retain their significance as limiting laws.

THE DEVELOPMENT OF THE SPACE-TIME VIEW OF QUANTUM ELECTRODYNAMICS*

by

Richard P. Feynman

California Institute of Technology, Pasadena, California

Nobel Lecture, December 11, 1965.

We have a habit in writing articles published in scientific journals to make the work as finished as possible, to cover all the tracks, to not worry about the blind alleys or to describe how you had the wrong idea first, and so on. So there isn't any place to publish, in a dignified manner, what you actually did in order to get to do the work, although, there has been in these days, some interest in this kind of thing. Since winning the prize is a personal thing, I thought I could be excused in this particular situation, if I were to talk personally about my relationship to quantum electrodynamics, rather than to discuss the subject itself in a refined and finished fashion. Furthermore, since there are three people who have won the prize in physics, if they are all going to be talking about quantum electrodynamics itself, one might become bored with the subject. So, what I would like to tell you about today are the sequence of events, really the sequence of ideas, which occurred, and by which I finally came out the other end with an unsolved problem for which I ultimately received a prize.

I realize that a truly scientific paper would be of greater value, but such a paper I could publish in regular journals. So, I shall use this Nobel Lecture as an opportunity to do something of less value, but which I cannot do elsewhere. I ask your indulgence in another manner. I shall include details of anecdotes which are of no value either scientifically, nor for understanding the development of ideas. They are included only to make the lecture more entertaining.

I worked on this problem about eight years until the final publication in 1947. The beginning of the thing was at the Massachusetts Institute of Technology, when I was an undergraduate student reading about the known physics, learning slowly about all these things that people were worrying about, and realizing ultimately that the fundamental problem of the day was that the quantum theory of electricity and magnetism was not completely satisfactory. This I gathered from books like those of Heitler and Dirac. I was inspired by the remarks in these books; not by the parts in which everything was proved and demonstrated carefully and calculated, because I couldn't understand those very well. At the young age what I could understand were

* This document is a revised version of Feynman's Lecture, with amendments made by Michael D. Godfrey and Michael A. Gottlieb (email: godfrey@isl.stanford.edu and codelieb@caltech.edu). © The Nobel Foundation, 1965.

the remarks about the fact that this doesn't make any sense, and the last sentence of the book of Dirac I can still remember, "It seems that some essentially new physical ideas are here needed." So, I had this as a challenge and an inspiration. I also had a personal feeling, that since they didn't get a satisfactory answer to the problem I wanted to solve, I don't have to pay a lot of attention to what they did do.

I did gather from my readings, however, that two things were the source of the difficulties with the quantum electrodynamical theories. The first was an infinite energy of interaction of the electron with itself. And this difficulty existed even in the classical theory. The other difficulty came from some infinities which had to do with the infinite number of degrees of freedom in the field. As I understood it at the time (as nearly as I can remember) this was simply the difficulty that if you quantized the harmonic oscillators of the field (say in a box) each oscillator has a ground state energy of $\frac{1}{2}\hbar\omega$ and there is an infinite number of modes in a box of ever increasing frequency ω , and therefore there is an infinite energy in the box. I now realize that that wasn't a completely correct statement of the central problem; it can be removed simply by changing the zero from which energy is measured. At any rate, I believed that the difficulty arose somehow from a combination of the electron acting on itself and the infinite number of degrees of freedom of the field.

Well, it seemed to me quite evident that the idea that a particle acts on itself, that the electrical force acts on the same particle that generates it, is not a necessary one—it is sort of a silly one, as a matter of fact. And, so I suggested to myself, that electrons cannot act on themselves, they can only act on other electrons. That means there is no field at all. You see, if all charges contribute to making a single common field, and if that common field acts back on all the charges, then each charge must act back on itself. Well, that was where the mistake was, there was no field. It was just that when you shook one charge, another would shake later. There was a direct interaction between charges, albeit with a delay. The law of force connecting the motion of one charge with another would just involve a delay. Shake this one, that one shakes later. The sun atom shakes; my eye electron shakes eight minutes later, because of a direct interaction across.

Now, this has the attractive feature that it solves both problems at once. First, I can say immediately, I don't let the electron act on itself, I just let this act on that, hence, no self-energy! Secondly, there is not an infinite number of degrees of freedom in the field. There is no field at all; or if you insist on thinking in terms of ideas like that of a field, this field is always completely determined by the action of the particles which produce it. You shake this particle, it shakes that one, but if you want to think in a field way, the field, if it's there, would be entirely determined by the matter which generates it, and therefore, the field does not have any *independent* degrees of freedom and the infinities from the degrees of freedom would then be removed. As a matter of fact, when we look out anywhere and see light, we can always "see" some matter as the source of the light. We don't just see light (except recently some radio reception has been found with no apparent material source).

You see then that my general plan was to first solve the classical problem, to get rid of the infinite self-energies in the classical theory, and to hope that when I made a quantum theory of it, everything would just be fine.

That was the beginning, and the idea seemed so obvious to me and so elegant that I fell deeply in love with it. And, like falling in love with a woman, it is only possible if you do not know much about her, so you cannot see her faults. The faults will become apparent later, but after the love is strong enough to hold you to her. So, I was held to this theory, in spite of all difficulties, by my youthful enthusiasm.

Then I went to graduate school and somewhere along the line I learned what was wrong with the idea that an electron does not act on itself. When you accelerate an electron it radiates energy and you have to do extra work to account for that energy. The extra force against which this work is done is called the force of radiation resistance. The origin of this extra force was identified in those days, following Lorentz, as the action of the electron itself. The first term of this action, of the electron on itself, gave a kind of inertia (not quite relativistically satisfactory). But that inertia-like term was infinite for a point-charge. Yet the next term in the sequence gave an energy loss rate, which for a point-charge agrees exactly with the rate you get by calculating how much energy is radiated. So, the force of radiation resistance, which is absolutely necessary for the conservation of energy, would disappear if I said that a charge could not act on itself.

So, I learned in the interim when I went to graduate school the glaringly obvious fault of my own theory. But, I was still in love with the original theory, and was still thinking that with it lay the solution to the difficulties of quantum electrodynamics. So, I continued to try on and off to save it somehow. I must have some action develop on a given electron when I accelerate it to account for radiation resistance. But, if I let electrons only act on other electrons the only possible source for this action is another electron in the world. So, one day, when I was working for Professor Wheeler and could no longer solve the problem that he had given me, I thought about this again and I calculated the following: Suppose I have two charges—I shake the first charge, which I think of as a source and this makes the second one shake, but the second one shaking produces an effect back on the source. And so, I calculated how much that effect back on the first charge was, hoping it might add up to the force of radiation resistance. It didn't come out right, of course, but I went to Professor Wheeler and told him my ideas. He said,—yes, but the answer you get for the problem with the two charges that you just mentioned will, unfortunately, depend upon the charge and the mass of the second charge and will vary inversely as the square of the distance R , between the charges, while the force of radiation resistance depends on none of these things. I thought, surely, he had computed it himself, but now having become a professor, I know that one can be wise enough to see immediately what some graduate student takes several weeks to develop. He also pointed out something else that bothered me, that if we had a situation with many charges all around the original source at roughly uniform density and if we added the effect of all the surrounding charges the inverse R^2 would be compensated by the R^2 in the volume element and we would get a result proportional to the thickness of the layer, which would go to infinity. That is, one would have an infinite total effect back at the source. And, finally he said to me, and you forgot something else, when you accelerate the first charge, the second acts later, and then the reaction back here at the source would be still later. In other words, the action occurs at the wrong time. I

suddenly realized what a stupid fellow I am, for what I had described and calculated was just ordinary reflected light, not radiation reaction.

But, as I was stupid, so was Professor Wheeler that much more clever. For he then went on to give a lecture as though he had worked this all out before and was completely prepared, but he had not, he worked it out as he went along. First, he said, let us suppose that the return action by the charges in the absorber reaches the source by advanced waves as well as by the ordinary retarded waves of reflected light; so that the law of interaction acts backward in time, as well as forward in time. I was enough of a physicist at that time not to say, "Oh, no, how could that be?" For today all physicists know from studying Einstein and Bohr, that sometimes an idea which looks completely paradoxical at first, if analyzed to completion in all detail and in experimental situations, may, in fact, not be paradoxical. So, it did not bother me any more than it bothered Professor Wheeler to use advance waves for the back reaction—a solution of Maxwell's equations, which previously had not been physically used.

Professor Wheeler used advanced waves to get the reaction back at the right time and then he suggested this: If there were lots of electrons in the absorber, there would be an index of refraction n , so, the retarded waves coming from the source would have their wavelengths slightly modified in going through the absorber. Now, if we shall assume that the advanced waves come back from the absorber without an index—why? I don't know, let's assume they come back without an index—then, there will be a gradual shifting in phase between the return and the original signal so that we would only have to figure that the contributions act as if they come from only a finite thickness, that of the first wave zone. (More specifically, up to that depth where the phase in the medium is shifted appreciably from what it would be in vacuum, a thickness proportional to $\lambda/(n - 1)$.) Now, the less the number of electrons in here, the less each contributes, but the thicker will be the layer that effectively contributes because with less electrons, the index differs less from 1. The higher the charges of these electrons, the more each contributes, but the thinner the effective layer, because the index would be higher. And when we estimated it, (calculated without being careful to keep the correct numerical factor) sure enough, it came out that the action back at the source was completely independent of the properties of the charges that were in the surrounding absorber. Further, it was of just the right character to represent radiation resistance, but we were unable to see if it was just exactly the right size. He sent me home with orders to figure out exactly how much advanced and how much retarded wave we need to get the thing to come out numerically right, and after that, figure out what happens to the advanced effects that you would expect if you put a test charge here close to the source? For if all charges generate advanced, as well as retarded effects, why would that test charge not be affected by the advanced waves from the source?

I found that you get the right answer if you use half-advanced and half-retarded as the field generated by each charge. That is, one is to use the solution of Maxwell's equation which is symmetrical in time and that the reason we got no advanced effects at a point close to the source in spite of the fact that the source was producing an advanced field is this: Suppose the source is surrounded by a spherical absorbing wall

ten light seconds away, and that the test charge is one second to the right of the source. Then the source is as much as eleven seconds away from some parts of the wall and only nine seconds away from other parts. The source acting at time $t = 0$ induces motions in the wall at time $t = +10$. Advanced effects from this can act on the test charge as early as eleven seconds earlier, or at $t = -1$. This is just at the time that the direct advanced waves from the source should reach the test charge, and it turns out the two effects are exactly equal and opposite and cancel out! At the later time $t = +1$ effects on the test charge from the source and from the walls are again equal, but this time are of the same sign and add to convert the half-retarded wave of the source to full retarded strength.

Thus, it became clear that there was the possibility that if we assume all actions are via half-advanced and half-retarded solutions of Maxwell's equations and that all sources are surrounded by material absorbing all the light which is emitted, then we could account for radiation resistance as a direct action of the charges of the absorber acting back by advanced waves on the source.

Many months were devoted to checking all these points. I worked to show that everything is independent of the shape of the container, and so on, that the laws are exactly right, and that the advanced effects really cancel in every case. We always tried to increase the efficiency of our demonstrations, and to see with more and more clarity why it works. I won't bore you by going through the details of this. Because of our using advanced waves we also had many apparent paradoxes, which we gradually reduced one by one, and saw that there was in fact no logical difficulty with the theory. It was perfectly satisfactory.

We also found that we could reformulate this thing in another way, and that is by a principle of least action. Since my original plan was to describe everything directly in terms of particle motions, it was my desire to represent this new theory without saying anything about fields. It turned out that we found a form for an action directly involving the motions of the charges only, which upon variation would give the equations of motion of these charges. The expression for this action A is

$$A = \sum_i m_i \int \left(\dot{X}_\mu^i \dot{X}_\mu^i \right)^{\frac{1}{2}} d\alpha_i + \frac{1}{2} \sum_{i,j} e_i e_j \int \int \delta(I_{ij}^2) \dot{X}_\mu^i(\alpha_i) \dot{X}_\mu^j(\alpha_j) d\alpha_i d\alpha_j \quad (1)$$

where

$$I_{ij}^2 = [X_\mu^i(\alpha_i) - X_\mu^j(\alpha_j)][X_\mu^i(\alpha_i) - X_\mu^j(\alpha_j)]$$

and where $X_\mu^i(\alpha_i)$ is the four-vector position of the i^{th} particle as a function of some parameter α_i , and $\dot{X}_\mu^i(\alpha_i)$ is $dX_\mu^i(\alpha_i)/d\alpha_i$. The first term is the integral of proper time, the ordinary action of relativistic mechanics of free particles of mass m_i . (We sum in the usual way on the repeated index μ .) The second term represents the electrical interaction of the charges. It is summed over each pair of charges (the factor $\frac{1}{2}$ is to count each pair once, the term $i = j$ is omitted to avoid self-action). The interaction is a double integral over a δ -function of the square of the space-time interval I^2 between two points on the paths. Thus, interaction occurs only when this interval vanishes, that is, along light cones.

The fact that the interaction is exactly one-half advanced and one-half retarded meant that we could write such a principle of least action, whereas interaction via retarded waves alone cannot be written in such a way.

So, all of classical electrodynamics was contained in this very simple form. It looked good, and therefore, it was undoubtedly true, at least to the beginner. It automatically gave half-advanced and half-retarded effects and it was without fields. By omitting the term in the sum when $i = j$, I omitted self-interaction and no longer had any infinite self-energy. This then was the hoped-for solution to the problem of ridding classical electrodynamics of the infinities.

It turns out, of course, that you can reinstate fields if you wish to, but you have to keep track of the field produced by each particle separately. This is because to find the right field to act on a given particle, you must exclude the field that it creates itself. A single universal field to which all contribute will not do. This idea had been suggested earlier by Frenkel and so we called these Frenkel fields. This theory which allowed only particles to act on each other was equivalent to Frenkel's fields using half-advanced and half-retarded solutions.

There were several suggestions for interesting modifications of electrodynamics. We discussed lots of them, but I shall report on only one. It was to replace this δ -function in the interaction by another function, say, $f(I_{ij}^2)$, which is not infinitely sharp. Instead of having the action occur only when the interval between the two charges is exactly zero, we would replace the δ -function of I^2 by a narrow peaked thing. Let's say that $f(Z)$ is large only near $Z = 0$ and has width of order a^2 . Interactions will now occur when $T^2 - R^2$ is roughly of order a^2 , where T is the time difference and R is the separation of the charges. This might look like it disagrees with experience, but if a is some small distance, like 10^{-13} cm, it says that the time delay T in action is roughly $\sqrt{R^2 \pm a^2}$ or approximately, if R is much larger than a , $T = R \pm a^2/2R$. This means that the deviation of time T from the ideal theoretical time R of Maxwell, gets smaller and smaller, the further the pieces are apart. Therefore, all theories involved in analyzing generators, motors, etc., in fact, all of the tests of electrodynamics that have been available since Maxwell's time, would be adequately satisfied if a were 10^{-13} cm. If R is of the order of a centimeter this deviation in T is only 10^{-26} seconds. So, it was possible, also, to change the theory in a simple manner and to still agree with all observations of classical electrodynamics. You have no clue of precisely what function to put in for f , but it was an interesting possibility to keep in mind when developing quantum electrodynamics.

It also occurred to us that if we did that (replace δ by f) we could reinstate the term $i = j$ in the sum because this would now represent, in a relativistically invariant fashion, a finite action of a charge on itself. In fact, it was possible to prove that if we did do such a thing, the main effect of the self-action (for not too rapid accelerations) would be to produce a modification of the mass. In fact, there need be no mass m_i term; all the mechanical mass could be electromagnetic self-action. So, if you would like, we could also have another theory with a still simpler expression for the action A . In expression (1) only the second term is kept, the sum extended over all i and j , and some function f replaces δ . Such a simple form could represent all of classical electrodynamics, which aside from gravitation is essentially all of classical physics.

Although it may sound confusing, I am describing several different alternative theories at once. The important thing to note is that at this time we had all these in mind as different possibilities. There were several possible solutions of the difficulty of classical electrodynamics, any one of which might serve as a good starting point to the solution of the difficulties of quantum electrodynamics.

I would also like to emphasize that by this time I was becoming used to a physical point of view different from the more customary point of view. In the customary view, things are discussed as a function of time in very great detail. For example, you have the field at this moment, a differential equation gives you the field at the next moment and so on; a method, which I shall call the Hamiltonian method, the time differential method. We have, instead (in (1) say) a thing that describes the character of the path throughout all of space and time. The behavior of nature is determined by saying her whole space-time path has a certain character. For an action like (1) the equations obtained by variation of $X_\mu^i(\alpha_i)$ are no longer at all easy to get back into Hamiltonian form. If you wish to use as variables only the coordinates of particles, then you can talk about the property of the paths—but the path of one particle at a given time is affected by the path of another at a different time. If you try to describe, therefore, things differentially, telling what the present conditions of the particles are, and how these present conditions will affect the future, you see, it is impossible with particles alone, because something the particle did in the past is going to affect the future.

Therefore, you need a lot of bookkeeping variables to keep track of what the particle did in the past. These are called field variables. You will, also, have to tell what the field is at this present moment, if you are to be able to see later what is going to happen. From the overall space-time view of the least action principle, the field disappears as nothing but bookkeeping variables insisted on by the Hamiltonian method.

As a by-product of this same view, I received a telephone call one day at the graduate college at Princeton from Professor Wheeler, in which he said, “Feynman, I know why all electrons have the same charge and the same mass.” “Why?” “Because, they are all the same electron!” And, then he explained on the telephone, “suppose that the world lines which we were ordinarily considering before in time and space—instead of only going up in time were a tremendous knot, and then, when we cut through the knot, by the plane corresponding to a fixed time, we would see many, many world lines and that would represent many electrons, except for one thing. If in one section this is an ordinary electron world line, in the section in which it reversed itself and is coming back from the future we have the wrong sign to the proper time—to the proper four velocities—and that’s equivalent to changing the sign of the charge, and, therefore, that part of a path would act like a positron.” “But, Professor,” I said, “there aren’t as many positrons as electrons.” “Well, maybe they are hidden in the protons or something,” he said. I did not take the idea that all the electrons were the same one from him as seriously as I took the observation that positrons could simply be represented as electrons going from the future to the past in a back section of their world lines. That, I stole!

To summarize, when I was done with this, as a physicist I had gained two things. One, I knew many different ways of formulating classical electrodynamics, with many

different mathematical forms. I got to know how to express the subject every which way. Second, I had a point of view—the overall space-time point of view—and a disrespect for the Hamiltonian method of describing physics.

I would like to interrupt here to make a remark. The fact that electrodynamics can be written in so many ways—the differential equations of Maxwell, various minimum principles with fields, minimum principles without fields, all different kinds of ways, was something I knew, but I have never understood. It always seems odd to me that the fundamental laws of physics, when discovered, can appear in so many different forms that are not apparently identical at first, but, with a little mathematical fiddling you can show the relationship. An example of that is the Schrödinger equation and the Heisenberg formulation of quantum mechanics. I don't know why this is—it remains a mystery, but it was something I learned from experience. There is always another way to say the same thing that doesn't look at all like the way you said it before. I don't know what the reason for this is. I think it is somehow a representation of the simplicity of nature. A thing like the inverse square law is just right to be represented by the solution of Poisson's equation, which, therefore, is a very different way to say the same thing that doesn't look at all like the way you said it before. I don't know what it means, that nature chooses these curious forms, but maybe that is a way of defining simplicity. Perhaps a thing is simple if you can describe it fully in several different ways without immediately knowing that you are describing the same thing.

I was now convinced that since we had solved the problem of classical electrodynamics (and completely in accordance with my program from M.I.T., only direct interaction between particles, in a way that made fields unnecessary) that everything was definitely going to be all right. I was convinced that all I had to do was make a quantum theory analogous to the classical one and everything would be solved.

So, the problem is only to make a quantum theory which has as its classical analog this expression (1). Now, there is no unique way to make a quantum theory from classical mechanics, although all the textbooks make believe there is. What they would tell you to do, was find the momentum variables and replace them by $(\hbar/i)(\partial/\partial x)$: but I couldn't find a momentum variable, as there wasn't any.

The character of quantum mechanics of the day was to write things in the famous Hamiltonian way—in the form of a differential equation, which described how the wave function changes from instant to instant, and in terms of an operator, H . If the classical physics could be reduced to a Hamiltonian form, everything was all right. Now, least action does not imply a Hamiltonian form if the action is a function of anything more than positions and velocities at the same moment. If the action is of the form of the integral of a function, (usually called the Lagrangian) of the velocities and positions at the same time

$$S = \int L(\dot{x}, x) dt \tag{2}$$

then you can start with the Lagrangian and then create a Hamiltonian and work out the quantum mechanics, more or less uniquely. But this thing (1) involves the key variables, positions, at different times and therefore, it was not obvious what to do to make the quantum-mechanical analog.

I tried—I would struggle in various ways. One of them was this: if I had harmonic oscillators interacting with a delay in time, I could work out what the normal modes were and guess that the quantum theory of the normal modes was the same as for simple oscillators and kind of work my way back in terms of the original variables. I succeeded in doing that, and I hoped then to generalize to other than a harmonic oscillator, but I learned to my regret something, which many people have learned. The harmonic oscillator is too simple; very often you can work out what it should do in quantum theory without getting much of a clue as to how to generalize your results to other systems.

So that didn't help me very much, but when I was struggling with this problem, I went to a beer party in the Nassau Tavern in Princeton. There was a gentleman, newly arrived from Europe (Herbert Jehle) who came and sat next to me. Europeans are much more serious than we are in America because they think that a good place to discuss intellectual matters is a beer party. So, he sat by me and asked, "what are you doing" and so on, and I said, "I'm drinking beer." Then I realized that he wanted to know what work I was doing and I told him I was struggling with this problem, and I simply turned to him and said, "listen, do you know any way of doing quantum mechanics, starting with action—where the action integral comes into the quantum mechanics?" "No," he said, "but Dirac has a paper in which the Lagrangian, at least, comes into quantum mechanics. I will show it to you tomorrow."

Next day we went to the Princeton Library—they have little rooms on the side to discuss things—and he showed me this paper. What Dirac said was the following: There is in quantum mechanics a very important quantity which carries the wave function from one time to another, besides the differential equation but equivalent to it, a kind of a kernel, which we might call $K(x', x)$, which carries the wave function $\psi(x)$ known at time t , to the wave function $\psi(x')$ at time, $t + \varepsilon$. Dirac points out that this function K was *analogous* to the quantity in classical mechanics that you would calculate if you took the exponential of $i\varepsilon$, multiplied by the Lagrangian $L(\dot{x}, x)$ imagining that these two positions x, x' corresponded to t and $t + \varepsilon$. In other words,

$$K(x', x) \text{ is analogous to } e^{i\varepsilon L\left(\frac{x'-x}{\varepsilon}, x\right)/\hbar}.$$

Professor Jehle showed me this, I read it, he explained it to me, and I said, "what does he mean, they are analogous; what does that mean, *analogous*? What is the use of that?" He said, "you Americans! You always want to find a use for everything!" I said, that I thought that Dirac must mean that they were equal. "No," he explained, "he doesn't mean they are equal." "Well," I said, "let's see what happens if we make them equal."

So I simply put them equal, taking the simplest example where the Lagrangian is $\frac{1}{2}M\dot{x}^2 - V(x)$ but soon found I had to put a constant of proportionality A in, suitably adjusted. When I substituted $Ae^{i\varepsilon L/\hbar}$ for K to get

$$\psi(x', t + \varepsilon) = \int A \exp\left[\frac{i\varepsilon}{\hbar} L\left(\frac{x' - x}{\varepsilon}, x\right)\right] \psi(x, t) dx \quad (3)$$

and just calculated things out by Taylor series expansion, out came the Schrödinger equation. So, I turned to Professor Jehle, not really understanding, and said, "well,

you see Professor Dirac meant that they were proportional." Professor Jehle's eyes were bugging out—he had taken out a little notebook and was rapidly copying it down from the blackboard, and said, "no, no, this is an important discovery. You Americans are always trying to find out how something can be used. That's a good way to discover things!" So, I thought I was finding out what Dirac meant, but, as a matter of fact, had made the discovery that what Dirac thought was analogous, was, in fact, equal. I had then, at least, the connection between the Lagrangian and quantum mechanics, but still with wave functions and infinitesimal times.

It must have been a day or so later, when I was lying in bed thinking about these things, that I imagined what would happen if I wanted to calculate the wave function at a finite interval later.

I would put one of these factors $e^{i\varepsilon L}$ in here, and that would give me the wave functions the next moment, $t + \varepsilon$, and then I could substitute that back into (3) to get another factor of $e^{i\varepsilon L}$ and give me the wave function the next moment, $t + 2\varepsilon$, and so on and so on. In that way I found myself thinking of a large number of integrals, one after the other in sequence. In the integrand was the product of the exponentials, which, of course, was the exponential of the sum of terms like εL . Now, L is the Lagrangian and ε is like the time interval dt , so that if you took a sum of such terms, that's exactly like an integral. That's like Riemann's formula for the integral $\int L dt$; you just take the value at each point and add them together. We are to take the limit as $\varepsilon \rightarrow 0$, of course. Therefore, the connection between the wave function of one instant and the wave function of another instant a finite time later could be obtained by an infinite number of integrals, (because ε goes to zero, of course) of exponential (iS/\hbar) , where S is the action expression (2). At last, I had succeeded in representing quantum mechanics directly in terms of the action S .

This led later on to the idea of the amplitude for a path; that for each possible way that the particle can go from one point to another in space-time, there's an amplitude. That amplitude is e to the (i/\hbar) times the action for the path. Amplitudes from various paths superpose by addition. This then is another, a third, way of describing quantum mechanics, which looks quite different than that of Schrödinger or Heisenberg, but which is equivalent to them.

Now immediately after making a few checks on this thing, what I wanted to do, of course, was to substitute the action (1) for the other (2). The first trouble was that I could not get the thing to work with the relativistic case of spin one-half. However, although I could deal with the matter only non-relativistically, I could deal with the light or the photon interactions perfectly well by just putting the interaction terms of (1) into any action, replacing the mass terms by the non-relativistic $(M\dot{x}^2/2)dt$. When the action has a delay, as it now had, and involved more than one time, I had to lose the idea of a wave function. That is, I could no longer describe the program as; given the amplitude for all positions at a certain time, compute the amplitude at another time. However, that didn't cause very much trouble. It just meant developing a new idea. Instead of wave functions we could talk about this: that if a source of a certain kind emits a particle, and a detector is there to receive it, we can give the amplitude that the source will emit and the detector receive. We do this without specifying the

exact instant that the source emits or the exact instant that any detector receives, without trying to specify the state of anything at any particular time in between, but by just finding the amplitude for the complete experiment. And, then we could discuss how that amplitude would change if you had a scattering sample in between, as you rotated and changed angles, and so on, without really having any wave functions.

It was also possible to discover what the old concepts of energy and momentum would mean with this generalized action. And, so I believed that I had a quantum theory of classical electrodynamics—or rather of this new classical electrodynamics described by action (1). I made a number of checks. If I took the Frenkel field point of view, which you remember was more differential, I could convert it directly to quantum mechanics in a more conventional way. The only problem was how to specify in quantum mechanics the classical boundary conditions to use only half-advanced and half-retarded solutions. By some ingenuity in defining what that meant, I found that the quantum mechanics with Frenkel fields, plus a special boundary condition, gave me back this action (1) in the new form of quantum mechanics with a delay. So, various things indicated that there wasn't any doubt I had everything straightened out.

It was also easy to guess how to modify the electrodynamics, if anybody ever wanted to modify it. I just changed the δ to an f , just as I would for the classical case. So, it was very easy, a simple thing. To describe the old retarded theory without explicit mention of fields I would have to write probabilities, not just amplitudes. I would have to square my amplitudes and that would involve double path integrals in which there are two S 's and so forth. Yet, as I worked out many of these things and studied different forms and different boundary conditions, I got a kind of funny feeling that things weren't exactly right. I could not clearly identify the difficulty and in one of the short periods during which I imagined I had laid it to rest, I published a thesis and received my Ph. D.

During the war, I didn't have time to work on these things very extensively, but wandered about on buses and so forth, with little pieces of paper, and struggled to work on it and discovered indeed that there was something wrong, something terribly wrong. I found that if one generalized the action from the nice Lagrangian forms (2) to these forms (1) then the quantities which I defined as energy, and so on, would be complex. The energy values of stationary states wouldn't be real and probabilities of events wouldn't add up to 100%. That is, if you took the probability that this would happen and that would happen—everything you could think of would happen—it would not add up to one.

Another problem on which I struggled very hard, was to represent relativistic electrons with this new quantum mechanics. I wanted to do a unique and different way—and not just by copying the operators of Dirac into some kind of an expression and using some kind of Dirac algebra instead of ordinary complex numbers. I was very much encouraged by the fact that in one space dimension I did find a way of giving an amplitude to every path by limiting myself to paths that only went back and forth at the speed of light. The amplitude was simply $(i\varepsilon)$ to a power equal to the number of velocity reversals, where I have divided the time into steps ε and

I am allowed to reverse velocity only at such a time. This gives (as ε approaches zero) Dirac's equation in two dimensions—one dimension of space and one of time ($\hbar = m = e = 1$).

Dirac's wave function has four components in four dimensions, but in this case, it has only two components and this rule for the amplitude of a path automatically generates the need for two components. Because if this is the formula for the amplitude of a path, it will not do you any good to know the total amplitude of all paths which come into a given point, to find the amplitude to reach the next point. This is because for the next time, if it came in from the right, there is no new factor $i\varepsilon$ if it goes out to the right, whereas, if it came in from the left there was a new factor $i\varepsilon$. So, to continue this same information forward to the next moment, it was not sufficient information to know the total amplitude to arrive, but you had to know the amplitude to arrive from the right and the amplitude to arrive from the left independently. If you did, however, you could then compute both of those again independently and thus you had to carry two amplitudes to form a differential equation (first order in time).

And, so I dreamed that if I were clever, I would find a formula for the amplitude of a path that was beautiful and simple for three dimensions of space and one of time, which would be equivalent to the Dirac equation, and for which the four components, matrices, and all those other mathematical funny things would come out as a simple consequence—I have never succeeded in that either. But, I did want to mention some of the unsuccessful things on which I spent almost as much effort as on the things that did work.

To summarize the situation a few years after the war, I would say I had much experience with quantum electrodynamics, at least in the knowledge of many different ways of formulating it in terms of path integrals of actions and in other forms. One of the important by-products, for example, of much experience in these simple forms, was that it was easy to see how to combine together what was in those days called the longitudinal and transverse fields and, in general, to see clearly the relativistic invariance of the theory. Because of the need to do things differentially there had been, in the standard quantum electrodynamics, a complete split of the field into two parts, one of which is called the longitudinal part and the other mediated by the photons, or transverse waves. The longitudinal part was described by a Coulomb potential acting instantaneously in the Schrödinger equation, while the transverse part had an entirely different description in terms of quantization of the transverse waves. This separation depended upon the relativistic tilt of your axes in space-time. People moving at different velocities would separate the same field into longitudinal and transverse fields in a different way. Furthermore, the entire formulation of quantum mechanics insisting, as it did, on the wave function at a given time, was hard to analyze relativistically. Somebody else in a different coordinate system would calculate the succession of events in terms of wave functions on differently cut slices of space-time, and with a different separation of longitudinal and transverse parts. The Hamiltonian theory did not look relativistically invariant, although, of course, it was. One of the great advantages of the overall point of view was that you could see the relativistic invariance right away—or as Schwinger would say—the covariance

was manifest. I had the advantage, therefore, of having a manifestly covariant form for quantum electrodynamics with suggestions for modifications and so on. I had the disadvantage that if I took it too seriously—I mean, if I took it seriously at all in this form,—I got into trouble with these complex energies and the failure of probabilities adding to one and so on. I was unsuccessfully struggling with that.

Then Lamb did his experiment, measuring the separation of the $^2S_{\frac{1}{2}}$ and $^2P_{\frac{1}{2}}$ levels of hydrogen, finding it to be about 1000 megacycles of frequency difference. Professor Bethe, with whom I was then associated at Cornell, is a man who has this characteristic: If there's a good experimental number you've got to figure it out from theory. So, he forced the quantum electrodynamics of the day to give him an answer to the separation of these two levels. He pointed out that the self-energy of an electron itself is infinite, so that the calculated energy of a bound electron should also come out infinite. But, when you calculated the separation of the two energy levels in terms of the corrected mass instead of the old mass, it would turn out, he thought, that the theory would give convergent finite answers. He made an estimate of the splitting that way and found out that it was still divergent, but he guessed that was probably due to the fact that he used an unrelativistic theory of the matter. Assuming it would be convergent if relativistically treated, he estimated he would get about a thousand megacycles for the Lamb-shift, and thus, made the most important discovery in the history of the theory of quantum electrodynamics. He worked this out on the train from Ithaca, New York, to Schenectady and telephoned me excitedly from Schenectady to tell me the result, which I don't remember fully appreciating at the time.

Returning to Cornell, he gave a lecture on the subject, which I attended. He explained that it gets very confusing to figure out exactly which infinite term corresponds to what in trying to make the correction for the infinite change in mass. If there were any modifications whatever, he said, even though not physically correct, (that is not necessarily the way nature actually works) but any modification whatever at high frequencies, which would make this correction finite, then there would be no problem at all to figuring out how to keep track of everything. You just calculate the finite mass correction Δm to the electron mass m , substitute the numerical values of $m + \Delta m$ for m in the results for any other problem and all these ambiguities would be resolved. If, in addition, this method were relativistically invariant, then we would be absolutely sure how to do it without destroying relativistic invariance.

After the lecture, I went up to him and told him, "I can do that for you, I'll bring it in for you tomorrow." I guess I knew every way to modify quantum electrodynamics known to man, at the time. So, I went in next day, and explained what would correspond to the modification of the δ -function to f and asked him to explain to me how you calculate the self-energy of an electron, for instance, so we can figure out if it's finite.

I want you to see an interesting point. I did not take the advice of Professor Jehle to find out how it was useful. I never used all that machinery which I had cooked up to solve a single relativistic problem. I hadn't even calculated the self-energy of an electron up to that moment, and was studying the difficulties with the conservation of

probability, and so on, without actually doing anything, except discussing the general properties of the theory.

But now I went to Professor Bethe, who explained to me on the blackboard, as we worked together, how to calculate the self-energy of an electron. Up to that time when you did the integrals they had been logarithmically divergent. I told him how to make the relativistically invariant modifications that I thought would make everything all right. We set up the integral which then diverged at the sixth power of the frequency instead of logarithmically!

So, I went back to my room and worried about this thing and went around in circles trying to figure out what was wrong. Because I was sure physically everything had to come out finite, I couldn't understand how it came out infinite. I became more and more interested and finally realized I had to learn how to make a calculation. So, ultimately, I taught myself how to calculate the self-energy of an electron working my patient way through the terrible confusion of those days of negative energy states and holes and longitudinal contributions and so on. When I finally found out how to do it and did it with the modifications I wanted to suggest, it turned out that it was nicely convergent and finite, just as I had expected. Professor Bethe and I have never been able to discover what we did wrong on that blackboard two months before, but apparently we just went off somewhere and we have never been able to figure out where. It turned out that what I had proposed, if we had carried it out without making a mistake, would have been all right and would have given a finite correction. Anyway, it forced me to go back over all this and to convince myself physically that nothing can go wrong. At any rate, the correction to mass was now finite and proportional to $\ln(a)$ where a is the width of the function f which was substituted for δ . If you wanted an unmodified electrodynamics, you would have to take a equal to zero, getting an infinite mass correction. But, that wasn't the point. Keeping a finite, I simply followed the program outlined by Professor Bethe and showed how to calculate all the various things, the scatterings of electrons from atoms without radiation, the shifts of levels and so forth, calculating everything in terms of the experimental mass, and noting that the results as Bethe suggested, were not sensitive to a in this form and even had a definite limit as $a \rightarrow 0$.

The rest of my work was simply to improve the techniques then available for calculations, making diagrams to help analyze perturbation theory quicker. Most of this was first worked out by guessing—you see, I didn't have the relativistic theory of matter. For example, it seemed to me obvious that the velocities in non-relativistic formulas have to be replaced by Dirac's α matrices or in the more relativistic forms by the operators γ_μ . I just took my guesses from the forms that I had worked out using path integrals for non-relativistic matter, but relativistic light. It was easy to develop rules of what to substitute to get the relativistic case. I was very surprised to discover that it was not known at that time that every one of the formulas that had been worked out so patiently by separating longitudinal and transverse waves could be obtained from the formula for the transverse waves alone, if instead of summing over only the two perpendicular polarization directions you would sum over all four possible directions of polarization. It was so obvious from the action (1) that I thought it was general knowledge and would do it all the time. I would get into arguments with

people because I didn't realize they didn't know that; but, it turned out that all their patient work with the longitudinal waves was always equivalent to just extending the sum on the two transverse directions of polarization over all four directions. This was one of the amusing advantages of the method. In addition, I included diagrams for the various terms of the perturbation series, improved notations to be used, worked out easy ways to evaluate integrals which occurred in these problems, and so on, and made a kind of handbook on how to do quantum electrodynamics.

But one step of importance that was physically new was involved with the negative energy sea of Dirac, which caused me so much logical difficulty. I got so confused that I remembered Wheeler's old idea about the positron being, maybe, the electron going backward in time. Therefore, in the time dependent perturbation theory that was usual for getting self-energy, I simply supposed that for a while we could go backward in the time, and looked at what terms I got by running the time variables backward. They were the same as the terms that other people got when they did the problem a more complicated way, using holes in the sea, except, possibly, for some signs. These I, at first, determined empirically by inventing and trying some rules.

I have tried to explain that all the improvements of relativistic theory were at first more or less straightforward, semi-empirical shenanigans. Each time I would discover something, however, I would go back and I would check it so many ways, compare it to every problem that had been done previously in electrodynamics (and later, in weak coupling meson theory) to see if it would always agree, and so on, until I was absolutely convinced of the truth of the various rules and regulations that I concocted to simplify all the work.

During this time, people had been developing meson theory, a subject I had not studied in any detail. I became interested in the possible application of my methods to perturbation calculations in meson theory. But, what was meson theory? All I knew was that meson theory was something analogous to electrodynamics, except that particles corresponding to the photon had a mass. It was easy to guess the δ -function in (1), which was a solution of d'Alembertian equals zero, was to be changed to the corresponding solution of d'Alembertian equals m^2 . Next, there were different kinds of mesons—the one in closest analogy to photons, coupled via $\gamma_\mu \gamma_\mu$, are called vector mesons—there were also scalar mesons. Well, maybe that corresponds to putting unity in place of the γ_μ , perhaps what they called “pseudo vector coupling” and I would guess what that probably was. I didn't have the knowledge to understand the way these were defined in the conventional papers because they were expressed at that time in terms of creation and annihilation operators, and so on, which, I had not successfully learned. I remember that when someone had started to teach me about creation and annihilation operators, that this operator creates an electron, I said, “how do you create an electron? It disagrees with the conservation of charge,” and in that way, I blocked my mind from learning a very practical scheme of calculation. Therefore, I had to find as many opportunities as possible to test whether I guessed right as to what the various theories were.

One day a dispute arose at a Physical Society meeting as to the correctness of a calculation by Slotnick of the interaction of an electron with a neutron using pseudo scalar theory with pseudo vector coupling and also pseudo scalar theory with pseudo

scalar coupling. He had found that the answers were not the same. In fact, by one theory, the result was divergent, although convergent with the other. Some people believed that the two theories must give the same answer for the problem. This was a welcome opportunity to test my guesses as to whether I really did understand what these two couplings were. So, I went home, and during the evening I worked out the electron neutron scattering for the pseudo scalar and pseudo vector coupling, saw they were not equal and subtracted them, and worked out the difference in detail. The next day at the meeting, I saw Slotnick and said, "Slotnick, I worked it out last night, I wanted to see if I got the same answers you do. I got a different answer for each coupling—but, I would like to check in detail with you because I want to make sure of my methods." And, he said, "what do you mean you worked it out last night, it took me six months!" And, when we compared the answers he looked at mine and he asked, "what is that Q in there, that variable Q ?" (I had expressions like $(\tan^{-1} Q)/Q$, etc.) I said, "that's the momentum transferred by the electron, the electron deflected by different angles." "Oh," he said, "no, I only have the limiting value as Q approaches zero; the forward scattering." Well, it was easy enough to just substitute Q equals zero in my form and I then got the same answers as he did. But, it took him six months to do the case of zero momentum transfer, whereas, during one evening I had done the finite and arbitrary momentum transfer. That was a thrilling moment for me, like receiving the Nobel Prize, because that convinced me, at last, I did have some kind of method and technique and understood how to do something that other people did not know how to do. That was my moment of triumph in which I realized I really had succeeded in working out something worthwhile.

At this stage, I was urged to publish this because everybody said it looks like an easy way to make calculations, and wanted to know how to do it. I had to publish it missing two things; one was proof of every statement in a mathematically conventional sense. Often, even in a physicist's sense, I did not have a demonstration of how to get all of these rules and equations from conventional electrodynamics. But, I did know from experience, from fooling around, that everything was, in fact, equivalent to the regular electrodynamics and had partial proofs of many pieces, although, I never really sat down, like Euclid did for the geometers of Greece, and made sure that you could get it all from a single simple set of axioms. As a result, the work was criticized, I don't know whether favorably or unfavorably, and the "method" was called the "intuitive method." For those who do not realize it, however, I should like to emphasize that there is a lot of work involved in using this "intuitive method" successfully. Because no simple clear proof of the formula or idea presents itself, it is necessary to do an unusually great amount of checking and rechecking for consistency and correctness in terms of what is known, by comparing to other analogous examples, limiting cases, etc. In the face of the lack of direct mathematical demonstration, one must be careful and thorough to make sure of the point, and one should make a perpetual attempt to demonstrate as much of the formula as possible. Nevertheless, a very great deal more truth can become known than can be proven.

It must be clearly understood that in all this work, I was representing the conventional electrodynamics with retarded interaction, and not my half-advanced and half-retarded theory corresponding to (1). I merely used (1) to guess at forms. And,

one of the forms I guessed at corresponded to changing δ to a function f of width a^2 , so that I could calculate finite results for all of the problems. This brings me to the second thing that was missing when I published the paper, an unresolved difficulty. With δ replaced by f the calculations would give results which were not “unitary,” that is, for which the sum of the probabilities of all alternatives was not unity. The deviation from unity was very small, in practice, if a was very small. In the limit that I took a very tiny, it might not make any difference. And, so the process of the renormalization could be made, you could calculate everything in terms of the experimental mass and then take the limit and the apparent difficulty that the unitarity is violated temporarily seems to disappear. I was unable to demonstrate that, as a matter of fact, it does.

It is lucky that I did not wait to straighten out that point, for as far as I know, nobody has yet been able to resolve this question. Experience with meson theories with stronger couplings and with strongly coupled vector mesons, although not proving anything, convinces me that if the coupling were stronger, or if you went to a higher order (137th order of perturbation theory for electrodynamics), this difficulty would remain in the limit and there would be real trouble. That is, I believe there is really no satisfactory quantum electrodynamics, but I'm not sure. And, I believe, that one of the reasons for the slowness of present-day progress in understanding the strong interactions is that there isn't any relativistic theoretical model, from which you can really calculate everything. Although it is usually said that the difficulty lies in the fact that strong interactions are too hard to calculate, I believe it is really because strong interactions in field theory have no solution, have no sense—they're either infinite, or, if you try to modify them, the modification destroys the unitarity. I don't think we have a completely satisfactory relativistic quantum-mechanical model, not even one that doesn't agree with nature, but, at least, agrees with the logic that the sum of probability of all alternatives has to be 100%. Therefore, I think that the renormalization theory is simply a way to sweep the difficulties of the divergences of electrodynamics under the rug. I am, of course, not sure of that.

This completes the story of the development of the space-time view of quantum electrodynamics. I wonder if anything can be learned from it. I doubt it. It is most striking that most of the ideas developed in the course of this research were not ultimately used in the final result. For example, the half-advanced and half-retarded potential was not finally used, the action expression (1) was not used, the idea that charges do not act on themselves was abandoned. The path-integral formulation of quantum mechanics was useful for guessing at final expressions and at formulating the general theory of electrodynamics in new ways—although, strictly it was not absolutely necessary. The same goes for the idea of the positron being a backward moving electron, it was very convenient, but not strictly necessary for the theory because it is exactly equivalent to the negative energy sea point of view.

We are struck by the very large number of different physical viewpoints and widely different mathematical formulations that are all equivalent to one another. The method used here, of reasoning in physical terms, therefore, appears to be extremely inefficient. On looking back over the work, I can only feel a kind of regret for the enormous amount of physical reasoning and mathematical re-expression which ends

by merely re-expressing what was previously known, although in a form which is much more efficient for the calculation of specific problems. Would it not have been much easier to simply work entirely in the mathematical framework to elaborate a more efficient expression? This would certainly seem to be the case, but it must be remarked that although the problem actually solved was only such a reformulation, the problem originally tackled was the (possibly still unsolved) problem of avoidance of the infinities of the usual theory. Therefore, a new theory was sought, not just a modification of the old. Although the quest was unsuccessful, we should look at the question of the value of physical ideas in developing a *new* theory.

Many different physical ideas can describe the same physical reality. Thus, classical electrodynamics can be described by a field view, or an action at a distance view, etc. Originally, Maxwell filled space with idler wheels, and Faraday with fields lines, but somehow the Maxwell equations themselves are pristine and independent of the elaboration of words attempting a physical description. The only true physical description is that describing the experimental meaning of the quantities in the equation—or better, the way the equations are to be used in describing experimental observations. This being the case perhaps the best way to proceed is to try to guess equations, and disregard physical models or descriptions. For example, McCullough guessed the correct equations for light propagation in a crystal long before his colleagues using elastic models could make head or tail of the phenomena, or again, Dirac obtained his equation for the description of the electron by an almost purely mathematical proposition. A simple physical view by which all the contents of this equation can be seen is still lacking.

Therefore, I think equation guessing might be the best method to proceed to obtain the laws for the part of physics which is presently unknown. Yet, when I was much younger, I tried this equation guessing and I have seen many students try this, but it is very easy to go off in wildly incorrect and impossible directions. I think the problem is not to find the *best* or most efficient method to proceed to a discovery, but to find any method at all. Physical reasoning does help some people to generate suggestions as to how the unknown may be related to the known. Theories of the known, which are described by different physical ideas may be equivalent in all their predictions and are hence scientifically indistinguishable. However, they are not psychologically identical when trying to move from that base into the unknown. For different views suggest different kinds of modifications which might be made and hence are not equivalent in the hypotheses one generates from them in ones attempt to understand what is not yet understood. I, therefore, think that a good theoretical physicist today might find it useful to have a wide range of physical viewpoints and mathematical expressions of the same theory (for example, of quantum electrodynamics) available to him. This may be asking too much of one man. Then new students should as a class have this. If every individual student follows the same current fashion in expressing and thinking about electrodynamics or field theory, then the variety of hypotheses being generated to understand strong interactions, say, is limited. Perhaps rightly so, for possibly the chance is high that the truth lies in the fashionable direction. But, on the off-chance that it is in another direction—a direction obvious from an unfashionable view of field theory—who will find it? Only someone who has

sacrificed himself by teaching himself quantum electrodynamics from a peculiar and unusual point of view, one that he may have to invent for himself. I say sacrificed himself because he most likely will get nothing from it, because the truth may lie in another direction, perhaps even the fashionable one.

But, if my own experience is any guide, the sacrifice is really not great because if the peculiar viewpoint taken is truly experimentally equivalent to the usual in the realm of the known there is always a range of applications and problems in this realm for which the special viewpoint gives one a special power and clarity of thought, which is valuable in itself. Furthermore, in the search for new laws, you always have the psychological excitement of feeling that possibly nobody has yet thought of the crazy possibility you are looking at right now.

So what happened to the old theory that I fell in love with as a youth? Well, I would say it's become an old lady, that has very little attractive left in her and the young today will not have their hearts pound when they look at her anymore. But, we can say the best we can for any old woman, that she has been a very good mother and she has given birth to some very good children. And, I thank the Swedish Academy of Sciences for complimenting one of them. Thank you.

Postscript

The records at Caltech indicate that the written version of Feynman's Nobel Lecture originates from a transcript of the lecture delivered by Feynman at Caltech, some time after it was given at the Nobel ceremonies. Copies of the transcript were provided to the Nobel Foundation and to the editors of *Science and Physics Today*. The Nobel Foundation published the lecture in *Les Prix Nobel en 1965*, Norstedt, 1966, in *Nobel Lectures, Physics, 1963-1970*, Elsevier, 1972, and it appears in *The Selected Papers of Richard Feynman*, World Scientific Press, 2000. In addition, the lecture is posted at the Nobel Foundation web site:

http://nobelprize.org/nobel_prizes/physics/laurates/1965/feynmanlecture.html.

This version of Feynman's Nobel Lecture was prepared to improve the readability of the text by correcting many small errors that appear in the previously published versions.

Michael D. Godfrey
Stanford University

Michael A. Gottlieb
California Institute of Technology

March 2008

20080317.1

JAMES FRANCK

Transformations of kinetic energy of free electrons into excitation energy of atoms by impacts

Nobel Lecture, December 11, 1926

Ladies and gentlemen!

The exceptional distinction conferred upon our work on electron impacts by the Royal Swedish Academy of Sciences requires that my friend Hertz and I have the honour of reporting to you on current problems within this province:

The division of the material between us left me with the task of presenting, in a historical setting, the development of these projects which have led to an association with Bohr's atomic theory.

Investigations of collision processes between electrons, atoms and molecules have already got well under way. Practically all investigations into the discharge of electricity through gases can be considered under this heading. An enormous amount of knowledge, decisive for the whole development of modern physics, has been gained, but it is just in this gathering that I feel it is unnecessary for me to make any special comment, since the lists of the men whom the Swedish Academy of Sciences have deemed worthy of the Nobel Prize contain a large number of names of research workers who have made their most significant discoveries in these fields.

Attracted by the complex problems of gas discharges and inspired particularly by the investigations of my distinguished teacher E. Warburg, our interest turned in this direction. A starting-point was provided by the observation that in inert gases (and as found later, also in metal vapour) no negative ions were formed by the attachment of free electrons to an atom. The electrons remained rather as free ones, even if they were moving slowly in a dense gas of this type, which can be inferred from their mobility in an electric field. Even the slightest pollution with normal gases produced, at once, a material attachment of the electrons and thus the appearance of normal negative ions.

As a result, one can perhaps divide gases somewhat more clearly than has been the case up to now from the observations described in the literature, into one class with, and one class without, an electron affinity. It was to be

expected that the motion of electrons in gases of the latter kind would obey laws of a particularly simple kind. These gases have exhibited special behaviour during investigations of other kinds into gas discharges. For instance, according to Ramsay and Collie, they have a specially low dielectric strength, and this was, further, extremely dependent upon the degree of purity of the gas (see, for example, Warburg's experiments). The important theory of the dielectric strength of gases, founded by Townsend, the equations of which even today, when used formally, still form the basic foundation of this field failed in these cases. The reason for this seemed likely to be that Townsend's hypothesis on the kind of collisions between slow electrons and atoms, particularly inert-gas atoms, differed from the reality, and it seemed promising to arrive at a kinetic theory of electrons in gases by a systematic examination of the elementary processes occurring when collisions took place between slow electrons and atoms and molecules. We had the experiences and techniques to support us, which men like J. J. Thomson, Stark, Townsend, and in particular, however, Lenard, had created, and also had their concept of the free path-lengths of electrons and the ionization energy, etc., to make use of.

The free path-lengths in the light inert gases were examined first. By "free path" in this connection is to be understood that path which, on the average, is that which an electron traces between two collisions with atoms along a straight track. The distance is measurable as soon as the number of atoms per unit volume is sufficiently small, this being attained by taking a low gas pressure. The method of measurement itself differed but slightly from that developed by Lenard. It is unnecessary to go into closer detail since the results gave the same order of values for the free path-length as Lenard obtained for slow electrons in other gases. The value is of that order which is obtained by calculation if the formulas of the kinetic gas theory are used for the free path-length, taking for the impact radius of the electron a value which is very small compared with the gas-kinetic atom radii. With this assumption, the electrons behave, to a first approximation, like a gaseous impurity in the inert gas, not reacting chemically with it - an impurity, however, which has the special quality of consisting of electrically charged particles and having a vanishingly small impact radius. As a result of significant experiences, we know, today, from the work of Ramsauer and others on the free path-lengths of electrons in heavy inert gases that the picture we had formed at that time was a very rough one, and that for collisions of slow electrons the laws of quantum theory are of far more significance than the

mechanical diameter, but as a first approximation for the establishment of the kinetics it suffices. Further, it also sufficed, as it turned out, to gain an understanding of the energy conversion on the occurrence of a collision between the slow electrons and the atoms of the inert gases and metal vapours. Since the mass of the electron is 1800 times smaller than that of the lightest atom we know, the hydrogen atom, the transfer of momentum from the light electron to the heavy atom during customary gas-kinetic collisions, i.e. collisions such as between two elastic balls, must be exceptionally small according to the laws of momentum. A slow electron with a given amount of kinetic energy, meeting an atom at rest, ought to be reflected without practically any energy loss, much the same as a rubber ball against a heavy wall. These elastic collisions can now be pursued by measurements.

I will pass over the detection of the single reflection and mention in more detail a simple experimental arrangement which, by means of an accumulation of collisions, enables us to measure the energy loss which is otherwise too small to measure in one elementary process. The mode of action might well be clear from a schematic layout (Fig. 1).

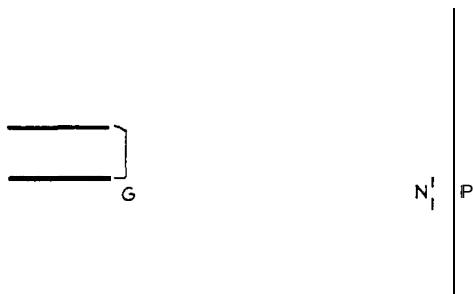


Fig. 1.

G indicates the electron source. It consists of a tungsten wire, heated to a bright-red glow by an electric current. That such a glowing wire is a source of electrons can, I think, be taken as read in this age of radio. A few centimetres away is a wire-screen electrode N. If we now charge the screen positively with respect to the glowing wire, by means of an accumulator, the electrons emitted by the wire towards the screen will be accelerated. The kinetic energy which the electrons must gain through this acceleration can easily be found for the case where no gas exists between G and N, that is, when the electrons fall through the field of force freely without collisions. We have the relationship:

$$\frac{1}{2} m v^2 = e \cdot V$$

Here, $\frac{1}{2}mv^2$ is the kinetic energy of each electron, e is its electrical elementary charge, and V the applied potential difference. If the latter is measured in volts, then, for instance, the kinetic energy of an electron which has fallen through 10 volts is approximately 10^{11} ergs. We have become accustomed to speak of x-volt electrons, and to simply denote the acceleration voltage (**x volts**) as a measure of energy. Thus in our arrangement the electrons fall upon the screen with an energy of x volts (the potential difference between G and N). Some of the electrons are caught by the screen, some fly through the mesh. The latter, assuming no field between N and P which would throw the electrons back, all reach the electrode P and produce a negative current which flows to earth through a galvanometer. By introducing an electric field between N and P the energy distribution of those electrons passing through the screen can be determined. If, for example, we take only 4-volt beams, which pass perpendicularly through the screen, then the electron current measured at the galvanometer as a function of a decelerating potential difference applied between N and P, must be constant, until P becomes 4 volts more negative than N. At this point the current must become suddenly zero since henceforth all electrons will be so repelled from P that they return to N. If now we introduce an inert gas such as helium or a metal vapour between the three electrodes and choose such a pressure as will ensure that the electrons between G and N will make many impacts upon atoms, whilst passing freely through the space between N and P, we can determine, by plotting the energy distribution of the electrons arriving at P, whether the electrons have lost energy by impacts on the atoms. In discussing the resulting current-voltage curve it should be noted that the electrons no longer pass through the screen mesh perpendicularly, but are scattered in all directions due to reflection from the atoms. As a result of this, there is an easily calculable change in shape of the curve, and this holds, too, for uniform kinetic energy of the electrons. From a consideration of the resulting curves it was found that for not too high pressures, particularly for monatomic gases of high atomic weight, the kinetic energy of slow electrons was the same as for those in vacuum under the same acceleration voltage. The gas complicates the trajectory of the electrons in the same way that a ball's trajectory is affected by rolling down a sloping board bedecked with a large number of nails, but the energy (because of the large mass of the atom compared with that of the electron) is practically the same as for conditions of free fall. Only for high pressures, that is, with the occurrence of many thousands of collisions, can the energy loss corresponding to elastic

collision be demonstrated.* A calculation of the number of collisions was later carried out by Hertz. Taking this as a basis and evaluating the curves measured for higher pressures accordingly, it emerges that, for example, energy is transferred to a helium atom amounting to $1.2-3.0 \times 10^{-4}$ of the energy of the electron prior to the collision, whilst the calculated value for the mass ratio under conditions of pure mechanical elastic impact is 2.9×10^{-4} . We may therefore, with close approximation to reality, speak of elastic collisions.

For polyatomic gases a significantly greater average energy loss was determined. Using the methods available at that time, it was not possible to distinguish whether this latter effect was contingent upon attachment of the electrons to the molecule, that is, the formation of negative ions, or whether a transfer of the kinetic energy of the striking electrons into vibrational and rotational degrees of freedom of the molecules was taking place. An investigation just carried out in my institute by Mr. Harries shows that the latter elementary process, even though at a low level, does occur, and is important in the explanation of the energy losses.

Can the principles of action found for slow electrons in the case of elastic collisions hold good for higher electron velocities? Apparently not, for the elementary knowledge of gas discharges teaches us that with faster electrons, i.e. with cathode rays, the impacted atoms are excited to luminescence or become ionized. Here, energy of the impacting electrons must be transferred into internal energy of the impacted atoms, the electrons must henceforth collide inelastically and give up greater amounts of energy. The determination of the least amount of energy which an electron must possess in order to ionize an atom was therefore of interest. Measured in volts, this energy is called the ionization voltage. Calculations of this value of energy by Townsend were available for some gases and these were based upon the validity of his assumptions about the course of the elementary action on collision. I mentioned already the reasons for doubting the correctness of these indirectly determined values. A direct method had been given by Lenard, but it gave the same ionization voltage for all gases. Other writers had obtained the same results within the range of measurement. We therefore repeated Lenard's investigations, using the improved pumping techniques which had become available in the meantime, and obtained characteristic, marked differences in values for the various gases. The method used by

* It is better to use here the experimental arrangements indicated later by Compton and Benade, Hertz, and others.

Lenard was as follows. Electrons, from a glowing wire, for example, were accelerated by a suitable electric field and allowed to pass through a screen grid into a space in which they suffered collisions with atoms. By means of a strong screening field these particular electrons were prevented from reaching an electrode to which was connected a measuring instrument. Atoms ionized by the impact resulted in the newly formed positive ions being accelerated through the screening field, which repelled the electrons, towards the negatively charged electrode. A positive current was thus obtained as soon as the energy of the electrons was sufficient for ionization to take place. I will talk later about the fact that a positive charge appears if the impacted atoms are excited to emit ultraviolet light, and that, as shown later, the charges measured at that time are to be attributed to this process and not to ionization, as we formerly supposed.

In any case, as already discussed, inelastic collisions were to be expected between electrons and atoms for the characteristic critical voltages appertaining to each kind of atom. And it proved easy to demonstrate this fact with the same apparatus as was used for the work on elastic collisions. Measurement of the energy distribution of the electrons, on increasing the accelerating voltage above the critical value, showed that electrons endowed with the critical translation energy could give up their entire kinetic energy on collision, and that electrons whose energy exceeded the critical by a fraction, likewise gave up the same significant amount of energy, the rest being retained as kinetic energy. A simple modification of the electric circuit diagram of our apparatus produced a significantly sharper measurement of the critical voltage and a visual proof of the discontinuously occurring release of energy from the electrons on collision. The measurement method consisted of measurements of the number of those electrons (possessing markedly different energies from zero after many collisions) as a function of the accelerating voltage.

The graph (Fig. 2) shows the results of measurements of electron current in mercury vapour. In this case, all electrons whose energy is greater than the energy of $\frac{1}{2}$ -volt beams were measured. It can be seen that in Hg vapour this partial electron current increases with increasing acceleration, similar to the characteristic of « glow-electron » current in vacuum, until the critical energy stage is reached when the current falls suddenly to almost zero. Since the electrons cannot lose more or less than the critical amount of energy, the cycle begins anew with further increase of voltage. The number of electrons whose velocity is greater than $\frac{1}{2}$ volt, again climbs up until the critical value is

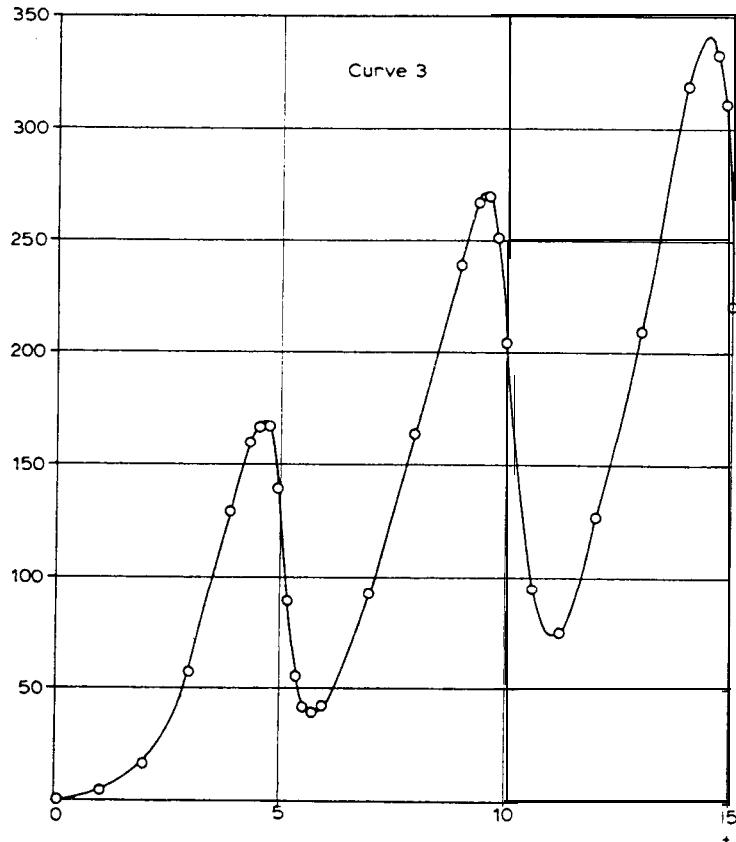


Fig. 2.

reached, the current again falls away. The process repeats itself periodically as soon as the accelerating voltage overreaches a multiple of the critical voltage. The distance between the succeeding maxima gives an exact value of the critical voltage. This is 4.9 V for mercury vapour.

As already mentioned we took this value to be the ionization voltage (the same applied to He which was determined by the same method and was about 20 V). Nevertheless, the quanta-like character of the energy transfer could not help but remind us - who practically from the start could witness from nearby the developments of Planck's quantum theory - to the use of the theory made by Einstein to explain the facts of the photoelectric effect! Since here, light energy is converted into the kinetic energy of electrons, could not perhaps, in our case, kinetic energy from electrons be converted into light energy? If that were the case, it should be easy to prove in the case of mercury; for the equation $\frac{1}{2}mv^2 = h\nu$ referred to a line of 2,537 Å which is

easily accessible in the ultraviolet region. This line is the longest wavelength absorption line of Hg vapour. It is often cited as Hg-resonance line since R.W. Wood has carried out with it his important experiments on resonance fluorescence. If the conjectured conversion of kinetic energy into light on impact should take place, 'then on bombardment with 4.9 eV electrons, the line 2,537 Å, and only this line out of the complete line spectrum of mercury should appear.'

Fig. 3 shows the result of the experiment. Actually, only the 2,537 Å line appears in the spectrogram next to a continuous spectrum in the long-wave region emitted by the red-glowing filament. (The second spectrogram shows the arc spectrum of mercury for comparison.) The first works of Niels Bohr on his atomic theory appeared half a year before the completion of this work. Let us compare, in a few words, the basic hypothesis of this theory with our results.

According to Bohr an atom can absorb as internal energy only discrete quantities of energy, namely those quantities which transfer the atom from one stationary state to another stationary state. If following on energy supply an excited state results from a transfer to a stationary state of higher energy, then the energy so taken up will be radiated in quanta fashion according to the hv relationship. The frequency of the absorption line having the longest wavelength, the resonance line, multiplied by Planck's constant, gives the energy required to reach the first state of excitation. These basic concepts agree in very particular with our results. The elastic collisions at low electron velocities show that for these impacts no energy is taken up as inner energy, and the first critical energy step results in just that amount of energy required for the excitation of the longest wave absorption line of Hg. Subse-

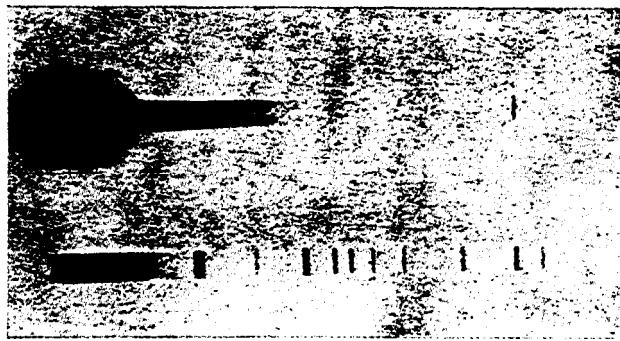


Fig. 3.

quently it appeared to me to be completely incomprehensible that we had failed to recognize the fundamental significance of Bohr's theory, so much so, that we never even mentioned it once in the relevant paper. It was unfortunate that we could not rectify our error (due in part to external circumstances) ourselves by clearing up the still existing uncertainties experimentally. The proof that only monochromatic light was radiated at the first excitation step, as Bohr's theory required, and that the gas is not simultaneously ionized (as we were also obliged to think for reasons other than those mentioned) came about instead during the war period through suggestions from Bohr himself and from van der Bijl. The appearance of positive charge at the first excitation step in Lenard's arrangement was explained by them on the basis of a photoelectric effect at the collector electrode, an hypothesis which was substantiated by Davis and Goucher.

Time does not allow me to describe how our further difficulties were clarified in the sense of Bohr's theory. And in regard to further development, too, I would like to devote only a few words, particularly since my friend Hertz's lecture covers it more closely. The actual ionization voltage of mercury was for the first time determined by Tate as being 10.3 volts, a value which agreed exceptionally well with that resulting, according to Bohr, from the limit of the absorption series. A great number of important, elegantly carried out, determinations of the first excitation level and the ionization voltage of many kinds of atoms was made during the war years and also in the following years, above all by American scientists; research workers such as Foote and Mohler, K. T. Compton and others are to be thanked for extensive clarification in this field.

Without going into details of the experimental arrangements, I should like to mention that it later proved successful, by the choice of suitable experimental conditions, to demonstrate also, from the current-voltage curves, the stepwise excitation of a great number of quantum transitions, lying between the first excitation level and ionization. A curve plotted for mercury vapour might well serve again as an example. It shows the quantum-like appearance of higher excitation levels by kinks in the curve (Fig. 4). It is noteworthy that, in addition, transitions which under the influence of light according to Bohr's correspondence principle do not appear, manifest themselves clearly. When, as is the case with mercury, and still more decidedly so with helium, the first transition is such that it cannot be achieved by light, we have excited atoms in a so-called metastable state. The discovery of a metastable state by means of the electron-impact method was first suc-

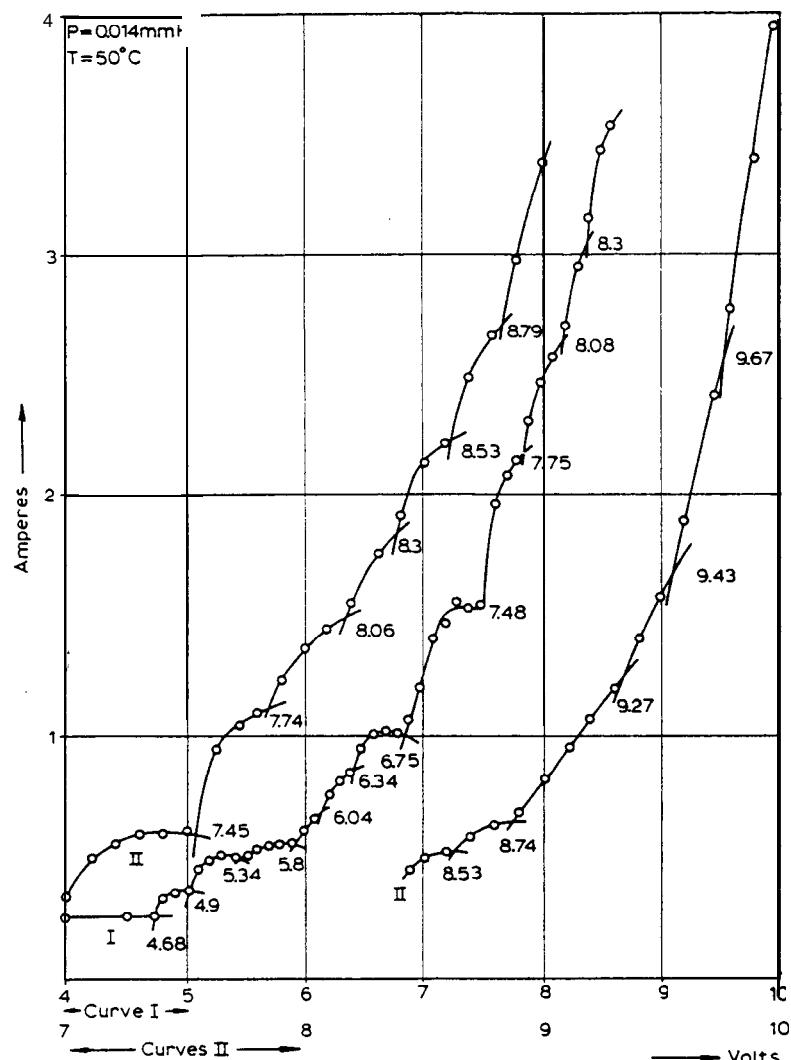


Fig. 4.

cessful with helium. Since helium is a gas in which the absorption series lies in the far ultraviolet-it was later found optically by Lyman - and on the other side, helium, apart from hydrogen, is the most simply constructed atom, the approximate determination of the energy levels of helium and perhaps too, the appearance, in particular, of the metastable level has proved useful for the development of Bohr's theory.

Much more could be said, but I think I have given you the main outline as far as is possible within the framework of a short survey, and must there-

fore draw to a close. The desire to describe, historically, our part in the development of the investigations leading to the establishment of the quantum transfer of energy to the atom by impacting electrons has forced me to take up your time with the description of many a false trail and roundabout path which we took in a field in which the direct path has now been opened by Bohr's theory. Only later, as we came to have confidence in his leadership, did all difficulties disappear. We know only too well that we owe the wide recognition that our work has received to contact with the great concepts and ideas of M. Planck and particularly of N. Bohr.

ON SUPERCONDUCTIVITY AND SUPERFLUIDITY

Nobel Lecture, December 8, 2003

by

VITALY L. GINZBURG

P. N. Lebedev Physics Institute, Russian Academy of Sciences, Moscow, Russia.

INTRODUCTION

First of all I would like to express my heartfelt gratitude to the Royal Swedish Academy of Sciences and its Nobel Committee for physics for awarding me the 2003 Nobel Prize in physics. I am well aware of how difficult it is to select no more than three Laureates out of the far greater number of nominees. So all the more valuable is this award. Personally, I have two additional motives for appreciating the award of the Prize. First, I am already 87, the Nobel Prize is not awarded posthumously, and posthumous recognition is not all that significant to me since I am an atheist. Second, the 1958 and 1962 Nobel Prizes were awarded respectively to Igor' Evgen'evich Tamm and Lev Davidovich Landau. Outside of high school, the notion of a teacher is very relative and is quite often applied by formal criteria: for instance, it is applied to the supervisor in the preparation of a thesis. But I believe that the title real teacher can appropriately be given only to those who have made the greatest impact on your work and whose example you have followed. Tamm and Landau were precisely these kind of people for me. I feel particularly pleased, because in a sense I have justified their good attitude toward me. Of course, the reason lies not with the Prize itself, but with the fact that my receiving the award after them signifies following their path.

Now about the Nobel Lecture. It is the custom, I do not know whether by rule or natural tradition, that the Nobel Lecture is concerned with the work for which the Prize was awarded. But I am aware of at least one exception. P.L. Kapitza was awarded the 1978 Prize for "his basic inventions and discoveries in the area of low-temperature physics". But Kapitza's Lecture was entitled "Plasma and the Controlled Thermonuclear Reactions". He justified his choice of the topic as follows: he had worked in the field of low-temperature physics many years before he had been awarded the Prize and he believed it would be more interesting to speak of what he was currently engaged in. That is why P.L. Kapitza spoke of his efforts to develop a fusion reactor employing high-frequency electromagnetic fields. By the way, this path has not led to success, which is insignificant in the present context.

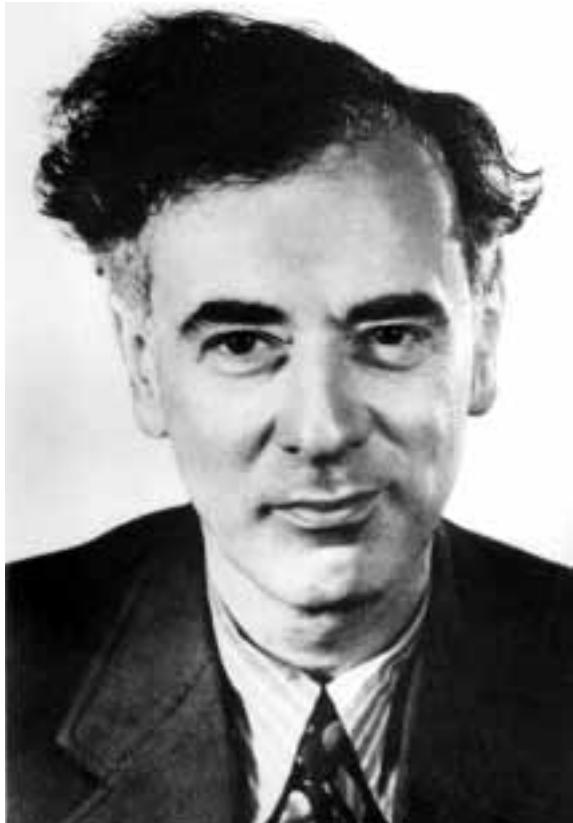
I have not forgotten my "pioneering contributions to the theory of superconductors and superfluids" for which I have received the Prize, but I would



I.E. Tamm.

like not to dwell on them. The point is that in 1997 I decided to sum up my activities in the corresponding field, and I wrote a paper entitled “Superconductivity and superfluidity (what was and what was not done)” [1, 2]. In particular, this article set out in detail the story of quasi-phenomenological superconductivity theory constructed jointly with Landau [3]. Under the circumstances, it would be unnecessary, and above all tedious to repeat all that. Furthermore, the Ginzburg–Landau theory of superconductivity, which I call the Ψ -theory of superconductivity, is employed in the work of A.A. Abrikosov [4], and he will supposedly dwell on it in his Nobel Lecture. This is to say nothing of the fact that the Ψ -theory of superconductivity has been covered in many books (see, for instance, Refs [5, 6]). At the same time, there are several problems bearing on the field of superconductivity and superfluidity which I have taken up and which have not been adequately investigated. This is why I decided to dwell on these two most important problems in my lecture.

The case in point is thermoelectric effects in the superconducting state and the Ψ -theory of superfluidity. However, before I turn to these issues, I will nevertheless cover briefly the entire story of my activities in the field of superconductivity. At the end of the lecture I will allow myself to touch on some educational program for physicists (the issue of a ‘physical minimum’), which has been of interest to me for more than thirty years.

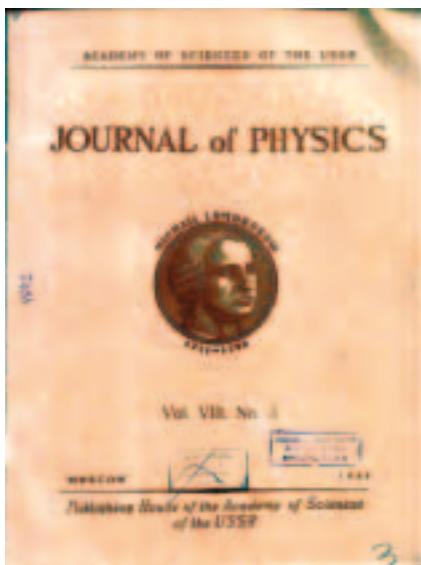


L.D. Landau.

BRIEF ACCOUNT OF MY ACTIVITY IN THE FIELD OF SUPERCONDUCTIVITY PRIOR TO THE ADVENT OF HIGH-TEMPERATURE SUPERCONDUCTORS.

Lev Landau was in prison for exactly one year and was released on April 28, 1939 primarily due to the efforts of Kapitza, who became his ‘personal guarantee’.¹ Landau resided in this state until his premature death in 1968. The Landau ‘case’ was officially discharged by virtue of “*corpus delicti*” (“absence of a basis of a crime”) only in 1990 (!). The imprisonment had a strong effect on Landau, but fortunately it did not bereave him of his outstanding capabilities as a physicist. That is why he ‘justified the confidence’, as they said at that time, of those who released him on bail instead of shooting him or leaving him to rot in jail (Landau personally told me that he had not been far from death) by constructing his superfluidity theory [7]. I was present at his report on this topic in 1940 or maybe in 1941 (the paper was submitted for publication on May 15, 1941). Also considered at the end of this paper was super-

¹ For more details, see for example article 10 in the book [2].



No. 111, No. 4

GENERAL & PHYSICS

1-14

OF THE PHYSICO-MATHEMATICAL INSTITUTE OF CYBERNETICS

By V. L. GOLDBERG

Institute of Cybernetics, Academy of Sciences of the USSR
Bulvar Dzerzhinskogo, 22, Kiev

Numerous papers in the literature have shown that the temperature dependence of the superconducting properties of various materials is determined by the temperature dependence of the superconducting gap $\Delta(T)$. It is also known that the gap $\Delta(T)$ is proportional to the temperature T for temperatures above the superconducting transition temperature T_c , and it is zero at T_c . It is also known that the gap $\Delta(T)$ is zero at $T = 0$ for all superconductors except for the so-called "superconductors with finite gap" (such as the superconductors of the Bardeen-Cooper-Schrieffer type). In this paper we consider the question of the temperature dependence of the gap $\Delta(T)$ for the superconductors with finite gap.

The temperature dependence of the gap $\Delta(T)$ has been theoretically studied in many papers [1-4]. A short review of these papers can be found in [5]. In this paper we shall consider the case of the gap $\Delta(T)$ being proportional to T^{α} for $T \ll T_c$, where α is a coefficient depending on the temperature T .

In our calculations we take into account the fact that the gap $\Delta(T)$ is proportional to the temperature T only for temperatures above T_c , and below T_c the gap $\Delta(T)$ is zero.

We shall assume that the gap $\Delta(T)$ is zero at $T = 0$ and is proportional to T^{α} for $T \ll T_c$, where α is a coefficient, $0 < \alpha < 1$, and is determined by the formula

$$\alpha = 1 - (B - B_c)/T_c. \quad (1)$$

where B is a coefficient depending on the temperature T . We shall also assume that equation (1) is valid for temperatures above T_c and below T_c . We shall also assume that the gap $\Delta(T)$ is zero at $T = 0$ and is proportional to T^{α} for $T \ll T_c$, where α is a coefficient, $0 < \alpha < 1$, and is determined by the formula

$$\alpha = 1 - (B - B_c)/T_c. \quad (2)$$

conductivity, which was treated as the superfluidity of electron liquid in metals.

That work impressed me, of course, but at that time I was enthusiastic about quite a different range of questions, namely, the theory of higher spin particles. That is why I did not take up the low temperature subject right away, and shortly after our lives radically changed when the war broke out (as is well known, for the USSR it began on June 22, 1941). The Physical Institute of the USSR Academy of Sciences, where I was working and still work, was evacuated from Moscow to the town of Kazan, where many difficulties were encountered, which I describe in my autobiography. In any case, it was not until 1943 that I made an attempt to do, in the spirit of the Landau theory of superfluidity [7], something of the kind as applied to superconductivity.² That work [9] is of no great value today, but I believe there were some interesting points in it, for Bardeen considered it at length in his famous review [10]. Even at that time I was aware the work was poor and therefore did not submit it to a journal in English, which we would normally have done at that time (the journal – *Journal of Physics USSR* – was terminated in 1947 during the cold war). My next paper was concerned with thermoelectric effects in the superconducting state [11], and its destiny seems to be unusual and strange. The point is that 60 years have passed, but some predictions made in that work have never been verified and thermoelectric effects in the superconducting state have not been adequately investigated. I myself returned to these problems more than once, but made no significant progress. Appeals

² True, somewhat earlier I had considered the problem of light scattering in Helium II [8] on the basis of the Landau theory [7].

addressed to other physicists have only a minor effect, for the problem is out of fashion. Here I would like to take advantage of my last opportunity to bring it to the attention of physicists. Section 4 below is concerned with this question.

However, the question of thermoelectric effects in superconductors, while interesting, is still a particular problem, which evidently emerges only in the presence of a temperature gradient. Furthermore, at that time there existed no thorough theory of superconductivity even under thermodynamic equilibrium. The fact is that the well-known London theory advanced in 1935 [12] (it will also be discussed in Section 4 of this lecture) yielded much, and is widely employed under certain conditions even nowadays [5, 6, 13], but it is absolutely insufficient. The last-mentioned circumstance was largely elucidated in my next work performed as far back as 1944 [14]. Specifically, the London theory is inapplicable in a strong magnetic field (in the theory of superconductivity, the field is termed strong when it is on the order of the critical magnetic field H_c ; we are dealing with type-I superconductors). From the London theory it follows also that the surface energy at the interface between the normal and superconducting phases is negative, and to attain positiveness one is forced to baselessly introduce some additional and, moreover, high surface energy of non-electromagnetic origin. Therefore, it became evident that the London theory had to be generalized. This problem was solved in 1950 in the Ψ -theory of superconductivity [3].³ This brings up the question, which has been repeatedly addressed to me: why did it take five years after the work in Ref. [14], in which the necessity of generalizing the London theory was recognized, to construct the Ψ -theory? Of course, I cannot answer this question as regards other physicists. As to myself, to some extent I was nearing my objective, as described in the article [1]. But, I believe, the main reason for the slowness of this process lay with the fact that I did not focus my attention on the theory of superconductivity. Theoretical physicists have the good fortune to be able to work almost simultaneously in different directions and in general to move from one subject to another. Specifically, in the period from 1944 to 1950, apart from superconductivity and superfluidity, I was engaged in radio wave propagation in the ionosphere (plasma), radiofrequency solar radiation, light scattering in liquids, the theory of transition radiation (I.M. Frank and I jointly called attention to the existence of this effect), the relativistic theory of higher-spin particles (in part jointly with Tamm), undulator radiation, the theory of ferroelectrics, and other things. Of special note is the fact that ferroelectric effects (as applied primarily to BaTiO_3) were considered [15] on the basis of the Landau theory of phase

³ As already mentioned, this theory is commonly referred to as the Ginzburg–Landau theory. But I resort to the term Ψ -theory of superconductivity, because it seems to me that using one's own name rings, at least in Russian, somewhat pretentiously. Furthermore, a similar theory, as applied to superfluidity, was jointly elaborated in my work not with Landau, but with L.P. Pitaevskii and A.A. Sobyanin.

transitions, and this direction subsequently progressed (see article 5 in the collection [2]).⁴

The Ψ -theory of superconductivity [3] is, if you like, an application of the Landau theory of phase transitions to superconductivity. In this case, some scalar complex Ψ function fulfills the role of the order parameter. By virtue of the foregoing I restrict myself to giving the equations employed for Ψ and the vector electromagnetic field potential \mathbf{A} (as is generally known, $\text{rot } \mathbf{A} = \mathbf{H}$, where \mathbf{H} is the magnetic field strength, which does not differ from the magnetic induction \mathbf{B} in this case; furthermore, advantage is taken of the gauge $\text{div } \mathbf{A} = 0$):

$$\frac{1}{2m^*} \left(-i\hbar\nabla - \frac{e^*}{c} \mathbf{A} \right)^2 \Psi + \alpha\Psi + \beta |\Psi|^2 \Psi = 0, \quad (1)$$

$$\begin{aligned} \Delta\mathbf{A} &= -\frac{4\pi}{c} \mathbf{j}_s, \\ \mathbf{j}_s &= -\frac{ie^*\hbar}{2m^*} (\Psi^* \nabla \Psi - \Psi \nabla \Psi^*) - \frac{(e^*)^2}{m^* c} |\Psi|^2 \mathbf{A}. \end{aligned} \quad (2)$$

We consider an equilibrium or, in any case, a stationary state, and we assume that the normal current density in the superconductor is $\mathbf{j}_n = 0$ (the total current density is $\mathbf{j} = \mathbf{j}_s + \mathbf{j}_n$, where \mathbf{j}_s is the superconducting current density). Furthermore, at the superconductor-vacuum interface we impose the boundary condition

$$\mathbf{n} \left(-i\hbar\nabla - \frac{e^*}{c} \mathbf{A} \right) \Psi = 0, \quad (3)$$

where \mathbf{n} is the normal to the interface.

In the vicinity of the critical temperature T_c , at which there occurs the normal-to-superconducting phase transition in the equilibrium case, in the Ψ -theory it can (and even must) be assumed that

$$\alpha = \alpha'_c(T - T_c), \quad \beta = \beta(T_c) \equiv \beta_c > 0, \quad \alpha'_c > 0 \quad (4)$$

and the superconductor behavior is determined by the parameters

$$\delta_0 = \sqrt{\frac{m^* c^2 \beta_c}{4\pi(e^*)^2 |\alpha|}}, \quad \alpha = \frac{m^* c}{e^* \hbar} \sqrt{\frac{\beta_c}{2\pi}} = \frac{\sqrt{2} e^*}{\hbar c} H_{cm} \delta_0^2. \quad (5)$$

Here, δ_0 is the depth of penetration of the weak magnetic field $H \ll H_{cm}$ and H_{cm} is the critical magnetic field for massive samples (earlier, mention was made of the critical field H_c , which, say, for films is stronger than H_{cm}).

Since the Ψ -theory is phenomenological, the values of mass m^* and charge

⁴ For more details on the above-mentioned and other works of mine, see the article “A Scientific Autobiography – an Attempt” in the book [16].

e^* are beforehand unknown. In this case, since Ψ is not an observable quantity (among the observable quantities are, in particular, the δ_0 and H_{cm} quantities), the mass can be arbitrarily selected: it is not among the measurable (observable) quantities. The question of choice of the e^* value is very interesting and intriguing. It seemed to me from the outset that e^* is some effective charge, which may be different from the electron charge or, as is said on occasion, the free-electron charge e . However, Landau did not see why e^* should be different from e , and in our paper [3] it is written as some compromise that “there are no grounds to believe that the charge e^* is different from the electron charge”. I remained of my opinion and saw the way to solve this question was to compare the theory with experiment. Specifically, the charge e^* enters in expression (5) for α , where δ_0 and H_{cm} are measured by experiment; at the same time, α enters into the expression for the surface energy σ_{ns} , for the depth of penetration in the strong field (the field $H \gtrsim H_{cm}$), and for the limiting fields of the overcooling and overheating of superconducting samples. Following the path of comparing the theory with experiment, I arrived at the conclusion [17] that $e^* = (2-3) e$. When discussed this result with Landau, he raised an objection, which he had evidently been guided by before, though had not advanced it. Specifically, with the charge e^* assumed to be an effective quantity like, say, the effective mass m_{eff} in the theory of metals and semiconductors, the effective charge may and, generally speaking, will depend on the coordinates, because the parameters that characterize the semiconductor are functions of the temperature, the pressure, and the composition, which in turn may depend on the coordinates \mathbf{r} . If $e^*(\mathbf{r})$, the gauge (gradient) invariance of equations (1)–(2) of the Ψ -theory is lost. I did not find objections to this remark, and in article [17] outlined the situation (reporting Landau’s opinion, naturally, with his permission). The solution, however, was quite simple. After the advent of the Bardeen–Cooper–Schrieffer (BCS) theory in 1957 [18], it became clear that in superconductors there occurs ‘pairing’ of electrons with opposite momenta and spins (I imply the simplest case). The resultant ‘pairs’, which are sometimes referred to as the Cooper pairs, possess zero spin and are Bose particles or, to be more precise, quasi-particles. The Bose-Einstein condensation of these pairs is responsible for the origin of superconductivity. By the way, as early as 1952 I noted [19] that the charged Bose gas would behave like a superconductor, but did not arrive at the idea of pairing. Interestingly, it had been advanced [20, 21] even before Cooper [22]. It is immediately apparent from the BCS theory that the role of charge in the theory of superconductivity should supposedly be played by the pair charge, i.e., $2e$. This fact was proved by Gor’kov [23], who derived the Ψ -theory equations from the BCS theory. Therefore, Landau was right in the sense that the charge e^* should be universal and I was right in that it is not equal to e . However, the seemingly simple idea that both requirements are compatible and $e^* = 2e$ occurred to none of us. After the event one may be ashamed of this blindness, but this is by no means a rare occasion in science, and it is not that I am ashamed of this blindness, but I am rather disappointed that it did take place.

Many results were obtained in our work [3]. For small values of the parameter κ we calculated the surface energy σ_{ns} and pointed out that it lowers with increasing κ and vanishes when $\kappa = \kappa_c = 1/\sqrt{2}$. Relying on the available experimental data we believed that for pure superconductors $\kappa < \kappa_c$, and this is generally correct. In any case, we considered in detail only the superconductors with $\kappa < \kappa_c$, which now are termed type-I superconductors. Subsequently I would also restrict myself to the investigation of type-I superconductors (a certain exception is Ref. [24]). In 1950, as well as previously, the superconducting alloys were known to usually behave in a significantly different manner than pure superconductors. Particularly clear data concerning alloys were obtained by L.V. Shubnikov⁵ and his collaborators in Kharkov in the mid-30s (see references and the results in [25]; this material was also touched upon in [26]; for more details see [27]). In [27], use is made of the term ‘Shubnikov phase’ for the alloys investigated by Shubnikov. However, an understanding of the situation was lacking, and Landau and I, like many others, believed that alloys are an ‘unsavory business’, and did not take an interest in them, restricting ourselves to the materials with $\kappa < \kappa_c$ for which $\sigma_{ns} > 0$, i.e., type-I superconductors. True, as noted in A. Abrikosov’s paper [4] and in [5], Landau hypothesized that alloys are the ones where $\kappa > 1/\sqrt{2}$, i.e., they are type-II superconductors according to present-day concepts.

The solution of different problems on the basis of Ψ -theory equations was our concern in the bulk of our work [3]. Apart from the above-mentioned question of the energy σ_{ns} , we considered primarily the behavior of superconducting plates and films in the external magnetic field and in some cases in the presence of current, and in doing this compared the theory with experiment. Subsequently, Landau took no interest in such calculations and in general in the development of the Ψ -theory. My own effort made in this direction is described in [1]. Here, I restrict myself to the mention of a fairly evident yet important generalization of the Ψ -theory [3], in which superconductors were assumed to be isotropic, to the anisotropic case [28]. Furthermore, investigations were made of the overheating and overcooling of superconductors in the magnetic field [29] and of the quantization of magnetic flux in the case of a superconducting cylinder with an arbitrary wall thickness [30], and the Ψ -theory was compared with experiment after the construction of the BCS theory [31]. Of special note is Ref. [32], which was developed in [33], had little bearing on the Ψ -theory, and applied to ferromagnetic superconductors. Such superconductors had not been discovered by that time, and Ref. [32] put forward the explanation for this fact related to the inclusion of magnetic energy. Subsequently (after the construction of the BCS theory), it became clear that the emergence of superconductivity in ferromagnetics is also hampered due to spin interaction. I was not engaged in that problem, but would like to call attention to the following. Certain considerations were given in [32], which allowed changing the role of the magnetic factor (the use

⁵ In 1937, when Stalin’s terror was in full swing, L.V. Shubnikov was arrested and shot.

of thin films and materials with a relatively strong coercive force). I do not think that anyone has given attention to these possibilities, for old papers are seldom read. Of course, I do not feel sure that at the present stage one can find something of interest in [32, 33] – I would just like these papers to be looked at.

In long ago 1943, I engaged in the study of superconductivity because at that time this phenomenon appeared to be the most mysterious one in the physics of the condensed state. But after the construction of the Ψ -theory, and especially of the BCS theory, the picture generally became clear as regards the materials known at that time. That is why I lost particular interest in superconductivity, though I worked in this area episodically (see, for instance, [30, 34]). My interest was rekindled in 1964 in connection with the formulation of the problem of the feasibility of high temperature superconductors (HTSCs). Mercury – the first superconductor discovered in 1911 – possesses $T_c = 4.15$ K, while the boiling temperature of ^4He at atmospheric pressure is $T_b, {}^4\text{He} = 4.2$ K. By the way, from 1908 to 1923, for fifteen long years, liquid helium was obtained only in Leiden, and low-temperature physics research was pursued on a very small scale, judged by present-day standards. For the example it would suffice to note that the bibliography given at the end of monograph [26] contains about 450 references to the papers on superconductivity (or, sometimes, related problems) over the period from 1911 to 1944; among them, only 35 references fall within the 1911–1925 period. Meanwhile, after 1986–1987, when high-temperature superconductivity was discovered, during the 10 subsequent years approximately 50,000 papers were published, i.e., about 15 papers per day (!).

There can be no doubt that immediately after the discovery and first investigations of superconductivity the question arose of why this phenomenon is observed only at low temperatures or, in other words, helium temperatures. Naturally, there was no way to provide the answer until the nature of superconductivity was understood, i.e., till the construction of the BCS theory in 1957 [18]. The following expression was derived for the critical temperature in this theory:

$$T_c = \theta \exp \left(-\frac{1}{\lambda_{\text{eff}}} \right), \quad (6)$$

where $k_B\theta$ is the energy range near the Fermi energy $E_F = k_B\theta_F$, in which the conduction electrons (more precisely, the corresponding quasi-particles) are attracted together, which is responsible for pair production and the instability of the normal state; furthermore, in the simplest case, $\lambda_{\text{eff}} = \lambda = N(0)V$, where $N(0)$ is the electronic level density near the Fermi surface in the normal state and V is some average matrix element of electron interaction which corresponds to the attraction. In the BCS theory, in its initial form, the ‘coupling constant’ λ_{eff} and, specifically, λ is assumed to be small (‘weak coupling’), i.e.,

$$\lambda \ll 1. \quad (7)$$

As regards the temperature θ , in the BCS theory it was assumed that

$$\theta \sim \theta_D , \quad (8)$$

where θ_D is the Debye temperature of the metal, for the interelectron attraction was thought to be due to electronphonon interaction (as is generally known, the highest phonon energy in a solid is of the order of $k_B\theta_D$). Typically, $\theta_D \lesssim 500$ K and $\lambda \lesssim 1/3$; whence it follows, according to (6), that $T_c \lesssim 500 \exp(-3) = 25$ K or more generally

$$T_c \lesssim 30 - 40 \text{ K} . \quad (9)$$

Defining all this more precisely would be out of place here. But it seems to me that the aforesaid will suffice to understand why the condition (9) is fulfilled for typical metals, and even safely fulfilled. In particular, prior to the discovery of high temperature superconductivity in 1986–1987, all attempts to discover or produce a superconductor with the highest possible critical temperature had led in 1973 to the production of only the Nb_3Ge compound with $T_c = 23 - 24$ K (of course, I do not endeavor to find the exact values of various parameters; they depend on the purity and processing of samples, etc.).

ON HIGH-TEMPERATURE AND ROOM-TEMPERATURE SUPERCONDUCTORS (HTSC AND RTSC)

The advent of the BCS theory made it possible to envisage the feasibility of a radical elevation of the critical temperature. It may be that I am not familiar with some facts, but to my knowledge this question was clearly and constructively posed for the first time by Little in 1964 [35]. Being forced to outline the following part of this section quite schematically owing to the lack of space, I can mention that Little proposed considering the possibility of replacing the phonon mechanism of attraction between conduction electrons with the same attraction arising from the interaction with bound electrons present in the same system. I call this mechanism excitonic or electron-excitonic; to state it in plain terms, we are dealing with the replacement of phonons with excitons – excitations in the system of bound electrons. True, this term is not universally used in the literature. In his case, Little employed a quasi-one-dimensional model, in which some conducting ‘spine’ was surrounded by side ‘polarizers’, say, organic molecules. For electronic excitons or, in other words, for the excited states of bound electrons, the characteristic temperatures $\theta_{ex} = E_{ex}/k_B \lesssim \theta_F \sim 10^4 - 10^5$ K and, in any case, the values $\theta_{ex} \sim 10^4$ K are quite realistic. It is therefore evident that replacing $\theta \sim \theta_D$ in (6) with $\theta \sim \theta_{ex}$ gives us the values $T_c \lesssim 10^3$ K (when, say, $\lambda \sim 1/3$). Of course, these are no more than words, for it is still unclear how to realize the Little model, and this has never been accomplished. Furthermore, it became clear that the fluctuations in quasi-one-dimensional systems are so strong that the transition to the superconducting state is unlikely to occur. However, having

familiarized myself with the paper [35], I put forward straight away [36] a quasi-two-dimensional model, wherein a plane conductor is in contact with a dielectric, say, a dielectric film. We termed the development of this version – the alternation of thin conducting layers with dielectric layers – a ‘sandwich’. Going over from the quasi-one-dimensional model to the quasi-two-dimensional model was not accidental, for immediately before the work [36] D.A. Kirzhnits⁶ and I had considered [37], not in connection with the high-temperature superconductivity problem, the problem of two-dimensional (surface) superconductivity. By the way, this problem is still of interest in itself, but I cannot enlarge on it for the lack of space and I will restrict myself to giving references [37, 38].

Compared to quasi-one-dimensional systems, quasi-twodimensional systems have the advantage that they exhibit significantly weaker fluctuations that destroy superconductivity. We took up the quasi-two-dimensional version [36, 39]. More precisely, at FIAN (the P.N. Lebedev Physical Institute of the USSR Academy of Sciences) a group of theorists turned to the high-temperature superconductivity problem in the broad sense, considering all issues and possibilities known to us. The fruits of this labor were represented in the monograph [40]; even its English version (1982) appeared 4–5 years before the experimental realization of high-temperature superconductors [41, 42] in 1986–1987. If the consideration of different models and possibilities is omitted, the most significant quantitative finding of our work, which is primarily due to Kirzhnits, is the crystal stability condition. The point is that the main objection against the possibility of developing a high-temperature superconductor was the anxiety that the crystal lattice will be unstable for the metal parameter values required to obtain a high-temperature superconductor, i.e., for a material with $T_c > T_{b, N_2} = 77.4 \text{ K}$ ⁷. When the problem is formulated in terms of the longitudinal material permittivity $\varepsilon(\omega, \mathbf{q})$, where ω is the frequency and \mathbf{q} is the wave vector (we restrict our consideration to an isotropic body here), the production of electron pairs necessitates, roughly speaking,

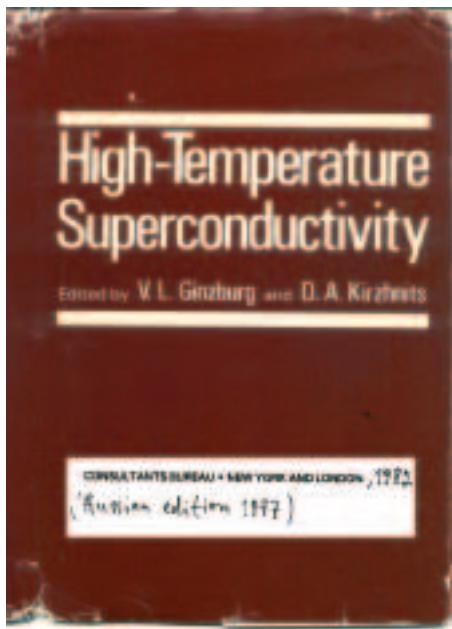
that the interelectron interaction $V = \frac{e^2}{\varepsilon(0, q) r}$ should be negative, i.e., should correspond to attraction. But this corresponds to the requirement that $\varepsilon(0, q) < 0$. Meanwhile, on the ground of some considerations it was believed that the lattice would be stable when

$$\varepsilon(0, q) > 0. \quad (10)$$

True, on closer examination (see [1, 40]) it was found that superconductivity is also possible under the condition (10), but the T_c values would turn out

⁶ Unfortunately, the outstanding theoretical physicist D.A. Kirzhnits deceased untimely in 1998.

⁷ I do not know whether there exists the commonly accepted definition of what superconductor can be regarded as a high-temperature one. In my opinion, HTSC takes place when $T_c > 77.4 \text{ K}$, i.e. is higher than the boiling temperature of nitrogen at atmospheric pressure.



to be moderate, even below the estimate (9). In [40] and references therein it was found that the correct stability condition for $q \neq 0$ is of the form

$$\frac{1}{\varepsilon(0, q)} \leq 1, \quad (11)$$

i.e. is fulfilled when either of two inequalities

$$\varepsilon(0, q) > 1, \quad \varepsilon(0, q) < 0. \quad (12)$$

takes place. In other words, any negative values of $\varepsilon(0, q)$ are admissible from the standpoint of stability and there are no limitations on T_c . To be more precise, up to now we do not know of such limitations. The following conclusion was drawn from our work, which is contained in Chapter 1 in book [40] written by me:

"On the basis of general theoretical considerations, we believe at present that the most reasonable estimate is $T_c \lesssim 300$ K; this estimate being, of course, for materials and systems under more or less normal conditions (equilibrium or quasi-equilibrium metallic systems in the absence of pressure or under relatively low pressures, etc.). In this case, if we exclude from consideration metallic hydrogen and, perhaps, organic metals, as well as semimetals in states near the region of electronic phase transitions, then it is suggested that we should use the exciton mechanism of attraction between the conduction electrons.

In this scheme, the most promising materials – from the point of view of the possibility of raising T_c – are, apparently, layered compounds and dielec-

tric–metal– dielectric sandwiches. However, the state of the theory, let alone the experiment, is still far from being such as to allow us to regard other possible directions as being closed, in particular, the use of filamentary compounds. Furthermore, for the present state of the problem of high-temperature superconductivity, the most sound and fruitful approach will be one that is not preconceived, in which attempts are made to move forward in the most diverse directions.

The investigation of the problem of high-temperature superconductivity is entering into the second decade of its history (if we are talking about the conscious search for materials with $T_c \geq 90$ K with the use of the exciton and other mechanisms). Supposably, there begins at the same time a new phase of these investigations, which is characterized not only by greater scope and diversity, but also by a significantly deeper understanding of the problems that arise. There is still no guarantee whatsoever that the efforts being made will lead to significant success, but a number of new superconducting materials have already been produced and are being investigated. Therefore, it is, in any case, difficult to doubt that further investigations of the problem of high-temperature superconductivity will yield many interesting results for physics and technology, even if materials that remain superconducting at liquid nitrogen (or even room) temperatures will not be produced. Besides, as has been emphasized, this ultimate aim does not seem to us to have been discredited in any way. As may be inferred, the next decade will be crucial for the problem of high-temperature superconductivity.” This was written in 1976. Time passed, but the multiple attempts to find a reliable and reproducible way of creating a high-temperature superconductor have been unsuccessful. As a result, after the flash of activity came a slackening which gave cause for me to characterize the situation in a popular paper [43] published in 1984 as follows:

“It somehow happened that research into high-temperature superconductivity became unfashionable (there is good reason to speak of fashion in this context since fashion sometimes plays a significant part in research work and in the scientific community). It is hard to achieve anything by making admonitions. Typically it is some obvious success (or reports of success, even if erroneous) that can radically and rapidly reverse attitudes. When they smell success, the former doubters, and even dedicated critics, are capable of turning coat and becoming ardent supporters of the new work. But this subject belongs to the psychology and sociology of science and technology.

In short, the search for high-temperature superconductivity can readily lead to unexpected results and discoveries, especially since the predictions of the existing theory are rather vague.”

I did not expect, of course, that this ‘prediction’ would come true in two years [41, 42]. It came true not only in the sense that high-temperature superconductors with $T_c > T_{b, N_2} = 77.4$ K were obtained, but also, so-to-say, in the social aspect: as I have mentioned above, a real boom began and a ‘high-temperature superconductivity psychosis’ started. One of the manifestations of the boom and psychosis was the almost total oblivion of everything that

had been done before 1986, as if the discussion of the high-temperature superconductivity problem had not begun 22 years before [35, 36]. I have already dwelt on this subject above and in the papers [44, 45] and would not like to return to it here. I will only note that J. Bardeen, whom I have always respected, treated the high-temperature superconductivity problem with understanding both before and after 1986 (see [46]; this article was also published in book [16]).

The foregoing in no way implies that our group or I pretend to a practical contribution of great importance to the development of high-temperature superconductivity. At the same time I believe that Little's works and ours have played a significant role in the formulation of the problem and have drawn attention to it. The solution of the problem was obtained to a large measure accidentally. The proposal to employ layered compounds was reasonable and promising, but neither I nor, to my knowledge, anybody else proposed the use of precisely the cuprates. Other layered compounds investigated do not belong to high-temperature superconductors. The following fact serves to illustrate the accidental, to a certain measure, character of discovery of high-temperature superconductivity. As far back as 1979, in one of the institutes in Moscow they produced and investigated [47] a $\text{La}_{1.8}\text{Sr}_{0.2}\text{CuO}_4$ ceramic, which was close to that investigated by Bednorz and Muller, with $T_c \simeq 36$ K [48]. However, the authors of [47] measured the resistance of their samples at temperatures not lower than the liquid-nitrogen temperature and therefore did not discover their superconductivity. From the above one may draw a trivial conclusion that all newly produced materials should be 'tested' for superconductivity. Also evident is another conclusion, namely, that even nowadays it is possible to make a major discovery and next year be awarded a Nobel Prize for it without gigantic facilities and the work of a large group. This should be a source of inspiration, particularly for young people.

The present situation in solid-state theory does not allow us to calculate the value of T_c nor of other superconductor parameters, with the possible exception of a metallic hydrogen yet to be produced. Moreover, for more than 15 years the mechanism of superconductivity in cuprates has remained obscure. I should remark that, despite the fact that I counted on the excitonic mechanism in high-temperature superconductivity research, the role of this mechanism in the known high-temperature superconductors is still completely unclear. In this case, in high-temperature superconductors (in cuprates) with $T_c < 170$ K (the highest-known value $T_c \simeq 165$ K was attained back in 1994 in the $\text{HgBa}_2\text{Ca}_2\text{Cu}_3\text{O}_{8+x}$ cuprate under high pressure), as I see it, the electron-phonon mechanism of pairing may prove to be the dominant one. This possibility has previously been underestimated (in particular, by me), since the estimate (9) has served as a guide. But it is valid only for a weak coupling (7). For a strong coupling (i.e., when $\lambda_{\text{eff}} \geq 1$), formula (6) is no longer applicable, but even from this formula it is clear that T_c increases with λ_{eff} . The generalization of the BCS theory [18] to the strong-coupling case [49] enables us to investigate the corresponding possibilities. Their analysis (see particularly [50] and references therein and in [1]) suggests that the elec-

tronphonon mechanism in cuprates may well ensure superconductivity with $T_c \lesssim 200$ K owing to the high θ_D and λ_{eff} values. At the same time, the electron-phonon interaction alone is supposedly insufficient in the context of so-called *d* pairing and maybe other special features of superconductivity in cuprates. However, the role of other possibilities (spin interactions, excitonic interaction) is unclear. Of course, it would be out of place to discuss this vital topical problem here. I only want, on the one hand, to emphasize that the longstanding disregard of electron-phonon interaction in cuprates has always seemed and now seems unjustified to me (see [51]). On the other hand, the likelihood of attaining, on the basis of the electron – phonon mechanism, the values $T_c \sim 300$ K, and this is room-temperature superconductivity (RTSC), appears to be small, as with the use of the spin mechanism. At the same time, the excitonic mechanism, as far as I know, does not provoke objections for $T_c \sim 300$ K, either. That is why I pin my hopes on precisely this mechanism for the attainment of room-temperature superconductivity. However, all this is no more than an intuitive judgment.

The creation of high-temperature superconductivity had been my dream for 22 years, even with no guarantee that the goal was at all attainable and, in particular, attainable in the foreseeable future. In my view, obtaining room-temperature superconductivity now occupies the same place.

THERMOELECTRIC PHENOMENA IN THE SUPERCONDUCTING STATE

The first attempt to observe thermoelectric phenomena and, specifically, thermoelectric current or thermal electromotive force in a nonuniformly heated circuit of two superconductors, to my knowledge, was made by Meissner [52] in 1927. He arrived at the conclusion that the thermoelectric effect is completely absent for superconductors. When I took an interest in this problem in 1943, this viewpoint was generally accepted (see, for instance, [53] and especially the first and later editions of the book [25]). However, I encountered this statement more recently as well. Meanwhile, this conclusion is erroneous, which was pointed out in my work [11] published as far back as 1944.

The point is that the superconducting state can carry, apart from a superconducting current \mathbf{j}_s , a normal current \mathbf{j}_n as well. This normal current is carried by ‘normal electrons’, i.e., electron- or hole-type quasi-particles present in the metal in both the normal and superconducting states. In the superconducting state, the density of such normal quasiparticles depends strongly on the temperature and, generally, tends to zero as $T \rightarrow 0$. These notions, which are sometimes referred to as the two-liquid model, can be traced back to paper [54]. An isotropic non-superconductor or, more precisely, an isotropic metal residing in a normal state, can carry only the current with a density

$$\mathbf{j} = \sigma \left(\mathbf{E} - \frac{\nabla \mu}{e} \right) + b \nabla T, \quad (13)$$

where μ is the chemical potential of the electrons and \mathbf{E} is the electric field. In the superconducting state, for a normal current we have (for more details, see [55])

$$\mathbf{j}_n = \sigma_n \left(\mathbf{E} - \frac{\nabla \mu}{e} \right) + b_n \nabla T. \quad (14)$$

At the same time, the superconducting current density \mathbf{j}_s in the London theory [12] approximation, to which we restrict ourselves here (naturally, this is precisely the approximation used in [11]), obeys the equations

$$\text{rot}(\Lambda \mathbf{j}_s) = -\frac{1}{c} \mathbf{H}, \quad (15)$$

$$\frac{\partial(\Lambda \mathbf{j}_s)}{\partial t} = \mathbf{E} - \frac{\nabla \mu}{e}, \quad (16)$$

where $\Lambda = m/(e^2 n_s)$ is somewhat a constant, with n_s being the ‘superconducting electron’ density (so that $\mathbf{j}_s = e n_s \mathbf{v}_s$, where \mathbf{v}_s is the velocity); in this scheme, the field penetration depth is

$$\delta\lambda = \sqrt{\frac{\Lambda c^2}{4\pi}} = \sqrt{\frac{mc^2}{4\pi e^2 n_s}}.$$

Notice that this is some simplification, for different chemical potentials μ_n and μ_s should in fact be introduced in Eqns (14) and (16), respectively, for the normal and superconducting electrons. In addition, yet another term (generally, not large) proportional to ∇j_s^2 (see [55]) figures in Eqn (16). When the superconductor is nonuniform, the parameter Λ depends on the coordinates.

As is clear from Eqn (16), in the stationary case, in the superconductor

$$\mathbf{E} - \frac{\nabla \mu}{e} = 0 \quad (17)$$

and, in view of Eqn (14),

$$\mathbf{j}_n = b_n(T) \nabla T. \quad (18)$$

Therefore, the thermoelectric current \mathbf{j}_n in no way vanishes in the superconducting state. However, this current is not directly observable in the simplest case, because it is compensated for by the superconducting current \mathbf{j}_s . Let us consider a uniform superconducting rod, one end of the rod residing at a temperature T_2 and the other at a temperature $T_1 < T_2$ (Fig. 1). Then, in the normal state (i.e., when $T_1 > T_c$), since there is no closed circuit, from Eqn (13) we have (see Fig. 1a)

$$\mathbf{j} = 0, \quad \mathbf{E} - \frac{\nabla\mu}{e} = -\frac{b}{\sigma} \nabla T. \quad (19)$$

In the superconducting state (for $T_2 < T_c$),

$$\begin{aligned} \mathbf{j} &= \mathbf{j}_s + \mathbf{j}_n = 0, \quad \mathbf{j}_s = -\mathbf{j}_n = -b_n \nabla T, \\ \mathbf{H} &= 0, \quad \mathbf{E} - \frac{\nabla\mu}{e} = 0. \end{aligned} \quad (20)$$

True, near the rod ends, where \mathbf{j}_s transforms to \mathbf{j}_n or vice versa, uncompensated charges (charge imbalance effect) emerge, and therefore the field \mathbf{E} is not equal to $\nabla\mu/e$; in what follows I ignore this feature.

An important point is that the thermoelectric current \mathbf{j}_n exists in the uniform case in the superconducting state (Fig. 1b), but the field $H = 0$. When the superconductor is nonuniform or anisotropic, the currents \mathbf{j}_s and \mathbf{j}_n do not in general compensate each other completely, and an observable thermoelectric magnetic field emerges, which was noted in [11]. In days of old (60 years ago!), as noted above, the case of alloys was considered to be unsavory and it was even unclear whether the Londons equation could be applied to alloys. That is why I restricted myself to a brief consideration of a bimetallic plate (say, of two different superconductors fused or soldered together: this juncture is the alloy) in the presence of a temperature gradient (see also § 16 in [26] and [55]). In this case, because the parameter Λ depends on the coordinates (evidently, the Λ parameter is different for different metals), along the junction line there emerges an uncompensated current \mathbf{j} and hence the magnetic field \mathbf{H} , which is perpendicular to the plate and the junction line (Fig. 2). Considered in greater detail in [11] and [26] was the case of an anisotropic superconductor. To this end, the Londons equations were generalized in a rather trivial way by replacing the scalar Λ with the tensor Λ_{ik} (for isotropic and cubic metals, $\Lambda_{ik} = \Lambda\delta_{ik}$). When the temperature gradient ∇T in a plateshaped noncubic superconducting crystal is not directed along the symmetry axis, there emerges a current \mathbf{j} flowing around the plate and a magnetic field \mathbf{H}_T transverse to the plate and proportional to $(\nabla T)^2$. In principle, this field is not difficult to observe with modern techniques. Curiously enough, this is an interesting effect, which in addition makes it possible to measure the thermoelectric coefficient $b_n(T)$ or, more precisely, the components of its generalized tensor $b_{n,ik}(T)$. More than 30 years ago I managed to convince W. Fairbank to stage the corresponding experiment, and its results remain, as far as I know, the only ones on this subject [56]. Unfortunately, this work did not make things clear [55, 57]. I am amazed by the fact that nobody has taken an interest in this question even after the fabrication of strongly anisotropic high-temperature superconductors. Evidently, such is the force of fashion in science, too.

True, a certain interest was attracted precisely by the isotropic superconductors, in essence, as applied to a more or less conventional thermoelectric current (Fig. 3a). For this circuit is equivalent to the ‘circuit’ of Fig. 3b. For

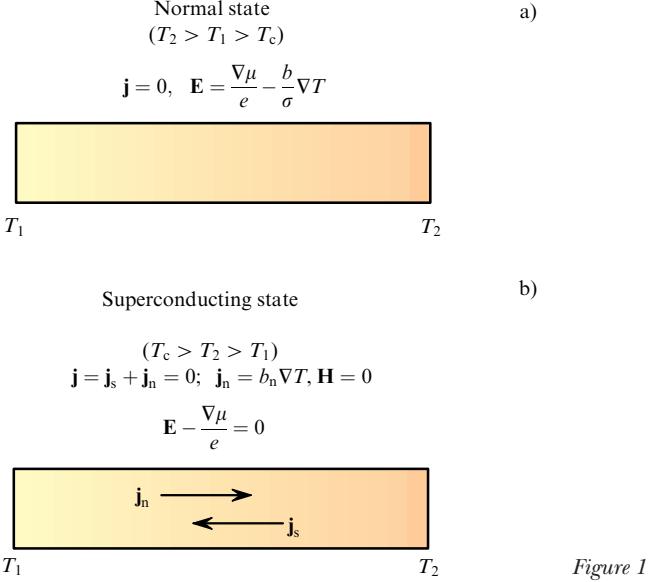


Figure 1.

this circuit it is easy to show [58, 59] (the derivation is also given in [55]) that the magnetic flux $\Phi = \int \mathbf{H} d\mathbf{S}$ through the opening is

$$\Phi = k\Phi_0 + \Phi_T, \quad \Phi_T = \frac{4\pi}{c} \int_{T_1}^{T_2} (b_{n,\text{II}} \delta_{\text{II}}^2 - b_{n,\text{I}} \delta_{\text{I}}^2) dT,$$

$$\Phi_0 = \frac{hc}{2e} = 2 \times 10^{-7} \text{ G cm}^2, \quad k = 0, 1, 2, 3 \dots . \quad (21)$$

Here, the indices I and II refer to the superconducting metals I and II, δ_{I} and δ_{II} are the field penetration depths for these metals, $b_{n,\text{I}}$ and $b_{n,\text{II}}$ are the corresponding coefficients $b_n(T)$ in formula (18), and Φ_0 is the so-called flux quantum. The configuration in Fig. 3b is essentially equivalent to the bimetallic plate in Fig. 2 with $k = 0$, i.e., without an opening. Unfortunately, I did not recognize this at the time (i.e., in [11, 26]).

If we assume for simplicity that $(b_n \delta^2)_{\text{II}} \gg (b_n \delta^2)_{\text{I}}$ and $\delta_{\text{II}}^2 = \delta_{\text{II}}^2(0)(1 - T/T_{c,\text{II}})^{-1}$, from expression (21) we obtain

$$\Phi_T = \frac{4\pi}{c} b_{n,\text{II}}(T_c) \delta_{\text{II}}^2(0) T_c \ln \left(\frac{T_c - T_1}{T_c - T_2} \right). \quad (22)$$

If we substitute the known values $b_n(T_c)$ and $\delta(0)$ for $\ln(T_c - T_1)/(T_c - T_2) \sim 1$ in expression (22) we arrive at an estimate $\Phi_T \sim 10^2 \Phi_0$. This flux is easy to measure, which was done in several papers (see [1, 55] and references therein). However, the flux Φ_T observed in some more complex configuration of the superconducting circuit was found to be orders of magnitude higher than the flux given by expressions (21)–(22) and to possess a different

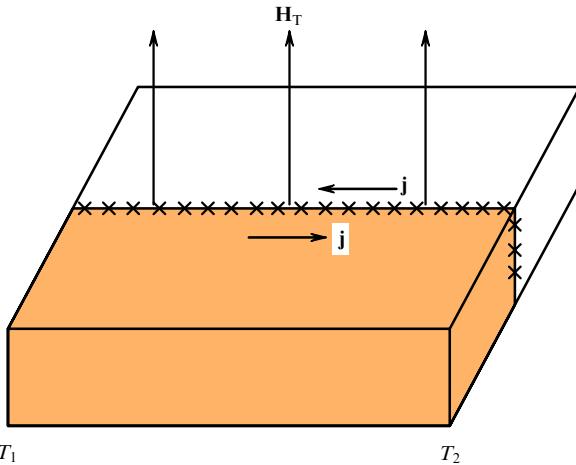


Figure 2.

temperature dependence [60]. The reason for this result has not been elucidated, and different assumptions have been made on that score [61, 62]; see also other references in [1].

It is also pertinent to note that expression (21) and the ensuing formula (22) are obtained under the assumption that the equality $\mathbf{j} = \mathbf{j}_s + \mathbf{j}_n = 0$ is fulfilled throughout the circuit depth (the current $\vec{\mathbf{j}}$ flows only near the surface). Meanwhile, as T_c is approached, the field penetration depth δ increases; as $T \rightarrow T_c$, the depth $\delta \rightarrow \infty$ and the current density \mathbf{j}_n tends to the thermoelectric current density in the normal state, i.e., for $T > T_c$. In these conditions, a more detailed analysis is required to include the charge imbalance effect. This interesting question has not been investigated (for more details, see [1]).

The aforesaid is not the whole story. Even in the simplest case of a uniform superconductor, the existence of a temperature gradient (see Fig. 1b) affects the thermal conduction: since $\mathbf{j}_n \neq 0$, there is bound to be an additional (convective) heat flux $\mathbf{q}_c = -\kappa_c \nabla T$ similar to that occurring in a superfluid liquid. This was noted even in [11] and was, in fact, the initial idea in this work.

The total heat flux in the superconducting state $\mathbf{q} = -\kappa_c \nabla T$, $\kappa = \kappa_{ph} + \kappa_e$, where κ_{ph} is the thermal conductivity coefficient related to the lattice (phonons), κ_e is the electron contribution in the absence of convection (circulation), i.e., subject to the condition $\mathbf{j}_n = 0$, and, as already noted, κ_c is the contribution of circulation. As is generally known, the thermal conductivity coefficient in the normal state is, by definition, measured for $\mathbf{j}=0$, and it is valid to say that $\kappa_c = 0$ (see ⁸). When estimating the κ_c coefficient, I, like others, got tangled up, and now I will restrict myself to a reference to paper [1] and

⁸ It is another matter that, for instance, a semiconductor subjected to the condition $\mathbf{j} = 0$ in the presence of electron and hole conduction can simultaneously carry electron \mathbf{j}_e and hole $\mathbf{j}_h = -\mathbf{j}_e$ currents; we ignore these possibilities.

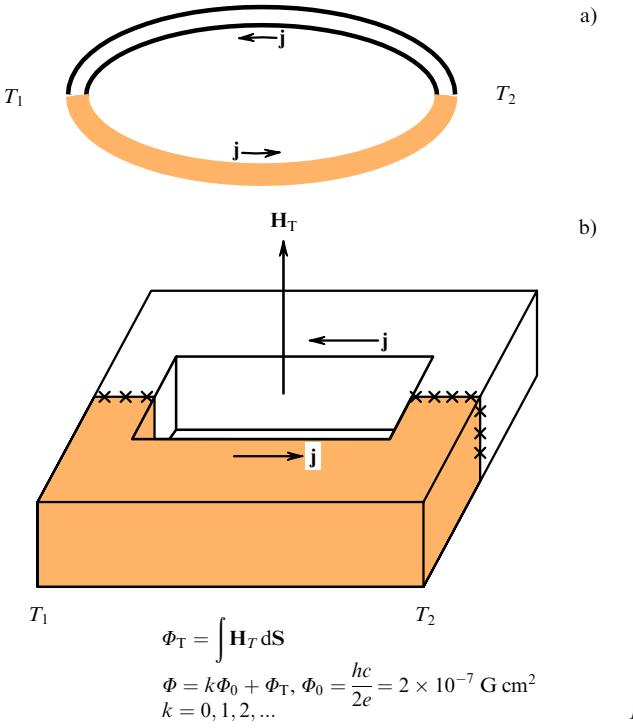


Figure 3.

a remark that in ordinary (not high-temperature) superconductors supposedly $\kappa_c \ll \kappa_e$. The role of κ_c in high-temperature superconductors is unclear to me. Most important of all, it is not clear how to extract κ_c , even if it were possible to determine separately κ_{ph} and $\kappa_{e, tot} = \kappa_e + \kappa_c$ (the total thermal conductivity coefficient κ is measured directly; on the separation of κ_{ph} from $\kappa_{e, tot}$, see [1]).

We have no way of dwelling on the thermoelectric effects in the superconducting state. My aim is to draw attention to this range of questions, which came under the scrutiny of science back in 1927 (see [52] as well as [25]) and under mine in 1944 [11], but which remains largely unclear to date. This is so in spite of a multitude of papers concerned with superconductivity.

SUPERFLUIDITY RESEARCH. Ψ -THEORY OF SUPERFLUIDITY

Superconductivity is, if you please, the superfluidity of a charged liquid or, equivalently, superfluidity is the superconductivity of a noncharged liquid. It is therefore natural that the investigations of both effects have been interrelated. My first work in this area [8], concerned with light scattering in Helium II, was already mentioned above. By the way, there is good reason to revert to this question in light of modern understanding of the fluctuations near the λ point. Several other papers were dealt with in [1]; here, I will consider only the Ψ -theory of superfluidity, albeit with one exception. Namely, I would like



E. F. Zharkov.



D. A. Kirzhnits.

to mention also a proposal made jointly with A.A. Sobyanin⁹ and partly with G.F. Zharkov [63, 64], and then mention the study of the feasibility of observing the thermomechanical circulation effect in a superfluid liquid.

In an annular vessel filled with a superfluid liquid (specifically, the case in point was Helium II), which has two different ‘bottlenecks’ (for instance, narrow capillaries), under a temperature gradient there is bound to emerge a circulation – a superfluid flow engulfing the entire vessel (Fig. 4). By the way, we made the inference about the existence of this effect [63] on the basis of analogy with the thermoelectric effect in a superconducting circuit. As to the inference about the existence of thermoelectric current in a superconducting circuit, I made it [11] at the time on the strength of analogy with the behavior of Helium II under a temperature gradient. The above thermocirculation effect in Helium II has been observed [65] and discussed [64], and, in my view, interesting possibilities were pointed out for future research [64]. However, nobody, as far as I know, has taken an interest in this question during the past 20 years.

After the development of the Ψ -theory of superconductivity [3], the transfer of something similar to the superfluidity case appeared to be rather obvious. At the same time, even before (see, for instance [9]) I was concerned about the behavior of Helium II near the λ point, and the question of the boundary condition for superfluid component velocity v_s was obscure. By the way, Landau, the originator of the theory of phase transitions and superfluidity, for some reason was never concerned with this range of questions, as far as I know. In the Landau theory of superfluidity [7], the velocity v_s along the

⁹ The talented theoretical physicist and public figure Aleksandr Sobyanin prematurely died at the age of 54 in 1997.



E. G. Maksimov.

wall (unlike the normal component velocity \mathbf{v}_n along the wall) does not vanish at the wall: there is some kind of discontinuity. But in this case, it seemed to me, this discontinuity was bound to be related to some surface energy σ_s [66]. However, dedicated experiments [67] showed that the σ_s energy is nonexistent or, in any case, is many orders of magnitude lower than the expected energy [66]. I saw a way out in the assumption that the superfluid component density at the wall $\rho_s(0)$ is zero. Then, the superfluid component flux $\mathbf{j}_s = \rho_s \mathbf{v}_s$ at the wall vanishes despite the fact that \mathbf{v}_s have a discontinuity at the wall. In the Ψ -theory of superfluidity, evidently,

$$\rho_s = m|\Psi|^2, \quad (23)$$

where it may be assumed that $m = m_{\text{He}}$ is the mass of a helium atom (we imply the superfluidity of Helium II) and, in view of the foregoing, the boundary condition at the wall is

$$\Psi(0) = 0, \quad (24)$$

instead of the condition (3) for superconductors. At this stage, as far as I remember, it turned out that L.P. Pitaevskii had independently taken up the Ψ -theory of superfluidity and, naturally, we combined efforts. As a result, the work [68] emerged; I speak of the Ψ -theory of superfluidity constructed in that work as ‘initial’ because I consider below the ‘generalized’ Ψ -theory of superfluidity elaborated together with Sobyanin [69, 70] (see also several other references in [1]).

The initial Ψ -theory of superfluidity [68] is quite similar to the Ψ -theory of superconductivity [3], of course, with the use of the boundary condition (24) and in the absence of the electric charge. In this case, the scalar complex function $\Psi = |\Psi| \exp(i\varphi)$ obeys the equation

$$\oint \mathbf{v}_s d\mathbf{l} = 2\pi \frac{\hbar}{m} k, \quad k = 0, \pm 1, \pm 2, \dots$$

$$\frac{2\pi\hbar}{m^4\text{He}} \approx 10^{-3} \text{ cm}^2 \text{ s}^{-1}$$

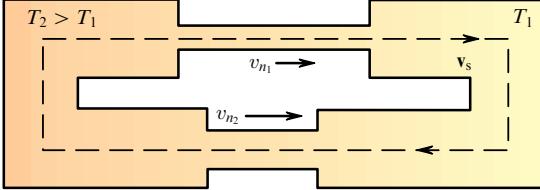


Figure 4.

$$-\frac{\hbar^2}{2m} \Delta \Psi + \alpha(T) \Psi + \beta_\lambda |\Psi|^2 \Psi = 0 \quad (25)$$

and

$$\mathbf{j}_s = \rho_s \mathbf{v}_s = -\frac{i\hbar}{2} (\Psi^* \nabla \Psi - \Psi \nabla \Psi^*) = \hbar |\Psi|^2 \nabla \varphi, \quad (26)$$

i.e., $\mathbf{v}_s = (\hbar/m) \nabla \varphi$, with $m = m_{\text{He}}$ irrespective of how Ψ is normalized (see [1, 68]).

Furthermore, the correlation length ξ denoted as l in [68] is (T_λ is the temperature of the λ point)

$$\xi(T) = \frac{\hbar}{\sqrt{2m|\alpha|}} = \xi(0) \tau^{-1/2}, \quad \tau = \frac{T_\lambda - T}{T_\lambda}. \quad (27)$$

The estimate of Ref. [68], based on experimental data, for ${}^4\text{He}$, i.e., for Helium II, leads to a value $\xi(0) \sim 3 \times 10^{-8}$ cm. At the same time, the Ψ -theory is applicable only when the macroscopic Ψ function varies only slightly over atomic-scale distances. Hence, there follows the condition $\xi(T) \gg a \sim 3 \times 10^{-8}$ cm (here a is the average interatomic distance in liquid helium). The Ψ -theory can therefore be adequate only near the λ point (for $\tau \ll 1$), say, for $(T_\lambda - T) < (0.1 - 0.2)$ K. A similar condition also takes place in the case of Ψ -theory of superconductivity, which is also appropriate, generally speaking, only near T_c . It is of prime importance that the Landau theory of phase transitions, which is a mean-field theory, for superconductors (i.e., the Ψ -theory of superconductivity) is also correct in the immediate vicinity of T_c . This is due to the relatively large value of $\xi(0)$ in superconductors (the length $\xi(0)$ is on the order of the dimension of the Cooper pairs, i.e. in ordinary superconductors is on the order of, say, 10^{-5} cm). The point is that the temperature range near T_c (or T_λ), in which fluctuations are already large and the mean field approximation is inappropriate, is proportional to $[\xi(0)]^6$ (see [1] and references therein, particularly [34]). In Helium II, the fluctuations near T_λ are relatively strong due to the smallness of $\xi(0)$, and the Ψ -theory [68] can

be used only for $(T_\lambda - T) \gg 10^3$ K [1]. Meanwhile, the temperature range significantly closer to T_λ is of special interest. That the meanfield theory is inapplicable in the region of the λ transition in ${}^4\text{He}$ is testified too by the very existence of the λ singularity in the temperature dependence of the heat capacity. This circumstance might not, at least on the face of it, be related too the temperature dependence of the density $\rho_s(T)$, which was proportional to $|\Psi|^2$ [see expression (23)]. That is why in 1957, when the work [68] was carried out, we did not see the drawbacks to our theory right away. However, this became clear somewhat later, when it was found out that in Helium II to a good approximation

$$\rho_s(\tau) = \rho_{s0}\tau^\zeta, \quad \zeta = \frac{2}{3}. \quad (28)$$

In the mean-field theory,

$$\zeta = 1. \quad (29)$$

In experiment, by the way, the index ζ is not exactly equal to $2/3$ but is very close to it. For instance, according to [71], $\xi = 0.6705 \pm 0.0006$.

Therefore, the initial Ψ -theory of superfluidity [68] is poorly applicable to liquid ${}^4\text{He}$ in a quantitative sense. At the same time, several results based on it were obtained in [68], which were also of significance for Helium II in a qualitative sense. The case in point is the density distribution $\rho_s(z)$ near the solid wall and in films with a thickness d in relation to this thickness. Also solved were the problems of velocity v_s circulation about a vortex line at the axis of which $\Psi = 0$, of the energy of this filament, and of the surface energy at the interface between Helium II and the solid wall. No less significant is the fact that liquid ${}^4\text{He}$ is not the only existing superfluid liquid. Such a liquid is also encountered in the case of ${}^3\text{He} - {}^4\text{He}$ solutions, liquid ${}^3\text{He}$, neutron stars, and maybe in other cases. In these cases, however, the Ψ function may prove to be no longer scalar but, on the other hand, the length $\xi(0)$ is relatively large (in liquid ${}^3\text{He}$, for instance, $\xi(0) \sim 10^{-5}$ cm), and the fluctuation region is rather small. Finally, the theory of Ref. [68] had played, so far as I can judge, a significant role in the construction and elaboration of the Gross–Pitaevskii theory, which is widely used in the investigation of Bose–Einstein condensation (see [72]).

Liquid ${}^4\text{He}$, i.e., helium II, has always occupied and still occupies the leading position in the physics of superfluidity, both historically and regarding the scale of investigations. The Landau theory [7], which describes its behavior, is primarily macroscopic or, if you like, quasi-macroscopic. But it does not provide answers to several questions, particularly near the λ point. At the same time, a microtheory of the BCS type for superconductors does not exist for Helium II. On the other hand, Helium II near the λ point is interesting from various viewpoints, in particular, in the investigation of two-liquid hydrodynamics near the λ point, in the modeling of some cosmological situations [73], etc. It is likely that the initial Ψ -theory of superfluidity [68, 74] can be used to some extent for the solution of these problems, though with the



L. P. Pitaevskii.



A. A. Sobyanyin.

above significant limitation arising from the inapplicability of the mean field approximation, i.e., from the neglect of fluctuations. The generalized Ψ -theory of superfluidity [69, 70] was intended to eliminate these drawbacks. It is based on some semiempirical generalization of the Landau theory of phase transitions (see, for instance, [75]). In the Landau theory of phase transitions and, in particular, in the Ψ -theory of superconductivity, i.e., when the Ψ function is selected as the order parameter, the free energy density of the ordered phase near the transition point T_λ is written in the form

$$F_{\text{II}} = F_{\text{I}} + \alpha |\Psi|^2 + \frac{\beta}{2} |\Psi|^4 + \frac{\gamma}{6} |\Psi|^6; \quad (30)$$

away from the tricritical point it being safe to assume that

$$\begin{aligned} \alpha &= \alpha'_\lambda(T - T_\lambda) = -a_0\tau, \quad \beta = \beta_\lambda, \\ \gamma &= 0, \quad \tau = \frac{T_\lambda - T}{T_\lambda}. \end{aligned} \quad (31)$$

In the generalized theory

$$F_{\text{II}} = F_{\text{I}} - a_0\tau|\tau|^{1/3}|\Psi|^2 + \frac{b_0}{2}\tau^{2/3}|\Psi|^4 + \frac{g_0}{3}|\Psi|^6. \quad (32)$$

When selecting expression (32), for small $|\Psi|^2$ in the equilibrium $|\Psi_0|^2 = -\alpha/\beta = (\alpha_0/b_0)\tau^{2/3}$, i.e., there occurs a temperature dependence which agrees with the observed one [see (28)]. Evidently, expression (32) is selected for precisely the attainment of this goal.

The generalized Ψ -theory of superfluidity [69, 70] formally differs from the initial theory [68, 74] just by the replacement of expressions (30)–(31)

with expression (32). Several expressions and inferences were derived on this basis. For instance, for a thin film of Helium II of thickness d , the λ -transition temperature is

$$T_\lambda(d) = T_\lambda - 2.53 \times 10^{-11} \left(\frac{3+M}{M} \right) d^{-3/2} K, \quad (33)$$

where $T_\lambda = T_\lambda(\infty)$ is the λ -transition temperature in massive helium (as is well known, $T_\lambda = 2.17$ K) and M is the parameter of the theory proportional to the g_0 coefficient in expression (32). When $M < 1$, the λ transition is of the second kind (by comparison with experiment, only a crude estimate was obtained for Helium II: $M = 0.5 \pm 0.3$). By the way, if we consider a cylindrical capillary of diameter d instead of a plane film, the coefficient 2.53 in formula (33) should be replaced with 4.76. Quite a number of other expressions were also derived [69, 70, 76].

Unfortunately, the generalized Ψ -theory of superfluidity has not come to the attention of either experimenters or theorists. True, some pessimistic judgments regarding it were expressed in the literature (they were mentioned in [1]). Sobyannik and I also abandoned the superfluidity research during the period of rapid changes in the USSR and Russia that set in after 1985–1988. Only in [1] did I review our work.

Undeniably the generalized Ψ -theory of superfluidity is not a lofty *ab initio* theory. At the same time, its simplicity (at least in comparison with other known methods) suggests that the Ψ -theory of superfluidity (initial as well as generalized) can still yield much in the study of superfluidity. In any case, the opposite opinion is not substantiated at all. This section of the lecture has been written precisely with the aim of attracting the attention of physicists engaged in the corresponding areas to the Ψ -theory of superfluidity. It may well be, in my view, that the lack of attention is a delusion. It is conceivable, on the contrary, that I am in error myself, though.

‘PHYSICAL MINIMUM’ – WHAT PROBLEMS OF PHYSICS AND ASTROPHYSICS SEEM NOW TO BE ESPECIALLY IMPORTANT AND INTERESTING IN THE BEGINNING OF THE XXI CENTURY?

I have encountered the viewpoint that my work in the area of superconductivity and superfluidity is a matter of the remote past. There is no question that the work of Ginzburg and Landau [3] performed back in 1950 stands out. But on the whole, as is clear from the foregoing and particularly from [1], I have been occupied with this field of physics since 1943 until the present time. In this case, it seems to me, several questions and problems have also been posed which have not been solved and which deserve attention. Of course, presently the most urgent problems in the area of superconductivity are the elucidation of the mechanism and several features of high-temperature superconductivity and the creation of room-temperature superconductivity. More precisely, what is wanted in the latter case is to elucidate the po-

tentialities and formation conditions of room-temperature superconductors. I am keenly aware that I will not be able to accomplish anything in the last two directions. I would like only to witness as many new findings as possible.

That is why in recent years I have been placing progressively stronger emphasis, as far as physics is concerned, on some educational program, which I conventionally call the ‘physical minimum’. As far as I know, many young scientists attend Nobel Lectures, and therefore I decided to enlarge on this ‘physical minimum’. I believe that this will be of greater interest to young people than to hear what was going on before they were born.

Physics has developed rapidly and fruitfully, especially in the past century. Its face changed radically even within a human life span. I myself was already 16 when the neutron and positron were discovered in 1932. And what would modern physics be without neutrons and positrons? As a result of so rapid a development, physics and their adjacent realms (for instance, astronomy) have enormously expanded, both as regards its basic contents and the body of information. In the recent past it was possible to be guided by the requirement “to know something about everything and to know everything about something” (say, in physics), but now, it seems to me, this is no longer possible. At the same time, I am startled and dispirited when young physicists (and sometimes not so young ones) restrict themselves to the knowledge in ‘their’ area and are not informed, if only in a general way, about the state of physics as a whole and its ‘hottest’ areas. This situation cannot be justified by alleging an absence of a pivot (keystone) in contemporary physics or its boundlessness. Quite the contrary. Physics does (maybe still does) have its pivot, which is represented by fundamental concepts and laws formulated in theoretical physics. It is possible, on the basis of theoretical physics studied during one’s student days, to understand all modern physics or, more precisely, to understand how matters stand everywhere in physics and be aware of the situation. Every physicist (naturally, this equally applies to other specialities, but I restrict myself to physicists for definitiveness) should simultaneously know, apart from theoretical physics, a wealth of facts from different branches of physics and be familiar with the newest notable accomplishments.

At the same time, we in Russia like to quote a certain Koz’má Prutkov, a fictitious character, who said pompously, in particular, that “there is no way of comprehending the incomprehensible”. So, one has to choose something. And so I took this path: I have made a ‘list’ of the top problems of the day. Any such ‘list’ is admittedly subjective. It is also clear that the ‘list’ should vary with time. Lastly, it is clear that subjects not included in the ‘list’ can in no way be regarded as unimportant or uninteresting. It is simply that many of them presently seem less pressing to me (or to the authors of other similar lists). But again, “one cannot comprehend the incomprehensible”. Those who know interesting subjects beyond the ‘list’ have no reason to be offended and should only supplement or change the ‘list’. I only suggest some enumeration of the questions that, in my view, every physicist should have at least a superficial idea of. Supposedly less trivial is the statement that this is not as difficult as it might seem at first glance. The time to be spent for this purpose is, I be-

lieve, no longer than the time a good student spends preparing for an examination, say, on electrodynamics. Acquaintance with all subjects included in this ‘list’ is what I call the ‘physical minimum’. Of course, this ‘minimum’ is the echo of the ‘theoretical minimum’ proposed by Landau in the 1930s. It is significant that there are many excellent textbooks on electrodynamics (or other subjects in the university curriculum), among which the corresponding volume of the “Course of Theoretical Physics” by L.D. Landau and E.M. Lifshitz ranks, in my view, highest. But a beginner needs help to get acquainted with the ‘physical minimum’. Working out this ‘list’, as well as commenting on it, has served and hopefully continues to serve precisely this purpose. In 1995, in the Russian edition of the book [16], I managed to work out a rather detailed commentary. But in the English translation [16] some was already out of date, which I failed to compensate for in full measure. Inserted at the beginning of the book [2] is an article also concerned with the ‘physical minimum’. Several additional remarks were introduced in the English translation of this book, which will hopefully be published soon. On the whole, should the proposal be taken advantage of and elaborated, the ‘physical minimum’ will meet with support and new books on this subject should appear. Unfortunately, I cannot set myself to this task.

In the context of this lecture it only remains for me to recall the well-known saying that the proof of the pudding is in the eating and give the above-mentioned ‘list’ for the beginning of the XXI century:

1. Controlled nuclear fusion.
2. High-temperature and room-temperature superconductivity (HTSC and RTSC).
3. Metallic hydrogen. Other exotic substances.
4. Two-dimensional electron liquid (anomalous Hall effect and other effects).
5. Some questions of solid-state physics (heterostructures in semiconductors, quantum wells and dots, metal – dielectric transitions, charge and spin density waves, mesoscopics).
6. Second-order and related phase transitions. Some examples of such transitions. Cooling (in particular, laser cooling) to superlow temperatures. Bose–Einstein condensation in gases.
7. Surface physics. Clusters.
8. Liquid crystals. Ferroelectrics. Ferrotoroics.
9. Fullerenes. Nanotubes.
10. The behavior of matter in superstrong magnetic fields.
11. Nonlinear physics. Turbulence. Solitons. Chaos. Strange attractors.
12. X-ray lasers, gamma-ray lasers, superhigh-power lasers.
13. Superheavy elements. Exotic nuclei.
14. Mass spectrum. Quarks and gluons. Quantum chromodynamics. Quark-gluon plasma.
15. Unified theory of weak and electromagnetic interactions. W^\pm and Z^0 -bosons. Leptons.

16. Standard model. Grand unification. Superunification. Proton decay.
Neutrino mass. Magnetic monopoles.
17. Fundamental length. Particle interaction at high and superhigh energies.
Colliders.
18. Nonconservation of CP-invariance.
19. Nonlinear phenomena in vacuum and in superstrong magnetic fields.
Phase transitions in a vacuum.
20. Strings. M-theory.
21. Experimental verification of the general theory of relativity.
22. Gravitational waves and their detection.
23. The cosmological problem. Inflation. Λ -term and ‘quintessence’.
Relationship between cosmology and high energy physics.
24. Neutron stars and pulsars. Supernova stars.
25. Black holes. Cosmic strings (?).
26. Quasars and galactic nuclei. Formation of galaxies.
27. The problem of dark matter (hidden mass) and its detection.
28. The origin of superhigh-energy cosmic rays.
29. Gamma-bursts. Hypernovae.
30. Neutrino physics and astronomy. Neutrino oscillations.

The singling out of 30 particular problems (more precisely, items in the ‘list’) is of course absolutely conditional. Moreover, some of them might be subdivided. In my first ‘list’ published in 1971 there were 17 problems [77]. Subsequently their number would grow (for some more details, see [2]). It would supposedly be well to add some new subjects to the ‘list’, given about, for instance, those concerning quantum computers and advances in optics. But I cannot do this with adequate comprehension.

Any ‘list’ is undoubtedly not a dogma, something can be discarded and something added, depending on the preferences of lecturers and authors of corresponding papers. More interesting is the question of the evolution of the ‘list’ with time, as it reflects the development of physics. In the ‘list’ of 1970–1971 [77] quarks were given only three lines in the enumeration of the attempts to explain the mass spectrum. This did not testify to my perspicacity. However, at that time (in 1970) quarks were only five or six years old (I mean the age of the corresponding hypothesis), and the fate of the concept of the quark was indeed vague. Now the situation is of course quite different. True, the heaviest t-quark was discovered only in 1994 (its mass, according to the data of 1999, is $m_t = 176 \pm 6$ GeV). The list [77] naturally contains no fullerenes, which were discovered in 1985, no gamma-bursts (the first report of their discovery was published in 1973). High-temperature superconductors were synthesized in 1986–1987, but in the list [77] this problem was nonetheless considered rather thoroughly for it had been discussed since 1964 (this was discussed in greater detail in the previous sections of the lecture). Generally, much has been done in physics over the past 30 or 35 years, but, I believe, not very much essentially new has appeared. In any case, the ‘lists’ in [77, 16], as well as that presented above, characterize to a certain extent the

development and the state of physical and astronomical problems from 1970–1971 to the present day.

It should be added that three ‘great problems’ of modern physics are also to be included in the ‘physics-minimum’, included in the sense that they should be singled out in some way and specially discussed, and development of these ‘great problems’ should be reviewed. This is discussed at some length in [2]. The ‘great problems’ are, first, the increase in entropy, time irreversibility, and the ‘time arrow’. Second is the problem of interpretation of nonrelativistic quantum mechanics and the possibility of learning something new even in the field of its applicability (I personally doubt this possibility but believe that one’s eyes should remain open). And third is the question of live-to-liveless reduction, i.e., the feasibility of explaining the origin of life and thought on the basis of physics alone. On the face of it, how could it be otherwise? But until the questions are elucidated, one cannot be quite sure of anything. I think that the problem of the origin of life will unreservedly be solved only after ‘life in a test-tube’ is created. Until then, this will be an open question.

One more concluding remark. In the past century, and even nowadays, one could encounter the opinion that in physics nearly everything had been done. There allegedly are only dim ‘cloudlets’ in the sky or theory, which will soon be eliminated to give rise to the ‘theory of everything’. I consider these views as some kind of blindness. The entire history of physics, as well as the state of present-day physics and, in particular, astrophysics, testifies to the opposite. In my view we are facing a boundless sea of unresolved problems. It only remains for me to envy the younger members of the audience, who will witness a great deal of new, important, and interesting things.

REFERENCES

1. Ginzburg V L *Usp. Fiz. Nauk* **167** 429 (1997); **168** 363 (1998) [*Phys. Usp.* **40** 407 (1997); **41** 307 (1998)].
2. Ginzburg V L *O Nauke, o Sebe i o Drugikh* (About Science, Myself, and Others) (Moscow: Fizmatlit, 2003)¹⁰ [Translated into English (Bristol: IOP Publ., 2004) Article 7 (to be published)].
3. Ginzburg V L, Landau L D *Zh. Eksp. Teor. Fiz.* **20** 1064 (1950); This paper was published in English in the volume: Landau L D *Collected Papers* (Oxford: Pergamon Press, 1965) p. 546.
4. Abrikosov A A *Zh. Eksp. Teor. Fiz.* **32** 1442 (1957) [*Sov. Phys. JETP* **5** 1174 (1957)].
5. Lifshitz E M, Pitaevskii L P *Statisticheskaya Fizika* (Statistical Physics) Pt. 2 *Teoriya Kondensirovannogo Sostoyaniya* (Theory of Condensed State) (Moscow: Nauka, 1978, 1999) [Translated into English (Oxford: Pergamon Press, 1980)].
6. Tinkham M *Introduction to Superconductivity* 2nd ed. (New York: McGraw Hill, 1996).
7. Landau L D *Zh. Eksp. Teor. Fiz.* **11** 592 (1941); *J. Phys. USSR* **5** 71 (1941).
8. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **13** 243 (1943); *J. Phys. USSR* **7** 305 (1943).
9. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **14** 134 (1944).

¹⁰ Article 7 published in this Collection is the somewhat edited article given above in Ref. [1].

10. Bardeen J, in *Kältephysik* (Handbuch der Physik, Bd. 15, Hrsg. S von Flügge) (Berlin: Springer-Verlag, 1956) p. 274 [Translated into Russian: Bardeen J, in *Fizika Nizkikh Temperatur* (Ed. A I Shal'nikov) (Moscow: IL, 1959) p. 679].
11. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **14** 177 (1944); *J. Phys. USSR* **8** 148 (1944).
12. London F, London H *Proc. R. Soc. London Ser. A* **149** 71 (1935); *Physica* **2** 341 (1935).
13. Waldrum J R *Superconductivity of Metals and Cuprates* (Bristol: Institute of Physics Publ., 1996).
14. Ginzburg VL *Zh. Eksp. Teor. Fiz.* **16** 87 (1946); *J. Phys. USSR* **9** 305 (1945).
15. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **15** 739 (1945); *J. Phys. USSR* **10** 107 (1946).
16. Ginzburg V L *The Physics of a Lifetime. Reflections on the Problems and Personalities of 20th Century Physics* (Berlin: Springer-Verlag, 2001).¹¹
17. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **29** 748 (1955) [*Sov. Phys. JETP* **2** 589 (1956)].
18. Bardeen J, Cooper L N, Schrieffer J R *Phys. Rev.* **108** 1175 (1957).
19. Ginzburg V L *Usp. Fiz. Nauk* **48** 25 (1952); *Fortschr. Phys.* **1** 101 (1953).
20. Ogg R A (Jr) *Phys. Rev.* **69** 243; **70** 93 (1946).
21. Schafroth MR *Phys. Rev.* **96** 1149 (1954); **100** 463 (1955).
22. Cooper L N *Phys. Rev.* **104** 1189 (1956).
23. Gor'kov L P *Zh. Eksp. Teor. Fiz.* **36** 1918; **37** 1407 (1959) [*Sov. Phys. JETP* **9** 1364 (1959); **10** 998 (1960)].
24. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **31** 541 (1956) [*Sov. Phys. JETP* **4** 594 (1957)].
25. Shoenberg D *Superconductivity* 3rd ed. (Cambridge: Cambridge Univ. Press, 1965) [Translated into Russian (Moscow: IL, 1955)].
26. Ginzburg V L *Sverkhprovodimost'* (Superconductivity) (Moscow-Leningrad: Izd. AN SSSR, 1946).
27. Buckel W *Supraleitung* (Weinheim, Bergster: Physik-Verlag, 1972) [Translated into English: Buckel W *Superconductivity: Fundamentals and Applications* (Weinheim: VCH, 1991); Translated into Russian (Moscow: Mir, 1975)].
28. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **23** 236 (1952).
29. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **34** 113 (1958) [*Sov. Phys. JETP* **7** 78 (1958)].
30. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **42** 299 (1962) [*Sov. Phys. JETP* **15** 207 (1962)].
31. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **36** 1930 (1959) [*Sov. Phys. JETP* **9** 1372 (1959)].
32. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **31** 202 (1956) [*Sov. Phys. JETP* **4** 153 (1957)].
33. Zharkov G F *Zh. Eksp. Teor. Fiz.* **34** 412 (1958); **37** 1784 (1959) [*Sov. Phys. JETP* **7** 278 (1958); **10** 1257 (1959)].
34. Ginzburg V L *Fiz. Tverd. Tela* **2** 2031 (1960) [*Sov. Phys. Solid State* **2** 1824 (1961)].
35. Little W A *Phys. Rev.* **134** A1416 (1964).
36. Ginzburg V L *Phys. Lett.* **13** 101 (1964); *Zh. Eksp. Teor. Fiz.* **47** 2318 (1964) [*Sov. Phys. JETP* **20** 1549 (1965)].
37. Ginzburg V L, Kirzhnits D A *Zh. Eksp. Teor. Fiz.* **46** 397 (1964) [*Sov. Phys. JETP* **19** 269 (1964)].
38. Ginzburg V L *Phys. Scripta* **T27** 76 (1989).
39. Ginzburg V L, Kirzhnits D A *Dokl. Akad. Nauk SSSR* **176** 553 (1967) [*Sov. Phys. Dokl.* **12** 880 (1968)].
40. Ginzburg V L, Kirzhnits D A (Eds) *Problema Vysokotemperaturnoi Sverkhprovodimosti* (The Problem of High-Temperature Superconductivity) (Moscow: Nauka, 1977) [Translated into English: Ginzburg V L, Kirzhnits D A (Eds) *High-Temperature Superconductivity* (New York: Consultants Bureau, 1982)].
41. Bednorz J G, Muller K A Z *Phys. B* **64** 189 (1986).
42. Wu M K, Ashburn J R, Torng C J, Hor P H, Meng R L, Gao L, Huang Z J, Wang Y Q, Chu C W *Phys. Rev. Lett.* **58** 908 (1987).

¹¹ This book is, for the most part, a translation of the book: Ginzburg V L *O Fizike i Astrofizike* (Moscow: Byuro Kvantum, 1995).

43. Ginzburg V L *Energiya* (a Scientific and Popular Journal) (9) 2 (1984).
44. Ginzburg V L *Prog. Low Temp. Phys.* **12** 1 (1989).
45. Ginzburg V L, in *From High-Temperature Superconductivity to Microminiature Refrigeration* (Eds B Cabrera, H Gutfreund, V Kresin) (New York: Plenum Press, 1996).
46. Ginzburg V L *J. Supercond.* **4** 327 (1986).
47. Shaplygin I S, Kakhan B G, Lazarev V B *Zh. Neorg. Khim.* **24** 1476 (1979).
48. Cava R J *et al. Phys. Rev. Lett.* **58** 408 (1987).
49. Eliashberg G M *Zh. Eksp. Teor. Fiz.* **38** 966; 39 1437 (1960) [*Sov. Phys. JETP* **11** 696 (1960); **12** 1000 (1961)].
50. Maksimov E G *Usp. Fiz. Nauk* **170** 1033 (2000) [*Phys. Usp.* **43** 965 (2000)].
51. Ginzburg V L, Maksimov E G *Sverkhprovodimosti: Fiz., Khim., Tekh.* **5** 1543 (1992) [*Superconductivity: Phys., Chem., Technol.* **5** 1505 (1992)].
52. Meissner W Z. *Ges. Kältenindustr.* **34** 197 (1927).
53. Burton E F, Smith GH, Wilhelm J O *Phenomena at the Temperature of Liquid Helium* (American Chemical Society: Monograph Ser., No. 83) (New York: Reinhold Publ. Corp., 1940).
54. Gorter C J, Casimir H *Phys. Z.* **35** 963 (1934).
55. Ginzburg V L, Zharkov G F *Usp. Fiz. Nauk* **125** 19 (1978) [*Sov. Phys. Usp.* **21** 381 (1978)].
56. Selzer P M, Fairbank W M *Phys. Lett. A* **48** 279 (1974).
57. Ginzburg V L, Zharkov G F *Pis'ma Zh. Eksp. Teor. Fiz.* **20** 658 (1974) [*JETP Lett.* **20** 302 (1974)].
58. Gal'perin Yu M, Gurevich V L, Kozub V N *Zh. Eksp. Teor. Fiz.* **66** 1387 (1974) [*Sov. Phys. JETP* **39** 680 (1974)].
59. Garland J C, Van Harlingen D *J. Phys. Lett. A* **47** 423 (1974).
60. Van Harlingen D *J. Physica B + C* **109–110** 1710 (1982).
61. Arutyunyan R M, Ginzburg V L, Zharkov G F *Zh. Eksp. Teor. Fiz.* **111** 2175 (1997) [*JETP* **84** 1186 (1997)]; *Usp. Fiz. Nauk* **167** 457 (1997) [*Phys. Usp.* **40** 435 (1997)].
62. Galperin YM *et al. Phys. Rev. B* **65** 064531 (2002).
63. Ginzburg V L, Zharkov G F, Sobyanin A A *Pis'ma Zh. Eksp. Teor. Fiz.* **20** 223 (1974) [*JETP Lett.* **20** 97 (1974)]; Ginzburg V L, Sobyanin A A, Zharkov G F *Phys. Lett. A* **87** 107 (1981).
64. Ginzburg V L, Sobyanin A A *Zh. Eksp. Teor. Fiz.* **85** 1606 (1983) [*Sov. Phys. JETP* **58** 934 (1983)].
65. Gamtselidze G A, Mirzoeva M I *Zh. Eksp. Teor. Fiz.* **79** 921 (1980); **84** 1725 (1983) [*Sov. Phys. JETP* **52** 468 (1980); **57** 1006 (1983)].
66. Ginzburg V L *Zh. Eksp. Teor. Fiz.* **29** 254 (1955) [*Sov. Phys. JETP* **2** 170 (1956)].
67. Gamtselidze G A *Zh. Eksp. Teor. Fiz.* **34** 1434 (1958) [*Sov. Phys. JETP* **7** 992 (1958)].
68. Ginzburg V L, Pitaevskii L P *Zh. Eksp. Teor. Fiz.* **34** 1240 (1958) [*Sov. Phys. JETP* **7** 858 (1958)].
69. Ginzburg V L, Sobyanin A A *Usp. Fiz. Nauk* **120** 153 (1976) [*Sov. Phys. Usp.* **19** 773 (1976)]; *J. Low Temp. Phys.* **49** 507 (1982).
70. Ginzburg V L, Sobyanin A A *Usp. Fiz. Nauk* **154** 545 (1988) [*Sov. Phys. Usp.* **31** 289 (1988)]; *Jpn. J. Appl. Phys.* **26** (Suppl. 26–3) 1785 (1987).
71. Golder L S, Mulders N, Ahlers G *J. Low. Temp. Phys.* **93** 131 (1992).
72. Pitaevskii L, Stringari S *Bose-Einstein Condensation* (Intern. Series of Monographs on Physics, Vol. 116) (Oxford: Clarendon Press, 2003).
73. Zurek W H *Nature* **382** 296 (1996).
74. Pitaevskii L P *Zh. Eksp. Teor. Fiz.* **35** 408 (1958) [*Sov. Phys. JETP* **8** 282 (1959)].
75. Mamaladze Yu G *Zh. Eksp. Teor. Fiz.* **52** 729 (1967) [*Sov. Phys. JETP* **25** 479 (1967)]; *Phys. Lett. A* **27** 322 (1968).
76. Ginzburg V L, Sobyanin A A, in *Superconductivity, Superdiamagnetism, Superfluidity* (Ed. V L Ginzburg) (Moscow: MIR Publ., 1987) p. 242.
77. Ginzburg V L *Usp. Fiz. Nauk* **103** 87 (1971) [*Sov. Phys. Usp.* **14** 21 (1971)].

FRITZ HABER

The synthesis of ammonia from its elements

Nobel Lecture, June 2, 1920

The Swedish Academy of Sciences has seen fit, by awarding the Nobel Prize, to honour the method of producing ammonia from nitrogen and hydrogen. This outstanding distinction puts upon me the obligation of explaining the position occupied by this reaction within the subject of chemistry as a whole, and to outline the road which led to it.

We are concerned with a chemical phenomenon of the simplest possible kind. Gaseous nitrogen combines with gaseous hydrogen in simple quantitative proportions to produce gaseous ammonia. The three substances involved have been well known to the chemist for over a hundred years. During the second half of the last century each of them has been studied hundreds of times in its behaviour under various conditions during a period in which a flood of new chemical knowledge became available. If it has not been until the present century that the production of ammonia from the elements has been discovered, this is due to the fact that very special equipment must be used and strict conditions must be adhered to if one is to succeed in obtaining spontaneous combination of nitrogen and hydrogen on a substantial scale, and that a combination of experimental success with thermodynamic considerations was needed.

It was particularly significant that earlier attempts had not succeeded, even fleetingly, in achieving with absolute certainty a spontaneous union of nitrogen and hydrogen to form ammonia. This gave rise to the prejudice that such a production of ammonia was impossible, and in the course of time this enjoyed considerable support in chemical circles. Such prejudice leads one to expect pitfalls which, far more than clearly-defined difficulties, deter one from becoming too deeply involved in the subject.

A narrow professional interest in the preparation of ammonia from the elements was based on the achievement of a simple result by means of special equipment. A more widespread interest was due to the fact that the synthesis of ammonia from its elements, if carried out on a large scale, would be a useful, at present perhaps the most useful, way of satisfying important national economic needs. Such practical uses were not the principal purpose of

my investigations. I was never in doubt that my laboratory work would produce no more than a scientific confirmation of basic principles and a criterion of experimental aids, and that much would need to be added to any success of mine to ensure economic success on an industrial scale. On the other hand I would hardly have concentrated so much on this problem had I not been convinced of the economic necessity of chemical progress in this field, and had I not shared to the full Fichte's conviction that while the immediate object of science lies in its own development, its ultimate aim must be bound up in the moulding influence which it exerts at the right time upon life in general and the whole human arrangement of things around us.

Since the middle of the last century it has become known that a supply of nitrogen is a basic necessity for the development of food crops; it was also recognized, however, that plants cannot absorb the elementary nitrogen which is the main constituent of the atmosphere, but need the nitrogen to be combined with oxygen in the form of nitrate in order to be able to assimilate it. This combination with oxygen can start with combination with hydrogen to form ammonia since ammonium nitrogen changes to saltpetre nitrogen in the soil.

Under natural conditions the soil does not lose its fixed nitrogen. Green plants use it to synthesize complicated substances without changing it into elementary nitrogen. Animals and humans ingest it with the plants and return it to the soil in fixed form in their excretions and finally with their deceased remains. Putrefaction and combustion does destroy a certain amount of fixed nitrogen, but Nature makes good the loss when, during thunderstorms, lightning combines nitrogen and oxygen in the upper layers of the atmosphere, which is then washed down by the rain. To this nitrogen-fixing action of electrical discharge as a source of bound nitrogen is added the activity of soil bacteria, some of which live free while others are to be found settled in the root nodules of many plants, converting free nitrogen into bound nitrogen.

Agricultural husbandry essentially maintains the balance of bound nitrogen. However, with the advent of the industrial age, the products of the soil are carried off from where the crops are grown to far-off places where they are consumed, with the result that the bound nitrogen is no longer returned to the earth from which it was taken.

This has caused the world-wide economic necessity of supplying bound nitrogen to the soil. This need is increased by national economic considerations, which, with the denser population of industrialized countries, call for

increased agricultural productivity at home, and it is yet further increased by the fact that expanding industry requires fixed nitrogen for many of its own chemical purposes. The demand for nitrogen, like that for coal, indicates how far removed our way of life has become from that of the people who ((themselves do fertilize the soil they cultivate)).

Agriculture, always the main consumer, is not satisfied with a supply of nitrogen alone - potash and phosphates are equally indispensable - but the world possesses far fewer natural resources for meeting nitrogen requirements. And so, naturally, concern over nitrogen supplies has become the first of the great obstacles that lie along the highway of world commerce upon which we have been travelling in recent decades.

Our way of thinking, so used to interpreting historical events in the context of man's unchangeable nature, easily misleads us into overlooking the enormous turning-point in the history of mankind represented by the last hundred years. In earlier periods the need for energy was satisfied by men's physical labour and by the use of wind and sun, which are older than ourselves and will outlive our life conditions. The past century has opened the floodgates for the energy stored in coal, and has introduced ways of life in industrialized countries in which the physical labour of men merely operates a relay to release the hundred times more powerful energy of coal into the lifestream of international commerce. Technical needs have arisen for which we only too easily find ourselves unprepared through a lack of adequate scientific development. The present state of affairs in the world, with the after-effects of the War in Central Europe placing an overwhelming load on our scientific work, makes this only too plain.

The need for opening up new sources of nitrogen became clearly apparent at the turn of the century. Since the middle of the last century we had been drawing upon the supply of saltpetre nitrogen which Nature had deposited in the high-mountain deserts of Chile. By comparing the fast-rising requirements with the calculated extent of these deposits it became clear that towards the middle of the present century a major emergency would be unavoidable, unless the chemistry found a way out.

The problem was not a new one to the chemists. When they began to distil coal they came across ammonia among the distillation products and this, in the form of ammonium sulphate, found application in agriculture. While in 1870 ammonia was a tiresome by-product of the gas industry, by 1900 it had become a very valued companion to combustible gases and the coke industry was in full swing everywhere to adapt furnaces to its by-pro-

duction. Its origin from the fixed nitrogen of coal was understood; an improvement in its yield, which by the normal process was hardly more than 1/5 of the nitrogen content in the coal, had been widely studied. But no satisfactory solution seemed likely in that direction.

With an average content of about 1% of fixed nitrogen, coal could not be processed for obtaining nitrogen only. The delivery of nitrogen as a by-product set limits to its production which made it impossible to make good a future deficiency of saltpetre from this source. It was clear that the demand for fixed nitrogen, which at the beginning of this century could be satisfied with a few hundred thousand tons a year, must increase to millions of tons. A demand of this order could only be met from *one* source—from the immense supply of elementary nitrogen available in our atmosphere—and the binding would have to be achieved by chemical means to the simplest and most widely available chemical elements, if the solution was to measure up to the demand. Just as the raw-material situation of our Earth indicates elementary nitrogen as the starting material, so ammonia or nitric acid are indicated as end products by the requirements of plants. The task thus became the combining of elementary nitrogen with oxygen or water.

This again was not a new or untried chemical problem. The combining of nitrogen with hydrogen to form ammonia as with oxygen to produce nitrates had already occupied science and, to some extent, technology.

Combination with hydrogen directly from the elements had been induced by various forms of electrical discharge, which of course resulted in an energy consumption of alarming proportions. Indirect combination, on the other hand, had been developed with remarkable technical results; the nitrogen was combined with other elements and this combination was then hydrolysed with water whereby ammonia was split off. Only the spontaneous association of the elements was unknown when, in 1904, I began to occupy myself with the subject; it was held to be impossible after pressure, heat, and the catalytic action of platinum sponge had been found unable to produce the effect.

The indirect method has occupied the attention of scientists and technologists ever since Margueritte and Sourdeval, basing themselves on earlier work by Bunsen and Playfair, developed it to the stage of sample production in 1860. Caustic baryte and coal at high temperatures with nitrogen yielded barium cyanide. At lower temperatures this combination broke down in the presence of water vapour, yielding ammonia and creating barium hydroxide which returned to the process. Thus, during alternate for-

mation and breaking down of barium cyanide, a continuous yield of ammonia and carbon dioxide was obtained from coal, water and elementary nitrogen. In the half-century following the publication by Margueritte and Sourdeval, this indirect method, the early technical execution of which made excessive demands on the reaction vessels, has been studied afresh in many modified forms.

Barytes could be replaced by heat-resistant oxides of other metals or semi-metals. The process of nitrogen fixation could be broken down into partial steps, first forming, by reduction, the metal, semi-metal or metal carbide which would, in a subsequent reaction, take up the nitrogen. As a solution to the problem of ammonia synthesis the result has never been entirely satisfactory.

If the reduction of oxide and the fixation of nitrogen took place in a single process then this required an extremely high temperature. If the process were split up, intermediate products were obtained which reacted more easily with nitrogen. But the intermediate product-metal, semi-metal, or carbide - then demanded, for its own production from the massive reserves of natural products, precisely those conditions which led to an uneconomical consumption of electrical energy, either by electrolytic or electrothermal means.

The more tightly knit nitrogen molecule does not break down as easily as oxygen, the next element in the periodic system. The abundant examples we have of autoxidation are thus matched by a complete lack of spontaneous reaction of elementary nitrogen in the inanimate world at normal temperatures. The inaccessibility of nitrogen nullified all the many efforts made to develop a technical ammonia process.

In only one respect has the study of indirect methods of synthesizing ammonia from the elements been able to get round the difficulties. Frank and Caro obtained the important calcium cyanamide through the action of nitrogen on calcium carbide obtained from lime and coal in the electric arc. Splitting the calcium cyanamide with water produces ammonia, and the process takes place in the soil without any particular help from us, once the cyanamide has been added to the soil as fertilizer. The saving in factory processing achieved by this, plus the fact that the only raw materials required are lime, coal and nitrogen, have been important factors in the establishment of the process.

Efforts to combine nitrogen with oxygen go back further than those aimed at combining it with hydrogen. The basic fact of the combination of ni-

trogen with oxygen during sparking had already been observed by Cavendish and Priestley. In this case the first product is nitric oxide, which converts to nitric acid in a spontaneous reaction with oxygen and water. The nitric oxide synthesis is a process requiring heat, and unless energy is supplied can, for thermodynamic reasons, only occur spontaneously to any appreciable extent at extremely high temperatures. However, the supply of energy required at normal temperatures is so small that disadvantage of having to provide it is outweighed by the advantage of needing only air and water as raw materials. No better and more economical process for the binding of nitrogen could therefore be devised if some means could be found for converting electrical energy into this kind of chemical energy without waste.

The example of Nature, which produces the reaction via lightning and Cavendish's earlier successful imitation of this with electric sparks, coupled with the outstanding electrotechnical developments of the final decades of previous century, increasingly brought this method of solving the nitrogen problem to the fore, as professional circles became less and less satisfied with the progress achieved through combining nitrogen with hydrogen. The brilliant developments which these efforts produced in the early years of this century are general knowledge. The main outlines of the technical design coupled particularly with the names of Birkeland and Eyde, of Schoenherr and of Pauling, have for years been the object of a great deal of interest among experts.

Installations on a considerable scale were built in a number of places and the method was evidently well suited to making use of the vast supply of energy which could be derived from waterfalls for chemical purposes; but this method of synthesizing nitrogen has still not reached the levels of production which it appeared to promise. Its progress is limited by the fact that with a consumption of one kilowatt-hour no more than 16 grams of nitrogen are converted into nitric acid, whilst a complete conversion of electrical to chemical energy ought to yield 30 times as much. An explanation of this has been given by Muthmann and Hofer, who have demonstrated that the high-tension arc used in this process, acts as a Deville's heat evaporation chamber.

The formation of nitric oxide is determined, and limited, by thermal conditions in the arc and its surroundings. Determination of the thermodynamic equilibrium of nitric oxide synthesis by Nernst confirmed this explanation. An extrapolation of his experimental results and the best figures for the specific heat of the gases involved up to the temperature of 3,000°C or 4,000°C

led to the remarkable conclusion that more than 1½ times or twice the technical yield per kilowatt-hour could still not be achieved when no re-forma-tion of nitric oxide in the cooling circuit occurred at all. The source of the low yield lay in the fact that the heating of a large air mass at very high tem-peratures enabled only a small fraction to convert thermodynamically to nitric oxide. In spite of the fact that, for a variety of reasons, this calculation cannot pretend to considerable accuracy, its result obviously approaches the truth. Practical experience has shown that no worthwhile saving of energy can be achieved by heat regeneration, manifestly because the deterioration of the quenching action involved militates against this. It is impossible to do away with the arc discharge without deviating from the basic processes which comply with the requirements of mass production.

However, it was perhaps not entirely impossible with a discharge arc to get away from the temperature range in which rapid adjustment of the thermodynamic balance covered every more favourable possibility of chang-ing electrical into chemical energy. After all, the arc existed by virtue of the constant production of units of higher energy in the form of gas ions caused by the electrical energy of electronic impacts and it was not a priori evident that the subsequent dissipation of energy in the form of heat precluded everything else than the thermal result of nitric oxide synthesis, particularly because Warburg and Leithaeuser had shown non-thermal synthesis of the oxide by means of corona discharge.

This possibility aroused much interest during the first ten years of this century and from 1907 led me to start investigations which I pursued over a number of years. Development has so changed opinions during those short ten years, that today it is already difficult to think oneself back into the views then generally held ; yet it is indicative that so experienced and professional a judge of chemico-technical possibilities as the "Badische Anilin- und Soda-fabrik" thought so highly of my efforts to obtain improved efficiency from electrical energy in the combining of nitrogen and oxygen, as to get in touch with me in 1908 and - by providing their resources - to facilitate my work on the subject; whereas they agreed with every caution to the proposal to back me in the high-pressure synthesis of ammonia as well, approving it only with hesitation.

In fact, even in my later judgement, the question of whether technical re-search should be concentrated on the direct synthesis of ammonia from the elements really depended on whether the consumption of energy during the combining of nitrogen and oxygen could be considerably reduced. In tech-

nical questions, where the scales oscillate between success and failure, the borderline between the two extremes is usually defined by modest differences in the consumption of energy and materials, and variations in these values which lie within one decimal power will determine the result.

With a number of excellent assistants I therefore studied for some long time the synthesis of nitric oxide by electrical discharge. I have searched through the pressure range from 12 atm. to 25 mm mercury, cooled the arc both from the wall and from the anode, and studied the relationship between energy consumption and frequency up to about 50,000 cycles per second. We obtained nitric oxide concentrations of 10% in air at decreased pressure which indicated a deviation from the thermodynamic balance. Yields of bound nitrogen were also noted for the same consumption in kilowatt-hours which exceeded the earlier-mentioned value of 16 grams by 10-15%. But in themselves these advantages were not conclusive, being moreover achieved by methods which were hardly suited to adaptation to mass-production. This series of investigations accordingly led to a strengthening of the view that the technical solution was to be sought in the direct combination of nitrogen with hydrogen.

A study of nitric-oxide synthesis in pressure flames led to the same result. It had been known since the days of Bunsen that the explosion of combustible gas with nitrogen and oxygen gives rise to the formation of nitrous products, and Liveing and Dewar had described the formation of nitric acid in a hydrogen flame under pressure. It appeared desirable to me also to familiarize myself with this source of nitric oxide, in which heat was used as the source of energy under conditions easily available in industry.

There were proposals to utilize the explosive reactions simultaneously in a motor and as a source for the synthesis of nitric oxide. I myself placed no faith in the linking of two such widely-differing functions. Yet the utilization of the heat of flame gases appeared to me to be not incompatible with the formation of nitric oxide, and worthy of closer investigation. This has been extended over the flames of carbon monoxide, hydrogen and acetylene. It was found that corresponding to 100 molecules of the main products of combustion, carbon dioxide and hydrogen, 3 to 6 molecules of nitric acid could be obtained. In the case of carbon monoxide and hydrogen this required increased pressure. Carbon monoxide had the advantage over the hydrogenated gases, since the presence of water vapour in the hot products of combustion favoured the reversion of the nitric oxide in the elements along the cooling circuit. With this gas the molecular nitric oxide: carbon

monoxide ratio could easily be brought, with air, to 3 : 100 and with a mixture rich in oxygen to double that ratio. For technical utilization however, these values were not sufficient incentive; the weight, which declined on the direct combining of nitrogen with hydrogen, therefore again underwent an increase.

I have not pursued further the combining of nitrogen and hydrogen by corona discharge and by sparking. It seemed certain to me that this method would not prove itself to be the most advantageous. In the final analysis the assessment of each method rests upon the ratio between the energy consumed and the yield, in other words, between coal consumption and nitrogen yield (the consumption of hydraulic power being reckoned as the equivalent consumption of coal).

Nothing seemed less hopeful, though, than the thought that the enforced combining of nitrogen with hydrogen could be achieved with so little energy that one would have spare energy left over for the production of hydrogen. There remained merely the possibility of discovering the requirements for spontaneous formation of ammonia from the elements. The positive heat of formation of ammonia indicated that such a synthesis might be achieved without the assistance of electrical energy. Against this there was the fact that neither Deville nor Ramsay and Young had obtained ammonia by heating nitrogen and hydrogen.

Ramsay and Young who, in 1884, during their study of the decomposition of the gas in the neighbourhood of 800°C had consistently observed a trace of undecomposed ammonia, made great efforts to obtain this trace from the elements at this temperature using iron as a carrier. But with pure gases the experiment was unsuccessful. There was a point of uncertainty here, and if this could be cleared up it would indicate the possibility of a direct synthesis of ammonia from the elements.

I therefore began tentatively to determine the approximate position of the ammonia equilibrium in the vicinity of 1000°C. It now transpired that earlier trials had only proved negative by accident; it was easy, in the vicinity of 1000°C and using iron as a catalyst, to obtain the same ammonia content from both approaches. The results of individual experiments fluctuated between 1/200% and 1/80%, and some discrepant values seemed to me to point to the upper limit as the probable value; later more precise data proved the lower limit to be the correct figure and showed the origin of the higher values to be in the properties of the catalysts, which when fresh temporarily bring about the synthesis of surplus ammonia.

It was further shown that the same results could be obtained with nickel as with iron, and it was found that calcium and in particular manganese were catalysts which would bring about a combination of the elements even at lower temperatures. At 1,000°C the rate of reaction was adequate with a small amount to produce continuously a comparatively large quantity of ammonia. By having a circulation system which alternately brought the gas at high temperature in contact with the metal and then washed out the ammonia at normal temperature, the conversion of a given mass of gas to ammonia could proceed stage by stage.

By determining results at a given pressure, temperature and initial mixture of nitrogen and hydrogen, the state of the theory allowed obtainable results to be approximately predicted for optional temperatures, pressures and mixtures of nitrogen and hydrogen. In the light of the formula, it was possible at once to foresee the increase of attainable maximum content with decreasing temperature, its proportional relationship with the gas pressure, and the fact that a mixture of 3 parts of hydrogen to 1 part of nitrogen must result in the highest ammonia content.

The most important point realized at that time was that from the beginning of red heat onwards no catalyst will produce more than a trace of ammonia from the most favourable gas mixture at normal pressure, and that even at greatly increased pressure the point of equilibrium must continue very unfavourable. If one wished to obtain practical results with a catalyst at normal pressure, then the temperature must not be allowed to rise much beyond 300°.

At that point it seemed to me, in 1905, useless to pursue the problem further. A combination of the elements had certainly been achieved, and the requirements for large-scale synthesis had been outlined; but these requirements appeared so unfavourable that they deterred one from a deeper study of the problem. The discovery of catalysts which would provide a rapid adjustment of the point of equilibrium in the vicinity of 300° and at normal pressure seemed to me quite unlikely: and they have not been found anywhere in the 15 years that have since elapsed.

The synthesis of ammonia which had been demonstrated at normal pressure could be carried out at high pressure on a laboratory scale without any great difficulties. It needed only a slight modification of the pressure oven, such as that used by Hempel 15 years earlier to carry out nitrogen absorption in the case of indirect ammonia synthesis under pressures of up to 66 atmospheres. But I did not think it worth the trouble; at that time I supported the

widely-held opinion that a technical realization of a gas reaction at the beginning of red heat under high pressure was impossible. Here the matter rested for the next three years.

Already in 1906, on the other hand, a new determination of the ammonia equilibrium proved necessary. In the course of his investigations into the heat theorem which has been named after him, Nernst succeeded in finding an approximate formula which permitted a prediction of the equilibria based on the values of the heat effect and the so-called chemical constants. In the case of ammonia this gave a deviation from the values obtained at my first measurements which, as later became apparent, was caused by the original value of the conventional chemical constant of hydrogen then used. This deviation led to fresh determinations of the equilibrium which Nernst had carried out at his Institute in a pressure oven indicated by him while I, in collaboration with Robert le Rossignol, repeated the determinations at normal pressure with greater care than before.

Further work followed, devoted to determining the equilibrium at normal pressure and at 30 atmospheres over an extended range of temperatures, to calculating the heat of formation of ammonia from the elements at normal temperature and at the threshold of red heat, and finally to obtaining knowledge of its specific heat at increased temperature. (See Annotation on p.340.)

<i>t</i> (°C)	<i>T</i> (degr. abs.)	$\frac{P_{NH_3}}{P_{N_2}^{\frac{1}{2}}P_{H_2}^{\frac{1}{2}}}$	$-\log \frac{P_{NH_3}}{P_{N_2}^{\frac{1}{2}}P_{H_2}^{\frac{1}{2}}}$	Percentage of NH_3 at equilibrium			
				at 1 atm	at 30 atm	at 100 atm	at 200 atm
200	473	0.1807	0.660	15.3	67.6	80.6	85.8
300	573	1.1543	0.070	2.18	31.8	52.1	62.8
400	673	1.8608	0.0138	0.44	10.7	25.1	36.3
500	773	2.3983	0.0040	0.129	3.62	10.4	17.6
600	873	2.8211	0.00151	0.049	1.43	4.47	8.25
700	973	3.1621	0.00069	0.0223	0.66	2.14	4.11
800	1,073	3.4417	0.00036	0.0117	0.35	1.15	2.24
900	1,173	3.6736	0.000212	0.0069	0.21	0.68	1.34
1,000	1,273	3.8679	0.000136	0.0044	0.13	0.44	0.87

During the course of these investigations, together with my young friend and co-worker Robert le Rossignol, whose work I would like to mention here with particular sincerity and gratitude, I took up once again, in 1908,

the problem of ammonia synthesis abandoned three years earlier. Immediately prior to this I had become acquainted with the technical processes in the liquefaction of air, and had simultaneously caught a glimpse of the formate industry, which caused flowing carbon monoxide to act upon alkali under heat and increased pressure, and I no longer considered it impossible to produce ammonia on a technical scale under high pressure and at high temperature. But the unfavourable opinion of colleagues taught me that an impressive advance would be needed to arouse technical interest in the subject.

To begin with, it was clear that a change to the use of maximum pressure would be advantageous. It would improve the point of equilibrium and probably the rate of reaction as well. The compressor which we then possessed allowed gas to be compressed to 200 atmospheres, and thus determined our working pressure which could not easily be exceeded for any very large series of experiments. In the neighbourhood of this pressure, the catalysts, with which we had become familiar in the course of our equilibrium determinations, very easily provided a rapid combination of nitrogen and hydrogen at above 700°C; this applied notably to manganese, followed by iron.

To achieve impressive results, however, we needed to discover catalysts which would induce rapid conversion at between 500° and 600°C. We hit upon the idea of searching the sixth, seventh and eighth groups in the Periodic System, whose principal metals chromium, manganese, iron and nickel possessed very definite catalytic properties, for metals which acted even more favourably; these we found in uranium and osmium. At the same time we discovered in osmium an excellent example of the extent to which the performance of a catalyst depends on its composition. When used at 200 atmospheres, both requirements which we deemed necessary to a technically-convincing conduct of the experiment, were met; the first concerned the ammonia content, the second the amount of ammonia produced per cubic centimetre of contact space per hour.

With a content of about 5% the circulation process described in 1905 was no longer a description of a method of synthesis, but a means of manufacture. With a yield of several grams of ammonia per hour per cubic centimetre of heated high-pressure chamber the dimensions of the chamber could be made so small that we felt the objections from the industry must disappear.

Finally we needed an improvement in the circulation system which could act as model for technical realization; separating the synthesis of ammonia

and its removal from the flow of gas by means of reducing the pressure was not a suitable method. The cycle of ammonia production and removal must clearly be achieved by the simplest possible means at a constant high pressure. It seemed essential that the heat produced during the synthesis of ammonia should be removed from exhaust gases, where it had only a deleterious effect, and be led to the fresh gas so that the process itself yielded the heat required for its operation. The construction and operation (carried out in collaboration with Robert le Rossignol) of a small-scale plant which suited these requirements, together with the performance of the new catalysts mentioned, was indeed sufficient to persuade the "Badische Anilin- und Soda-fabrik" which thus far had devoted its attention to the indirect method of producing ammonia by means of the nitrides of aluminium, silicium and titanium, to undertake high-pressure synthesis from the elements.

The company then studied the catalysts on a large scale, using far more substantial means, and discovered ways, in the temperature employed in their production plant and particularly in the deliberate use of inert materials, of improving the performance of poor catalysts to the level of osmium and uranium. Their results were, indeed, important in the case of the classic ammonia catalyst employed by Ramsay and Young, namely iron. They also discovered an improvement in the design of the oven which overcame the effect of hydrogen on the carbon content of steel which they had observed over a long period of operation.

The main work of the company however, was in substituting electrolytic hydrogen, with which we conducted our experiments, for water-gas hydrogen which introduced impurities. The difficulties encountered by the Technical Director Dr. Bosch resembled those which his predecessor Knietsch had overcome with equal success in the course of his technical application of the sulphuric acid contact process. Dr. Bosch has made a large-scale industry of ammonia synthesis.

Present-day industrial working pressures in the vicinity of 200 atmospheres, a working temperature of about 500-600°C, circulation under constant high pressure, and the method of heat exchange from exhaust to inlet gas are all main features of laboratory work which have been retained.

Recently Claude has announced an improvement of the process in the application of higher pressures. The pressure range around 200 atmospheres was originally chosen since it represented the limit of easily attainable levels at the current stage of development in compressor technique. In subsequent experiments Mr. Greenwood and I have gone as far as 370 atmospheres. An

increase in pressure is basically only of interest if it considerably reduces the temperature of rapid conversion without creating fresh technical difficulties.

From the tabulated equilibria (p. 336), it can be seen that the change from normal pressure to 200 atmospheres creates favourable equilibrium conditions - existing between 200° and 300°C - at a temperature 300°C higher, which stimulates more greatly the activity of the catalysts. Why a higher temperature is needed is a question which we must leave to a more enlightened period of science to answer. The heterogenous catalysis of the gas reactions is a process which in the initial phase apparently represents an electrodynamic distortion of the molecule by the atomic fields at the boundary of the solid catalyst material with the gas space; it is thus a phenomenon from a field of molecular physics into which Stark's discovery had just given us a first glimpse.

The synthesis of ammonia from the elements is a result which physical chemistry was bound to reach. The notion of the reversibility of the breakdown of ammonia was already held by Deville, Ramsay and Young, and by 1901 Le Chatelier had already given thought to the effects of temperature and pressure. Failure of the first attempts at synthesis however led him to abandon the matter and to publish his deliberations only in the obscurity of a French patent taken out under a foreign name. This only came to my notice a long time after the successful conclusion of my experiments.

The solution to the problem which has been found assumes its importance from the fact that very high temperature levels are not employed and that this makes the ratio of coal consumption to nitrogen production more favourable than is the case with other processes. Results are enough to show that, in combination with other methods of nitrogen fixation which I have mentioned, they relieve us of future worries caused by the exhaustion of the saltpetre deposits that has threatened us these 20 years.

It may be that this solution is not the final one. Nitrogen bacteria teach us that Nature, with her sophisticated forms of the chemistry of living matter, still understands and utilizes methods which we do not as yet know how to imitate. Let it suffice that in the meantime improved nitrogen fertilization of the soil brings new nutritive riches to mankind and that the chemical industry comes to the aid of the farmer who, in the good earth, changes stones into bread.

Annotation to p.336

The results, in brief, were as follows:

- (a) Actual specific heat C_p of the ammonia gas per mol at constant pressure between 309°C and 523°C is:

$$C_p = 8.62 + 3.5 \times 10^{-3}t + 5.1 \times 10^{-6}t^2.$$

- (b) Heat of formation Q of the ammonia gas at constant pressure in gramcalories per mol from the elements at $t^\circ\text{C}$ is:

$$Q = 10,950 + 4.85t - 0.93 \times 10^{-3}t^2 - 1.7 \times 10^{-6}t^3$$

- (c) Percent content of ammonia in equilibrium with nitrogen-hydrogen mixture

$$(3 \text{ Vol. H}_2 + 1 \text{ Vol. N}_2):$$

The following expression has been used for the calculation:

$$\log \frac{P_{NH_3}}{P_{N_2}^{\frac{3}{2}} P_{H_2}^{\frac{1}{2}}} = \frac{9.591}{4.571 T} - \frac{498}{1.985} \log T - \frac{0.00046 T}{4.571} + \frac{0.85 \times 10^{-6}}{4.571} T^2 + 2.10$$

Also expressions with higher temperature links may be adapted to the observations. A rational expression can only then be postulated when a rational statement concerning the specific heat of all three participant gases has succeeded.

B. KATZ

On the quantal mechanism of neural transmitter release

Nobel Lecture, December 12, 1970

I have been asked on more than one occasion to explain the common denominator between the three of us who are sharing this year's award in physiology or medicine. I think the answer is quite simple: the work of all three has a single source, namely the "discoveries relating to chemical transmission of nerve impulses" for which Henry Dale and Otto Loewi received a previous award in 1936. Dale and his colleagues, W. Feldberg, Marthe Vogt and G. L. Brown, had shown that in spite of the rapid and unfailing nature of neuromuscular transmission, the motor nerve impulse is not simply passed on to the muscle fibre, by a continuous process of electric excitation, but that there is intervention of a chemical mediator, involving the release from the nerve and the subsequent action on the muscle, of a specific transmitter substance, acetylcholine. This concept is summarized in the following scheme



It was only to be expected that on closer examination this intermediate process would resolve itself into a sequence of reactions made up of a number of discrete steps, each of which calls for experimental study. What I should like to do in this lecture is to deal briefly with certain advances that have been made during the last 20 years in the investigation of the first stage of the transmission process, namely the mechanism by which arrival of an impulse enables the motor nerve ending to release the transmitter substance. I shall concentrate on studies made by a micro-electrophysiological approach and shall refer in particular to the work carried out together with my colleagues Paul Fatt, José del Castillo and Ricardo Miledi with all of whom I had the privilege to collaborate (for refs. see refs. 1-4).

It had been known for many years^{5,6} that the end-plate region of the muscle fibre, that is the surface area contacted by the motor nerve, serves as a sensitive chemo-detector for a variety of cholinesters and especially for acetylcholine. When acetylcholine is applied in small amounts to the outer surface, it opens

up ionic channels in the membrane through which ambient cations can pass⁷. This allows sufficient current flow to produce a measurable discharge, or local "depolarization" (*i.e.* lowering of the normal membrane potential) of the muscle fibre. The end-plate surface of the muscle acts, in effect, as a chemo-electric transducer which enables us to register impacts of small quantities of acetylcholine in the form of local potential changes.

Normally, the nerve impulse liberates an amount of acetylcholine which is sufficient to produce a very large local depolarization, often of more than 50 mV, the so-called end-plate potential. This rises quickly above the <firing threshold> of the muscle fibre and so initiates a new propagating wave of membrane excitation.

Some 20 years ago, using the method of intracellular recording, Paul Fatt and I came across something quite unexpected. In the absence of any form of stimulation, the end-plate region of the muscle fibre is not completely at rest, but displays electric activity in the form of discrete, randomly recurring <miniature> end-plate potentials. Each is only of the order of 0.5 mV in amplitude, but in other respects resembles the much larger end-plate potential evoked by the nerve impulse: it shows the same sharp rise and slow decay, and has the character of a discrete all-or-none phenomenon though on a much smaller amplitude scale (see Fig. 1).

Numerous experiments have shown that each miniature end-plate potential arises from the synchronous impact of a large multi-molecular quantum of acetylcholine spontaneously discharged by the adjacent nerve terminal. Each event is highly localized and involves only a very small portion of the synaptic axon surface; successive discharges form a random sequence in temporal as well as in spatial distribution along the motor nerve ending.

One of the unanswered questions concerns the precise number of acetylcholine molecules which make up each quantal unit of discharge. This is still uncertain; on present estimates one may assume that at least a thousand molecules are contained within an elementary packet, possibly many more. As an upper limit R. Miledi⁸ gave a figure of 10^5 molecules; this was based on the minimum quantity of applied acetylcholine which was needed to produce an equivalent effect.

One of the difficulties in answering this question is that neither the chemosensitivity of the end-plate nor the resolving power of our recording system is high enough to enable us to detect the effects of individual or of pauci-molecular reactions of acetylcholine directly. An indirect approach was chosen quite recently by R. Miledi and me⁹. We found that a steady dose of acetyl-

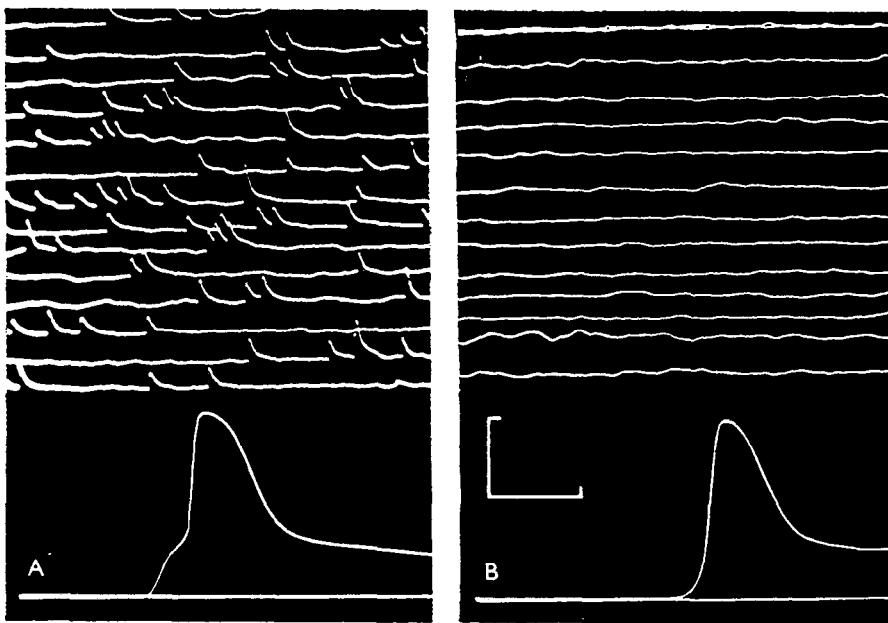


Fig. 1. Spontaneous "miniature end-plate potentials". (From ref. 1) A: intracellular recording at an end-plate. B: recorded 2 mm away in same muscle fibre. Upper portions were recorded at low speed and high amplification (calibrations 3.6 mV and 46 msec); they show the localized spontaneous activity at the junctional region. Lower records show the electric response to a nerve impulse, taken at high speed and lower gain (calibrations 50 mV and 2 msec). The stimulus was applied to the nerve at the beginning of the trace; response A shows step-like end-plate potential leading to a propagating muscle spike; in B, the spike alone is recorded after additional delay due to conduction along 2 mm of muscle fibre. In all figures, unless otherwise stated, upward deflexion means positive-going potential change at the micro-electrode.

choline produces not only a measurable depolarization, but also a measurable excess of voltage noise in the end-plate zone. These experiments are still in progress; a preliminary analysis indicates that the elementary "shot effects" which underlie, and statistically build up, the steady acetylcholine-potential amount to a fraction of a microvolt in amplitude. This elementary voltage change is presumably due to the transient opening of an ionic channel in the muscle membrane, by the action of one or several acetylcholine molecules. The miniature potential is a thousand times larger and would thus require the opening of one or a few thousand "ionic gates", but this figure still leaves us in doubt about the number of acetylcholine molecules which are discharged in a single quantal unit from the nerve terminal.

The discovery that acetylcholine is spontaneously released from nerve endings in the form of large multi-molecular packets acquired further significance as a result of the following findings.

(i) The presence of miniature potentials is not peculiar to the neuromuscular junction nor is it confined to cholinergic systems, but has been found to occur at diverse kind of synapses, in the peripheral and central nervous system, and throughout the animal kingdom. It appears to be a characteristic property of many, maybe of all, those neuronal and neuro-effector junctions at which chemical mediation occurs.

(ii) An important structural correlate has been found by De Robertis and Bennett¹⁰ and by Palade and Palay¹¹ in the synaptic vesicles which form a distinct population of intra-axonal organelles clustered together near the pre-synaptic release sites. Recent biochemical studies, especially on the electric organ of *Torpedo*^{1,2} have shown that the major part of the acetylcholine stores of a cholinergic axon is parcelled up within these vesicular organelles. Electron-microscopic studies on the same tissue, using the freeze-etching technique¹³ have revealed many instances in which such vesicles are found to be attached to the presynaptic axon membrane and to have opened into the synaptic cleft.

(iii) One of the most important findings was that the frequency of the miniature end-plate potentials, that is the rate of acetylcholine secretion, is controlled by the membrane potential of the axon terminals: depolarization of the presynaptic membrane causes the rate of the discharge to increase in a graded manner without change in the size of the individual <blips> (ref. 14, see Fig. 2). This is, in fact, what occurs after the arrival of an action potential : after a short delay the frequency of miniature end-plate potentials rises by several orders of magnitude for a brief period and then rapidly falls again towards the resting level¹⁵. The result is a very large synchronous end-plate potential which exceeds the <firing threshold> of the muscle fibre.

It has been well established that the normal end-plate potential is made up of a statistical fusion of quantal components which are identical with the spontaneously occurring units. In effect, the nerve impulse does not start up a new secretory process, but it facilitates or, in statistical terms, raises the probability of events that occur all the time at a low rate, so that - instead of an average of one packet per second - we obtain a few hundred packets of transmitter substance released within a millisecond.

How does the nerve impulse, or more generally, how does a depolarization raise the probability of occurrence of this quanta1 event? The membrane po-

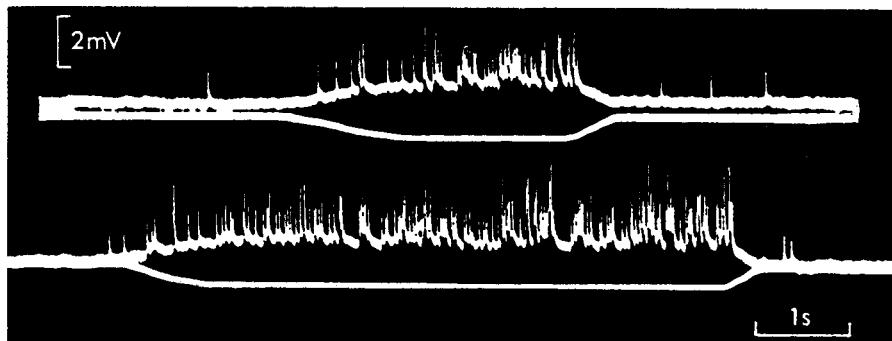


Fig. 2. Electrical control of the frequency of m.e.p.p.'s (cf. ref. 14). In each pair of traces, the upper shows miniature potentials, the lower indicates current flowing through the terminal part of the motor axon. The cathode was placed near the junction so as to depolarize the nerve endings. This caused the frequency of the discharge to increase dramatically.

tential change alone is not sufficient. The presence of calcium ions in the external medium is required to make depolarization effective (see Fig. 3). If one progressively reduces the calcium concentration, or adds a <competitive> ion such as magnesium or manganese in increasing amounts, depolarization becomes less and less capable of accelerating the discharge rate of miniature potentials above their resting frequency. Over the last six years, experiments by Miledi and myself have led us to conclude that external calcium is the only immediate ionic requirement for depolarization to evoke transmitter release¹⁶⁻¹⁸.

On our present evidence, the sequence of events may be described as follows: depolarization opens specific <calcium gates> in the-terminal axon membrane-this leads to an influx of calcium ions (provided the membrane potential has not been displaced excessively, *i.e.* to or beyond the calcium equilibrium level, see ref. 19). Having reached the internal surface of the axon membrane, calcium ions then initiate the "quantal release reaction". Up to this point of the argument we are on reasonably firm ground established by electrophysiological experiments. Beyond this point, I must draw on converging observations from ultrastructural and biochemical studies, all of which go to make up a plausible and, I think, very strong hypothesis, namely that the quanta of transmitter molecules are enclosed within synaptic vesicles which undergo frequent transient collisions with the axon membrane, that calcium brings about attachment and local fusion between vesicular and axon membranes, and that this is followed by all-or-none discharge of the vesicu-

lar content into the synaptic cleft. To students of neurosecretory process the postulate of <evacuation> of vesicles or of dense-cored granules from the cell

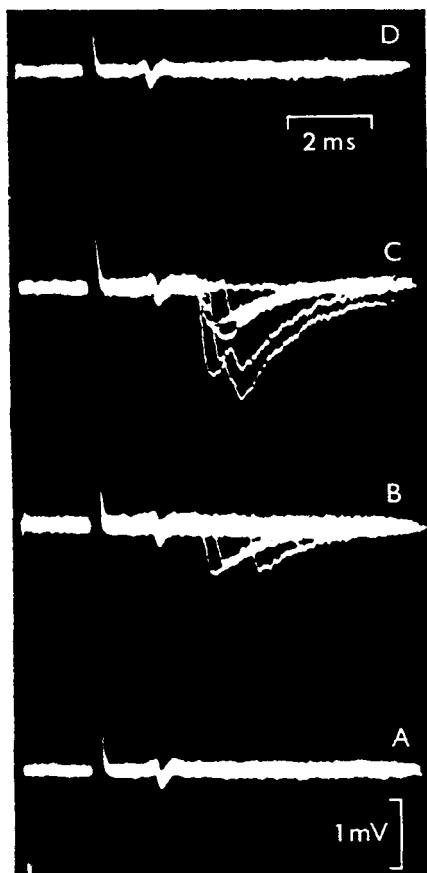


Fig. 3. Exploration of the nerve-muscle junction with a "calcium micropipette". (From ref. 29.) A frog sartorius was immersed in Ca-free solution containing 0.84 mM Mg. A micropipette filled with 0.5 M CaCl_2 was used to record focal external potentials from a localized junctional spot. Efflux of Ca was controlled electrophoretically. In the 4 records, A to D (each obtained by superimposing six traces), Ca-efflux was stopped initially by applying sufficient negative voltage to the pipette (in A); the bias was then reduced in two steps (in B and C) allowing increasing amounts of Ca to escape; finally the negative bias was re-applied (in D) to stop Ca-efflux again. In A (no Ca), each stimulus is followed at constant interval by a small spike in the nerve terminal, but there is no transmitter release. In B (little Ca), single transmitter quanta were released, after variable "synaptic delay", in 3 out of the 6 trials. In C (more Ca), single or multiple transmitter quanta were now released in 5 out of 6 trials, while in D (no Ca) transmitter release was stopped again.

surface is a familiar and well-documented event; in present-day terminology it is usually described as <exocytosis>.

If I am asked how all this research, which is developing rapidly along with biochemical, physiological and cytological micro-techniques, is going to influence our understanding of the operation of the central nervous system and of its functional defects, I am on much more tenuous ground. There is some evidence that in certain peripheral neuromuscular syndromes the quantal release efficacy of the motor impulse is impaired²⁰, resembling somewhat the experimental condition of low Ca/high Mg, while in other forms of myoneural disease a <packaging failure>, *i.e.* insufficient accumulation of transmitter by individual vesicles, has been suggested²¹. It would clearly be of interest to pursue this line and to find out whether there is similarly localized involvement in some central-nervous lesions.

Finally, there is the more general question, whether the statistical fluctuations and <uncertainties> which are inevitably associated with the quantal nature of transmitter release, play any recognizable role in the organized function of the nervous system. That there are large quantal fluctuations in the response of unitary synapses in the central nervous system, has been shown very clearly by Kuno²² and others. In many instances, the number of packets released by an impulse impinging on a spinal motoneurone was found to be small, and the predictability of the synaptic response appeared to follow the statistical rules of Poisson's law. In other instances, the fluctuations were much smaller, indicating a greater efficacy of the afferent impulse, so that either a large average number of packets was being discharged from the terminal²³⁻²⁵, or a small number was being released with high probability. One would presume that in a <fully trained> neuronal pathway, quantal fluctuations become unimportant because of simultaneous involvement of a large population of synaptic transfer sites. The larger the average number, the smoother and more predictable becomes the synaptic performance. However, large numbers and smooth performance may not be the rule at *all* times and in *all* pathways. Experiments on the neuromuscular junction have shown that certain processes of synaptic modification during and after prolonged activity, are associated with quantal recruitment, that is with an increase in number of packages delivered per impulse²⁶⁻²⁸. Similar changes would be expected to occur, and make synaptic performance more predictable, during development and <training>, while the opposite trend might underlie some forms of pathological and degenerative impairment.

1. P. Fatt and B. Katz, *J. Physiol. (London)*, 117(1952) 109.
2. J. del Castillo and B. Katz, *Progr. Biophys. Biophys. Chem.*, 6 (1956) 121.
3. B. Katz and R. Miledi, *Studies in Physiology*, Springer, Berlin, 1965, pp. 118 ff.
4. B. Katz, *The Release of Neural Transmitter Substances*, Liverpool University Press, 1969.
5. J.N. Langley, *J. Physiol. (London)*, 36 (1907) 347.
6. S.W. Kuffler, *J. Neurophysiol.*, 6 (1943) 99.
7. A. Takeuchi and N. Takeuchi, *J. Physiol. (London)*, 154 (1960) 52.
8. R. Miledi, *Discovery*, 22 (1961) 442.
9. B. Katz and R. Miledi, *Nature*, 226 (1970) 962.
10. E.D.P. deRobertis and H.S. Bennett, *Federation Proc.*, 13 (1954) 35.
11. G.E. Palade and S.L. Palay, *Anat. Record*, 118 (1954) 335.
12. M. Israel, J. Gautron and B. Lesbats, *Compt. Rend.*, 266 (1968) 273.
13. E. Nickel and L.T. Potter, *Brain Res.*, 23(1970) 95.
14. J. del Castillo and B. Katz, *J. Physiol. (London)*, 124 (1954) 586.
15. B. Katz and R. Miledi, *Proc. Roy. Soc. (London), Ser.B* 161 (1965) 483.
16. B. Katz and R. Miledi, *J. Physiol. (London)*, 203 (1969) 459.
17. B. Katz and R. Miledi, *J. Physiol. (London)*, 203 (1969) 689.
18. B. Katz and R. Miledi, *J. Physiol. (London)*, 207 (1970) 789.
19. B. Katz and R. Miledi, *J. Physiol. (London)*, 192 (1967) 407.
20. D. Elmqvist and E.H. Lambert, *Mayo Clin. Proc.*, 43 (1968) 689.
21. D. Elmqvist, W.W. Hofmann, J. Kugelberg and D.M.J. Quastel, *J. Physiol. (London)*, 174 (1964) 417.
22. M. Kuno, *J. Physiol. (London)*, 175 (1964) 81.
23. R.E. Burke and P.G. Nelson, *Science*, 151 (1966) 1088.
24. E. Eide, L. Fedina, J. Jansen, A. Lundberg and L. Vyklicky, *Nature*, 215 (1967) 1176.
25. M. Kuno and J. T. Miyahara, *J. Physiol. (London)*, 201 (1969) 465.
26. J. del Castillo and B. Katz, *J. Physiol. (London)*, 124 (1954) 574.
27. R. Miledi and R.E. Thies, *J. Physiol. (London)*, 192 (1967) 54P.
28. B. Katz and R. Miledi, *J. Physiol. (London)*, 195 (1968) 481.
29. B. Katz and R. Miledi, *Proc. Roy. Soc. (London), Ser.B* 161 (1965) 496.

F R I T Z L I P M A N N

Development of the acetylation problem: a personal account

Nobel Lecture, December 11, 1953

The fact that my Swedish colleagues have honored me with the Nobel Prize gives me some confidence to consider my own effort more seriously as a part in the general effort of biochemistry of today. I therefore thought of tracing, in the segment of my interest, the recent development of facts and ideas which led, it seems, to a fuller understanding of the chemical functioning of the organism. When I started out in the middle twenties, biochemistry was just trying to break away from the major concern with breakdown processes and procedures. With the slowly increasing comprehension of biosynthetic mechanisms, a rather radical change of attitude ensued which is, I feel, not quite fully realized even at the present time. Out of the early, justifiably stubborn empiricism grew up a definite rational structure. Process patterns emerged and it became important to recognize certain rules and introduce new terms, thereby emphasizing the fact that biochemistry was now developing into an adult science, best characterized, may be, as organismic technology.

In my development, the recognition of facts and the rationalization of these facts into a unified picture, have interplayed continuously. After my apprenticeship with Otto Meyerhof, a first interest on my own became the phenomenon we call the Pasteur effect, this peculiar depression of the wasteful fermentation in the respiring cell. By looking for a chemical explanation of this economy measure on the cellular level, I was prompted into a study of the mechanism of pyruvic acid oxidation, since it is at the pyruvic stage where respiration branches off from fermentation. For this study I chose as a promising system a relatively simple looking pyruvic acid oxidation enzyme in a certain strain of *Lactobacillus delbrueckii*. The decision to explore this particular reaction started me on a rather continuous journey into partly virgin territory to meet with some unexpected discoveries, but also to encounter quite a few nagging disappointments.

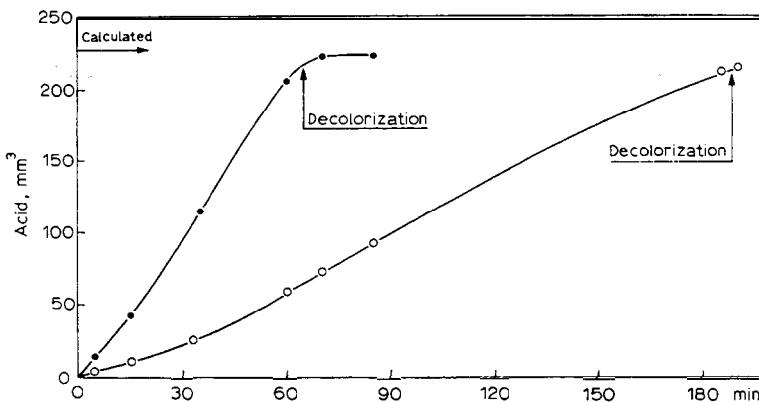


Fig. 1. Formation of acid in the methylene blue reduction. Action of phosphate.
 ●—● Addition of phosphate, equivalent to a total concentration of $4 \cdot 10^{-3}$ M.
 ○—○ Without phosphate.

Discovery of acetyl phosphate

The most important event during this whole period, I now feel, was the accidental observation that in the *L. delbrueckii* system, pyruvic acid oxidation was completely dependent on the presence of inorganic phosphate. This observation was made in the course of attempts to replace oxygen by methylene blue. To measure the methylene blue reduction manometrically, I had to switch to a bicarbonate buffer instead of the otherwise routinely used phosphate. In bicarbonate, to my surprise, as shown in Fig. 1, pyruvate oxidation was very slow, but the addition of a little phosphate caused a remarkable increase in rate. The next figure, Fig. 2, shows the phosphate effect more drastically, using a preparation from which all phosphate was removed by washing with acetate buffer. Then it appeared that the reaction was really fully dependent on phosphate.

In spite of such a phosphate dependence, the phosphate balance measured by the ordinary Fiske-Subbarow procedure did not at first indicate any phosphorylative step. Nevertheless, the suspicion remained that phosphate in some manner was entering into the reaction and that a phosphorylated intermediary was formed. As a first approximation, a coupling of this pyruvate oxidation with adenylic acid phosphorylation was attempted. And, indeed, addition of adenylic acid to the pyruvic oxidation system brought out a net disappearance of inorganic phosphate, accounted for as adenosine triphosphate (Table 11). In parallel with the then just developing fermentation pic-

Table 1. Disappearance of inorganic phosphate with adenylic acid.

<i>Initial value</i>	<i>0.125 M</i> <i>pyruvate</i>	<i>0.125 M</i> <i>pyruvate, 0.03 M adenylic acid</i>	<i>0.03 M adenylic acid</i>
Inorganic P (mg)	0.59	0.53	0.32
Easily hydrolyzed P (mg)	0	0.06	0.28
O ₂ (μ l)	—	490	474
			58

ture, I now concluded that the missing link in the reaction chain was acetyl phosphate. In partial confirmation it was shown that a crude preparation of acetyl phosphate, synthesized by the old method of Kämmerer and Carius² would transfer phosphate to adenylic acid (Table 2). However, it still took quite some time from then on to identify acetyl phosphate definitely as the initial product of the pyruvic oxidation in this system^{3,4}. Most important

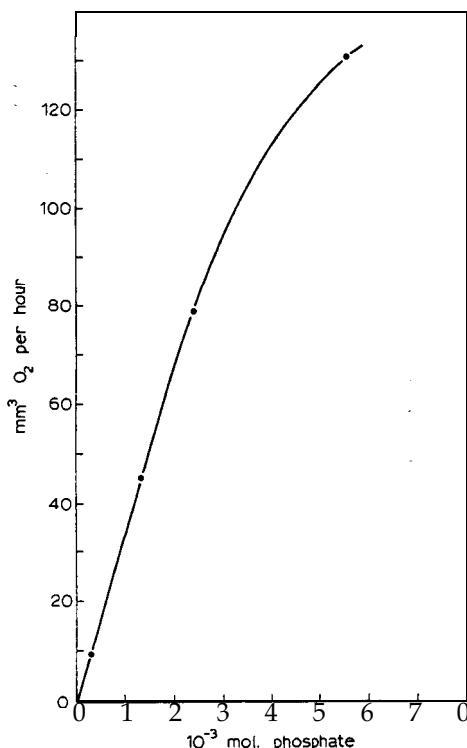


Fig. 2. Phosphate dependence of pyruvate oxidation.

Table 2. Transfer of phosphate from acetyl phosphate to adenylic acid with bacterial preparations.

<i>Adenylic acid (mg)</i>	<i>4</i>	<i>0.1</i>	—
P (mg), inorg. + labile (acetyl P)	1.12	1.39	1.37
P (mg), after 7 min hydrolysis at 100° in normal HCl	1.47	1.48	1.47
P (mg) formed ⁷	0.35	0.09	0.10

(Fresh solution, containing 0.75 mg acid-labile P in 0.5 cc., 46 mg dry bacteria, total volume 1.25 cc., with 0.04 M in NaF.)

during this and later work became the development of procedures⁵ and in particular of the very handy hydroxamic acid method⁶ for the determination of acyl phosphates and other reactive acyl derivatives.

At the time when these observations were made, about a dozen years ago, there was, to say the least, a tendency to believe that phosphorylation was rather specifically coupled with the glycolytic reaction. Here, however, we had found a coupling of phosphorylation with a respiratory system. This observation immediately suggested a rather sweeping biochemical significance, of transformations of electron transfer potential, respiratory or fermentative, to phosphate bond energy and therefrom to a wide range of biosynthetic reactions⁷.

There was a further unusual feature in this pyruvate oxidation system in that the product emerging from the process not only carried an energy-rich phosphoryl radical such as already known, but the acetyl phosphate was even more impressive through its energy-rich acetyl. It rather naturally became a contender for the role of "active" acetate, for the widespread existence of which the isotope experience had already furnished extensive evidence. I became, therefore, quite attracted by the possibility that acetyl phosphate could serve two rather different purposes, either to transfer its phosphoryl group into the phosphate pool, or to supply its active acetyl for biosynthesis of carbon structures. Thus acetyl phosphate should be able to serve as acetyl donor as well as phosphoryl donor, transferring, as shown in Fig. 3, on either side of the oxygen center, such as indicated by Bentley's early experiments on cleavage^{7a} of acetyl phosphate in H₂¹⁸O.

These two novel aspects of the energy problem, namely (1) the emergence of an energy-rich phosphate bond from a purely respiratory reaction; and (2) the presumed derivation of a metabolic building-block through this same reaction, prompted me to propose not only the generalization of the phosphate bond as a versatile energy distributing system, but also to aim from

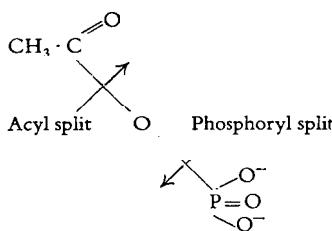


Fig. 3. Acetyl phosphate as acetyl and phosphoryl donor.

there towards a general concept of transfer of activated groupings by carrier as the fundamental reaction in biosynthesis^{8,9}. Although in the related manner the appearance of acetyl phosphate as a metabolic intermediary first focussed attention to possible mechanisms for the metabolic elaboration of group activation, it soon turned out that the relationship between acetyl phosphate and acetyl transfer was much more complicated than anticipated.

Since a better understanding of the mechanisms of group activation seemed to become a most urgent problem in biosynthesis, I now set out to find a suitable system to check on the assumption that acetyl phosphate represented active acetate. After working out a relatively easy method to prepare the compound^{5,10}, a first unexpected difficulty arose when it appeared that animal tissues contain rather generally a very active, specific and heat-stable acetyl phosphatase^{11,9}. In crude preparations of muscle, liver and other tissues the half life of acetyl phosphate is only a few minutes. This strange activity in animal tissues made tests with this substance very difficult.

In looking for a sensitive method to study acetyl transfer, the acetylation of aromatic amines was chosen eventually as a most promising and technically easy procedure. We were furthermore quite confident that any results obtained with this method could be generalized over the whole metabolic territory concerning the transfer of active acetate including such reactions as citrate, acetoacetate and lipid synthesis. Acetylation of sulfonamide had been found to occur in rabbit liver slices¹². However, for our purposes, we had to eliminate cell membrane barriers to test for the activity of complex intermediary metabolites. Although acetylation was found with rabbit liver homogenate, the reaction was rather weak. In search of a more active system, pigeon liver homogenate was tried and found to harbour an exceedingly potent acetylation system (Ref. 11, cf. also Ref. 12). This finding of a particularly active acetylation reaction in cell-free pigeon liver preparations was most fortunate and played a quite important part in the development of the acetylation problem.

We had now eventually arrived at the point where the desired test for acetyl phosphate as an acetyl precursor could be performed. Although the acetyl phosphatase activity of the pigeon liver homogenate was considerable and, to some extent, obscured the test with acetyl phosphate, it became, nevertheless, clear to us that in this preparation, acetyl phosphate did not furnish active acetate¹¹. Under anaerobic conditions with massive concentrations of acetyl phosphate, no acetyl groups for the acetylation of sulfonamide could be derived under conditions where an easy acetylation occurred with a respiring homogenate.

It furthermore appeared that as an energy source the particle bound oxidative phosphorylation of the kind observed first by Herman Kalckar¹⁴ could be replaced by ATP, as had first been observed with the acetylation of choline in brain preparations by Nachmansohn and his group^{15,16}. Using ATP and acetate as precursors, it was possible to set up a homogeneous particle-free acetylation system obtained by extraction of acetone pigeon liver. In this extract likewise acetyl phosphate was unable to replace the ATP-acetate as acetyl precursor.

Discovery and identification of coenzyme A

In spite of this disappointment with acetyl phosphate, our decision to turn to a study of acetylation started then to be rewarding in another way. During these studies we became aware of the participation of a heat-stable factor which disappeared from our enzyme extracts on aging or dialysis. This co-factor was present in boiled extracts of all organs, as well as in micro-organisms and yeast. It could not be replaced by any other known cofactor. Therefore, it was suspected that we were dealing with a new coenzyme. From then on, for a number of years, the isolation and identification of this coenzyme became the prominent task of our laboratory. The problem now increased in volume and I had the very good fortune that a group of exceedingly able people were attracted to the laboratory; first Constance Tuttle, then Nathan O. Kaplan and shortly afterwards, G. David Novelli. More recently, Morris Soodak and John Gregory, and quite a few others have made here most important contributions to the advance of this problem.

Early data on the replacement of this heat-stable factor by boiled extracts are shown in the next table (Table 3). The pigeon liver acetylation system proved to be a very convenient assay system for the new coenzyme¹⁷ since

Table 3. Reversible inactivation through dialysis or autolysis.

<i>Treatment of extract</i>	<i>Filtrate of boiled organ added, corresponding to gram fresh weight</i>	<i>Sulfanilamide conjugated (γ)</i>	<i>Incuba-tion time (min)</i>
Untreated		69	65
Kept 16 hours, 7°		7	40
	0.2 g rat liver	58	
Dialyzed 16 hours, 7°		0	65
	0.2 g rat liver	42	
Untreated		59	50
Kept 16 hours, 7-10°		0	
	0.4 g rat liver	28	

(1 ml of extract in a total volume of 2 ml; magnesium chloride and sodium acetate were present in 0.02 M concentration. The experiment was started through addition of a mixture of 0.32 mg of adenyl polyphosphate P, 88 μ of sulfanilamide, and fluoride to 0.05 M final concentration.)

on aging for 4 hours at room temperature, the cofactor was completely auto-lyzed. Fortunately, on the other hand, the enzyme responsible for the decomposition of this factor was quite unstable and faded out during the aging, while the acetylation apoenzymes were unaffected.

The next figure, Fig. 4, shows coenzyme A (CoA) assay curves obtained

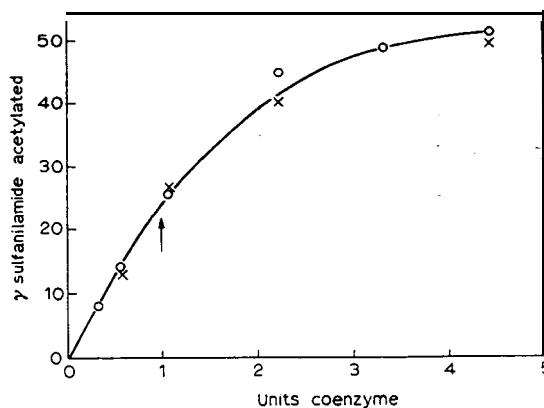


Fig. 4. Concentration-activity curves for coenzyme A preparations of different purity. The arrow indicates the point of 1 unit on the curve. (o) crude coenzyme, 0.25 unit per mg; (x) purified coenzyme, 130 units per mg.

with acetone pigeon liver extract. Finding pig liver a good source for the coenzyme, we set out to collect a reasonably large quantity of a highly purified preparation and then to concentrate on the chemistry with this material. In this analysis we paid particular attention to the possibility of finding in this obviously novel cofactor one of the vitamins, then not as yet metabolically identified. In this task we were very fortunate to have the help of the great experience of Dr. Roger Williams' laboratory. Dr. Beverly Guirard, who occupied herself with this preparation, at first seemed not to find any appreciable amounts of the known vitamins. However, she became aware of the fact that on prolonged enzymatic treatment, the value of pantothenic acid, as determined microbiologically, did slightly increase. This gave the hint that the coenzyme may not release the pantothenic acid so easily, a fact well known from experience with pantothenic acid assay in tissue extracts. In confirmation, she found on acid hydrolysis of the coenzyme, considerable amounts of β -alanine, corresponding to 11 per cent of pantothenic acid in this preparation which, as we now know, was 40 per cent pure. The results of Dr. Guirard's vitamin survey, which gave us the practical assurance of the presence of pantothenic acid in the new coenzyme¹⁸, are shown in Table 4.

Table 4. Vitamin content of preparation A.

<i>Vitamins</i>	<i>Per cent</i>
Nicotinic acid	0.06
Folic acid	0.0002
Riboflavin	0.006
Inositol	0.05
Pyridoxine	0.03
Biotin, thiamine	not detectable
Pantothenic acid	
Direct	0.085
Incubated with papain-clarase, 1 week	0.16
From β -alanine, after acid hydrolysis	11.0

The appearance of a B-vitamin in the preparation was of course a most exciting event for our group and gave us further confidence that we were dealing here with a key substance. We still felt, however, slightly dissatisfied with the proof for pantothenic acid. Therefore, to liberate the chemically rather unstable pantothenic acid from CoA, we made use of observations on

enzymatic cleavage of the coenzyme. Two enzyme preparations, intestinal phosphatase and an enzyme in pigeon liver extract, had caused independent inactivation. It then was found that through combined action of these two enzymes, pantothenic acid was liberated^{18,19}.

The two independent enzymatic cleavages indicated early that in CoA existed two independent sites of attachment to the pantothenic acid molecule. One of these obviously was a phosphate link, linking presumably to one of a hydroxyl group in pantothenic acid. The other moiety attached to pantothenic acid, which, cleaved off by liver enzyme, remained unidentified for a long time. In addition to pantothenic acid, our sample of 40 per cent purity had been found to contain about 2 per cent sulfur by elementary analysis and identified by cyanide-nitroprusside test as a potential SH-grouping^{20,21}. Furthermore, the coenzyme preparation contained large amounts of adenylic acid²¹.

In the subsequent elaboration of the structure, the indications by enzyme analysis for the two sites of attachment to pantothenic acid have been most helpful. The phosphate link was soon identified as a pyrophosphate bridge²²; 5-adenylic acid was identified by Novelli²³ as enzymatic split product and by Baddiley²⁴, through chemical cleavage. At the same time, Novelli made observations which indicated the presence of a third phosphate in addition to the pyrophosphate bridge. These indications were confirmed by analysis of a nearly pure preparation which was obtained by Gregoryas from *Streptomyces fradiae* in collaboration with the research group at the Upjohn Company²⁵. The generous help of the Upjohn Laboratories has been of great

Table 5. Composition of best preparation²¹ of coenzyme A.²¹

	<i>Calculated*</i> (%)	<i>Found</i> (%)	<i>Ratio</i>
Pantothenic acid	28.6	26.8 (enzymatic assay) 25.6 (microbiological)	1
Adenine	17.6	17.0 (spectrophotometric)	1.05
Phosphorus (total)	12.12	10.6	2.83
Monoester phosphorus**	—	3.6	0.96
Sulfur	4.18	4.13	1.07

* Pantothenic acid, 2-mercaptopethylamine, 3 phosphoric acid, adenosine, - 5H₂O; molecular weight 767.

** Liberated by prostate phosphomonoesterase.

importance for the final identification of the structure of CoA. The analysis of this practically pure preparation is presented in Table 5.

It was at this period that we started to pay more and more attention to the sulfur in the coenzyme. As shown in Table 5, our purest preparation contained 4.13 per cent sulfur corresponding to one mole per mole of pantothenate. We also found²⁶ that dephosphorylation of CoA yielded a compound containing pantothenic acid and the sulfur carrying moiety, which we suspected as bound through the carboxyl. Through the work of Snell and his group²⁷, the sulfur-containing moiety proved to be attached to pantothenic acid through a link broken by our liver enzyme. It was identified as thioethanolamine by Snell and his group, linked peptidically to pantothenic acid.

Through analysis and synthesis, Baddiley now identified the point of attachment of the phosphate bridge to pantothenic acid in 4-position²⁴ and Novelli *et al.*²⁸ completed the structure analysis by enzymatic synthesis of "dephospho-CoA" from pantetheine-4'-phosphates and ATP. Furthermore, the attachment of the third phosphate was identified by Kaplan²⁹ to attach in s-position on the ribose of the 5-adenylic acid (while in triphosphopyridine nucleotide it happens to be in 2-position). Therefore, the structure was now established, as shown in Fig. 5.

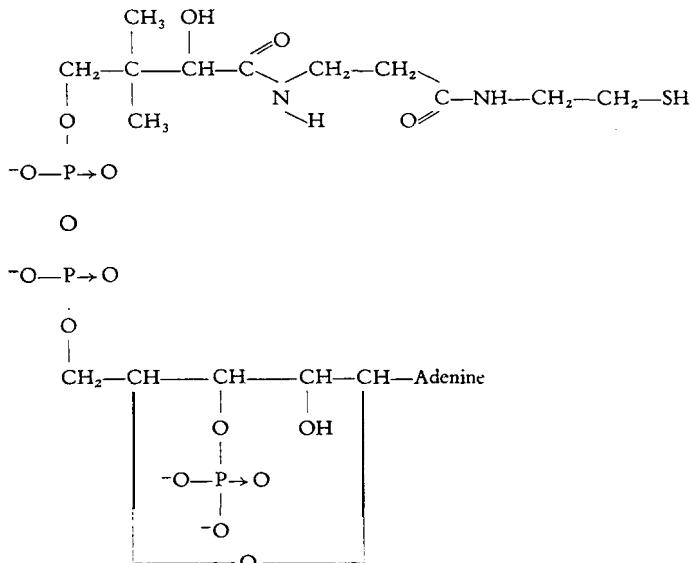


Fig. 5. Structure of coenzyme A.

The metabolic function of CoA

Parallel with this slow but steady elaboration of the structure, all the time we explored intensively metabolic mechanisms in the acetylation field. By use of the enzymatic assay, as shown in Tables 6, 7, 8, and 9, CoA was found present in all living cells, animals, plants and microorganisms¹⁷. Furthermore, the finding that all cellular pantothenic acid could be accounted for by CoA¹⁷

Table 6. Coenzyme A in animal tissues.

(All values are given in units of coenzyme A per g of fresh tissue.)

	<i>Human</i>	<i>Rabbit</i>	<i>Rat</i>	<i>Pigeon</i>
Liver		112	—	132
Adrenal		65	91	
Adrenal, demedullated*			79	
Kidney		50	74	
Brain		40 (cortex)	28	40
Heart		26	42	45
Testes		26		
Intestine			26	
Thymus			20	
Skeletal muscle		6		
Blood plasma	0			
Red blood cells	3-4			

* We wish to thank Dr. H. W. Deane and Dr. R. O. Greep for the demedullated glands.

Table 7. Coenzyme A in micro-organisms.*

<i>Micro-organism</i>	<i>Coenzyme A per g dry weight (units)</i>	<i>Preparation used for assay</i>
<i>Proteus morganii</i>	572	Freshly grown organism, boiled
<i>Lactobacillus arabinosus</i>	150	
<i>Lactobacillus delbrueckii</i>	40	Dry preparation, suspended in water and boiled
Tray-dried yeast	72	
Tray-dried yeast	41	
<i>Escherichia coli</i>	320	
Propionic acid bacteria	330	
<i>Clostridium butylicum</i> (dried extract)	2,000	

* We wish to thank Mr. G. D. Novelli for collaboration in these experiments.

made it clear that CoA represented the only functional form of this vitamin. The finding of the vitamin furnished great impetus; nevertheless, a temptation to connect the pantothenic acid with the acetyl transfer function has blinded us for a long time to other possibilities.

The first attempts to further explore the function of CoA were made with

Table 8. Coenzyme A in plant material.

	<i>Coenzyme A per g fresh weight (units)</i>
Spinach	0.74
Tomato	1.3
Frozen peas	4.5 —
Wheat germ (commercial sample)	30
Royal jelly (bee)*	0

* Kindly supplied by Dr. Thomas S. Gardner.

Table 9. Citric acid synthesis in dialyzed extract of *E. coli*.

<i>Additions</i>	<i>Citric acid synthesized per ml extract (γM)</i>
None	0
Acetate, ATP	0.23
Acetate, ATP, coenzyme A	1.30
Acetyl phosphate	0.25
Acetyl phosphate, coenzyme A	4.0

(All tubes contained 1.0 ml of extract, 0.025 M oxalacetic acid, 0.0016 M $NaHCO_3$, 0.02 M $MgCl_2$, and 0.01 M cysteine in a final volume of 2.5 ml. The concentrations of the additions were as follows: sodium acetate 0.05 M, sodium ATP 0.02 M, lithium acetyl phosphate 0.004 M, and coenzyme A 17 units.)

pantothenic acid-deficient cells and tissues. A deficiency of pyruvate oxidation in pantothenic acid-deficient *Proteus morganii*, an early isolated observation by Dorfman³⁰ and Hills³¹, now fitted rather well into the picture. We soon became quite interested in this effect, taking it as an indication for participation of CoA in citric acid synthesis. A parallel between CoA levels and pyruvate oxidation in *Proteus morganii* was demonstrated³². Using panto-

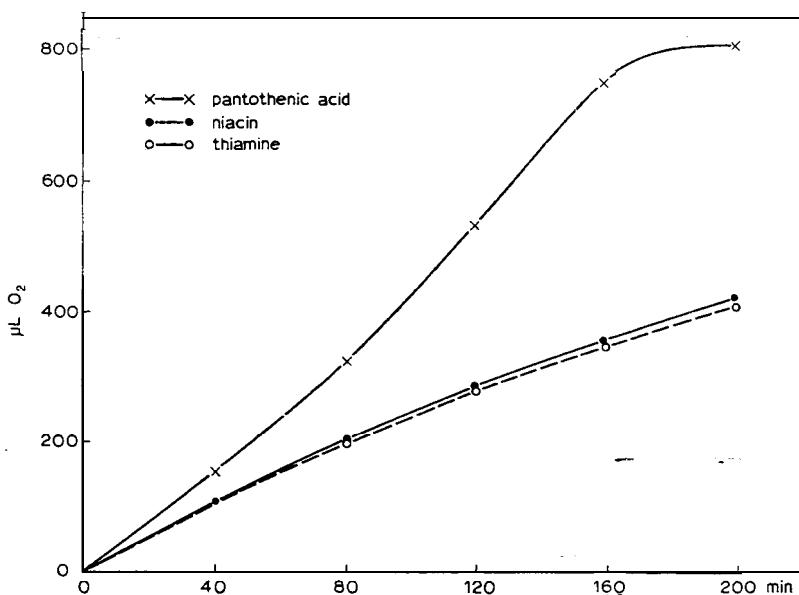


Fig. 5a. Effect of coenzyme A on acetate oxidation in yeast. Pantothenate-deficient yeast was preincubated in glucose-phosphate medium with 50 γ of pantothenate, 100 γ of thiamine, and 100 γ of niacin in separate flasks; 5.6 mg of dry weight of each suspension were added to individual Warburg vessels. Total fluid volume 3.0 ml of 0.06 M KH_2PO_4 , and 0.01 M acetic acid; temperature 37°; gas phase air.

thenic aciddeficient yeast, Novelli *et al.*³³ demonstrated a CoA-dependence of acetate oxidation (Fig. 5a) and Olson and Kaplan³⁴ found with duck liver a striking parallel between CoA content and pyruvic utilization, which is shown in Fig. 6.

But more important information was being gathered on -the enzymatic level. The first example of a generality of function was obtained by comparing the activation of apoenzymes for choline- and sulfonamide-acetylation respectively, using our highly purified preparations⁹ of CoA. As shown in Fig. 7, similar activation curves obtained for the two respective enzymes. Through these experiments, the heat-stable factor for choline acetylation that had been found by Nachmansohn and Berman³⁵ and by Feldberg and Mann³⁶ was identified with CoA.

The next most significant step toward a generalization of CoA function for acetyl transfer was made by demonstrating its functioning in the enzymatic synthesis of acetoacetate. The CoA effect in acetoacetate synthesis was studied by Morris Soodak³⁷, who obtained for this reaction a reactivation

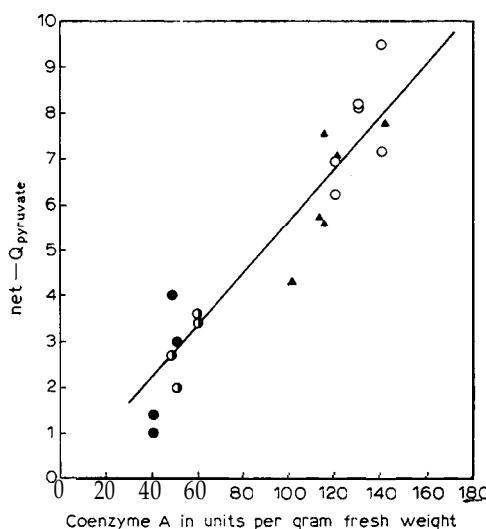


Fig. 6. Relationship between net pyruvate utilization (net- Qpyruvate) and coenzyme A content of liver slices from deficient, pantothenic acid-treated, and normal ducks. Net - Qpyruvate values for individual liver slices are plotted against their respective coenzyme A values in units per g of fresh weight of slice. The following symbols represent the various groups: (1) deficient, (●) deficient treated *in vitro*, (▲) deficient treated *in vivo* by intraperitoneal injection of 10 m of calcium pantothenate per 100 g of body weight, 1 to 2 hours before observation, and (○) normal controls fed *ad libitum*.

curve quite similar to those for enzymatic acetylation, as shown in Fig. 8. Soon afterwards Stern and Ochoa³⁸ showed a CoA-dependent citrate synthesis with a pigeon liver fraction similar to the one used by Soodak for acetooacetate synthesis. In our laboratory, Novelli et al. confirmed and extended this observation with extracts of *Escherichia coli*³⁹.

In the course of this work, which more and more clearly defined the acetyl transfer function of CoA, Novelli once more tried acetyl phosphate. To our surprise and satisfaction, it then appeared, as shown in Table 9, that in *Escherichia coli* extracts in contrast to the animal tissue, acetyl phosphate was more than twice as active as acetyl donor for citrate synthesis than ATP-acetate³⁹. Acetyl phosphate, therefore, functioned as a potent microbial acetyl donor. Acetyl transfer from acetyl phosphate, like that from ATP-acetate, was CoA-dependent, as shown in Table 9. Furthermore, a small amount of "microbial conversion factor", as we called it first, primed acetyl phosphate for activity with pigeon liver acetylation systems⁴⁰, as shown in Table 10.

Table 10. Acetylation of p-aminobenzoic acid (PABA) by pigeon-liver enzyme, acetyl phosphate, and *C. kluyveri* extract.

System	Acetyl PABA, Bratton and Marshall ²³ (μM)
Acetyl~P + liver fraction, A-60	0
Acetyl~P + transacetylase	0
Transacetylase + liver fraction	0
Acetyl~P + liver fraction + bacterial transacetylase	0.92

(Conditions: tris(hydroxymethyl)aminomethane buffer (pH 8.1), 0.2 M; cysteine, 0.01 M; acetyl-P, 0.025 M; PABA, 0.001 M; CoA (67 units per mg), 10 units; bacterial transacetylase (acid ammonium sulfate fraction, 43 to 86 per cent saturation), 0.3 mg; pigeon-liver fraction (A-60-4)⁸, 0.3 ml. Final volume 1.0 ml, 28°, 60 minutes.)

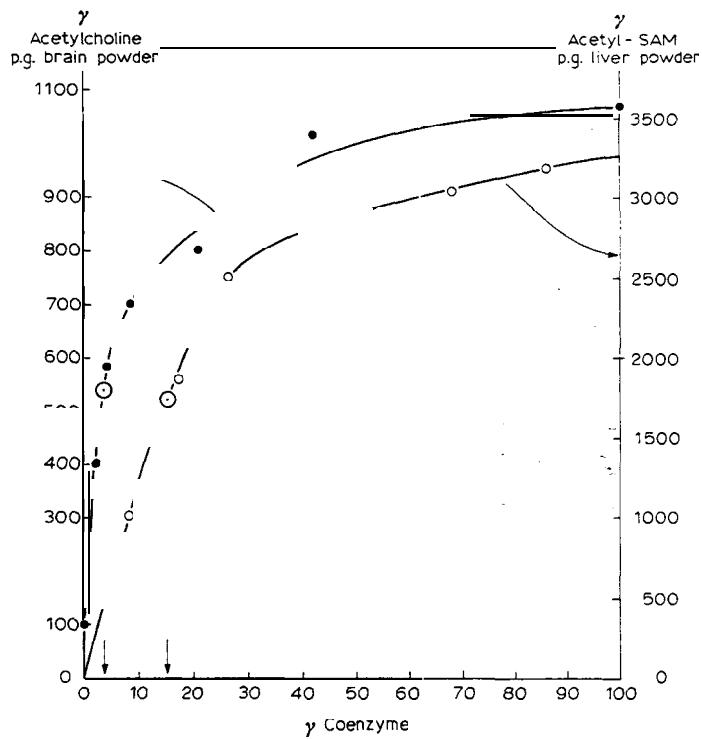


Fig. 7. Activation curves for acetocholinekinase and acetoarylaminekinase by purified coenzyme A preparations.

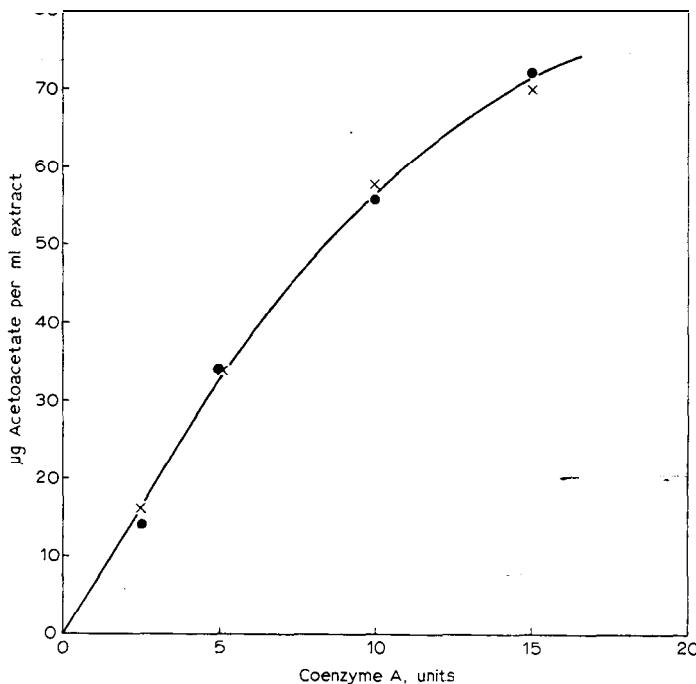


Fig. 8. Coenzyme A dependence of acetoacetate formation from acetate + ATP with ammonium sulfate pigeon-liver fraction.

Eventually the microbial conversion factor was identified by Stadtman *et al.*⁴⁰ with the transacetylase first encountered by Stadtman and Barker in extracts of *Clostridium kluyveri*⁴¹ and likewise, although not clearly defined as such, in extracts of *Escherichia coli* and *Clostridium butylicum* by Lipmann and Tuttle⁴².

The definition of such a function was based on the work of Doudoroff *et al.*⁴³ on transglucosidation with sucrose phosphorylase. Their imaginative use of isotope exchange for closer definition of enzyme mechanisms has been most influential. Like glucose-I-phosphate with sucrose phosphorylase, acetyl phosphate with these various microbial preparations equilibrates its phosphate rapidly with the inorganic phosphate of the solution. As in Doudoroff *et al.* experiments, first a covalent substrate enzyme derivative had been proposed⁴³. However, then Stadtman *et al.*⁴⁰, with the new experience of CoA-dependent acetyl transfer, could implicate CoA in this equilibration between acetyl- and inorganic phosphate and thus could define the transacetylase as an enzyme equilibrating acetyl between phosphate and CoA:

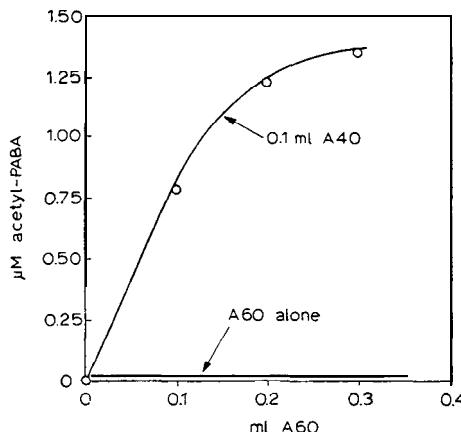
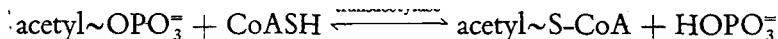
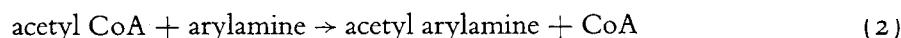
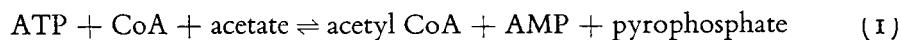


Fig. 9. Acetylation of PABA with Fractions A-40 and A-60 added in various proportions. The system contains *M* tris(hydroxymethyl)aminomethane buffer (pH 8.3). 0.2 ml; 0.1 M cysteine, 0.1 ml; 0.05 M sodium ATP, 0.1 ml; *M* sodium acetate, 0.1 ml; 0.1 M MgCl₂, 0.1 ml; *M* NaF, 0.1 ml; 0.01 M PAPA, 0.2 ml; and CoA, 11 units in 0.1 ml. Total volume 1.4 ml. Incubated at 30° for 90 minutes.



In the course of these various observations, it became quite clear that there existed in cellular metabolism an acetyl distribution system centering around CoA as the acetyl carrier which was rather similar to the ATP-centered phosphoryl distribution system. The general pattern of group transfer became recognizable, with donor and acceptor enzymes being connected through the CoA ---- acetyl CoA shuttle.

A clearer definition of the donor-acceptor enzyme scheme was obtained through acetone fractionation of our standard system for acetylation of sulfonamide into two separate enzyme fractions, which were inactive separately but showed the acetylation effect when combined. A fraction, A-40, separating out with 40 per cent acetone, was shown by Chou⁴⁴ to contain the donor enzyme responsible for the ATP-CoA-acetate reaction, while with more acetone precipitated, the acceptor function, A-60, the acetoarylamine-kinase as we propose to call this type of enzyme. The need for a combination of the two for overall acetyl transfer is shown in Fig. 9. This showed that a separate system was responsible for acetyl CoA formation through interaction of ATP, CoA and acetate (cf. below) and that the overall acetylation was a two-step reaction:



These observations crystallized into the definition of a metabolic acetyl transfer territory as pictured in Fig. 10.

This picture had developed from the growing understanding of enzymatic interplay involving metabolic generation of acyl CoA and transfer of the active acyl to various acceptor systems. A most important, then still missing link in the picture was supplied through the brilliant work of Feodor Lynen⁴⁵ who chemically identified acetyl CoA as the thioester of CoA. Therewith the thioester link was introduced as a new energy-rich bond and this discovery added a very novel facet to our understanding of the mechanisms of metabolic energy transformation.

The carboxyl and the methyl activation in acetyl CoA

In spite of many similarities between the general aspects of group transfer involving phosphoryl and acetyl groupings, there is a considerable difference insofar as the grouping transferred in the acetyl territory is an organic grouping and displays a quite different versatility for condensation reactions, yielding eventually large and complex carbon structures. There is one feature in this picture which always has attracted our particular attention: the two-

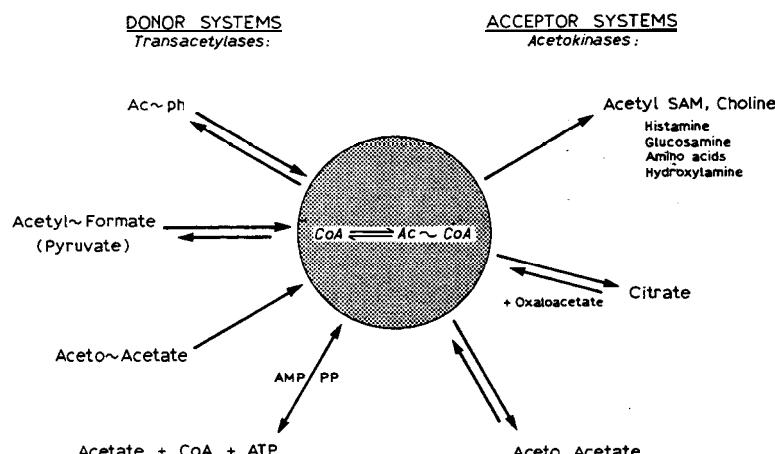


Fig. 10. Acetyl transfer scheme.

fold type of activation involving (1) the carboxyl end or the "head" of the acetyl; and (2) its methyl or "tail" end.

The definition of the head reaction is relatively simple. Acetylation of arylamine or choline is a typical head reaction. There is to be mentioned, furthermore, the observation by Chantrenne⁴⁶, introducing CoA as a rather general catalyst of acyl activation. He demonstrated the activity of CoA in benzoyl transfer such as hippuric acid synthesis. The mechanism of this synthesis was elaborated recently by Taggart⁴⁷, who clearly defined benzoyl CoA as the benzoyl donor in this reaction. An even greater and more prominent generalization is offered through the more and more developing importance of succinyl CoA in intermediary metabolism.

The second type, the methyl or tail activation, is not as well understood. In citric acid synthesis, as may be seen from Fig. 11, the methyl end engages in an aldolase type of condensation with the carbonyl group of the oxalacetate as acceptor. This condensation requires an energy input which must be

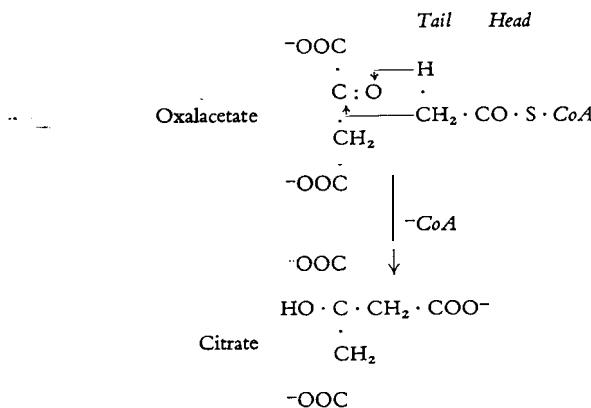
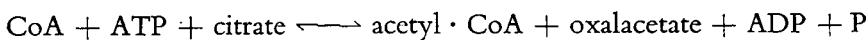


Fig.11. Tail reaction. Citric acid synthesis.

derived from the thioester link and at the end of the reaction CoA appears to be liberated in some manner.

The complexity of the citrate condensation is emphasized through the existence of an ATP-CoA-citrate reaction recently observed by Srere *et al.*⁴⁸, which results in the disruption of citrate to oxalacetate and acetyl CoA.



The mechanism remains still to be understood in greater detail. The reaction is mentioned here because it introduced the new variety into the citric acid cycle through a conversion of phosphoryl via citrate into acetyl.

For a long time the citrate reaction was the only known tail condensation. However, recently another interesting example has developed in the study of the precursors in steroid and isoprene synthesis. The initial condensation product in this series appears to be β -methyl, β -hydroxyglutarate (dicrotolic acid), formed through condensation of acetoacetate with acetyl CoA. The striking analogy between this and the citrate condensation appears on Fig. 12. This initial condensation seems to be followed by decarboxylation and dehydration to β -methyl crotonic acid first demonstrated by Bonner *et al.*⁴⁹ as intermediary in rubber synthesis.

In a third type, a combination of head *and* tail reaction takes place with two acyl CoA's reacting with each other in a head-to-tail condensation. When studying acetoacetate synthesis, we were at first not quite aware of its belonging to this third type of reaction. The calculation of the energy re-

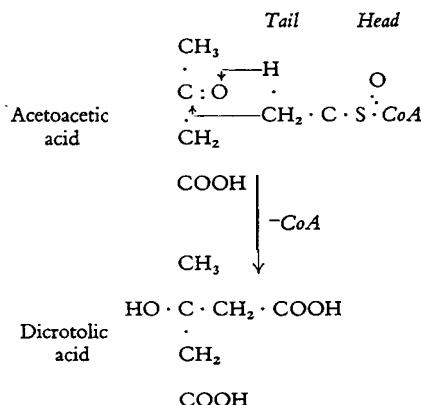


Fig. 12. Tail reaction. Synthesis of β -hydroxy, β -methyl glutaric acid (dicrotolic acid). (Note the similarity to citric acid synthesis.)

quired had yielded a figure of around 15 kilocalories which could be covered by one energy-rich bonds. However, by using as acetyl donor carboxyl-labeled acetyl phosphate, fed through transacetylase, the marker appeared in the carbonyl as well as the carboxyl part of acetoacetate (Table 11). This demonstrated a head-tail character for the reaction. The finer mechanism of this reaction between two acetyl CoA's, as shown in Fig. 13, has been more

Table 11. Synthesis of acetoacetate from $\text{CH}_3\text{C}^{14}\text{OPO}_3=$.

	<i>Start</i>	<i>End</i>
	<i>c.p.m. per μM</i>	
Acetate	0	2,900
Acetyl phosphate	23,000	22,400
Acetoacetate	—	35,200
Carboxyl-C	—	15,700
Carbonyl-C	—	19,500

recently elaborated in particular by Lynen's group⁵⁰, by Green's⁵¹ laboratory and by Ochoa and his group⁵². Presumably in the building up of longer terpene chains, a head-tail condensation may occur between two β -methyl crotonyl CoA's, followed by hydrogenation and dehydration-(Fig. 14).

A quite new type of head-tail condensation, presumably between succinyl CoA and glycyl CoA was recently suggested by Shemin's work on hemin synthesis⁵³. The thus primarily formed keto, amino dicarboxylic acid then appears to be decarboxylated to δ -amino levulinic acid. Shemin synthesized the latter marked with ^{14}C and showed its incorporation into the heme molecule.

The mechanism of the ATP-CoA-acetate reaction

In the recent past we have been mostly occupied with the mechanism through which the phosphate bond in ATP converts to acetyl bond in acetyl CoA. In

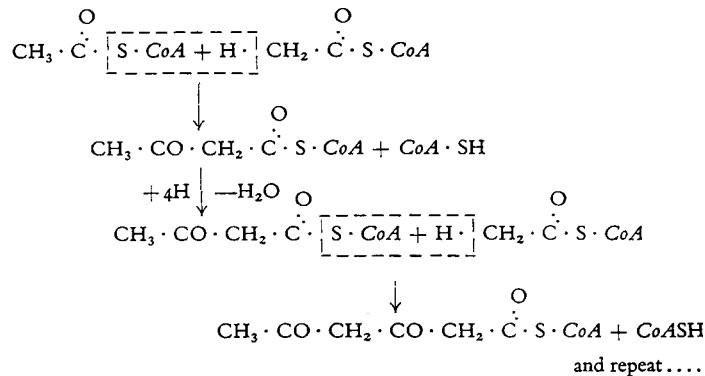


Fig. 13. Head-to-tail condensation. Acetoacetate synthesis and follow reactions in straight-chain fatty acid synthesis.

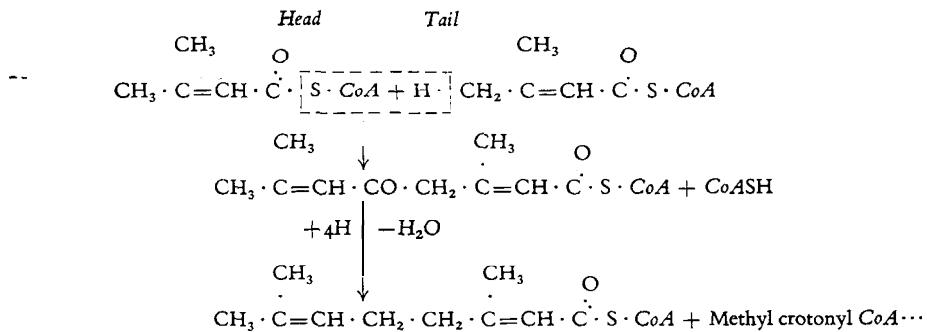
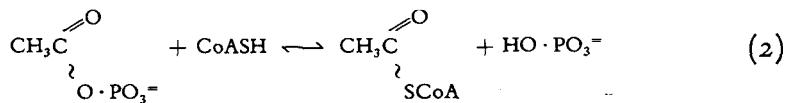
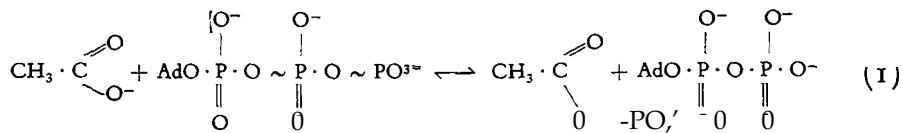


Fig. 14. Tentative scheme for terpene synthesis. Head-to-tail condensation of methylcrotonyl CoA's.

animal tissue where acetyl phosphate appeared not to be an intermediate, the conversion mechanism remained for a long time very puzzling. Before considering this reaction, it will be advantageous to review first the microbiological mechanism of such interconversion and in particular the role of acetyl phosphate as an intermediary. This transformation is rather straightforward: a sequence of two independent enzymatic reactions, the first a transphosphorylation from ATP to acetate and the second, as discussed, a transacylation from acetyl phosphate to CoA:



It should be noted that in the first transphosphorylation step the acetyl phosphate cleaves and condenses between O and P. In the second transacylation reaction, however, acetyl phosphate cleaves and condenses between C and O. Thus the same molecule reacts on either side of the oxygen bridge between the carbon and the phosphorus. This shift of the site of cleavage in the sequence is significant. This possibility attracted my early attention (see Fig. 3) and was one of the reasons that prompted me into this whole exploration. A shift from P~O.C to P.O~C should actually be a feature of many condensations initiated by a phosphoryl split from ATP. These be-

come increasingly numerous such as in glutamine, glutathione, pantothenate and seemingly in protein synthesis.

The finer mechanism generally is obscured by enzyme bound steps. In all these reactions, however, somewhere along the line a shift from transphosphorylation to transacylation seems to be inherent. This shift stands out very clearly in the microbial two-step reaction. But in animal tissue, the energy transmission from phosphoryl to acetyl occurs through a continuous enzyme bound reaction chain which is more difficult to elucidate. Nevertheless, some progress has been made, which also starts to reflect on other mechanisms of this type.

Jones *et al.*^{54,55,56} have explored the reaction with liver and with yeast and a surprising feature was uncovered, namely, that the initial phosphoryl split of ATP occurs between the pyrophosphoryl group and-AMP. The cleavage products of ATP were identified as inorganic pyrophosphate (PP) and adenylic acid (AMP). The mechanism was obscured by the presence of pyrophosphatase which, however, could be suppressed with fluoride. Table 12

Table 12. Effect of fluoride on pyrophosphate formation.

Fluoride added (μM per ml)	CoA* added (μM per ml)	Acetyl CoA (μM per ml)	PP (μM per ml)	P_i (μM per ml)
0	0	0	0	7.4
0	2.9	2.72	0	13.2
50	0	0	0	1.75
50	2.9	2.88	3.10** 3.16***	2.10

(Each vessel contained in 1 ml: 12 μM ATP; 10 μM potassium acetate; 10 μM $MgCl_2$; 20 μM H_2S ; 200 μM tris(hydroxymethyl)aminomethane buffer, pH 7.5 and 0.02 ml (20 units) of yeast enzyme fraction 4. Vessels were incubated at 37° for 30 minutes.)

* 1 μM CoA = 310 units.

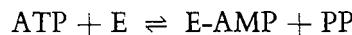
** Determined by colour increase.

*** Determined by Mn precipitation method.

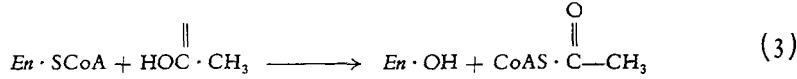
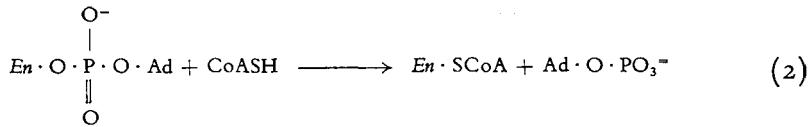
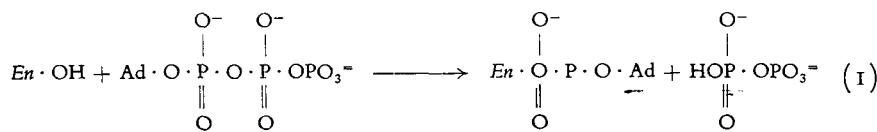
shows the course of reaction in the presence and the absence of fluoride, using the hydroxamate formation as an index.

Some further rather revealing information was obtained by the use of isotopes. This was suggested by Lynen during his visit to our laboratory⁵⁰. It was found that ATP and radioactive inorganic pyrophosphate exchange

in the absence of CoA or acetate. Such an exchange is best compatible with an initial reaction between the enzyme and ATP, resulting in a covalent binding of AMP to the enzyme, E,



It furthermore was found that acetyl CoA exchanges with radioactive free acetate in the absence of ATP or pyrophosphate. This exchange would indicate an exchange of acetyl for enzyme in the final step. Therefore, an overall sequence (*En* standing for enzyme) was proposed as follows*:



The middle step, that is the substitution of enzyme-bound AMP by CoA, is the most problematic but also the most interesting one since it may foreshadow mechanisms implicating nucleotide activation for polynucleotide formation. No indications were found for the identity of the grouping on the enzyme through which the initial binding of AMP and further exchange with CoA might occur. A further purification of the enzyme would be necessary before obtaining fuller information.

A pyrophosphoryl split of ATP also was found by Maass^{57,58} to initiate pantothenate synthesis. In contrast to the ATP-CoA-acetate reaction this peptide synthesis appears to occur by way of a pyrophosphoryl-enzyme instead of AMP-enzyme intermediary. This phenomenon of a priming of an

* Note added in proof in 1963: The liberation of pyrophosphate in the ATP-CoA-acetate reaction has meanwhile been resolved into an initial formation of acetyl adenylate, followed by a transacetylation to CoA to yield acetyl CoA as the end product (P. Berg, *J. Biol. Chem.*, 222 (1956) 991; W.P. Jencks and F. Lipmann, *J. Biol. Chem.*, 225(1957)207).

enzyme for peptidic synthesis by phosphorylation has been very tentatively used as a start for developing possible schemes for polypeptide synthesis⁵⁹.

Altogether, in this area, a diversified picture is rapidly developing. There is good reason to hope that in the not too distant future, out of the fair confusion of the present, a clearer understanding will eventually evolve. A new level of complexity seems slowly to unravel and the gap between the biochemical and biological approach further narrows down.

1. F. Lipmann, *Cold Spring Harbor Symp. Quant. Biol.*, 7 (1939) 248.
2. H. Kämmerer and L. Carius, *Ann.*, 131 (1864) 153.
3. F. Lipmann, *J. Biol. Chem.*, 134 (1940) 463.
4. F. Lipmann, *J. Biol. Chem.*, 155 (1944) 55.
5. F. Lipmann and L. C. Tuttle, *J. Biol. Chem.*, 153 (1944) 571.
6. F. Lipmann and L. C. Tuttie, *J. Biol. Chem.*, 159 (1945) 21.
7. F. Lipmann, *Advan. Enzymol.*, 1(1941) 99.
- 7a. R. Bentley, *Cold Spring Harbor Symp. Quant. Biol.*, 8 (1948) 11.
8. F. Lipmann, in *Currents in Biochemical Research*, D. E. Green (Ed.), Interscience Publishers, N.Y., 1946, p. 137.
- 9.. F. Lipmann, *Advan. Enzymol.*, 6 (1946) 231.
10. E. R. Stadtman and F. Lipmann, *J. Biol. Chem.*, 185 (1950) 549.
11. F. Lipmann, *J. Biol. Chem.*, 160 (1945) 173.
12. W. O. Sykes, *Biockem. J.*, 38 (1944) xxix.
13. J. R. Klein and J. S. Harris, *J. Biol. Chem.*, 124 (1938) 613.
14. H. Kalckar, *Biochem. J.*, 33 (1939) 631.
15. D. Nachmansohn, H. M. John, and H. Waelsch, *J. Biol. Chem.*, 150 (1943) 485.
16. D. Nachmansohn and A. L. Machado, *J. Neurophysiol.*, 6 (1943) 397.
17. N. O. Kaplan and F. Lipmann, *J. Biol. Chem.*, 174 (1948) 37.
18. F. Lipmann, N. O. Kaplan, G. D. Novelli, L. C. Tuttle, and B. M. Guirard, *J. Biol. Chem.*, 167 (1947) 869.
19. G. D. Novelli, N. O. Kaplan, and F. Lipmann, *J. Biol. Chem.*, 177 (1949) 97.
20. F. Lipmann, N. O. Kaplan, and G. D. Novelli, *Federation Proc.*, 6 (1947) 272.
21. F. Lipmann, N. O. Kaplan, G. D. Novelli, L. C. Tuttle, and B. M. Guirard, *J. Biol. Chem.*, 186 (1950) 235.
22. G. D. Novelli, N. O. Kaplan, and F. Lipmann, *Federation Proc.*, 9 (1950) 209.
23. G. D. Novelli, *Physiol. Rev.*, 33 (1953) 525.
24. J. Baddiley and E. M. Thain, *J. Chem. Soc.*, Sept. (1951) 2253.
25. J. D. Gregory, G. D. Novelli, and F. Lipmann, *J. Am. Chem. Soc.*, 74 (1952) 854.
26. W. H. de Vries, W. M. Govier, J. S. Evans, J. D. Gregory, G. D. Novelli, M. Soodak, and F. Lipmann, *J. Am. Chem. Soc.*, 72 (1950) 4838.
27. E. E. Snell, G. M. Brown, V. J. Peters, J. A. Craig, E. L. Wittle, J. A. Moore, V. M. McGlohon, and O. D. Bird, *J. Am. Chem. Soc.*, 72 (1950) 5349.

28. J. Baddiley, E. M. Thain, G. D. Novelli, and F. Lipmann, *Nature*, 171 (1953) 76.
29. L. S. Shuster and N. O. Kaplan, *J. Biol. Chem.*, 201 (1953) 535.
30. A. Dorfman, A. Berkman, and S. A. Koser, *J. Biol. Chem.*, 144 (1942) 393.
31. G. M. Hills, *Biochem. J.*, 37 (1943) 418.
32. G. D. Novelli and F. Lipmann, *Arch. Biochem.*, 14 (1947) 23.
33. G. D. Novelli and F. Lipmann, *J. Biol. Chem.*, 171 (1947) 833.
34. R. E. Olson and N. O. Kaplan, *J. Biol. Chem.*, 175 (1948) 515.
35. D. Nachmansohn and M. Berman, *J. Biol. Chem.*, 165 (1946) 551.
36. W. Feldberg and T. Mann, *J. Physiol. London*, 104 (1946) 411.
37. M. Soodak and F. Lipmann, *J. Biol. Chem.*, 175 (1948) 999.
38. J. R. Stem and S. Ochoa, *J. Biol. Chem.*, 179 (1949) 491.
39. G. D. Novelli and F. Lipmann, *J. Biol. Chem.*, 182 (1950) 213.
40. E. R. Stadtman, G. D. Novelli, and F. Lipmann, *J. Biol. Chem.*, 191(1951) 365.
41. E. R. Stadtman and H. A. Barker, *J. Biol. Chem.*, 184 (1950) 769.
42. F. Lipmann and L. C. Tuttle, *J. Biol. Chem.*, 158 (1945) 505.
43. M. Doudoroff, H. A. Barker, and W. Z. Hassid, *J. Biol. Chem.*, 168 (1947) 725.
44. T. C. Chou and F. Lipmarm, *J. Biol. Chem.*, 196 (1952) 89.
45. F. Lynen, E. Reichert, and L. Rueff, *Ann.*, 574 (1951) 1.
46. H. Chantrenne, *J. Biol. Chem.*, 189 (1951) 227.
47. D. Schachter and J. V. Taggart, *J. Biol. Chem.*, 203 (1953) 925.
48. P. A. Srere and F. Lipmanu, *J. Am. Chem. Soc.*, 75 (1953) 4874.
49. J. Bonner and B. Arreguin, *Arch. Biochem.*, 21 (1949) 109.
50. F. Lynen, *Harvey Lectures*, 48 (1954) 212.
51. D. E. Green, S. Mii, H. R. Mahler, and R. M. Bock, *J. Biol. Chem.*, 206 (1954) 1.
52. S. Ochoa, *Advan. Enzymol.*, 15 (1954) 183.
53. D. Shemin and C. S. Russel, *J Am. Chem. Soc.*, 75 (1953) 4873.
54. M. E. Jones, F. Lipmann, H. Hibbs, and F. Lynen, *J. Am. Chem. Soc.*, 75 (1953) 3285.
55. M. E. Jones, S. Black, R. M. Flynn, and F. Lipmann, *Biochem. Biophys. Acta*, 12 (1953) 141.
56. M. E. Jones, *Federation Proc.*, 12 (1953) 708.
57. W. K. Maas, *Federation Proc.*, 12 (1953) 241.
58. W. K. Maas and G. D. Novelli, *Arch. Biochem. Biophys.*, 43 (1953) 236.
59. F. Lipmarm, in *Mechanism of Enzyme Action*, W. D. McElroy and B. Glass (Eds.), Johns Hopkins Press, Baltimore, 1954, p. 599.

OTTO F. MEYERHOF

Energy conversions in muscle

Nobel Lecture, December 12, 1923

This highest scientific honour, in the form of the Nobel Prize, which has been awarded to me for my investigations into the conversions of energy in muscle, gives me the pleasant duty of reporting to you on this problem and upon the results which my work has achieved. It is especially gratifying to me that this recognition is in part shared with me by my distinguished friend, the previous speaker, Professor A.V. Hill from London, with whom my work has had so many close points of contact and with whom, in spite of the present political unrest, I have worked in cooperation towards the mutual goal of explaining the process of muscle contraction.

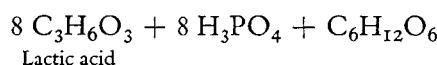
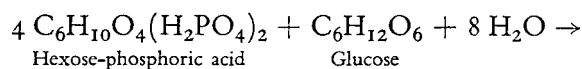
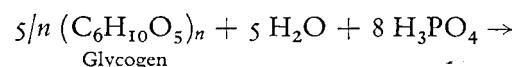
The fact that chemical processes must be involved as a source of energy for muscle performance was already accepted as a necessary deduction from their own thesis by the discoverers of the law of the conservation of energy. In fact, the young Helmholtz had already made certain observations concerning the conversion of matter in muscles during activity which in themselves were correct but, on account of their incompleteness and the lack of knowledge concerning the chemical nature of the relevant substances, served no useful purpose. The previous speaker has already told you about the considerable progress achieved by the English scientists Fletcher and Hopkins by their recognition of the fact that lactic acid formation in the muscle is closely connected with the contraction process. These investigations were the first to throw light upon the highly paradoxical fact, already established by the physiologist Hermann, that the muscle can perform a considerable part of its external function in the complete absence of oxygen. As, on the other hand, it was indisputable that in the last resort the energy for muscle activity comes from the oxidation of nutriment, the connection between activity and combustion clearly had to be an indirect one. In fact, Fletcher and Hopkins observed that in the absence of oxygen in the muscle, lactic acid appears, slowly in the relaxed state and rapidly in the active state, and that this lactic acid disappears again in the presence of oxygen. Obviously, then, oxygen is involved not while the muscle is active, but only when it is in the relaxed state, and this assumption has been supported by further

research on the part of Parnas and Verzar. In what relation the lactic acid stands to muscle performance, where it comes from and what becomes of it when it disappears in the presence of oxygen, was completely obscure. In fact, there were several different, irreconcilable interpretations current, all of which appeared nevertheless to be supported by experiment. It was at this point that I started to work on the problem. A bright light in the midst of this obscurity appeared when Professor Hill made the important discovery, about which he has just spoken to you, that the contraction heat of the muscle occurs in two distinct phases of approximately the same extent - one phase which is directly connected with the work and is the same in presence or absence of oxygen, which he called the "initial heat"; and a second phase, which basically only occurs in the presence of oxygen, and which he called "delayed heat" and quite rightly connected with the - disappearance of the lactic acid. Apart from the pioneer work of Fletcher and Hopkins, it was this discovery above all which, shining out like a beacon light through a sea mist, made it possible for me to steer a safe course through the shallows.

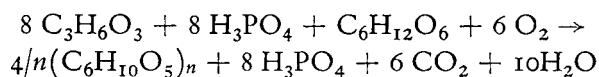
If we now observe an excised frog muscle operating under maximum oxygen supply, chemical analysis will only prove that a certain quantity of glycogen in the muscle disappears, whereas an exactly sufficient quantity of oxygen necessary for its oxidation is assimilated, and the corresponding amount of carbon dioxide is given off. The connection between these processes can be more exactly analysed if the muscle is first allowed to work under anaerobic conditions, and subsequently brought out into oxygen. During the anaerobic phase, in fact, lactic acid accumulates in the muscle approximately in proportion to the amount of work performed. At the same time a corresponding quantity of glycogen disappears, while the quantity of lower carbohydrates, particularly free glucose and the hexose-phosphoric acid discovered in the muscle by Embden, is not noticeably altered. In the second, oxidative, phase the lactic acid which has formed disappears, while a specific quantity of extra oxygen is assimilated. In fact, the disappearance of lactic acid during this period is in exact proportion to the increased consumption of oxygen. However, the oxygen is only sufficient to oxidize a fraction of the disappearing lactic acid; the remainder, which in the case of complete fatigue is about three-quarters of the total lactic acid, is quantitatively reconverted into glycogen. I must state already here that this ratio of the lactic acid which disappears altogether to that burnt is not always constant under all conditions, and from the energetic point of view this is important, to which I must return later. To start with, however, we

will concern ourselves with this figure obtained under suitable conditions of extreme anaerobic fatigue, and subsequent recuperation in oxygen. Of four molecules of lactic acid which disappear, three are then converted back into glycogen and one is oxidized. To be exact, we cannot even maintain with certainty that the lactic acid itself is burnt. We find only an oxidized carbohydrate-equivalent with the respiratory quotient 1. Whether this is sugar or lactic acid we cannot be certain. I have, therefore, chosen the formulation for the two phases which you can see on this board.

Anaerobic fatigue phase



Oxidative recovery phase



In the anaerobic, active phase the glycogen is broken down into lactic acid via glucose and, I assume in agreement with Embden, by way of hexose-di-phosphoric acid. On the board the decomposition of five sugar-equivalents of glycogen is assumed, of which four are esterified with phosphoric acid and form eight molecules of lactic acid.

In the second, aerobic phase these eight molecules of lactic acid disappear, while two of them, or alternatively, as we might equally well assume, one molecule of sugar, are burnt. The importance of this strangely coupled reaction can only be understood after a study of energetics. But before I turn to this, it is important to stress that this activity metabolism in the muscle is not a separate phenomenon, but is no more than an increase of the metabolism in the resting state. For even in the resting state the glycogen in an isolated muscle in oxygen disappears directly by way of oxidation into carbon dioxide and water. If, however, we keep a resting muscle in nitrogen for a considerable time, lactic acid is constantly accumulating in it during the

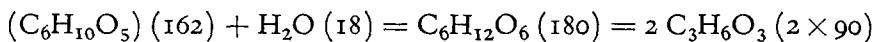
anaerobiosis. If we now compare this lactic acid accumulation with the quantity of oxygen which the muscle would have assimilated in the same time under aerobic conditions, we find that approximately three times the amount of lactic acid has accumulated as could have been consumed by the oxygen in the same amount of time. Here also, then, the lactic acid is not just a simple intermediate product of the decomposition of the sugar. In fact, if we bring the muscle back into the air after extended anaerobiosis, it will assimilate a certain quantity of extra oxygen, approximately equivalent to the amount previously lost. At the same time, the lactic acid disappears once again in such a way that most of it is reconverted into glycogen, whereas only a fraction, or the corresponding quantity, of carbohydrate is consumed. The process is, therefore, exactly the same as when the muscle is active, only the accumulation of lactic acid progresses much more slowly. We can directly see from this the importance of muscle respiration in the resting state, in that it maintains a labile condition of lactic acid production and removal, which can be accelerated instantly on stimulation. Probably this explosive release of lactic acid during contraction occurs, because stimulation suddenly increases the permeability of membranes which have previously to a certain extent acted as a barrier between the participants in the reaction. Respiration in the muscle in the resting state can, therefore, be said to keep them in a state of readiness for activity.

We can establish that the lactic acid is directly associated with muscle contraction by an exact comparison of the work performed under anaerobic conditions with the formation of lactic acid. As the best expression of the activity potential of muscle we may choose here, following Fick and Professor Hill, the tension which the muscle develops on stimulation when prevented from shortening, i.e. the so-called isometric contraction. If we allow the muscle to go on working under anaerobic conditions until it is exhausted it produces a certain quantity of lactic acid and develops a degree of tension in proportion to this quantity. This total anaerobic work can be very considerable - for instance, a frog muscle of 1 g in weight in N_2 can produce 160 kg of tension in 1000 contractions.

I found that there is a very simple reason why there is any limit at all to this and why activity does not in fact continue until the available glycogen is used up. It was thought earlier, and in particular by Fletcher and Hopkins themselves, who were the first to become aware of the so-called fatigue-maximum, that this was conditioned by the exhaustion of an immediate preliminary stage of the lactic acid. This is, however, not the case - it stems

rather from the accumulation of the acid in the muscle itself. If we remove a large part of the acid from the muscle by placing it in a Ringer's solution particularly rich in bicarbonate, it produces before total exhaustion not only very much more lactic acid, but also correspondingly more work. By the addition of various buffer mixtures to the muscle it was proved that the increase in performance due to this admixture corresponded almost exactly to the percentage of lactic acid which escaped from the muscle into the surrounding solution.

The significance of these chemical reactions only becomes clear when we consider the energetic conditions. In the anaerobic active phase lactic acid is formed from glycogen, at the rate of 1 g lactic acid from 0.9 g glycogen, since during the formation of every 180 g of lactic acid 18 g of water are absorbed



The combustion heat of glycogen, according to Stohmann's readings, is 4191 cal/g - that is, 3772 per 0.9 g. As these readings were made about thirty years ago with still somewhat primitive instruments, a revised determination seemed desirable, especially as the American scientists Emery and Benedict had found a rather higher value of 4227 cal/g. This new determination was made at my suggestion in Germany by a pupil of Professor Roth in Brunswick, and at the same time in Manchester by Mr. Slater. In the first case Stohmann's readings were completely confirmed, resulting in 4188 cal/g or 3769 cal/0.9 g. Slater, however, using a differently produced glycogen, obtained a very much higher value, i.e. in relation to the above glycogen formula he obtained 3883 cal/0.9 g. I will come later to the reasons which for the time being have caused me to regard the values of Stohmann and Roth to be more accurate. The combustion heat of lactic acid I determined anew, since the values given in the literature appeared unreliable, and I obtained 3601 cal for dilute lactic acid, a value which was confirmed in Roth's Institute and which agreed also with the Americans Emery and Benedict. If the chemical process during contraction turned out to be as one was only recently tempted to imagine it - that is to say, if during activity lactic acid is formed from glycogen and this is evenly consumed during relaxation - then only the difference in combustion heat between 0.9 g of glycogen and 1 g of lactic acid - i.e. 170 cal - would be released by activity in the muscle. On the other hand, the combustion of the lactic acid at 3601 cal would take

place in the oxidative recuperative period. Such a process, in which only 5% of the heat released would occur in the contractive phase, would appear to be extremely doubtful theoretically and would into the bargain contradict the fact established by Professor Hill that heat quantities in the active and recuperative phases are approximately the same. In fact, this consideration was really the beginning of my preoccupation with the problem of muscle. The process is actually quite different. It soon became evident to me from a great number of determinations that during the formation in the muscle of 1 g of lactic acid, not 170 cal but 380-390 cal were released, a figure which was not very far removed from older, slightly less accurate, results which Peters, a pupil of Professor Hill, had obtained. Before we discuss the reason for this very big divergence of the contraction heat from the difference between the combustion heats, we must first calculate the energy balance of the recovery phase. If, as is represented in the above equation and has on average been proved by my experiments, of a total of 4 molecules of lactic acid which disappear, one is burnt (or, which comes to the same thing, a carbohydrate equivalent of it), then altogether, for 1 g of sugar taking part in the reaction, or 0.9 g of glycogen, $3772/4 = 943$ cal must be released. As we have measured 385 cal in the active phase, the remainder - i.e. approx. 560 cal - must be expected during the recuperative phase. According to this, 40% of the heat must occur in the active phase, 60% in the recovery phase. In fact this was very well confirmed, at least with regard to the order of magnitude, by measurement of the total heat production in the recovery phase and comparison with the oxygen consumption. According to this: (1) the increase in heat produced in the oxidative recuperative phase was approximately as great as, or only slightly greater than, the anaerobic heat of the exhaustion state of the muscle; (2) this heat, reckoned according to oxygen assimilation, was smaller than the corresponding carbohydrate consumption which took place simultaneously. For every 1 c.c. of oxygen during carbohydrate oxidation 5 cal should have appeared; but there were only 3.5 cal, and altogether had vanished about the same amount of heat in the recovery phase as had appeared in the anaerobic phase.

We now find that this result agrees very well with Professor Hill's findings, about which he has just spoken to you: the heats of the active and the recovery phases are equal. But, as he went on to establish, this result can be confirmed even more exactly by the more accurate analysis of heat formation which is made possible by the myothermic method developed by Hill and Hartree, also by the study of the oxygen consumption in relation

to the disappearance of lactic acid during recovery under various conditions. The result of these experiments shows that the quotient

$$\frac{\text{total disappearing lactic acid}}{\text{lactic acid burnt}}$$

is not constant - it is greater in completely fresh muscles, and can in fact amount to as much as 5:1 - 6:1, and it is of approximately the same size in live humans as in live frogs. In the case of humans this was proved indirectly in Professor Hill's laboratory. I myself obtained a similar result on the whole frog by using the same direct methods as would have been used on the isolated muscle.

Such an increase in the quotient means, however, that a smaller part of the heat occurs in the recovery phase. For if out of six molecules, for instance, only one is burnt then there must arise from the conversion of 1 g of sugar $3772/6 = 630$ cal, of which 385 are in the fatigue phase, so that 245 must be in the recovery phase. In this case 60% must already be released in the active phase, and only 40% in the recovery phase. The quotient appears to lie between these two amounts according to the degree of exhaustion and the condition of the muscle - between 6:1 and 4:1; and in unfavourable conditions it is even smaller. The muscular mechanism operates so much the more economically the more lactic acid molecules can be transformed back into glycogen through the oxidation of one of them. Thus this figure represents the efficiency of the recovery process. It is an expression of how much of the oxidative energy is used in endothermal processes for the conversion of the material in the preliminary stages. In the case of the above equation the efficiency would be 40% - under the more favourable conditions of a completely fresh frog muscle or of a living animal it would be 50-60%. Curiously enough, the ratio is smaller in the case of respiration in rest, i.e. of 2-3 molecules lactic acid which disappear 1 is burnt. My more recent experiments have, in fact, shown that many poisons, and also traumatic damages, experienced by the animal before death, will cause the ratio to deteriorate still further. All these circumstances result in a squandering of energy.

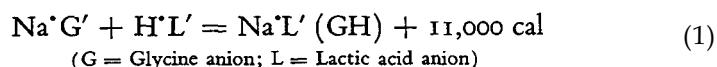
There is a very important problem connected with the size of the anaerobic contraction heat itself, which, as we have seen, is in the region of 385 cal, whereas thermochemical data only give a difference in the combustion heat of 170 cal for the conversion of glycogen into dilute lactic acid. How does this difference arise? The following has been established: if the forma-

tion of heat and of lactic acid are compared, not in the working muscle, but in crushed muscular tissue suspended in phosphate solution, we then obtain about 200 cal/g instead of 385 cal, and at the same time the lactic acid passes into the phosphate solution. The heat of neutralization of lactic acid with biphosphate is, however, 19 cal/g. Added to this is the heat of the cleavage of glycogen into lactic acid, 170 cal, and these taken together amount to 190 cal, which agrees, allowing a margin of error, with our measured value of 200 cal.

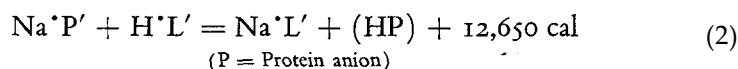
Similarly, the heat even in an intact muscle can be reduced if a considerable amount of the lactic acid passes into the surrounding solution. This can be brought about if lactic acid is allowed to form in a resting muscle suspended in a carbonated Ringer's solution. Half of the lactic acid can escape into the surrounding solution, and the ratio will be 280 cal per g lactic acid instead of 385 cal, and finally only 230 cal.

This particular heat formation which lactic acid produces in the living muscle is, however, bound up with the hydrogen ion. In fact, other acids which we allow to penetrate from outside into the frog muscle cause considerable heat production which is independent of lactic acid formation within the muscle. In this way I observed that the penetration of valeric acid into the muscle caused the release of up to 11,000 cal/mol of acid assimilated by the muscle. Reckoned in terms of lactic acid this corresponds to a heat production of 120 cal/g. This heat, as close analysis shows, is dependent upon the reaction of the acid with the tissue protein. This tissue protein acts as a buffer substance and keeps the reaction within the muscle always more or less constant, even during heavy lactic acid production. Even in the case of maximum exhaustion, when approx. 0.4% of lactic acid is produced in the muscle, the index of the hydrogen-ion concentration (pH) is only displaced from 7.5 to 6.8. With this buffer reaction there is a characteristic heat formation - an inverse protein dissociation heat - which is bound up with the deionization of the protein.

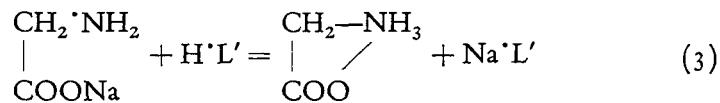
These conditions are very clearly seen in relation to the amino acids, which behave in principle in the same way as protein, which is in fact composed of amino acids. If we start, for instance, with a solution of glycine with the addition of caustic soda solution there will be in the solution, apart from other substances, the salt sodium glycine, which we can regard as being completely dissociated, and which will be abbreviated here as NaG.



If we now add hydrochloric acid or lactic acid or another not too weak acid, then a reaction will take place which is shown here in Formula (1). Out of the totally dissociated glycine-acid sodium there is formed the weak, non-dissociated glycine acid, and in this reaction a positive heat of approx. 11,000 cal can be measured. This is nothing else than the inverse heat from the electrolytic dissociation of the glycine. If a buffer solution is made of concentrated protein solution, free from basic salts, by the addition of caustic soda solution at about pH 8, and if lactic acid is now added in such a quantity that the H-ion concentration is barely altered, a corresponding reaction will obviously take place, which is shown here in the following equation:



Here we find an even greater heat production - with muscle proteinin the presence of ammonium salt it will be 12,650 cal. This dissociation heat of protein is the largest known dissociation heat of any acid. The reason for this may be connected with the fact that the deionization of amino acids and protein causes the production of internal ammonium salt, as shown in the following diagram:



We can check this supposition by means of formaldehyde. Formaldehyde causes the formation of a methylene compound from the amino acids with stronger acid properties. At the same time as the addition of acids the very high dissociation heat disappears almost completely,

Prom the dissociation heat of 12,600 cal per equivalent can be calculated a heat production of 140 cal per g lactic acid. As, however, the lactic acid in the muscle to a certain reacts with phosphate and carbonate this figure should be reduced a little. I have already shown that of the 385 cal formed per 1 g of lactic acid in the anaerobic contraction phase 170 are due to the splitting of the glycogen into dilute lactic acid. There remain 215 cal. Of these, up to 140 can be explained by the dissociation heat of the protein. There remain, over and above the margin of possible error, 70-80 cal for which up till now we can only provide a hypothetical explanation. At first

sight three such explanations appear possible: (1) the combustion heat of the glycogen may be higher than has been supposed; (2) secondary reactions about which we as yet know nothing may take place; (3) the dissociation heat of the protein may be greater in living muscle than in solution. The first possibility seemed to have found strong support from Slater's experiments, according to which glycogen is supposed to have a combustion heat 100 cal greater than the value I have given it. However, I was able to prove that this was unlikely, by bringing about a splitting of the glycogen into maltose and dextrose by means of a diastatic ferment which engendered combustion heats of 3 752 and 3 748 cal. To my surprise the heat of the cleavage of dissolved glycogen into dissolved maltose and dextrose was only about 10 cal/g of glucose. But these values in fact agree approximately with the glycogen combustion heat determined by Stohmann, as the solution heat of dextrose amounts to minus 12.5 cal/g. For the conversion of 0.9 g of glycogen into 1 g of solid dextrose the figure of $10 + 12.5 = 22.5$ cal was thus determined by experiment. From this the combustion heat of the dissolved glycogen could be calculated as about 3770 cal per 0.9 g. That of anhydrous glycogen is incidentally much higher, on account of a very considerable hydration heat, but this need not concern us here, since the glycogen in the muscle is in hydrated form. At the same time, there is a very interesting conclusion to be drawn in connection with the result of these readings. We have good reason to believe that the actual process of muscular work does not begin with the splitting of the glycogen but with that of the dextrose or the phosphoric acid hexose. The energy released by the conversion of the glycogen into hexose is then lost to the activity process. Up till now it has been possible to assume that more than 30 cal/g of sugar - that is to say, no less than 8% of the total energy - was squandered in this way. It is now apparent, however, that as a result of the negative solution heat of glucose, the conversion of dissolved glycogen into dissolved dextrose only requires approx. 10 cal - that is to say, barely 3% of the contraction heat. It is possibly even lower in the case of the conversion into phosphoric acid hexose. This result is in harmony with the splendid economy which is shown in the reactions of the living organisms.

Of the possibilities just mentioned for the explanation of the difference of about 70 cal there now remain, therefore, only the two last. Up till now there has been no reason to suppose that, side by side with the carbohydrate metabolism, a fat or protein metabolism also plays a part in the contraction mechanism, nor that inorganic compounds such as phosphoric acid undergo

permanent changes on account of anaerobic exhaustion. Therefore I incline to the third of the above hypotheses, namely that protein in living muscle has a higher dissociation heat than in solution: this supposition is supported by the consideration that the material of which the muscular machinery is composed is protein, whereas carbohydrates form the combustion material. Somehow the oxidation energy of the combustion material must play a part in the mechanism itself. This is in fact the case with the deionization of muscle protein. The oxidation energy first becomes active in the recovery phase; and then, as a result of the coupling of the oxidation with the re-synthesis of the lactic acid, not only is the endothermic process of rebuilding the glycogen brought about, but also the alkali is released from the vanishing alkali-lactate, and this leads to the endothermic, involuntary dissociation of the muscle protein. In this way the muscular apparatus is once again put into working order. We can compare this process to the charging of an accumulator, which Professor Hill has given as an image of the recovery reaction, or, if we prefer, to the winding-up of a watch, as I have more often described it. It would then appear theoretically quite logical to expect that a relatively large part of the total energy should be lost in this process. But we must leave the definitive explanation of this point to the future.

We can, however, say now that the deionization of the protein by the lactic acid produced by muscle activity plays, without any doubt, an important part in the contraction mechanism. It explains in the first place the flaccidity of the muscle which sets in after shortening under anaerobic conditions in spite of the presence of lactic acid. This flaccidity would be brought about by nothing more than the diminution of the lactic acid acidity, just as, conversely, we would hold the H-ions responsible for the release of the contraction. Certainly, the flaccidity heat discovered by Hill and Hartree, even though it must be considered to originate in the superimposition of various chemical and physical processes, is obviously the principal cause of the deionization heat of protein. On the other hand, the exhaustion maximum of lactic acid would be conditioned by the supply of alkali separable from the protein salt of the muscle.

I think, therefore, that in this way we have obtained a comparatively simple and satisfactory picture of energy conversions in muscle, the future shape of which will be of theoretical value and practical interest.

To remove the uncertainty concerning the combustion heat of glycogen mentioned in this lecture, I have since, with Dr. Meier, brought about the combustion of several glycogen preparations with various reservations. For

anhydrous glycogen ($C_6H_{10}O_5$) from frog muscles we obtained 3806 cal, for glycogen hydrate ($C_6H_{10}O_5 H_2O$), 3786 cal, from which, taking into account the solution heat of dissolved glycogen, 3775 cal was found, so that the value used above is almost exactly correct.

It also became clear that during deionization of saturated solutions of amino acids in non-aqueous media (alcohol-water mixture) besides a deionization heat, a precipitation heat occurs, which amounts to just about 70 cal/mol - a pointer to the possible explanation for the cause of the unexpected remainder in the contraction heat.

ALBERT A. MICHELSON

Recent advances in spectroscopy

Nobel Lecture, December 12, 1907

The fame of Newton rests chiefly on his epoch-making discovery of the laws of gravitational astronomy - by means of which the position of the moons, the planets, and the comets, and other members of our solar system can be calculated and verified with the utmost precision - and in many cases such calculation and verification may be extended to systems of suns and planets outside our own.

But in no less degree are we indebted to this monumental genius for that equally important branch of Astrophysics - in which the spectroscope plays so fundamental a role - by means of which we are enabled to discover the physical and chemical constitution of the heavenly bodies, as well as their positions and motions. As the number and intricacy of the wonderful systems of stellar worlds which the telescope can reveal increase with its power, so also do the evidences of the innermost molecular structure of matter increase with the power of the spectroscope. If Newton's fundamental experiment of separating the colors of sunlight had been made under conditions so slightly different from those in his actual experiment that in the present stage of experimental science, they would at once suggest themselves to the veriest tyro - the science of spectroscopy would have been founded.

So simple a matter as the narrowing of the aperture through which the sunlight streamed before it fell upon the prism which separates it into its constituent colors - would have sufficed to show that the spectrum was crossed by dark lines, named after their discoverer, the Fraunhofer lines of the Solar Spectrum. These may be readily enough observed, with no other appliances than a slit in a shutter which is observed through an ordinary prism of glass. Fraunhofer increased the power of the combination enormously by observing with a telescope - and this simple combination, omitting minor details, constitutes that wonder of modern science, the Spectroscope. As the power of a telescope is measured by the closeness of the double stars which it can "resolve", so that of the Spectroscope may be estimated by the closeness of the spectral lines which it can separate. In order to form an idea of the advance in the power of spectrosopes

let us for a moment consider the map of the Solar Spectrum (Figure 1).

The portion which is visible to the unaided eye extends from the Fraunhofer line A to H; but by photography it may be traced far into the ultra-violet region and by bolometric measurements it is found to extend enormously farther in the region beyond the red. In the yellow we observe a dark line marked *D*, which coincides in position with the bright light emitted by sodium - as when salt is placed in an alcohol flame. It may be readily shown by a prism of very moderate power that this line is double, and as the power of the instrument increases the distance apart or separation of this doublet furnishes a very convenient measure of its separating or resolving power. Of course this separation may be effected by simple magnification, but this would in itself be of no service, as the "lines" themselves would be broadened by the magnification in the same proportion. It can be shown that the effective resolving power depends on the material of the prism which must be as highly dispersive as possible and on the size, or number, of the prisms employed - and by increasing these it has been found possible to "resolve" double lines thirty or forty times as near together as are the sodium lines. It will be convenient to take the measure of the resolving power when just sufficient to separate the sodium lines as 1,000. Then the limit of resolving power of prism spectroscopes may be said not much to exceed 40,000.*

This value of resolving power is found in practice to obtain under average conditions. Theoretically there is no limit save that imposed by the optical conditions to be fulfilled - and especially by the difficulty in obtaining large masses of the refracting material of sufficient homogeneity and high dispersive power. It is very likely that this limit has not yet been reached.

Meanwhile another device for analysing light into its component parts has been found by Fraunhofer (1821) which at present has practically superseded the prism - namely, the diffraction grating. Fraunhofer's original grating consisted of a number of fine equidistant wires, but he afterwards made them by ruling fine lines on a glass plate covered with gold leaf and removing the alternate strips. They are now made by ruling upon a glass or a metal surface fine equidistant lines with a diamond point.

The separation of light into its elements by a grating depends on its action on the constituent light-waves.

Let Fig. 2 represent a highly magnified cross section of a diffraction grating with plane waves of light falling upon it normally, as indicated by the ar-

* Lord Rayleigh has obtained results with prism of carbon disulphide which promise a much higher resolving power.

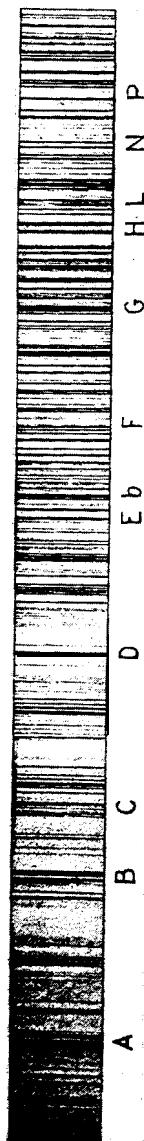


Fig. 1.

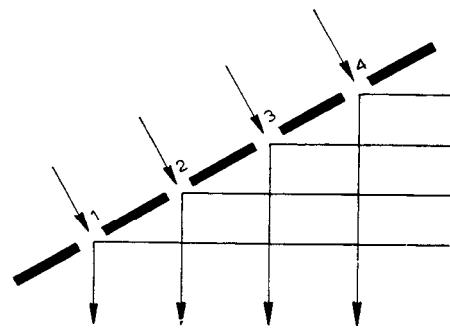


Fig. 2.

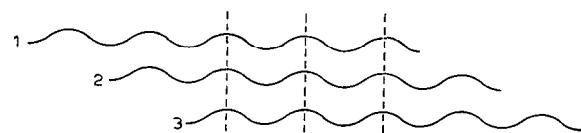


Fig. 3.

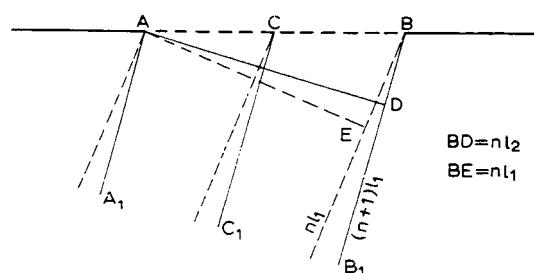


Fig. 4.

rows. The wave motion will pass through the apertures, and will continue as a series of plane waves; and if brought to a focus by a telescope will produce an image of the slit source just as if no grating were present (save that it is fainter, as some of the light is cut off by the opaque portions). This image may be considered as produced by the concurrence of all the elementary waves from the separate apertures meeting in the same phase of vibration, thus re-inforcing each other. But this may also be true in an oblique direction, as shown in the figure, if the retardation of the successive waves is just one whole wavelength (or any whole number) as is illustrated in Fig. 3, where the successive waves from apertures 1, 2, 3 ... are shown to reinforce each other just as if they all belong to a single wave-train. In this direction therefore there will also be an image of the slit source; and this direction is determined by the relation:

$$\sin \theta = \frac{m l}{s}$$

where l is the length of the light-wave of this particular color, s the distance between the apertures (the grating space), and m the number of waves in the common retardation (1,2,3, etc.). But even if the light thus diffracted be absolutely homogeneous (that is, consist of an infinite wave-train of constant wavelength) it does not follow that the light is all diffracted in the given direction; there will be some light in directions differing slightly from this - growing less until the extreme difference of path is (say) $n + 1$ waves, (instead of n when it is ml).

In fact, if we divide the pencil having this new direction into two equal parts AC and CB , the ray AA , will be $n + 1/2$ waves in advance of CC , and the two will be in opposite phases of vibration, and will therefore neutralize each other. The same will be true of each pair of rays taken in the same manner over the whole grating space, and the result is total darkness for this direction. Let us suppose we are examining the double sodium line. The difference between the components is about one thousandth of the wavelength. With a grating of n lines there will be total darkness in a direction corresponding to a retardation of $(n + 1)l$. Let this direction correspond to the brightest part of the image for the second sodium line 1_2 , so that $(n + 1)l = n l_2$ or $(l_2 - l_1)/s = l/n$. Under these conditions the two images are just <<resolved>>. But $(l_2 - l_1)/l = l/1,000$ for sodium lines, whence $n = 1,000$. That is, a grating of 1,000 lines will "resolve" the sodium lines in the first spectrum, or $R = 1,000$. In the second (where the common retardation

is two wavelengths) the resolving power is twice as great or $2n$, and in the m th spectrum, m times as great. The resolving power is therefore the product of the number of lines in the grating by the order of the spectrum, that is, $R = mn$.

In order, therefore, to obtain high resolving power the grating must have a great number of rulings and if possible a high order of spectrum should be used. The rulings need not be exceedingly close together, but it is found practically sufficient if there are from 500 to 1,000 lines per millimeter. The earlier gratings were relatively small and contained only a few thousand lines. The best of these were ruled by Nobert (1851). A very great advance was made by Rutherford of New York, who (1868) ruled gratings two inches long on speculum metal and containing about 20,000 lines. These gratings exceeded in resolving power the best prism trains in use at the time. The next advance was made by Rowland of the Johns Hopkins University, who succeeded in ruling gratings six inches long (by two to three inches stroke) having about one hundred thousand lines, and capable (theoretically, at least) of resolving in the spectrum, double lines whose distance apart was only one one-hundredth as great as that of the sodium lines. Practically this is about the limit of the power of the best Rowland grating which I have examined.

The difference between the theoretical and the actual performance is due to want of absolute uniformity in the grating space. This is due to the enormous difficulty in constructing a screw which shall be practically perfect throughout its whole length, a difficulty which increases very rapidly as the length of the screw increases; and it has been supposed that the limit of accuracy was reached in these gratings.

The great and rapidly increasing importance of spectrum analysis - especially in determining the distribution of light in so-called spectral lines under normal conditions, the resolution of complicated systems of lines, and in the investigation of the effects of temperature, of pressure, and especially of a magnetic field -justified the undertaking of much larger gratings than these. As an example of progress made in this direction, I have the honor of exhibiting a grating having a ruled surface nine inches long by four and one half inches stroke (220 x 110mm). This has one hundred and ten thousand lines and is nearly perfect in the second order; so that its resolving power is theoretically 220,000 and this is very nearly realized in actual experiments.

It will be observed that the effect produced at the focus of the telescope depends on the concurrence or opposition - in general on the *interference* of

the elementary trains of light-waves. We are again indebted to the genius of Newton for the first observation of such interference; and a comparatively slight modification of the celebrated experiment of "Newton's rings" leads to a third method of spectrum analysis which, if more indirect and less convenient than the methods just described, is far more powerful. If two plane surfaces (say the inner surfaces of two glass plates) are adjusted very accurately to parallelism, and sodium light fall on the combination at nearly normal incidence, the light reflected from the two surfaces will interfere, showing a series of concentric rings alternately bright and dark, according to the relative retardation of the two reflected light beams.

If this retardation changes (by slowly increasing the distance between the surfaces) the center of the ring system goes through alternations of light and darkness, the number of these alternations corresponding exactly to the number of light-waves in twice the increase in distance. Hence the measurement of the length of the waves of any monochromatic light may be obtained by counting the number of such alternations in a given distance. Such measurement of wavelengths constitutes one of the most important objects of spectroscopic research.

Another object accomplished by such measurement is the establishment of a natural standard of length in place of the arbitrary standard at present in use - the meter. Originally it was intended this should be the ten-millionth part of an earth-quadrant, but it was found that the results of measurements differed so much that this definition was abandoned. The proposition to make the ultimate standard the length of a pendulum which vibrates seconds at Paris met with a similar fate.

Shortly after the excellent gratings made by Rutherford appeared, it was proposed (by Dr. B. A. Gould) to make the length of a wave of sodium light the ultimate standard; but this idea was never carried out. It can be shown that it also is not susceptible of the requisite degree of accuracy, and in fact a number of measurements made with a Rowland grating have been shown to be in error by about one part in thirty thousand. But modern conditions require a much higher degree of accuracy. In fact, it is doubtful if any natural standard could replace the arbitrary standard meter, unless it can be shown that it admits of realization in the shape of a material standard which can not be distinguished from the original.

One of the most serious difficulties encountered in the attempt to carry into practice the method of counting the alternations of light and darkness in the interference method, is the defect in homogeneity of the light em-

ployed. This causes indistinctness of the interference rings when the distance is greater than a few centimeters. The light emitted by various kinds of gases and metallic vapors, when made luminous by the electric discharge, differ enormously in this respect. A systematic search showed that among some forty or more radiations nearly all were defective, some being represented by a spectrum of broad hazy "lines", others being double, triple, or even more highly complex. But the red light emitted by luminous vapor of metallic cadmium was found to be almost ideally adapted for the purpose. Accordingly this was employed: and the results of three independent measurements, made by different observers and at different times, of the number of light-waves of red cadmium light in the standard meter are as follows:

I	1,553,392.4
II	1,553,393.2
III	1,553,393.4.

It will be seen that the differences are less than half a millionth part, and this is about the limit of accuracy of the comparative measurements of the material standards. Within the last year a similar determination has been carried out by Perot and Fabry, with a result not to be distinguished from the above. It follows that we now have a natural standard of length, the length of a light-wave of incandescent cadmium vapor; by means of which a material standard can be realized, whose length can not be distinguished from the actual standard meter - so that if, through accident or in time, the actual standard meter should alter, or if it were lost or destroyed, it could be replaced so accurately that the difference could not be observed.

In the search for a radiation sufficiently homogeneous for the purpose of a standard it became evident that the interference method might be made to yield information concerning the distribution of light in an approximately homogeneous source when such observations would be entirely beyond the power of the best spectrosopes. To illustrate, suppose this source to be again the double radiation from sodium vapor. As the wavelengths of these two radiations differ by about one part in a thousand, then at a difference of path of five hundred waves (about 0.36 mm) the bright fringes of one wave-train would cover the dark fringes of the other, so that if the two radiations were of equal intensity, all traces of interference would vanish. At twice this distance they would reappear and so on indefinitely, if the separate radiations were absolutely homogeneous. As this is not the case, however, there would

be a gradual falling off in the clearness or visibility of the bands. Inversely, if such changes are observed in actual experiment, we infer that we are dealing with a double source. Further, from the distance between the maxima of distinctness, we may determine (and with extraordinary accuracy) the ratio of wavelengths of the components; from the ratio of maxima to minima we may infer the ratio of their intensities; and finally the gradual falling off when the distance becomes large gives accurate information of the "width" of the corresponding spectral lines.

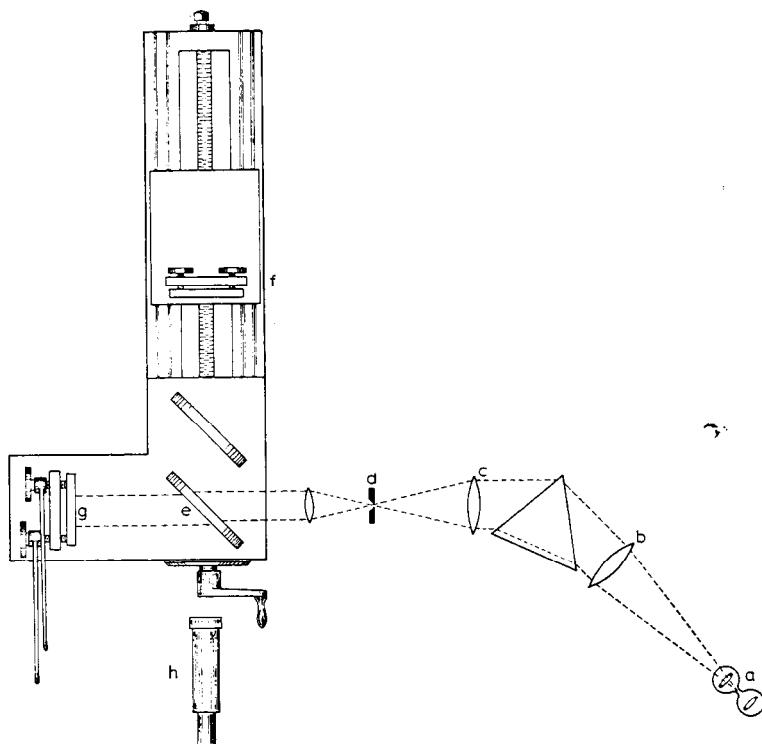


Fig. 5.

In this way it was found that the red line of hydrogen is a double with components about one fortieth of the distance apart of the sodium lines. Thallium has a brilliant green radiation which is also double, the distance being one sixtieth that of the sodium lines. Mercury shows a brilliant green line, which is highly complex, but whose chief component is a doublet, whose separation is only one seven-hundredth of that of sodium. The inter-

ference fringes are still visible when the difference of path is of the order of five hundred millimeters, corresponding to over a million light-waves; and the corresponding width of spectral line would be less than a thousandth part of that which separates the sodium lines.

Fig. 5 illustrates the arrangement of the apparatus as it is actually used. An ordinary prism spectroscope gives a preliminary analysis of the light from the source. This is necessary because the spectra of most substances consist of numerous lines. For example, the spectrum of mercury contains two yellow lines, a very brilliant green line, and a less brilliant violet line, so that if we pass all the light together into the interferometer, we have a combination of all four. It is usually better to separate the various radiations before they enter the interferometer. Accordingly, the light from the vacuum tube at *a* passes through an ordinary spectroscope *b c*, and the light from only one of the lines in the spectrum thus formed is allowed to pass through the slit *d* into the interferometer.

As explained above, the light divides at the plate *e*, part going to the mirror *f*, which is movable, and part passing through, to the mirror *g*. The first ray returns on the path *feh*. The second returns to *e*, is reflected, and passes into the telescope *h*.

The resolving power of the interferometer is measured by the number of light-waves in the difference of path of the two interfering pencils, and as this is unlimited, the interferometer furnishes the most powerful means for investigating the structure of spectral lines or groups. Its use is, however, somewhat handicapped by the fact that the examination of a single group of lines may require a considerable number of observations which take some time and during which it may be difficult to prevent changes in the light source. Nevertheless it was found possible by its means to investigate the wonderful discovery of Zeeman - of the effect of a magnetic field on the character of the radiation from a source subjected to its influence; and the results thus obtained have been since confirmed by methods which have been subsequently devised.

One of these is the application of the echelon. This is in effect a diffraction grating in which high resolving power is obtained by using a very high order of spectrum into which practically all, the light is concentrated. The number of elements may be quite moderate - since the resolving power is the product of the two. The order of the spectrum is the number of wavelengths in the retardation at each step. This retardation (which must be very accurately constant) is secured by allowing the incident light to fall upon a

pile of glass plates optically plane parallel and of the same thickness - each one a little wider than the preceding as in Fig.6.

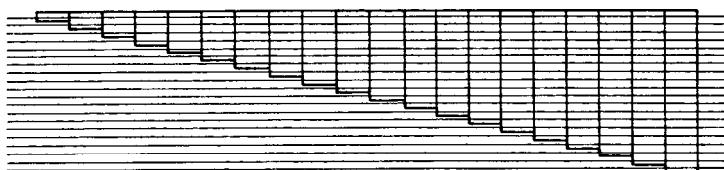


Fig.6.

Thus if the pile has forty plates, each one centimeter thick, the retardation will be about ten thousand light-waves; and the resolving power would be forty times this or four hundred thousand - which is about four times as that of a six-inch diffraction grating of the usual form. The number of elements might be increased till the absorption of the glass brought a limit. A difficulty, which appears long before this limit is reached, is due to the loss of light by repeated reflections between the many surfaces. This has been very ingeniously overcome by Mr. Twyman of the firm of Hilger & Company by pressing the plates together to actual contact - when the reflection vanishes. It is likely that the echelon under these conditions may be used by reflection instead of transmission (the plates being silvered for the purpose) with the advantage of quadrupling the resolving power for the same number of plates and eliminating the absorption.

An illustration of the efficiency of the echelon spectroscope is furnished by the following photographs of the spectrum of green radiations from mercury vapor. The first of the figures shows the spectrum of the second order of a diffraction grating whose ruled surface is nine inches by four and a half - the largest in existence. The second is by an echelon of thirty plates, seven millimeters thick, and in the third the echelon consisted of forty plates, each an inch and a fourth thick (30 mm). The corresponding lines are similarly lettered in the three figures. The scale is in Å.U.(Ångstrom units). It will be noted in the last of the three figures that the midth of the fainter companion is about one one-hundredth of an Å.U. The limit of resolution of the instrument is about half as much, or its resolving power is over a million (Figs. 7, 8 and 9).

It will be observed that the echelon spectra are repeated - thus a_1 and a_2 are two successive spectra of the same line. This is true of any grating spectrum, and the difficulties which arise from the overlapping of the suc-

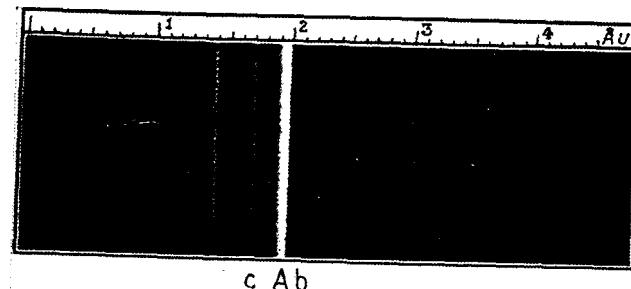


Fig.7.

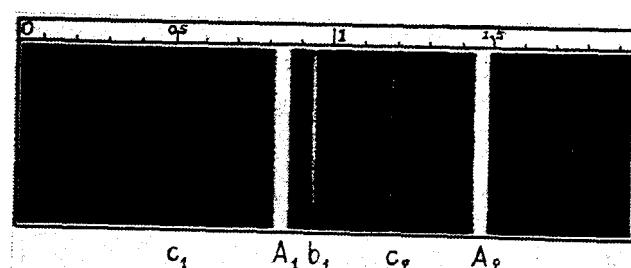


Fig.8.

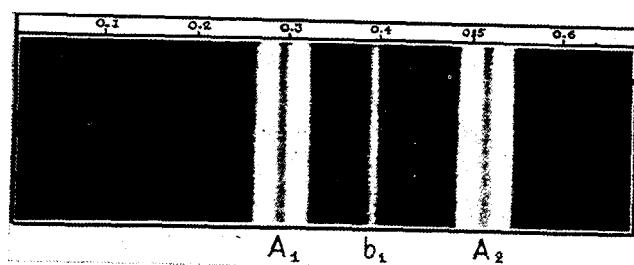


Fig.9.

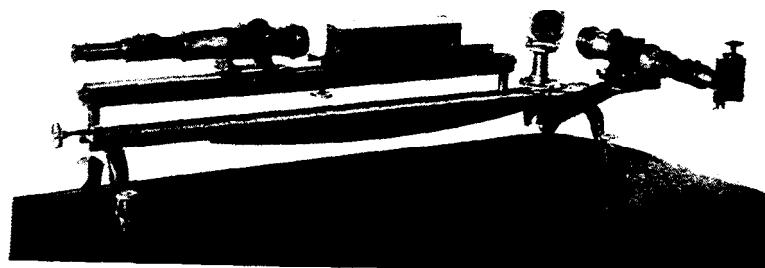


Fig.10.

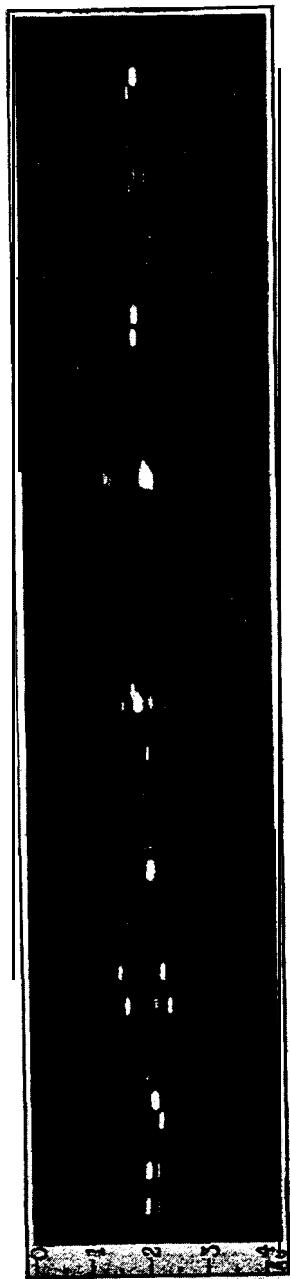


Fig.11.

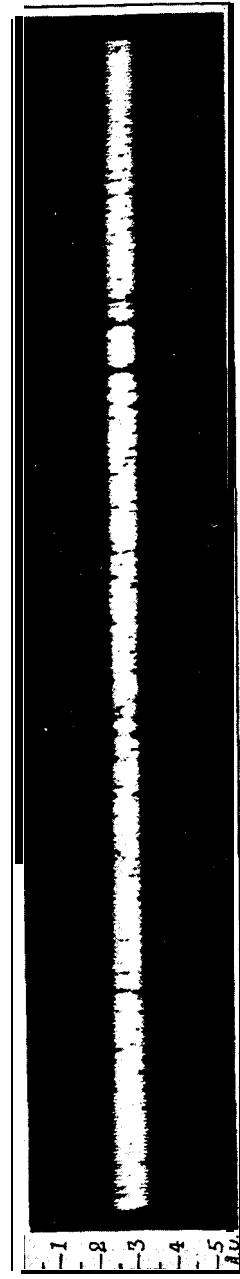


Fig.12.

cessive orders of spectrum may be overcome by separating these by a prism whose refracting edge is perpendicular to the lines of the grating. The same is true of the echelon spectrum - save that the order of the overlapping spectra is so high that a prism is hardly adequate and recourse must be had to a grating - with its plane of diffraction perpendicular to that of the echelon, as shown in Fig. 10.

With this arrangement it is possible to photograph a large part of the spectrum at once.*

Fig. 11 shows such a photograph of the iron spectrum, and it may be noted that this combination of grating and echelon makes it possible to observe absorption spectra as well as bright line spectra.

Fig. 12 shows a photograph of the solar spectrum taken in this way. It will be noted that the spectral "lines" are generally too broad to justify the use of so great a resolving power.

Finally it may be pointed out that this combination gives us the means of comparing the wavelengths of spectral lines with a degree of accuracy far superior to that of the grating.

* If the preliminary analysis has been made before the light entered the slit of echelon spectroscope, it would be possible to examine but one - at most a few - lines at a time.

PAUL EHRLICH

Partial cell functions

Nobel Lecture, December 11, 1908

The history of the knowledge of the phenomena of life and of the organized world can be divided into two main periods. For a long time anatomy, and particularly the anatomy of the human body, was the α and ω of scientific knowledge. Further progress only became possible with the discovery of the microscope. A long time had yet to pass until through Schwann the cell was established as the final biological unit. It would mean bringing coals to Newcastle were I to describe here the immeasurable progress which biology in all its branches owes to the introduction of this concept of the cell. For this concept is the axis around which the whole of the modern science of life revolves.

It is, I think, a generally acknowledged and undisputed fact that everything which happens in the body, assimilation, disassimilation, must ultimately be attributed to the cell alone; and furthermore, that the cells of different organs are differentiated from each other in a specific way and only perform their different functions by means of this differentiation.

The results produced here are mainly based on histological examinations of dead and living tissues, with, of course, most valuable contributions from the neighbouring sciences - physiology, toxicology, and particularly comparative anatomy and biology. Yet I am inclined to think that the limit of what the microscope could and has done for us is now approaching and that for a further penetration into the important, all-governing *problem of cell life* even the most highly refined optical aids will be of no use to us. Now, at this moment, the time has come to penetrate into the most *subtle chemism* of cell life and to break down the concept of the cell as a *unit* into that of a *great number* of individual specific *partial functions*. But since what happens in the cell is *chiefly* of a *chemical* nature and since the configuration of chemical structures lies beyond the limits of the eye's perception we shall have to find other methods of investigation for this. This approach is not only of great importance for a *real* understanding of the life processes, but also the basis for a truly rational use of medicinal substances.

The *first* advance in this complicated field came about, as so frequently happens, in a roundabout way. After Behring's renowned discovery of the

antitoxins I had set myself the task of penetrating further into the mysterious nature of this process, and after long labours I have succeeded in finding the key to it.

As you know, the function of stimulating the formation of antibodies belongs to one *particular* group of poisonous substances only, to the so-called *toxins*. These are metabolic products of animal and plant cells: diphtherial toxin, tetanus toxin, the phytotoxin of jequirity, ricin, snake venom, e tutti quanti. None of these substances can be made to crystallize, and they obviously belong to the class of protein substances. The toxin is generally characterized by two properties: (1) by its poisonousness; and (2) by its ability to stimulate the production of the specific antitoxin in the animal body.

My quantitative investigations of this process have shown that the toxins, especially the solutions of diphtherial toxin, will - either spontaneously if left standing for some time, or through the action of thermal influences or certain chemicals (iodine) - change in such a way that they are more or less deprived of their toxicity but retain their ability to produce antibodies. Furthermore, it has become obvious that the products of this transformation which I call *toxoids*, and which my honoured friend Professor Arrhenius has also encountered in his numerous experiments, have still retained the ability to neutralize the antitoxin in a specific way. Indeed, in favourable cases I and others have succeeded in proving that the transformation of toxin into toxoid can be a perfectly quantitative one, so that a given toxic solution will combine with exactly the same amount of antitoxin before and after the transformation into toxoid.

These facts permit only one explanation, namely that there must be two differently functioning groups present in the toxin. One of these, which has been preserved also in the "*toxoid*" and must therefore be considered the more stable one, must be allowed the ability on the one hand to stimulate the formation of antibodies in the animal body by immunization, and on the other to neutralize antibodies in the test tube and *in vivo*. The relations between toxin and its antitoxin are *strictly specific* - tetanus antitoxin neutralizes exclusively tetanus toxin, diphtheria serum only diphtherial toxin, snake serum only snake venom, to mention just a few examples out of hundreds. For this reason it must be assumed that the antipodes enter into a *chemical bond* which, in view of the *strict specificity* is most easily explained by the existence of two groups of distinctive configuration - of groups which according to the comparison made by Emil Fischer fit each other "like lock and key". Considering the stability of the bond on the one hand and the fact on the other that neu-

tralization occurs even in very great dilutions and without the help of chemical agents, it must be assumed that this process is to be attributed to a close chemical relationship and probably represents an analogue to actual chemical syntheses.

More recent investigations have in fact shown that it is possible by chemical actions to break up the product of the union, the neutral combination toxin-antitoxin, into its original components. For instance Morgenroth in particular has proved in the case of a number of toxins - I shall just mention snake venom and diphtherial toxin - that under the action of hydrochloric acid the compound can be separated again into its original constituents, in the same way that in pure chemistry stable compounds, such as the glycosides, can through the action of acids be broken down into their two components: sugar and the basal aromatic complex. These investigations have shown that the *stable* group of the toxin molecule, which I call *haptochrome*, can exercise great chemical activity of a specific kind, and thus the obvious assumption was that it must be precisely this group which causes the adhesion of the toxin to the cell. When we see how some bacterial poisons produce disturbances only after weeks of incubation and then damage the heart or kidney or nerves, when we see how animals suffering from tetanus present contractions and spasms for months, we are forced to the direct conclusion that all these phenomena can only be caused by the *adhesion of the toxic substance to quite definite cell complexes*.

I therefore assumed that the tetanus toxin for instance must unite with certain chemical groupings in the protoplasm of cells, particularly the motor ganglion cells, and that this chemical union represents the prerequisite and cause of the disease. I have therefore simply called such cell groupings "*poison receptors*" or just "*receptors*". Wassermann has been able to prove my view correct in every detail in his noted experiments in which he was the first to produce evidence that normal brain is able to render innocuous given quantities of tetanus toxin which are introduced. Many objections have been made to these experiments, but none have proved valid, and I believe that I may now pronounce it as a definite fact that certain specific groupings must in fact exist in the cells which fix the poison. That these, the cell's *receptors*, which *produce the fixation*, react to the *haptochrome* part of the toxin can be deduced from the immunizations through toxoids, where the haptochrome group is the only one which has been preserved. But since this haptochrome grouping of the toxin must have a highly complex and peculiar stereochemical structure, and since it reacts *simultaneously* and in *the same* sense to the cell receptors

and the antitoxin, it must be concluded from this that the group in the protoplasm, the *cell receptor*, must be identical with the "antitoxin" which is contained in solution in the serum of immunized animals, for a *really well-made key* will not open different locks at the same time. As the cell receptor is obviously pre-formed, and the artificially-produced antitoxin only the consequence, i.e. secondary, one can hardly fail to assume that the antitoxin is *nothing else* but *discharged components* of the cell, namely receptors discharged in excess. The explanation of this fact was a very obvious one. One only needs to assume that the various specific cell receptors which take up the snake venom, the diphtherial toxin, the tetanus toxin, the botulin poison, etc. are not *properly speaking* designed for the purpose of serving as *toxic receptors* for substances with which the animal under the normal conditions of its life might perhaps never come into contact, but that they exist, in actual fact, in order to combine chemically with *normal* products of metabolism, i.e. to *assimilate* them. As these receptors, which may be regarded as *lateral chains* ("Seitenketten") of the protoplasm, capable of assimilation, become occupied by the toxin, the relevant normal function of this group is eliminated. Now *that* element comes in which was to be expected from the *fundamental law* of tissue defect and its compensation, discovered by Karl Weigert - the deficiency is not merely *exactly compensated*, but *made up to excess*, i.e. there is hyperregeneration. Finally, if the injections are increased and repeated, so many such groupings are formed in the body of the cells that they inhibit as it were the normal functions and the cell gets rid of the disturbing excess by discharging them into the blood.

The colossal difference between the amount of poison injected and the antitoxin produced is probably the most characteristic feature of this process and this is best illuminated by Knorr's statement that *one part* of toxin produces an amount of antitoxin capable of neutralizing *millions of times* the amount of the poison which started the process.

There are however many minds which consider the process a much more simple one. Straub is of the opinion that it is on the whole analogous to simpler processes of vital detoxification, e.g. to the forming of a sulphuric acid ester from injected phenol, and that these processes only differ in that the phenol-sulphuric acid remains stable in the organism, whilst the toxin-antitoxin complex in the organism is not held but is partially destroyed. But only one component, the injected toxin, is said to perish, while the *other*, the product of the reaction of the organism - as something which has developed in the body and thus is not foreign to it - escapes elimination and remains

preserved in the blood and body fluids. By systematic repetition of the poisoning it would then be possible to accumulate protective power in the blood, so that when it is introduced into other organisms it can also protect these from toxic diseases and would thus be acting as a curative serum.

So far Straub. Faced with such a simple explanation it can only be surprising that this problem has occupied the great army of researchers studying immunity for so many years. But in fact the author has *completely* missed the *vital clue*, namely that according to his theory a certain amount of toxin would produce only exactly the *equivalent* amount of antitoxin! In actual fact this is fortunately not the case in immunization. On the contrary, it has been proved much more conclusively - and I refer to my statement about Knorr above - that one part of poison can produce so much antibody that a millionfold multiple of the equivalent is achieved. This should prove Straub's view untenable.

It is much more important that from the evidence of this hyperregeneration the *pre-formation and the chemical individuality of the toxin receptors concerned is proved*. That which can be constantly formed anew in the cell and mixed with the blood like a secretion *must* have a chemical "*individuality*", and with realization of *this* the *first step* had been taken which led to the differentiation of the concept of the cell into that of *a great number of separate, individual functions*. I had assumed right from the beginning that the toxin represents *nothing more than a nutritive substance capable of assimilation*, to which in addition - by some sort of accident - is attached a *lateral grouping*, usually of an *unstable nature*, which causes the toxic action as such.

This view, which I have held from the beginning, has subsequently very quickly found confirmation many times over. It has in fact been possible to prove the complete independence of the haptophore and toxophore groups, as substances were discovered which had the ability to produce antibodies, and therefore were antigens, without at the same time having a toxic effect. Perhaps I may remind you in the first place of the precipitins, which were first discovered by Kraus, Tschistowitsch and Bordet. Through the important discovery that even the genuine protein substances of animal and plant organism are able, irrespective of whether they have a toxic effect or not, to produce antibodies with a specific chemical reaction, an antigenic nature could be proved also of actual nutritive substances, just as could previously be expected after my observations. But even among the poisons produced by nature some have been found which will readily demonstrate the independence of the haptophore and the toxophore apparatus. These are the cytotoxins

which are normally found in the blood serum of higher animals or can be produced arbitrarily through immunization with any type of cell. They differ from all other poisons known to us in their extraordinary specificity, in their monotropic action, which so far distinguishes only these poisons which are fabricated in the living animal body. Because of the complexity of their constitution a differentiation between the haptophoric and the toxophoric principle is palpably obvious, so that here the distributive component, the amboceptor, is given the function of concentrating the actual active substances on the affected substratum, through the increase in avidity which follows localization. The fact that the animal cells are antigens, although they have no toxic action, proves simultaneously, not only the possibility of immunization with protein substances in solution, but also the sole responsibility on the part of the haptophore group for the formation of antibodies.

It is precisely this discovery and analysis of the specific relations between haptophore antibody groups and receptors which has become of the highest theoretical and practical importance for the more recent serum diagnosis. I mention only the determining of the agglutination titre which has found its most important use in Widal's typhoid reaction; the differentiation of proteins established by Wassermann and Uhlenhuth which is so important for forensic blood tests; the measuring of the opsonic index inaugurated by Wright, not to mention the manifold uses which have been found for the process of complement fixation - the scientific foundation of which likewise rests on the principle of the adhesion of the antibody to the haptophore group.

I will not go into this any further now and will only draw "*the*" conclusion from it that there are a series of nutritive substances, probably mostly of protein nature, which find *specific* receptors in the cells and that it is thus possible, *through immunization*, to lure into the blood these structures in great abundance and in the form of typical varieties - as represented by the agglutinins, the precipitins, the amboceptors, the opsonins on the one hand, and the antitoxins and antifermenents on the other. They can then be accumulated there to such an extent that a thorough study of these substances, which within the cell-formation is quite impossible, can now actually be undertaken. How far the analysis of such processes can be taken is shown by the study of the type of link between toxin and antitoxin and the discovery of the very complicated action of the amboceptors.

Of course this does not solve the secret of life itself, which may be compared with the complicated organism of a mechanical work of art, but nevertheless the possibility of taking out *individual* wheels and studying them exactly sig-

nifies an advance compared with the old method of *breaking into pieces* the whole work and then trying to deduce something from the mixture of broken pieces.

I describe all the receptors which are able to and designed to assimilate nutritive substances as "*nutriceptors*" and would regard these nutriceptors as the source of the *antibodies* which are theoretically and practically so important, and which I have enumerated above. Obviously anyone adhering to the pluralistic point of view - and considering the complicated system of the organism, the almost illimitable variety and specificity of cell functions, this seems to me absolutely inescapable - must assume that there exists a whole range of nutriceptors of different types. From the point of view of immunization these can be differentiated into three types:

(1) Those which do *not* enter the blood in the form of antibodies. It may be assumed that this will probably be the case with those nutriceptors which serve the *very simplest* functions, for instance the assimilation of simple fat substances or of types of sugar.

(2) Those which enter the blood in the form of the antibodies mentioned and characterized above, and the development of which corresponds to a *hyperregeneration*.

(3) The third form presents a contrast to this in so far as it is not a case of new formations, but of a *decrease* in receptors. Experimental proof of this occurrence has however so far been only very rare. The only known instance is probably the evidence produced by H. Kossel that after prolonged immunization of rabbits with the haemotoxic eel serum the blood corpuscles as such did finally become insensitive to this agent, as though they had lost the specific receptors.

Now I, in company with my colleagues, Dr. Röhl and Miss Gulbransen, have succeeded in penetrating further into the nature of the artificial loss of receptors and in illuminating the whole mechanism. Our work will shortly be published in a more extensive form; here I would like to emphasize that the experiments were done on trypanosomes. Franke had at one time infected a monkey at my Institute with a certain species of trypanosome, then brought about its cure through chemotherapeutic agents, and then again, in order to test the immunity of the animal, reinfected it with the original strain. But contrary to expectation it turned out that the monkey was not immune, but that it sickened again after a very prolonged period of incubation. If mice were treated with blood coming from the infected animal, i.e. containing trypanosomes, they fell ill and died. But if the trypanosomes were first re-

moved from the blood of the monkey it became apparent that the serum thus produced was capable of killing off the *original parasites*. This revealed that a variety of the parasites had developed in the monkey which in contrast to the original strain was no longer affected by the serum - a *serum-resistant strain*. Similar observations were at the same time recorded by Kleine and lately also by Mesnil.

Now if experimental animals which have been infected with a certain species of trypanosome are treated not with a full sterilizing dose of a suitable substance (arsanil, arsacetin, arsenophenylglycin), but with a somewhat smaller one, trypanosomes disappear from the blood for a greater or lesser period of time. The formation of antibodies has occurred in this case too, as can be easily proved. The few parasites which have escaped death now remain in the organs for a greater or lesser period of time, gradually adapt themselves to the anti-substances in the serum, and then, as soon as this has happened, return to the blood where they increase rapidly and lead to the death of the animal. If the trypanosomes obtained by this method are transferred to one group of mice which have been previously infected with the original strain, have been cured through the administration of suitable doses and have thus become carriers of the specific antibodies, and to a second group of normal mice, one becomes convinced that the parasites grow equally quickly in both groups. The parasites of the recidive strain have therefore undergone a biological change in that they have become *serum-resistant*.* The change which has thus been produced in the parasites is not a superficial one, but may *be reproduced unchanged for many months* by passage through normal animals. The recidive strain retains unchanged its property of being *resistant* to the antibodies produced by the *original strain* and can thus be identified with absolute *certainty*.

It was now our concern to obtain an insight into the nature of this process. The explanation for this which we have found after many and varied experiments is the following : the original strain contains an abundance of a certain uniform type of nutriceptor which we shall call group "A". When the parasites are killed and dissolved within the organism of the mice the "A" grouping acts as an antigen and now produces an antibody which originates by virtue

* It is, by the way, possible to get exactly the same strain in another, much simpler way which consists of infecting the mice with the original strain, fully curing them on the second day with a full dose, and then reinfecting them 2-3 days later with the same strain. After a greater or lesser period of time parasites will appear in the blood which fully correspond to those of the recidive strain.

of its relationship to group "A". If living parasites are now brought into contact with this antibody, either in the test tube or in vivo, it will be adhered to by the trypanosomes. The effect on the parasites in this way is that they undergo in vivo the biological change which leads to the development of the recidive strain. This change occurs in that in the new strain the original "A" grouping disappears and a new grouping, which we shall call "B", appears instead. That there is a new grouping in the recidive strain can be shown as follows : if two mice are infected with the recidive strain - carrier of the "B" grouping - and then completely healed; if one mouse is then infected with the original strain, and the other with the recidive strain itself, the reinoculation with the original strain - carrier of "A" grouping - proceeds smoothly, while reinfection with the recidive strain fails at first. This shows that the original strain and the recidive strain are not identical, or must possess *two differently functioning groups*. We therefore have a typical case of immunization producing loss of receptors while developing a completely new type of receptor.

Whether one calls this change a mutation or a variation is really of little significance; the main thing is that it can be produced intentionally and artificially and that it is hereditary. But in view of the great interest which this particular problem has for biology and the theory of evolution, we have tried to get a fuller understanding of the process.

First of all it was necessary to determine how the trypanosome-antibodies influence the parasites. In accordance with the assumption common in immunology it might be accepted, that these antibodies produce direct toxic actions, i.e. contain toxophoric or trypanolytic groups, and that therefore the adhesion as such would necessarily produce damage to or death of the cell. But my colleagues and I have become convinced that this is not the case. In contrast to the usual species of trypanosomes, which contain only one uniform grouping "A", "B", or "C", etc. and which may therefore be called "*unios*", other types present themselves, which have two groups in the protoplasm at the same time, e.g. "A" and "B", and may therefore be called "*binios*". If one such binio "A" - "B" is acted upon by the isolated antibody "A" or "B", this does not cause the slightest damage to growth. This arises only if the parasite is occupied by both anti-substances at the same time. It follows from this that the presence of antibodies does not have a direct toxic effect on the trypanosomes, and it seems to follow from this triple experiment that the antibody only has an effect in so far as it prevents the intake of nutritive substances through occupation of the group concerned. If in the binio "A" - "B" the grouping "A" is obstructed by the antibody, the parasite can continue

to vegetate through its grouping "B". This also proves that the groupings "A" and "B" must be chiefly regarded as nutriceptors.

If the amount of antibody is very large the parasite can no longer feed itself at all and dies. It is easiest to convince oneself of this by mixing parasites with varying amounts of antiserum in the test tube. With the high concentrations which stop the intake of food altogether the death of the parasites follows, while with weaker concentrations which permit a *vita minima* in which mutation is possible a recidive strain develops. This mutation must therefore be entirely due to *starvation of the protoplasm*, under the influence of which new potential structures of the trypanosome develop. Antibodies like those which we have just been considering, and which have a purely *anti-nutritive* action, I call "*atrepisks*" and I believe that these *probably play an extraordinarily important role* not only for bacteria, but in *biology in general*.

Most of my colleagues in this field will probably find it easy to accept the idea that there are certain chemical groupings in the cell for the reception of the various nutritive substances, once their existence has been definitely proved by the presence of the antibodies. But what is much more difficult is the question whether for the reception of other, less complicated substances too there are analogous functional groups. For the simplest further function of the cell, the *absorption of oxygen*, the problem is in my opinion already solved. We know that in the haemoglobin molecule it is exclusively the organically associated *iron residue* which provides the loose link between oxygen on *the one hand* and carbon dioxide and hydrocyanic acid on *the other*. It will therefore be necessary to assume certain groupings in the protoplasm of the red blood corpuscles, which have a maximal relationship to iron, and form a *complex compound* with it which has the characteristic functional properties. The protoplasm of the red blood corpuscles would thus be characterized by the abundant presence of "*ferroceptors*" which complemented with iron would lead to the finished haemoglobin molecule. In a similar way it will also be necessary to assume that the blue respiratory pigment of crayfish contains "*cuproceptors*", and others probably "*manganoreceptors*". Also, the localization of *iodine* in certain glandular systems, particularly the thyroid gland, and the evidence that iodine is arranged in certain aromatic lateral chains will have to be interpreted in this way.

Much more difficult, however, is the question whether such preformed chemoreceptors may also be assumed to exist in the cell in the case of the great number of actual *medicaments*. This question takes us into the important field of the relation between *constitution* and *action*, which represents the basis

for a rational development of therapy. Only when we really know the points where the parasites attack, only when we have established what I call the *therapeutic biology of the parasites*, will it be possible to combat the infective agents successfully.

I have therefore carried out these studies of mine on the detection of definite chemoreceptors, first on monocellular living beings - protists - because the *conditions* there are much more favourable to a clear understanding than those in the infinitely complicated machinery of the higher organisms. I therefore asked myself the question: do the *trypanosomes* possess in their protoplasm definite *groupings* which govern the captivation of definite chemical substances?

If a certain substance is able to kill trypanosomes or other parasites in the test tube or in the animal body, this can *only* happen because an accumulation of it takes place in these parasites, but the *process* itself is not explained by the establishment of these *bare facts*. There are *very many* explanations for this and only when it is possible to prove that we have here a *function* which is open to *specific* changes and variations, will we have proof of a *preformed* formation.

Unfortunately it appears that the way in which it was so easy to produce proof of preformation for the *nutriceptors*, namely by the *transfer* of the cast-off receptors into the blood, does not apply for the chemoreceptors, as they are much more simply constructed and remain attached to the cell - that is, they are *not rejected*.

Here it was only possible to see clearly in a *roundabout way*, which took us via the *drug-resistant* strains of the trypanosomes. Together with my faithful colleagues Franke, Browning, and Röhl, I have shown that it is possible to obtain by a systematic treatment trypanosome strains which are resistant to the three substances which so far are known to be inimical to trypanosomes: compounds of the arsenic series, fuchsine, and the acid azo-dye from the benzopurpurine series, trypan red. These resistant strains have the following characteristics:

(1) Stability of the acquired property. This is so great that for instance our arsenic strain, after it has passed in $2\frac{1}{2}$ years about 380 times through mice, is now even today *equally resistant to* drugs as the original strain.

(2) A principal characteristic of drug resistance is its *strict specificity* which is distinctive in that it relates not to one specific compound but to the *whole chemical grouping* to which this specific compound belongs. The strain resistant to *fuchsine*, for instance, is not only resistant to this, but also to a whole series of related triphenylmethane dyes, e.g. *malachite green*, *ethyl green*, *hexaethyl*

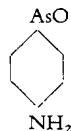
violet. On the other hand it has remained sensitive to both the other types, that is, to trypan red and an arsenical substance. A corresponding specificity is shown by the strain resistant to trypan red and also that resistant to arsenicals. That there are in fact three different functions here is furthermore apparent from the fact that by successive treatment of *one* and the same strain of trypanosomes with the three above-named substances it is possible to obtain a triple-resistant strain, i.e. a strain which is resistant to representatives of all three classes. Such a strain, assuming *maximal* stability, is extremely valuable for the discovery of new types of trypanocidal agents. If, for instance, some new substance is obtained which as such is capable of destroying the normal trypanosomes, it is only necessary to let this substance act upon the triple-resistant strain to find out whether it is a new type of remedial substance or not. If not, the triple-resistant parasites will *not disappear* with this treatment, but continue to flourish; but if they *do disappear* then the substance under test does not correspond to any of the three types of remedial substance mentioned, and a *representative of a new class of remedial substance* is being dealt with. The triple-resistant strain is therefore so to speak the *cribrum therapeuticum*, the *therapeutic sieve*, with the aid of which it is possible to recognize what is *homologous*, and separate what is *different*.

A further important question was then to determine in what way this *specific* drug resistance comes about. Here it was the atoxyl strain which I used for the experiments. To get an exact picture it seemed necessary to investigate the behaviour of the arsenic resistant parasites in the test tube, removed from all the disturbances and complications of the organism. In this a great difficulty soon arose, as the remedial substance used most often in therapy, atoxyl (*p*-aminophenyl arsenic acid) does not have the slightest lethal effect on trypanosomes in *the test tube*; even solutions of a higher percentage were not sufficient for this. This phenomenon was all the more striking since according to Koch's investigations the parasites could be made to vanish within the human body in a few hours after injections of 0.5 grams of atoxyl; a lethal effect had therefore been achieved with a concentration of 1:120,000.

This was a process which more recently has been named "indirect effect". It was not difficult for me to find the reason for this phenomenon as I had in previous years made a thorough study of the reducing power of the body. We know that arsenic acid in the body is reduced to arsensious acid; we also know that cacodylic acid is reduced to that *foul-smelling* cacodyl; it was therefore obvious to think of reduction first of all. In atoxyl, *p*-aminophenyl arsenic acid, the arsenic residue is pentavalent, while in the two products of

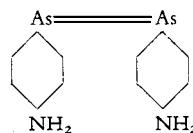
the reduction of it the arsenic residue only has a trivalent action - as in arsenious acid. We thus obtained two different products :

(1) monomolecular *p*-aminophenyl arsenic oxide



and

(2) arising from the reduction of the latter, the yellow diaminoarseno-benzene



In contrast to atoxyl, these substances proved to be highly trypanocidal both in the test tube and in the animal body. Even solutions of 1:1,000,000 of the arsenic oxide compound destroyed the trypanosomes within one hour. The closely related *p*-hydroxyphenyl arsenic oxide has an even stronger effect: 1:10,000,000.

Through this it was proved that the pentavalent arsenic residue releases no trypanocidal function whatsoever, but that this function is exclusively connected with the trivalent unsaturated state.

More than 60 years ago Bunsen, with prophetic clarity of perception, pointed out that cacodyl, the product of reduction, is so poisonous in comparison with the almost non-poisonous cacodylic acid, and deduced from this the chemical character of the binding of the cacodyl. It also tallies extraordinarily well with this that *unsaturated* carbon monoxide, for instance, and a number of other unsaturated compounds are so much more toxic than the corresponding saturated radicals. We shall therefore have to assume that the *arsenoceptor* of the cells is only able to take up the arsenic residue which is unsaturated and therefore eager to adhere.

With the aid of such reduced compounds it was now quite easy to examine the atoxyl strain in the test tube. It became apparent that suitable concentrations of the chemicals would still destroy it, i.e. that this was not a case of

loss of receptors, as we had proved with regard to the recidive strain. But a comparison of the *lethal* dose with *that* necessary to destroy the *normal* strain showed that the resistant strain required a much higher *concentration*, and that an amount which would destroy the *normal* strain at once did not, even after one hour, show the slightest effect on the *viability* of the resistant parasites.

These test-tube investigations seemed to indicate that the arsenoceptor had been preserved in the atoxyl-resistant trypanosome strain, but that its avidity had decreased, which could be seen from the fact that only through the use of much stronger solutions could the *toxic concentration* necessary for *lethal effects* be achieved; the normal arsenoceptor of the original strain will attract the same amount to itself from weaker solutions because of its initially higher avidity.

We have now been able to prove biologically quite clearly that the *arsenoceptor* does in fact represent a certain *function*, the avidity of which can be *systematically* and *successively* decreased through *immunization*. So far we have been able to reach *three* different stages of relationship. Stage I was achieved by subjecting the parasites systematically to the treatment of *p*-aminophenyl arsenic acid and its acetyl product. We continued the treatment *ad maximum* for years, until there was no further increase. The resistant strain thus obtained was *at the same time* also resistant to a whole series of other arsenic compounds, from among which I would particularly like to mention the *p*-oxide compound, the urea compound, the benzylidene compound, a number of acid derivatives, etc.

As there is the possibility - and in animal experiments this happens very frequently - that arsenic-resistant strains develop during therapeutic processes in animal and in man and these do of course completely prevent a successful continuation of therapy, it was now necessary to find substances which were still able to attack the resistant strain and combine with its receptors. After a long search we found altogether three compounds, the most important of which is arsenophenylglycine. With the help of this compound it was possible to bring even the *arsenic strain I* characterized above to a cure, which can only be thus explained that the substance seizes the avidity stump of the arsenoceptor like a *pair of pincers*. With this anchorage, however, the possibility opens to obtain a still higher resistance to arsenic. We did in fact succeed in this, though not without considerable trouble, and derived from the arsenic strain I at a higher level, *arsenic strain II*, which was completely resistant to *arsenophenylglycine*.

Now Plimmer has recently discovered a preparation, tartar emetic, which

in high dilutions also destroys trypanosomes. Tartar emetic is the salt of an antimony compound which is closely related chemically to arsenic. When we thereupon tested tartar emetic on the arsenic strain II we found that the latter was destroyed by the *tartar emetic*. Furthermore we succeeded in going a stage further by treating the arsenic strain II with *arsenious acid*, so that there now developed the third strain, *arsenic strain III*, which had now also become resistant to tartar emetic. I would like to emphasize particularly that this arsenic strain III, which was bred only under the *influence of arsenious acid*, was resistant to *tartar emetic*, but not to *arsenious acid*. This result can only be explained by the assumption that it is arsenious acid which, of all conceivable arsenic compounds, has the *maximal* relationship to the arseneceptor, and that it will probably require the greatest effort or even be entirely impossible to produce a strain - and this would be *arsenic strain IV* - which would be resistant to arsenious acid as well.

To support my view that under the influence and attack of selected compounds there is a successive avidity restriction of the same receptor, I could produce many additional interesting facts, as for instance the phenomenon that the trypanosome can of course also be made resistant directly, with a more strongly effective reagent, i.e. arsenophenylglycine. A strain produced in this way proved as expected resistant also to the class of less avid substances, that is, atoxyl, acetyl arsenilate, etc. A panresistant strain would thus be obtained if one were to start producing resistance with right away the most highly effective agents - and these are tartar emetic and arsenious acid. According to our researches, however, it seems unfortunately impossible to produce resistance directly with these substances; it is only possible to do this in the roundabout way via the previous treatment of strains with phenyl arsenic acid derivatives.

The restriction of avidity is of course a chemical process which obviously allows the interpretation that in the vicinity of the arsenic grouping concerned, *other groups develop or disappear* which reduce the capacity to react. Perhaps I may give a chemical example. Benzyl cyanide reacts to nitrosodimethyl-aniline. But in order that the reaction may take place, heat and a stronger condensation agent, the free alkali, are necessary. If on the other hand a nitro-group is introduced into the benzene nucleus, the reacting power of the methylen group is heightened tremendously: the two substances, nitro-benzyl cyanide and nitrosodimethylaniline, react even in the cold. The introduction of the nitro-group has therefore had an accelerating influence on the reaction. If the nitro-compound is reduced to *p*-aminobenzyl cyanide it is less capable

of reaction than the original material; the amino-group has therefore had a diminishing influence on the reaction, while the acetyl product of the amino-compound reacts more or less like the original material.

We can see from this simple example that three different groupings, attached to the benzene nucleus in the paraposition, will either have no influence whatsoever on the methylene group, or *strengthen* it or *weaken* it. The *weakening would in our case correspond to the restriction of avidity*.

In my opinion the protoplasm can therefore be divided into a large number of individual functions which are interspersed among the *nutriceptors* in the form of different *chemoceptors*. But in my opinion these two main groups must be closely interconnected. This becomes apparent from the following consideration:

Trypanosomes of different origin, bred in different laboratories, usually show a different behaviour right from the beginning towards a certain curative substance. For instance, the trypanosome strain Mal de Caderas which I tried first had no resistance to trypan red, and I was therefore able to get a cure with this substance. This is still possible even today. Jakimov has had similarly good cures in Russia, while Uhlenhuth could not observe any influence on his strains. These are therefore *innate* differences; but that these are not *wholly fortuitous* is obvious from the fact that even *today*, after it has passed through normal mice for *many years*, my strain shows the same *curability* through trypan red as before. In contrast to this my Nagana strain could not be cured by trypan red and is still the same today. But when we made this strain into a *recidive strain* then it became apparent that within *14 days* this property which had been continued and maintained for years had altered. This proves that the *chemoceptors* are connected with the constitution of the *protoplasm* and undergo *alterations* if we alter the constitution of the protoplasm by mutation.

The reverse, i.e. whether a change of the cell substance, and particularly its nutriceptors, can be achieved by influencing the chemoceptors, has on the other hand not yet been definitely established. Browning had indeed observed and reported that through the serum reaction the fuchsine and atoxyl strains differ from each other and from the original strain. But more detailed investigation has shown further that these were not specific changes, in connection with *fuchsine* or *arsenic*, but changes corresponding to the recidive mutation described above; changes which are due to the fact that during the treatment the mice have frequently undergone recidivations which then led to the development of recidive strains.

I have thus come to the end. I am conscious of the fact that there are gaps in the work I have presented. But how could this be otherwise with a subject a truly exhaustive study of which would require the recapitulation of long and wearisome labours? But I did want to show you that we are getting to grips with the problem of obtaining an insight into the nature of action of therapeutic substances, the conception of which must consist in the recognition de sedibus et causis pharmacorum. I also hope that if these aspects are followed up systematically, it will be easier than heretofore to develop a rational drug synthesis, and I may mention that in this respect arsenophenyl-glycine has so far proved an entirely ideal remedy in animal experiments.

For with the help of this substance it is really possible in every animal species and with every kind of trypanosome infection to achieve a *complete cure* with *one injection*, a result which corresponds to what I call *therapia sterilisans magna*.

W O L F G A N G P A U L I

Exclusion principle and quantum mechanics

Nobel Lecture, December 13, 1946

The history of the discovery of the « exclusion principle », for which I have received the honor of the Nobel Prize award in the year 1945, goes back to my students days in Munich. While, in school in Vienna, I had already obtained some knowledge of classical physics and the then new Einstein relativity theory, it was at the University of Munich that I was introduced by Sommerfeld to the structure of the atom - somewhat strange from the point of view of classical physics. I was not spared the shock which every physicist, accustomed to the classical way of thinking, experienced when he came to know of Bohr's « basic postulate of quantum theory » for the first time. At that time there were two approaches to the difficult problems connected with the quantum of action. One was an effort to bring abstract order to the new ideas by looking for a key to translate classical mechanics and electrodynamics into quantum language which would form a logical generalization of these. This was the direction which was taken by Bohr's « correspondence principle ». Sommerfeld, however, preferred, in view of the difficulties which blocked the use of the concepts of kinematical models, a direct interpretation, as independent of models as possible, of the laws of spectra in terms of integral numbers, following, as Kepler once did in his investigation of the planetary system, an inner feeling for harmony. Both methods, which did not appear to me irreconcilable, influenced me. The series of whole numbers 2, 8, 18, 32... giving the lengths of the periods in the natural system of chemical elements, was zealously discussed in Munich, including the remark of the Swedish physicist, Rydberg, that these numbers are of the simple form $2 n^2$, if n takes on all integer values. Sommerfeld tried especially to connect the number 8 and the number of corners of a cube.

A new phase of my scientific life began when I met Niels Bohr personally for the first time. This was in 1922, when he gave a series of guest lectures at Göttingen, in which he reported on his theoretical investigations on the Periodic System of Elements. I shall recall only briefly that the essential progress made by Bohr's considerations at that time was in explaining, by means of the spherically symmetric atomic model, the formation of the intermediate

shells of the atom and the general properties of the rare earths. The question, as to why all electrons for an atom in its ground state were not bound in the innermost shell, had already been emphasized by Bohr as a fundamental problem in his earlier works. In his Göttingen lectures he treated particularly the closing of this innermost K-shell in the helium atom and its essential connection with the two non-combining spectra of helium, the ortho- and para-helium spectra. However, no convincing explanation for this phenomenon could be given on the basis of classical mechanics. It made a strong impression on me that Bohr at that time and in later discussions was looking for a *general* explanation which should hold for the closing of *every* electron shell and in which the number 2 was considered to be as essential as 8 in contrast to Sommerfeld's approach.

Following Bohr's invitation, I went to Copenhagen in the autumn of 1922, where I made a serious effort to explain the so-called « anomalous Zeeman effect », as the spectroscopists called a type of splitting of the spectral lines in a magnetic field which is different from the normal triplet. On the one hand, the anomalous type of splitting exhibited beautiful and simple laws and Landé had already succeeded to find the simpler splitting of the spectroscopic terms from the observed splitting of the lines. The most fundamental of his results thereby was the use of half-integers as magnetic quantum numbers for the doublet-spectra of the alkali metals. On the other hand, the anomalous splitting was hardly understandable from the standpoint of the mechanical model of the atom, since very general assumptions concerning the electron, using classical theory as well as quantum theory, always led to the same triplet. A closer investigation of this problem left me with the feeling that it was even more unapproachable. We know now that at that time one was confronted with two logically different difficulties simultaneously. One was the absence of a general key to translate a given mechanical model into quantum theory which one tried in vain by using classical mechanics to describe the stationary quantum states themselves. The second difficulty was our ignorance concerning the proper classical model itself which could be suited to derive at all an anomalous splitting of spectral lines emitted by an atom in an external magnetic field. It is therefore not surprising that I could not find a satisfactory solution of the problem at that time. I succeeded, however, in generalizing Landé's term analysis for very strong magnetic fields², a case which, as a result of the magneto-optic transformation (Paschen-Back effect), is in many respects simpler. This early work

was of decisive importance for the finding of the exclusion principle.

Very soon after my return to the University of Hamburg, in 1923, I gave there my inaugural lecture as *Privatdozent* on the Periodic System of Elements. The contents of this lecture appeared very unsatisfactory to me, since the problem of the closing of the electronic shells had been clarified no further. The only thing that was clear was that a closer relation of this problem to the theory of multiplet structure must exist. I therefore tried to examine again critically the simplest case, the doublet structure of the alkali spectra. According to the point of view then orthodox, which was also taken over by Bohr in his already mentioned lectures in Göttingen, a non-vanishing angular momentum of the atomic core was supposed to be the cause of this doublet structure.

In the autumn of 1924 I published some arguments against this point of view, which I definitely rejected as incorrect and proposed instead of it the assumption of a new quantum theoretic property of the electron, which I called a « two-valuedness not describable classically »³. At this time a paper of the English physicist, Stoner, appeared⁴ which contained, besides improvements in the classification of electrons in subgroups, the following essential remark: For a given value of the principal quantum number is the number of energy levels of a single electron in the alkali metal spectra in an external magnetic field the same as the number of electrons in the closed shell of the rare gases which corresponds to this principal quantum number.

On the basis of my earlier results on the classification of spectral terms in a strong magnetic field the general formulation of the exclusion principle became clear to me. The fundamental idea can be stated in the following way: The complicated numbers of electrons in closed subgroups are reduced to the simple number *one* if the division of the groups by giving the values of the four quantum numbers of an electron is carried so far that every degeneracy is removed. An entirely non-degenerate energy level is already « closed », if it is occupied by a single electron; states in contradiction with this postulate have to be excluded. The exposition of this general formulation of the exclusion principle was made in Hamburg in the spring of 1925⁵, after I was able to verify some additional conclusions concerning the anomalous Zeeman effect of more complicated atoms during a visit to Tübingen with the help of the spectroscopic material assembled there.

With the exception of experts on the classification of spectral terms, the physicists found it difficult to understand the exclusion principle, since no meaning in terms of a model was given to the fourth degree of freedom of

the electron. The gap was filled by Uhlenbeck and Goudsmit's idea of electron spin⁶, which made it possible to understand the anomalous Zeeman effect simply by assuming that the spin quantum number of one electron is equal to $\frac{1}{2}$ and that the quotient of the magnetic moment to the mechanical angular moment has for the spin a value twice as large as for the ordinary orbit of the electron. Since that time, the exclusion principle has been closely connected with the idea of spin. Although at first I strongly doubted the correctness of this idea because of its classical-mechanical character, I was finally converted to it by Thomas' calculations⁷ on the magnitude of doublet splitting. On the other hand, my earlier doubts as well as the cautious expression « classically non-describable two-valuedness » experienced a certain verification during later developments, since Bohr was able to show on the basis of wave mechanics that the electron spin cannot be measured by classically describable experiments (as, for instance, deflection of molecular beams in external electromagnetic fields) and must therefore be considered as an essentially quantum-mechanical property of the electron^{8,9}.

The subsequent developments were determined by the occurrence of the new quantum mechanics. In 1925, the same year in which I published my paper on the exclusion principle, De Broglie formulated his idea of matter waves and Heisenberg the new matrix-mechanics, after which in the next year Schrödinger's wave mechanics quickly followed. It is at present unnecessary to stress the importance and the fundamental character of these discoveries, all the more as these physicists have themselves explained, here in Stockholm, the meaning of their leading ideas¹⁰. Nor does time permit me to illustrate in detail the general epistemological significance of the new discipline of quantum mechanics, which has been done, among others, in a number of articles by Bohr, using hereby the idea of « complementarity » as a new central concept¹¹. I shall only recall that the statements of quantum mechanics are dealing only with possibilities, not with actualities. They have the form « This is not possible » or « Either this or that is possible », but they can never say « That will actually happen then and there ». The actual observation appears as an event outside the range of a description by physical laws and brings forth in general a discontinuous selection out of the several possibilities foreseen by the statistical laws of the new theory. Only this renouncement concerning the old claims for an objective description of the physical phenomena, independent of the way in which they are observed, made it possible to reach again the self-consistency of quantum theory, which ac-

tually had been lost since Planck's discovery of the quantum of action. Without discussing further the change of the attitude of modern physics to such concepts as « causality » and « physical reality » in comparison with the older classical physics I shall discuss more particularly in the following the position of the exclusion principle on the new quantum mechanics.

As it was first shown by Heisenberg¹², wave mechanics leads to qualitatively different conclusions for particles of the same kind (for instance for electrons) than for particles of different kinds. As a consequence of the impossibility to distinguish one of several like particles from the other, the wave functions describing an ensemble of a given number of like particles in the configuration space are sharply separated into different classes of symmetry which can never be transformed into each other by external perturbations. In the term « configuration space » we are including here the spin degree of freedom, which is described in the wave function of a single particle by an index with only a finite number of possible values. For electrons this number is equal to two; the configuration space of N electrons has therefore $3 N$ space dimensions and N indices of « two-valuedness ». Among the different classes of symmetry, the most important ones (which moreover for two particles are the only ones) are the symmetrical class, in which the wave function does not change its value when the space and spin coordinates of two particles are permuted, and the antisymmetrical class, in which for such a permutation the wave function changes its sign. At this stage of the theory three different hypotheses turned out to be logically possible concerning the actual ensemble of several like particles in Nature.

- I. This ensemble is a mixture of all symmetry classes.
- II. Only the symmetrical class occurs.
- III. Only the antisymmetrical class occurs.

As we shall see, the first assumption is never realized in Nature. Moreover, it is only the third assumption that is in accordance with the exclusion principle, since an antisymmetrical function containing two particles in the same state is identically zero. The assumption III can therefore be considered as the correct and general wave mechanical formulation of the exclusion principle. It is this possibility which actually holds for electrons.

This situation appeared to me as disappointing in an important respect. Already in my original paper I stressed the circumstance that I was unable to give a logical reason for the exclusion principle or to deduce it from more

general assumptions. I had always the feeling and I still have it today, that this is a deficiency. Of course in the beginning I hoped that the new quantum mechanics, with the help of which it was possible to deduce so many half-empirical formal rules in use at that time, will also rigorously deduce the exclusion principle. Instead of it there was for electrons still an exclusion: not of particular states any longer, but of whole classes of states, namely the exclusion of all classes different from the antisymmetrical one. The impression that the shadow of some incompleteness fell here on the bright light of success of the new quantum mechanics seems to me unavoidable. We shall resume this problem when we discuss relativistic quantum mechanics but wish to give first an account of further results of the application of wave mechanics to systems of several like particles.

In the paper of Heisenberg, which we are discussing, he was also able to give a simple explanation of the existence of the two non-combining spectra of helium which I mentioned in the beginning of this lecture. Indeed, besides the rigorous separation of the wave functions into symmetry classes with respect to space-coordinates and spin indices together, there exists an approximate separation into symmetry classes with respect to space coordinates alone. The latter holds only so long as an interaction between the spin and the orbital motion of the electron can be neglected. In this way the para- and ortho-helium spectra could be interpreted as belonging to the class of symmetrical and antisymmetrical wave functions respectively in the space coordinates alone. It became clear that the energy difference between corresponding levels of the two classes has nothing to do with magnetic interactions but is of a new type of much larger order of magnitude, which one called exchange energy.

Of more fundamental significance is the connection of the symmetry classes with general problems of the statistical theory of heat. As is well known, this theory leads to the result that the entropy of a system is (apart from a constant factor) given by the logarithm of the number of quantum states of the whole system on a so-called energy shell. One might first expect that this number should be equal to the corresponding volume of the multi-dimensional phase space divided by h^f , where h is Planck's constant and f the number of degrees of freedom of the whole system. However, it turned out that for a system of N like particles, one had still to divide this quotient by $N!$ in order to get a value for the entropy in accordance with the usual postulate of homogeneity that the entropy has to be proportional to the mass for a given inner state of the substance. In this way a qualitative distinction between

like and unlike particles was already preconceived in the general statistical mechanics, a distinction which Gibbs tried to express with his concepts of a generic and a specific phase. In the light of the result of wave mechanics concerning the symmetry classes, this division by $N!$, which had caused already much discussion, can easily be interpreted by accepting one of our assumptions II and III, according to both of which only one class of symmetry occurs in Nature. The density of quantum states of the whole system then really becomes smaller by a factor $N!$ in comparison with the density which had to be expected according to an assumption of the type I admitting all symmetry classes.

Even for an ideal gas, in which the interaction energy between molecules can be neglected, deviations from the ordinary equation of state have to be expected for the reason that only one class of symmetry is possible as soon as the mean De Broglie wavelength of a gas molecule becomes of an order of magnitude comparable with the average distance between two molecules, that is, for small temperatures and large densities. For the antisymmetrical class the statistical consequences have been derived by Fermi and Dirac¹³, for the symmetrical class the same had been done already before the discovery of the new quantum mechanics by Einstein and Bose¹⁴. The former case could be applied to the electrons in a metal and could be used for the interpretation of magnetic and other properties of metals.

As soon as the symmetry classes for electrons were cleared, the question arose which are the symmetry classes for other particles. One example for particles with symmetrical wave functions only (assumption II) was already known long ago, namely the photons. This is not only an immediate consequence of Planck's derivation of the spectral distribution of the radiation energy in the thermodynamical equilibrium, but it is also necessary for the applicability of the classical field concepts to light waves in the limit where a large and not accurately fixed number of photons is present in a single quantum state. We note that the symmetrical class for photons occurs together with the integer value I for their spin, while the antisymmetrical class for the electron occurs together with the half-integer value $\frac{1}{2}$ for the spin.

The important question of the symmetry classes for nuclei, however, had still to be investigated. Of course the symmetry class refers here also to the permutation of both the space coordinates and the spin indices of two like nuclei. The spin index can assume $2I + 1$ values if I is the spin-quantum number of the nucleus which can be either an integer or a half-integer. I may include the historical remark that already in 1924, before the electron spin

was discovered, I proposed to use the assumption of a nuclear spin to interpret the hyperfine-structure of spectral lines¹⁵. This proposal met on the one hand strong opposition from many sides but influenced on the other hand Goudsmit and Uhlenbeck in their claim of an electron spin. It was only some years later that my attempt to interpret the hyperfine-structure could be definitely confirmed experimentally by investigations in which also Zeeman himself participated and which showed the existence of a magneto-optic transformation of the hyperfine-structure as I had predicted it. Since that time the hyperfine-structure of spectral lines became a general method of determining the nuclear spin.

In order to determine experimentally also the symmetry class of the nuclei, other methods were necessary. The most convenient, although not the only one, consists in the investigation of band spectra due to a molecule with two like atoms¹⁶. It could easily be derived that in the ground state of the electron configuration of such a molecule the states with even and odd values of the rotational quantum number are symmetric and antisymmetric respectively for a permutation of the space coordinates of the two nuclei. Further there exist among the $(2I+1)^2$ spin states of the pair of nuclei, $(2I+1)(I+1)$ states symmetrical and $(2I+1)I$ states antisymmetrical in the spins, since the $(2I+1)$ states with two spins in the same direction are necessarily symmetrical. Therefore the conclusion was reached: If the total wave function of space coordinates and spin indices of the nuclei is symmetrical, the ratio of the weight of states with an even rotational quantum number to the weight of states with an odd rotational quantum number is given by $(I+1) : I$. In the reverse case of an antisymmetrical total wave function of the nuclei, the same ratio is $I : (I+1)$. Transitions between one state with an even and another state with an odd rotational quantum number will be extremely rare as they can only be caused by an interaction between the orbital motions and the spins of the nuclei. Therefore the ratio of the weights of the rotational states with different parity will give rise to two different systems of band spectra with different intensities, the lines of which are alternating.

The first application of this method was the result that the protons have the spin $\frac{1}{2}$ and fulfill the exclusion principle just as the electrons. The initial difficulties to understand quantitatively the specific heat of hydrogen molecules at low temperatures were removed by Dennison's hypothesis¹⁷, that at this low temperature the thermal equilibrium between the two modifications of the hydrogen molecule (ortho-H₂: odd rotational quantum numbers,

parallel proton spins; para-H₂: even rotational quantum numbers, antiparallel spins) was not yet reached. As you know, this hypothesis was later, confirmed by the experiments of Bonhoeffer and Harteck and of Eucken, which showed the theoretically predicted slow transformation of one modification into the other.

Among the symmetry classes for other nuclei those with a different parity of their mass number M and their charge number Z are of a particular interest. If we consider a compound system consisting of numbers A_1, A_2, \dots of different constituents, each of which is fulfilling the exclusion principle, and a number S of constituents with symmetrical states, one has to expect symmetrical or antisymmetrical states if the sum $A_1 + A_2 + \dots$ is even or odd. This holds regardless of the parity of S . Earlier one tried the assumption that nuclei consist of protons and electrons, so that M is the number of protons, $M - Z$ the number of electrons in the nucleus. It had to be expected then that the parity of Z determines the symmetry class of the whole nucleus. Already for some time the counter-example of nitrogen has been known to have the spin 1 and symmetrical states¹⁸. After the discovery of the neutron, the nuclei have been considered, however, as composed of protons and neutrons in such a way that a nucleus with mass number M and charge number Z should consist of Z protons and $M - Z$ neutrons. In case the neutrons would have symmetrical states, one should again expect that the parity of the charge number Z determines the symmetry class of the nuclei. If, however, the neutrons fulfill the exclusion principle, it has to be expected that the parity of M determines the symmetry class : For an even M , one should always have symmetrical states, for an odd M , antisymmetrical ones. It was the latter rule that was confirmed by experiment without exception, thus proving that the neutrons fulfill the exclusion principle.

The most important and most simple crucial example for a nucleus with a different parity of M and Z is the heavy hydrogen or deuteron with $M = 2$ and $Z = 1$ which has symmetrical states and the spin $I = 1$, as could be proved by the investigation of the band spectra of a molecule with two deuterons¹⁹. From the spin value 1 of the deuteron can be concluded that the neutron must have a half-integer spin. The simplest possible assumption that this spin of the neutron is equal to $\frac{1}{2}$, just as the spin of the proton and of the electron, turned out to be correct.

There is hope, that further experiments with light nuclei, especially with protons, neutrons, and deuterons will give us further information about the nature of the forces between the constituents of the nuclei, which, at present,

is not yet sufficiently clear. Already now we can say, however, that these interactions are fundamentally different from electromagnetic interactions. The comparison between neutron-proton scattering and proton-proton scattering even showed that the forces between these particles are in good approximation the same, that means independent of their electric charge. If one had only to take into account the magnitude of the interaction energy, one should therefore expect a stable di-proton or ^2He ($M = 2, Z = 2$) with nearly the same binding energy as the deuteron. Such a state is, however, forbidden by the exclusion principle in accordance with experience, because this state would acquire a wave function symmetric with respect to the two protons. This is only the simplest example of the application of the exclusion principle to the structure of compound nuclei, for the understanding of which this principle is indispensable, because the constituents of these heavier nuclei, the protons and the neutrons, fulfil it.

In order to prepare for the discussion of more fundamental questions, we want to stress here a law of Nature which is generally valid, namely, the connection between spin and symmetry class. *A half-integer value of the spin quantum number is always connected with antisymmetrical states (exclusion principle), an integer spin with symmetrical states.* This law holds not only for protons and neutrons but also for protons and electrons. Moreover, it can easily be seen that it holds for compound systems, if it holds for all of its constituents. If we search for a theoretical explanation of this law, we must pass to the discussion of relativistic wave mechanics, since we saw that it can certainly not be explained by non-relativistic wave mechanics.

We first consider classical fields²⁰, which, like scalars, vectors, and tensors transform with respect to rotations in the ordinary space according to a one-valued representation of the rotation group. We may, in the following, call such fields briefly « one-valued » fields. So long as interactions of different kinds of field are not taken into account, we can assume that all field components will satisfy a second-order wave equation, permitting a superposition of plane waves as a general solution. Frequency and wave number of these plane waves are connected by a law which, in accordance with De Broglie's fundamental assumption, can be obtained from the relation between energy and momentum of a particle claimed in relativistic mechanics by division with the constant factor equal to Planck's constant divided by 2π . Therefore, there will appear in the classical field equations, in general, a new constant μ with the dimension of a reciprocal length, with which the

rest-mass m in the particle picture is connected by $m = h \mu/c$, where c is the vacuum-velocity of light. From the assumed property of one-valuedness of the field it can be concluded, that the number of possible plane waves for a given frequency, wave number and direction of propagation, is for a non-vanishing μ always odd. Without going into details of the general definition of spin, we can consider this property of the polarization of plane waves as characteristic for fields which, as a result of their quantization, give rise to integer spin values.

The simplest cases of one-valued fields are the scalar field and a field consisting of a four-vector and an antisymmetric tensor like the potentials and field strengths in Maxwell's theory. While the scalar field is simply fulfilling the usual wave equation of the second order in which the term proportional to μ^2 has to be included, the other field has to fulfill equations due to Proca which are a generalization of Maxwell's equations which become in the particular case $\mu = 0$. It is satisfactory that for these simplest cases of one-valued fields the energy density is a positive definite quadratic form of the field-quantities and their first derivatives at a certain point. For the general case of one-valued fields it can at least be achieved that the total energy after integration over space is always positive.

The field components can be assumed to be either real or complex. For a complex field, in addition to energy and momentum of the field, a four-vector can be defined which satisfies the continuity equation and can be interpreted as the four-vector of the electric current. Its fourth component determines the electric charge density and can assume both positive and negative values. It is possible that the charged mesons observed in cosmic rays have integral spins and thus can be described by such a complex field. In the particular case of real fields this four-vector of current vanishes identically.

Especially in view of the properties of the radiation in the thermodynamical equilibrium in which specific properties of the field sources do not play any role, it seemed to be justified first to disregard in the formal process of field quantization the interaction of the field with the sources. Dealing with this problem, one tried indeed to apply the same mathematical method of passing from a classical system to a corresponding system governed by the laws of quantum mechanics which has been so successful in passing from classical point mechanics to wave mechanics. It should not be forgotten, however, that a field can only be observed with help of its interaction with test bodies which are themselves again sources of the field.

The result of the formal process of field quantization were partly very

encouraging. The quantized wave fields can be characterized by a wave function which depends on an infinite sequence of (non-negative) integers as variables. As the total energy and the total momentum of the field and, in case of complex fields, also its total electric charge turn out to be linear functions of these numbers, they can be interpreted as the number of particles present in a specified state of a single particle. By using a sequence of configuration spaces with a different number of dimensions corresponding to the different possible values of the total number of particles present, it could easily be shown that this description of our system by a wave function depending on integers is equivalent to an ensemble of particles with wave functions symmetrical in their configuration spaces.

Moreover Bohr and Rosenfeld²¹ proved in the case of the electromagnetic field that the uncertainty relations which result for the average values of the field strengths over finite space-time regions from the formal commutation rules of this theory have a direct physical meaning so long as the sources can be treated classically and their atomistic structure can be disregarded. We emphasize the following property of these commutation rules: All physical quantities in two world points, for which the four-vector of their joining straight line is spacelike commute with each other. This is indeed necessary for physical reasons because any disturbance by measurements in a world point P_1 , can only reach such points P_2 , for which the vector P_1P_2 , is timelike, that is, for which $c(t_1 - t_2) > r_{12}$. The points P_2 with a spacelike vector P_1P_2 , for which $c(t_1 - t_2) < r_{12}$ cannot be reached by this disturbance and measurements in P_1 and P_2 can then never influence each other.

This consequence made it possible to investigate the logical possibility of particles with integer spin which would obey the exclusion principle. Such particles could be described by a sequence of configuration spaces with different dimensions and wave functions antisymmetrical in the coordinates of these spaces or also by a wave function depending on integers again to be interpreted as the number of particles present in specified states which now can only assume the values 0 or 1. Wigner and Jordan²² proved that also in this case operators can be defined which are functions of the ordinary space-time coordinates and which can be applied to such a wave function. These operators do not fulfil any longer commutation rules: instead of the difference, the sum of the two possible products of two operators, which are distinguished by the different order of its factors, is now fixed by the mathematical conditions the operators have to satisfy. The simple change of the sign in these conditions changes entirely the physical meaning of the for-

malism. In the case of the exclusion principle there can never exist a limiting case where such operators can be replaced by a classical field. Using this formalism of Wigner and Jordan I could prove under very general assumptions that a relativistic invariant theory describing systems of like particles with integer spin obeying the exclusion principle would always lead to the non-commutability of physical quantities joined by a spacelike vector²³. This would violate a reasonable physical principle which holds good for particles with symmetrical states. In this way, by combination of the claims of relativistic invariance and the properties of field quantization, one step in the direction of an understanding of the connection of spin and symmetry class could be made.

The quantization of one-valued complex fields with a non-vanishing four-vector of the electric current gives the further result that particles both with positive and negative electric charge should exist and that they can be annihilated and generated in external electromagnetic field²². This pair-generation and annihilation claimed by the theory makes it necessary to distinguish clearly the concept of charge density and of particle density. The latter concept does not occur in a relativistic wave theory either for fields carrying an electric charge or for neutral fields. This is satisfactory since the use of the particle picture and the uncertainty relations (for instance by analyzing imaginative experiments of the type of the γ -ray microscope) gives also the result that a localization of the particle is only possible with limited accuracy²⁴. This holds both for the particles with integer and with half-integer spins. In a state with a mean value E of its energy, described by a wave packet with a mean frequency $v = E/h$, a particle can only be localized with an error $\Delta x > hc/E$ or $\Delta x > c/v$. For photons, it follows that the limit for the localization is the wavelength; for a particle with a finite rest-mass m and a characteristic length $\mu^{-1} = \hbar/mc$, this limit is in the rest system of the center of the wave packet that describes the state of the particles given by $\Delta x > \hbar/mc$ or $\Delta x > \mu^{-1}$.

Until now I have mentioned only those results of the application of quantum mechanics to classical fields which are satisfactory. We saw that the statements of this theory about averages of field strength over finite space-time regions have a direct meaning while this is not so for the values of the field strength at a certain point. Unfortunately in the classical expression of the energy of the field there enter averages of the squares of the field strengths over such regions which cannot be expressed by the averages of the field strengths themselves. This has the consequence that the zero-point energy

of the vacuum derived from the quantized field becomes infinite, a result which is directly connected with the fact that the system considered has an infinite number of degrees of freedom. It is clear that this zero-point energy has no physical reality, for instance it is not the source of a gravitational field. Formally it is easy to subtract constant infinite terms which are independent of the state considered and never change; nevertheless it seems to me that already this result is an indication that a fundamental change in the concepts underlying the present theory of quantized fields will be necessary.

In order to clarify certain aspects of relativistic quantum theory I have discussed here, different from the historical order of events, the one-valued fields first. Already earlier Dirac²⁵ had formulated his relativistic wave equations corresponding to material particles with spin $\frac{1}{2}$ using a pair of so-called spinors with two components each. He applied these equations to the problem of one electron in an electromagnetic field. In spite of the great success of this theory in the quantitative explanation of the fine structure of the energy levels of the hydrogen atom and in the computation of the scattering cross section of one photon by a free electron, there was one consequence of this theory which was obviously in contradiction with experience. The energy of the electron can have, according to the theory, both positive and negative values, and, in external electromagnetic fields, transitions should occur from states with one sign of energy to states with the other sign. On the other hand there exists in this theory a four-vector satisfying the continuity equation with a fourth component corresponding to a density which is definitely positive.

It can be shown that there is a similar situation for all fields, which, like the spinors, transform for rotations in ordinary space according to two-valued representations, thus changing their sign for a full rotation. We shall call briefly such quantities « two-valued ». From the relativistic wave equations of such quantities one can always derive a four-vector bilinear in the field components which satisfies the continuity equation and for which the fourth component, at least after integration over the space, gives an essentially positive quantity. On the other hand, the expression for the total energy can have both the positive and the negative sign.

Is there any means to shift the minus sign from the energy back to the density of the four-vector? Then the latter could again be interpreted as charge density in contrast to particle density and the energy would become positive as it ought to be. You know that Dirac's answer was that this could actually be achieved by application of the exclusion principle. In his lecture

delivered here in Stockholm¹⁰ he himself explained his proposal of a new interpretation of his theory, according to which in the actual vacuum all the states of negative energy should be occupied and only deviations of this state of smallest energy, namely holes in the sea of these occupied states are assumed to be observable. It is the exclusion principle which guarantees the stability of the vacuum, in which all states of negative energy are occupied. Furthermore the holes have all properties of particles with positive energy and positive electric charge, which in external electromagnetic fields can be produced and annihilated in pairs. These predicted positrons, the exact mirror images of the electrons, have been actually discovered experimentally.

The new interpretation of the theory obviously abandons in principle the standpoint of the one-body problem and considers a many-body problem from the beginning. It cannot any longer be claimed that Dirac's relativistic wave equations are the only possible ones but if one wants to have relativistic field equations corresponding to particles, for which the value $\frac{1}{2}$ of their spin is known, one has certainly to assume the Dirac equations. Although it is logically possible to quantize these equations like classical fields, which would give symmetrical states of a system consisting of many such particles, this would be in contradiction with the postulate that the energy of the system has actually to be positive. This postulate is fulfilled on the other hand if we apply the exclusion principle and Dirac's interpretation of the vacuum and the holes, which at the same time substitutes the physical concept of charge density with values of both signs for the mathematical fiction of a positive particle density. A similar conclusion holds for all relativistic wave equations with two-valued quantities as field components. This is the other step (historically the earlier one) in the direction of an understanding of the connection between spin and symmetry class.

I can only shortly note that Dirac's new interpretation of empty and occupied states of negative energy can be formulated very elegantly with the help of the formalism of Jordan and Wigner mentioned before. The transition from the old to the new interpretation of the theory can indeed be carried through simply by interchanging the meaning of one of the operators with that of its hermitian conjugate if they are applied to states originally of negative energy. The infinite « zero charge » of the occupied states of negative energy is then formally analogous to the infinite zero-point energy of the quantized one-valued fields. The former has no physical reality either and is not the source of an electromagnetic field.

In spite of the formal analogy between the quantization of the one-valued fields leading to ensembles of like particles with symmetrical states and to particles fulfilling the exclusion principle described by two-valued operator quantities, depending on space and time coordinates, there is of course the fundamental difference that for the latter there is no limiting case, where the mathematical operators can be treated like classical fields. On the other hand we can expect that the possibilities and the limitations for the applications of the concepts of space and time, which find their expression in the different concepts of charge density and particle density, will be the same for charged particles with integer and with half-integer spins.

The difficulties of the present theory become much worse, if the interaction of the electromagnetic field with matter is taken into consideration, since the well-known infinities regarding the energy of an electron in its own field, the so-called self-energy, then occur as a result of the application of the usual perturbation formalism to this problem. The root of this difficulty seems to be the circumstance that the formalism of field quantization has only a direct meaning so long as the sources of the field can be treated as continuously distributed, obeying the laws of classical physics, and so long as only averages of field quantities over finite space-time regions are used. The electrons themselves, however, are essentially non-classical field sources.

At the end of this lecture I may express my critical opinion, that a correct theory should neither lead to infinite zero-point energies nor to infinite zero charges, that it should not use mathematical tricks to subtract infinities or singularities, nor should it invent a « hypothetical world » which is only a mathematical fiction before it is able to formulate the correct interpretation of the actual world of physics.

From the point of view of logic, my report on « Exclusion principle and quantum mechanics » has no conclusion. I believe that it will only be possible to write the conclusion if a theory will be established which will determine the value of the fine-structure constant and will thus explain the atomistic structure of electricity, which is such an essential quality of all atomic sources of electric fields actually occurring in Nature.

1. A. Landé, *Z. Physik*, 5 (1921) 231 and *Z. Physik*, 7 (1921) 398, *Physik. Z.*, 22 (1921) 417.
2. W. Pauli, *Z. Physik*, 16 (1923) 155.
3. W. Pauli, *Z. Physik*, 31 (1925) 373.
4. E. C. Stoner, *Phil. Mag.*, 48 (1924) 719.
5. W. Pauli, *Z. Physik*, 31 (1925) 765.
6. S. Goudsmit and G. Uhlenbeck, *Naturwiss.*, 13 (1925) 953, *Nature*, 117 (1926) 264.
7. L. H. Thomas, *Nature*, 117 (1926) 514, and *Phil. Mag.*, 3 (1927) 1. Compare also J. Frenkel, *Z. Physik*, 37 (1926) 243.
8. Compare *Rapport du Sixième Conseil Solvay de Physique, Paris*, 1932, pp. 217-225.
9. For this earlier stage of the history of the exclusion principle compare also the author's note in *Science*, 103 (1946) 213, which partly coincides with the first part of the present lecture.
10. The Nobel Lectures of W. Heisenberg, E. Schrödinger, and P. A. M. Dirac are collected in *Die moderne Atomtheorie*, Leipzig, 1934.
11. The articles of N. Bohr are collected in *Atomic Theory and the Description of Nature*, Cambridge University Press, 1934. See also his article « Light and Life », *Nature*, 131 (1933) 421, 457.
12. W. Heisenberg, *Z. Physik*, 38 (1926) 411 and 39 (1926) 499.
13. E. Fermi, *Z. Physik*, 36 (1926) 902.
P. A. M. Dirac, *Proc. Roy. Soc. London*, A 112 (1926) 661.
14. S. N. Bose, *Z. Physik*, 26 (1924) 178 and 27 (1924) 384.
A. Einstein, *Berl. Ber.*, (1924) 261; (1925) 1, 18.
15. W. Pauli, *Naturwiss.*, 12 (1924) 741.
16. W. Heisenberg, *Z. Physik*, 41 (1927) 239, F. Hund, *Z. Physik*, 42 (1927) 39.
17. D. M. Dennison, *Proc. Roy. Soc. London*, A 115 (1927) 483.
18. R. de L. Kronig, *Naturwiss.*, 16 (1928) 335.
W. Heitler und G. Herzberg, *Naturwiss.*, 17 (1929) 673.
19. G. N. Lewis and M. F. Ashley, *Phys. Rev.*, 43 (1933) 837.
G. M. Murphy and H. Johnston, *Phys. Rev.*, 45 (1934) 550 and 46 (1934) 95.
20. Compare for the following the author's report in *Rev. Mod. Phys.*, 13 (1941) 203, in which older literature is given. See also W. Pauli and V. Weisskopf, *Helv. Phys. Acta*, 7 (1934) 809.
21. N. Bohr and L. Rosenfeld, *Kgl. Danske Videnskab. Selskab. Mat. Fys. Medd.*, 12 [8] (1933).
22. P. Jordan and E. Wigner, *Z. Physik*, 47 (1928) 63I.
Compare also V. Fock, *Z. Physik*, 75 (1932) 622.
23. W. Pauli, *Ann. Inst. Poincaré*, 6 (1936) 137 and *Phys. Rev.*, 58 (1940) 716.
24. L. Landau and R. Peierls, *Z. Physik*, 69 (1931) 56.
Compare also the author's article in *Handbuch der Physik*, 24, Part 1, 1933, Chap. A, § 2.
25. P. A. M. Dirac, *Proc. Roy. Soc. London*, A 117 (1928) 610.

OTTO STERN

The method of molecular rays

Nobel Lecture, December 12, 1946

In the following lecture I shall try to analyze the method of molecular rays. My aim is to bring out its distinctive features, the points where it is different from other methods used in physics, for what kind of problems it is especially suited and why. Let me state from the beginning that I consider the directness and simplicity as the distinguishing properties of the molecular ray method. For this reason it is particularly well suited for shedding light on fundamental problems. I hope to make this clear by discussing the actual experiments.

Let us first consider the group of experiments which prove directly the fundamental assumptions of the kinetic theory. The existence of molecular rays in itself, the possibility of producing molecular rays, is a direct proof of one fundamental assumption of that theory. This assumption is that in gases the molecules move in straight lines until they collide with other molecules or the walls of the containing vessel. The usual arrangement for producing molecular rays is as follows (Fig. . 1): We have a vessel filled with gas or vapor, the oven. This vessel is closed except for a narrow slit, the oven slit. Through this slit the molecules escape into the surrounding larger vessel which is continually evacuated so that the escaping molecules do not suffer any collisions. Now we have another narrow slit, the collimating slit, opposite and parallel to the oven slit. If the molecules really move in straight lines then the collimating slit should cut out a narrow beam whose cross section by simple geometry can be calculated from the dimensions of the slits and their distance. That it is actually the case was proven first by Dunoyer in 1911. He used sodium vapor and condensed the beam molecules hitting the wall by cooling it with liquid air. The sodium deposit formed on the wall had exactly the shape calculated under the assumption that the molecules move in straight lines like rays of light. Therefore we call such a beam a «molecular ray» or «molecular beam».

The next step was the direct measurement of the velocity of the molecules. The kinetic theory gives quite definite numerical values for this velocity, depending on the temperature and the molecular weight. For example, for silver atoms of 1,000° the average velocity is about 600 m/sec (silver mole-

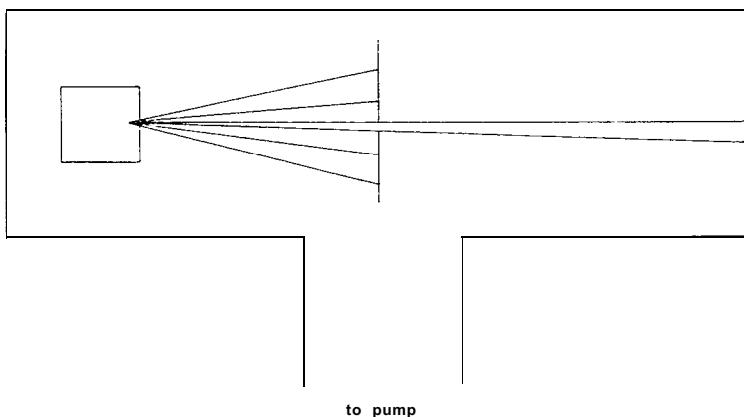


Fig. 1. Arrangement for producing molecular rays.

cules are monoatomic). We measured the velocity in different ways. One way - historically not the first one - was sending the molecular ray through a system of rotating tooth wheels, the method used by Fizeau to measure the velocity of light. We had two tooth wheels sitting on the same axis at a distance of several cm. When the wheels were at rest the molecular beam went through two corresponding gaps of the first and the second wheel. When the wheels rotated a molecule going through a gap in the first wheel could not go through the corresponding gap in the second wheel. The gap had moved during the time in which the molecule travelled from the first wheel to the second. However, under a certain condition the molecule could go through the next gap of the second wheel, the condition being that the travelling time for the molecule is just the time for the wheel to turn the distance between two neighboring gaps. By determining this time, that means the number of rotations per second for which the beam goes through both tooth wheels, we measure the velocity of the molecules. We found agreement with the theory with regard to the numerical values and to the velocity distribution according to Maxwell's law.

This method has the advantage of producing a beam of molecules with nearly uniform velocity. However, it is not very accurate.

As the last one in this group of experiments I want to report on experiments carried out in Pittsburgh by Drs. Estermann, Simpson, and myself before the War, which are now being published. In these experiments we used the free fall of molecules to measure their velocities.

In vacuo all bodies, large and small, fall equal distances in equal times, $s =$

$\frac{1}{2}gt^2$ (t = time, s = distance of fall, g = acceleration of gravity). We used a beam of cesium atoms about 2 m long. Since the average velocity of the atoms is about 200 m/sec the travel time is about 1/100 sec. During this time a body falls not quite a mm. So our cesium atoms did not travel exactly on the straight horizontal line through oven and collimating slit but arrived a little lower depending on their velocity. The fast ones fell less, the slow ones more. So by measuring the intensity (the number of cesium atoms arriving per second) at the end of the beam perpendicular to it as a function of the distance from the straight line, we get directly the distribution of velocities (Fig. 2). As you see the agreement with Maxwell's law is very good. I might mention that we measured the intensity not by condensation but by the so-called Taylor-Langmuir method worked out by Taylor in our Hamburg

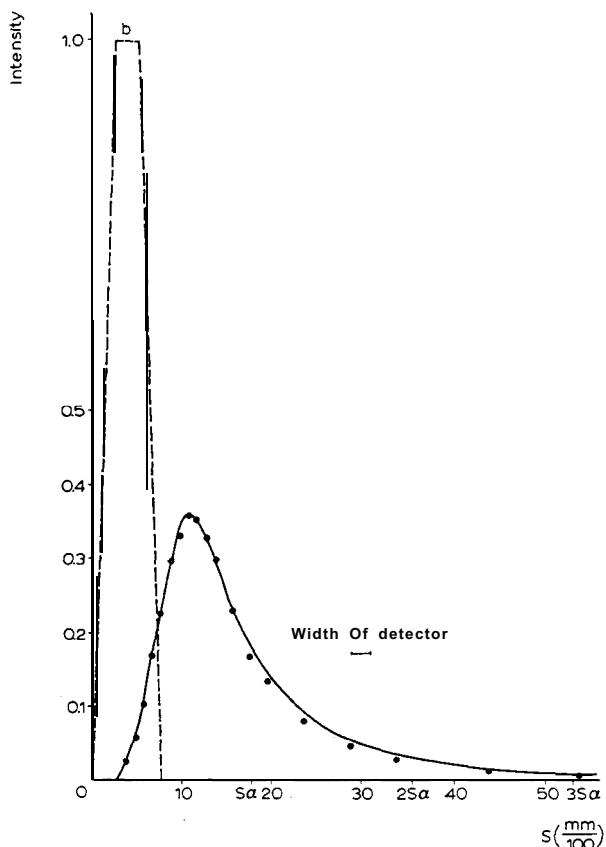


Fig. 2. Gravity deflection of a cesium beam. (Full line): calculated from Maxwell's law; (points): measurements; (pecked line *b*): undeflected beam.

laboratory in 1928. It is based on Langmuir's discovery that every alkali atom striking the surface of a hot tungsten wire (eventually oxygen-coated) goes away as an ion. By measuring the ion current outgoing from the wire we measured directly the number of atoms striking the wire.

What can we conclude about the method of molecular rays from the group of experiments we have considered so far? It gives us certainly a great satisfaction to demonstrate in such a simple direct manner the fundamentals of the kinetic theory. Furthermore, even if so many conclusions of the theory were checked by experiments that practically no reasonable doubt about the correctness of this part of the theory was possible, these experiments reinforced and strengthened the fundamentals beyond any doubt.

I said this part of the theory.

The classical theory is a grandiose conception. The same fundamental laws govern the movements of the stars, the fall of this piece of chalk, and the fall of molecules. But it turned out that the extrapolation to the molecules did not hold in some respects. The theory had to be changed in order to describe the laws governing the movements of the molecules and even more of the electrons. And it was at this point that the molecular ray method proved its value. Here the experiment did not just check the results of the theory on which there was practically no doubt anyway, but gave a decisive answer in cases where the theory was uncertain and even gave contradictory answers.

The best example is the experiment which Gerlach and I performed in 1922. It was known from spectroscopic experiments (Zeeman effect) that the atoms of hydrogen, the alkali metals, silver, and so on, were small magnets. Let us consider the case of the hydrogen atom as the simplest one even if our experiments were performed with silver atoms. There is no essential difference, and the results were checked with hydrogen atoms a few years later by one of my students in our Hamburg laboratory.

The essential point is that the classical theory and the quantum theory predict quite differently the behavior of the atomic magnets in a magnetic field. The classical theory predicts that the atomic magnets assume all possible directions with respect to the direction of the magnetic field. On the other hand, the quantum theory predicts that we shall find only two directions parallel and antiparallel to the field (new theory, the old one gave also the direction perpendicular to the field).

The contradiction I spoke of is this. At this time according to Bohr's theory one assumed that the magnetic moment of the hydrogen atom is produced by the movement of the electron around the nucleus in the same way

as a circular current in a wire is equivalent to a magnet. Then the statement of the quantum theory means that the electrons of all hydrogen atoms move in planes perpendicular to the direction of the magnetic field. In this case one should find optically a strong double refraction which was certainly not true. So there was a serious dilemma.

Our molecular ray experiment gave a definite answer. We sent a beam of silver atoms through a very inhomogeneous magnetic field. In such a field the magnets are deflected because the field strength on the place of one pole of the magnet is a little different from the field strength acting on the other pole. So in the balance a force is acting on the atom and it is deflected. A simple calculation shows that from the classical theory follows that we should find a broadening of the beam with the maximum intensity on the place of the beam without field. However, from the quantum theory follows that we should find there no intensity at all, and deflected molecules on both sides. The beam should split up in two beams corresponding to the two orientations of the magnet. The experiment decided in favor of the quantum theory (Fig. 3).



Fig. 3. Discrete beams of deflected molecules.

The contradiction with respect to the double refraction was solved about four years later through the new quantum mechanics in connection with the Uhlenbeck-Goudsmit hypothesis that the electron itself is a small magnet like a rotating charged sphere. But even before this explanation was given, the experiment verified directly the splitting in discrete beams as predicted by the quantum theory.

So again the directness stands out as characteristic for the molecular ray method. However, we can recognize another feature as essential in this experiment, namely that our measuring tool is a macroscopic one. I want to make this point clearer.

The first experiment which gave a direct proof of the fundamental hypoth-

esis of the quantum theory was the celebrated experiment of Franck and Hertz. These workers proved that the energy of one atom can be changed only by finite amounts. By bombarding mercury atoms with electrons they found that the electrons did lose energy only if their energy was higher than 4.7 eV. So they demonstrated directly that the energy of a mercury atom cannot be changed continuously but only by finite amounts, quanta of energy. As a tool for measuring the energy changes of the atom they used electrons, that means an atomic tool. In our experiment we used an electromagnet and slits, that means the same kind of tools we could use to measure the magnetic moment of an ordinary macroscopic magnet. Our experiment demonstrated in a special case a fact, which became fundamental for the new quantum mechanics, that the result of our measurements depends in a characteristic manner on the dimensions of the measured object and that quantum effects become perceptible when we make the object smaller and smaller.

We can see this better when we first consider a group of experiments which demonstrate the wave properties of rays of matter. In his famous theory which became the basis of the modern quantum theory, De Broglie stated that moving particles should also show properties of waves. The wavelength of these waves is given by the equation $\lambda = h/mv$ (h = Planck's constant; m = mass; v = velocity of the particle). The experimental proof was first given in 1927 by Davisson and Germer, and by Thomson for electrons. Some years later we succeeded in performing similar experiments with molecular rays of helium atoms and hydrogen molecules using the cleavage surface of a lithium fluoride crystal as diffraction grating. We could check the diffraction in every detail. The most convincing experiment is perhaps the one where we sent a beam of helium gas through the two rotating tooth wheels which I mentioned at the beginning, thus determining the velocity v in a primitive, purely mechanical, manner. The helium beam then impinged on the lithium fluoride crystal and by measuring the angle between the diffracted and the direct beam we determined the wavelength since we know the lattice constant of the lithium fluoride. We found agreement with De Broglie's formula within the accuracy of our experiments (about 2%). There is no doubt that these experiments could be carried out also by using a ruled grating instead of the crystal. In fact we found hints of a diffracted beam with a ruled grating already in 1928 and with the improved technique of today the experiment would not be too difficult.

With respect to the differences between the experiments with electrons and molecular rays, one can say that the molecular ray experiments go far-

ther. Also the mass of the moving particle is varied (He , H_2). But the main point is again that we work in such a direct primitive manner with neutral particles.

These experiments demonstrate clearly and directly the fundamental fact of the dual nature of rays of matter. It is no accident that in the development of the theory the molecular ray experiments played an important role. Not only the actual experiments were used, but also molecular ray experiments carried out only in thought. Bohr, Heisenberg, and Pauli used them in making clear their points on this direct simple example of an experiment. I want to mention only one consideration concerning the magnetic deflection experiment because it shows the fundamental limits of the experimental method.

First, it is clear that we cannot use too narrow slits, otherwise the diffraction on the slit will spread out the beam. This spreading out can roughly be described as the deflection of the molecules by an angle which is the larger the narrower the slit and the larger the De Broglie wavelength is. Therefore it causes a deflection of the molecules proportional to the distance which the molecule has traversed or to the length of the beam or to the time t since the molecule started from the collimating slit. The deflection by the magnetic force must be appreciably larger if we want to measure the magnetic moment. Fortunately this deflection is proportional to the square of the length of the beam or the time t , essentially as in the case of the gravity ($s = \frac{1}{2} gt^2$). Consequently it is always possible, by making the beam long enough, that means the time t large enough, to make the magnetic deflection larger than the deflection by diffraction. On the other hand it follows that a minimum time is necessary to measure the magnetic moment and this minimum time gets larger when the magnetic deflection, that means the magnetic moment, gets smaller. That is a special case of a general law of the new quantum mechanics. This law - applied to the measurement of moments - says that for *every* method using the same field strength the minimum time is the same. This circumstance was decisive in the group of experiments measuring the magnetic moment of the proton.

The theory predicted that the magnetic moments of electron and proton should be inversely proportional to the masses of those particles. Since the proton has about a two thousand times larger mass than the electron, its magnetic moment should be about two thousand times smaller. Such a small moment could not be measured by the spectroscopic method (Zeeman effect) but we (Frisch, Estermann, and myself) succeeded in measuring it by

using a beam of neutral hydrogen molecules. I do not have time to go into the details. The main point is that in measuring with molecular rays we use a longer time t . In the spectroscopic method this time is the lifetime of the excited atom which emits light by jumping into the normal state. Now this lifetime is generally of the order of magnitude of 10^{-8} sec. Working with molecular rays we use atoms (or molecules) in the normal state whose lifetime is infinite. The duration of our experiment is determined by the time t which the atom (or molecule) spends in the magnetic field. This time was of the order of magnitude of 10^{-4} sec (the length of the field about 10 cm and the velocity of the molecules about 1 km/sec). So our time is about 10,000 times larger and we can measure 10,000 times smaller moments with molecular rays than spectroscopically.

The result of our measurement was very interesting. The magnetic moment of the proton turned out to be about $2\frac{1}{2}$ times larger than the theory predicted. Since the proton is a fundamental particle - all nuclei are built up from protons and neutrons - this result is of great importance. Up to now the theory is not able to explain the result quantitatively.

It might seem now that the great sensitivity as shown in the last experiment is also a distinctive and characteristic property of the molecular ray method. However, that is not the case. The reason for the sensitivity as we have seen is that we make our measurements on atoms in the normal state. But of course many of the other experimental methods do that also.

We can see the situation clearly by considering the last achievement of the molecular ray method, the application of the resonance method by Rabi.

With the deflection method it is difficult to measure the moment to better than several per cent, mainly because of the difficulty of measuring the inhomogeneity in such small dimensions. With the resonance method, Rabi's accuracy is better than 1%, practically the theoretical limit given by the duration of about 10^{-4} sec of the measurement. Theoretically it would be possible to increase the accuracy simply by making this time longer. But that would mean making the beam longer and for practical reasons we cannot go much farther in this direction. In this connection it is significant that perhaps the best new measurements of the magnetic moments of the proton, neutron, and deuteron were made with the resonance method, however not using molecular rays but just liquid water with practically no limit for the duration of the measurement. So the sensitivity cannot be considered as a distinguishing property of the molecular ray method. However, that we have

such clear-cut simple conditions was the reason for applying the ultrasensitive resonance method first to molecular rays.

In conclusion I would like to summarize as follows: The most distinctive characteristic property of the molecular ray method is its simplicity and directness. It enables us to make measurements on isolated neutral atoms or molecules with macroscopic tools. For this reason it is especially valuable for testing and demonstrating directly fundamental assumptions of the theory.

E UGENE P. WIGNER

Events, laws of nature, and invariance principles

Nobel Lecture, December 12, 1963

It is a great and unexpected honor to have the opportunity to speak here today. Six years ago, Yang and Lee spoke here, reviewing symmetry principles in general and their discovery of the violation of the parity principle in particular¹. There is little point in repeating what they said, on the history of the invariance principles, or on my own contribution to these which they, naturally, exaggerated. What I would like to discuss instead is the general role of symmetry and invariance principles in physics, both modern and classical. More precisely, I would like to discuss the relation between three categories which play a fundamental role in all natural sciences: events, which are the raw materials for the second category, the laws of nature, and symmetry principles for which I would like to support the thesis that the laws of nature form the raw material.

Events and Laws of Nature

It is often said that the objective of physics is the explanation of nature, or at least of inanimate nature. What do we mean by explanation? It is the establishment of a few simple principles which describe the properties of what is to be explained. If we understand something, its behavior, that is the events which it presents, should not produce any surprises for us. We should always have the impression that it could not be otherwise.

It is clear that, in this sense, physics does not endeavor to explain nature. In fact, the great success of physics is due to a restriction of its objectives: it only endeavors to explain the *regularities* in the behavior of objects. This renunciation of the broader aim, and the specification of the domain for which an explanation can be sought, now appears to us an obvious necessity. In fact, the specification of the explainable may have been the greatest discovery of physics so far. It does not seem easy to find its inventor, or to give the exact date of its origin. Kepler still tried to find exact rules for the magnitude of the planetary orbits, similar to his laws of planetary motion. Newton already

realized that physics would deal, for a long time, only with the explanation of those of the regularities discovered by Kepler which we now call Kepler's laws².

The regularities in the phenomena which physical science endeavors to uncover are called the laws of nature. The name is actually very appropriate. Just as legal laws regulate actions and behavior under certain conditions, but do not try to regulate all actions and behavior, the laws of physics also determine the behavior of its objects of interest only under certain well defined conditions, but leave much freedom otherwise. The elements of the behavior which are not specified by the laws of nature are called initial conditions. These, then, together with the laws of nature, specify the behavior as far as it can be specified at all: if a further specification were possible, this specification would be considered as an added initial condition. As is well known, before the advent of quantum theory, it was believed that a complete description of the behavior of an object is possible so that, if classical theory were valid, the initial conditions and the laws of nature together would completely determine the behavior of an object.

The preceding statement is a definition of the term <<initial condition>>. Because of its somewhat unusual nature, it may be worthwhile to illustrate this on an example. Suppose we did not know Newton's equation for the motion of stars and planets

$$\ddot{\mathbf{r}}_i = G \Sigma' M_j \frac{\mathbf{r}_{ij}}{r_{ij}^3} \quad \mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i \quad (1)$$

but had found only the equation determining the third derivative of the position

$$\dddot{\mathbf{r}}_i = G \Sigma' M_j \frac{\dot{\mathbf{r}}_{ij}(\mathbf{r}_{ij} \cdot \dot{\mathbf{r}}_{ij}) - 3 \mathbf{r}_{ij}(\ddot{\mathbf{r}}_{ij} \cdot \mathbf{r}_{ij})}{r_{ij}^5} \quad (2)$$

More generally, if the forces F_i are non-gravitational, one would have written

$$M_i \ddot{\mathbf{r}}_i = \dot{\mathbf{r}}_i \cdot \text{grad}F_i + \dot{F}_i \quad (2a)$$

The initial conditions then would contain not only all the \mathbf{r}_i and $\dot{\mathbf{r}}_i$, but also the $\ddot{\mathbf{r}}_i$. These data, together with the <<equation of motion>> (2), would then determine the future behavior of the system just as $\mathbf{r}_i, \dot{\mathbf{r}}_i$ and (1) determines it. The fact that initial conditions and laws of nature completely determine the behavior is, similarly, true in any causal theory.

The surprising discovery of Newton's age is just the clear separation of laws

of nature on the one hand and initial conditions on the other. The former are precise beyond anything reasonable; we know virtually nothing about the latter. Let us pause for a minute at this last statement. Are there really no regularities concerning what we just called initial conditions?

The last statement would certainly not be true if the laws of nature (2), (2a) were adopted, *i.e.*, if we considered the \ddot{r}_i as part of the initial conditions. In this case, there would be a relation, in fact the precise relation (1), between the elements of the initial conditions. The question, therefore, can be only: are there any relations between what we really do consider as initial conditions. Formulated in a more constructive way: how can we ascertain that we know all the laws of nature relevant to a set of phenomena? If we do not, we would determine unnecessarily many initial conditions in order to specify the behavior of the object. One way to ascertain this would be to prove that all the initial conditions can be chosen arbitrarily - a procedure which is, however, impossible in the domain of the very large (we cannot change the orbits of the planets), or the very small (we cannot precisely control atomic particles). No other equally unambiguous criterion is known to me, but there is a distinguishing property of the correctly chosen, that is minimal set, of initial conditions which is worth mentioning.

The minimal set of initial conditions not only does not permit any exact relation between its elements, on the contrary, there is reason to contend that these are, as a rule, as random as the externally imposed, gross constraints allow. I wish to illustrate this point first on an example which, at first, seems to contradict it because this example shows its power, and also its weakness, best.

Let us consider for this purpose again our planetary system. It was mentioned before that the approximate regularities in the initial conditions, that is the determinants of the orbits, led Kepler to the considerations which were then left by the wayside by Newton. These regularities form the apparent counter-example to the aforementioned thesis. However, the existence of the regularities in the initial conditions is considered so unsatisfactory that it is felt necessary to show that the regularities are but a consequence of a situation in which there were no regularities. Perhaps v. Weizäcker's attempt in this direction³ is most interesting: he assumes that, originally, the solar system consisted of a central star, with a gas in rotation, but otherwise in random motion, around it. He then deduces the aforementioned regularities of the planetary system, now called Bode's law, from his assumption. More generally, one tries to deduce almost all «organized motion», even the existence of life, in a similar

fashion. It must be admitted that few of these explanations have been carried out in detail⁴ but the fact that such explanations are attempted remains significant.

The preceding paragraph dealt with cases in which there is at least an apparent evidence against the random nature of the uncontrolled initial conditions. It attempted to show that the apparently organized nature of these initial conditions was preceded by a state in which the uncontrolled initial conditions were random. These are, on the whole, exceptional situations. In most cases, there is no reason to question the random nature of the non-controlled, or non-specified, initial conditions and the random nature of these initial conditions is supported by the validity of the conclusions arrived at on the basis of the assumption of randomness. One encounters such situations in the kinetic theory of gases and, more generally, whenever one describes processes in which the entropy increases. Altogether, then, one obtains the impression that whereas the laws of nature codify beautifully simple regularities, the initial conditions exhibit, as far as they are not controlled, equally simple and beautiful irregularity. Hence, there is perhaps little chance that some of the former remain overlooked.

The preceding discussion characterized the laws of nature as regularities in the behavior of an object. In quantum theory, this is natural: the laws of quantum mechanics can be suitably formulated as correlations between subsequent observations on an object. These correlations are the regularities given by the laws of quantum mechanics⁵. The statement of classical theory, its equations of motion, are not customarily viewed as correlations between observations. It is true, however, that their purpose and function is to furnish such correlations and that they are, in essence, nothing but a shorthand expression for such correlations.

Laws of Nature and Invariance

We have ceased to expect from physics an explanation of all events, even of the gross structure of the universe, and we aim only at the discovery of the laws of nature, that is the regularities, of the events. The preceding section gives reason for the hope that the regularities form a sharply defined set, and are clearly separable from what we call initial conditions, in which there is a strong element of randomness. However, we are far from having found that set. In fact, if it is true that there are precise regularities, we have reason to

believe that we know only an infinitesimal fraction of these. The best evidence for this statement derives perhaps from a fact which was mentioned here by Yang six years ago: the multiplicity of the types of interactions. Yang mentioned four of them: gravitational, weak, electromagnetic, and strong, and it now seems that there are two types of strong interactions. All these play a role in every process, but it is hard, if not impossible, to believe that the laws of nature should have such complexity as implied by four or five different types of interactions between which no connection, no analogy, can be discovered.

It is natural, therefore, to ask for a superprinciple which is in a similar relation to the laws of nature as these are to the events. The laws of nature permit us to foresee events on the basis of the knowledge of other events; the principles of invariance should permit us to establish new correlations between events, on the basis of the knowledge of established correlations between events. This is exactly what they do. If it is established that the existence of the events A, B, C, . . . necessarily entails the occurrence of X, then the occurrence of the events A', B', C', . . . also necessarily entails X', if A', B', C', . . . and X' are obtained from A, B, C, . . . and X by one of the invariance transformations. There are three categories of such invariance transformations:

- (a) euclidean transformations: the primed events occur at a different location in space, but in the same relation to each other, as the unprimed events.
- (b) time displacements: the primed events occur at a different time, but separated by the same time intervals from each other, as the unprimed ones.
- (c) uniform motion: the primed events appear to be the same as the unprimed events from the point of view of a uniformly moving coordinate system.

The first two categories of invariance principles were always taken for granted. In fact, it may be argued that laws of nature could not have been recognized if they did not satisfy some elementary invariance principles such as those of Categories (a) and (b) - if they changed from place to place, or if they were also different at different times. The principle (c) is not so natural. In fact, it has often been questioned and it was an accomplishment of extraordinary magnitude, on the part of Einstein, to have reestablished it in his special theory of relativity. However, before discussing this point further, it may be useful to make a few general remarks.

The first remarkable characteristic of the invariance principles which were enumerated is that they are all geometric, at least if four-dimensional space-time is the underlying geometrical space. By this I mean that the invariance transformations do not change the events; they only change their location in space and time, and their state of motion. One could easily imagine a prin-

ciple in which, let us say, protons are replaced by electrons and *vice versa*, velocities by positions, and so on⁶.

The second remarkable characteristic of the preceding principles is that they are invariance rather than covariance principles. This means that they postulate the same conclusion for the primed premises as for the unprimed premises. It is quite conceivable that, if certain events A, B, . . . take place, the events X₁, X₂, X₃... will follow with certain probabilities p₁, p₂, p₃ . . . From the transformed events A', B', C', the transformed consequences X₁', X₂', X₃' . . . could follow with changed probabilities such as p₁'=p₁(1-p₁+Σpn²), p₂'=p₂(1-p₂+Σp_n²), . . . but this is not the case; we always had p_i' = p_i.

These two points are specifically mentioned because there are symmetry principles, the so-called crossing relations⁷, which *may be* precisely valid and which surely do not depend on specific types of interactions. In these regards they are, or may be, similar to the geometric invariance principles. They differ from these because they do change the events and they are covariance rather than invariance principles. Thus, from a full knowledge of the cross section for neutron-proton scattering, they permit one to obtain some of the neutron-antineutron collision cross sections. The former events are surely different from the neutron-antineutron collisions and the cross sections for the latter are not equal to the neutron-proton cross sections but are obtained from these by a rather complicated mathematical procedure. Hence, the crossing relations are not considered to be geometrical symmetry conditions and they will not be considered here. Similarly, we shall not be concerned with the dynamic symmetry principles which are symmetries of specific interactions, such as electromagnetic interactions or strong interactions, and are not formulated in terms of events⁷.

On the other hand, it should be noted that the invariance principles themselves depend on the dividing line between initial conditions and laws of nature. Thus, the law of nature (2) or (2a), obtained from Newton's principle by differentiation with respect to time, is invariant also under the transformation to a uniformly accelerated coordinate system

$$\mathbf{r}_i' = \mathbf{r}_i + t^2 \mathbf{a} \quad t' = t \quad (3)$$

where \mathbf{a} is an arbitrary vector. Naturally, this added principle can have no physical consequence because, if the initial conditions $\mathbf{r}_i, \dot{\mathbf{r}}_i, \ddot{\mathbf{r}}_i$ are realizable (*i.e.*, satisfy (I)), the transformed initial conditions $\mathbf{r}_i' = \mathbf{r}_i, \dot{\mathbf{r}}_i' = \dot{\mathbf{r}}_i, \ddot{\mathbf{r}}_i' = \ddot{\mathbf{r}}_i + 2\mathbf{a}$ cannot be realizable.

The symmetry principles of the preceding discussion are those of New-

tonian mechanics or the special theory of relativity. One may well wonder why the much more general, and apparently geometrical, principles of invariance of the general theory have not been discussed. The reason is that this writer believes, in conformity with the views expressed by V. Fock⁸, that the curvilinear coordinate transformations of the general theory of relativity are not invariance transformations in the sense considered here. These were so-called active transformations, replacing events A, B, C, . . . by events A', B', C', . . . and unless active transformations are possible, there is no physically meaningful invariance. However, the mere replacement of one curvilinear coordinate system by another is a «redescription» in the sense of Melvin⁹; it does not change the events and does not represent a structure in the laws of nature. This does not mean that the transformations of the general theory of relativity are not useful tools for finding the correct laws of gravitation; they evidently are. However, as I suggested elsewhere⁷, the principle which they serve to formulate is different from the geometrical invariance principles considered here.

The Use of Invariance Principles, Approximate Invariances

The preceding two sections emphasized the inherent nature of the invariance principles as being rigorous correlations between those correlations between events which are postulated by the laws of nature. This at once points to the use of the set of invariance principles which is surely most important at present: to be a touchstone for the validity of possible laws of nature. A law of nature can be accepted as valid only if the correlations which it postulates are consistent with the accepted invariance principles.

Incidentally, Einstein's original article which led to his formulation of the special theory of relativity illustrates the preceding point with greatest clarity¹⁰. He points out in this article that the correlations between events are the same in coordinate systems in uniform motion with respect to each other, even though the causes *attributed* to these correlations at that time did depend on the state of motion of the coordinate system. Similarly, Einstein made the most extensive use of invariance principles to guess the correct form of a law of nature, in this case that of the gravitational law, by postulating that this law conform with the invariance principles which he postulated¹¹. Equally remarkable is the present application of invariance principles in quantum electrodynamics. This is not a consistent theory - in fact, not a theory in the proper sense because its equations are in contradiction to each other. However, these

contradictions can be resolved with reasonable uniqueness by postulating that the conclusions conform to the theory of relativity¹². Another approach, even more fundamental, tries to axiomatize quantum field theories, the invariance principles forming the cornerstone of the axioms¹³. I will not further enlarge on this question because it has been discussed often and eloquently. In fact, I myself spoke about it but a short time ago⁷.

To be touchstones for the laws of nature is probably the most important function of invariance principles. It is not the only one. In many cases, consequences of the laws of nature can be derived from the character of the mathematical framework of the theory, together with the postulate that the law - the exact form of which need not be known - conform with invariance principles. The best known example heretofore is the derivation of the conservation laws for linear and angular momentum, and for energy, and the motion of the center of mass, either on the basis of the Lagrangian framework of classical mechanics, or the Hilbert space of quantum mechanics, by means of the geometrical invariance principles enumerated before¹⁴. Incidentally, conservation laws furnish at present the only generally valid correlations between observations with which we are familiar; for those which derive from the geometrical principles of invariance it is clear that their validity transcends that of any special theory - gravitational, electromagnetic, etc. - which are only loosely connected in present-day physics. Again, the connection between invariance principles and conservation laws - which in this context always include the law of the motion of the center of mass - has been discussed in the literature frequently and adequately.

In quantum theory, invariance principles permit even further reaching conclusions than in classical mechanics and, as a matter of fact, my original interest in invariance principles was due to this very fact. The reason for the increased effectiveness of invariance principles in quantum theory is due, essentially, to the linear nature of the underlying Hilbert spaces. As a result, from any two state vectors, Ψ_1 and Ψ_2 , an infinity of new state vectors

$$\Psi = a_1 \Psi_1 + a_2 \Psi_2 \quad (4)$$

can be formed, a_1 and a_2 being arbitrary numbers. Similarly, several, even infinitely many, states can be superimposed with arbitrary coefficients. This possibility of superposing states is by no means natural physically. In particular even if we know how to bring a system into the states Ψ_1 and Ψ_2 , we cannot give a prescription how to bring it into a superposition of these states. This prescription would have to depend, naturally, on the coefficients with which

the two states are superimposed and is simply unknown. Hence, the superposition principle is strictly an existence postulate-but a very effective and useful existence postulate.

To illustrate this point, let us note that in classical theory, if a state, such as a planetary orbit, is given, another state, that is another orbit, can be produced by rotating the initial orbit around the center of attraction. This is interesting but has no very surprising consequences. In quantum theory, the same is true. In addition, however, the states obtained from a given one by rotation can be superimposed as a result of the aforementioned principle. If the rotations to which the original state was subjected are uniformly distributed over all directions, and if the states so resulting are superimposed with equal coefficients, the resulting state has necessarily spherical symmetry. This is illustrated in the Fig. 1 in the plane case. This construction of a spherically symmetric state could fail only if the superposition resulted in the null-vector of Hilbert space in which case one would not obtain any state. In such a case, however, other coefficients could be chosen for the superposition-in the plane case the coefficients $e^{im\varphi}$ where φ is the angle of rotation of the original state-and the resulting state, though not spherically symmetric, or in the plane case axially symmetric - would still exhibit simple properties with respect to rotation. This possibility, the construction of states which have either full rotational symmetry, or at least some simple behavior with respect to rotations, is the one which is fundamentally new in quantum theory. It is also conceptually satisfying that simple systems, such as atoms, have states of high symmetry.

The superposition principle also permits the exploitation of reflection sym-

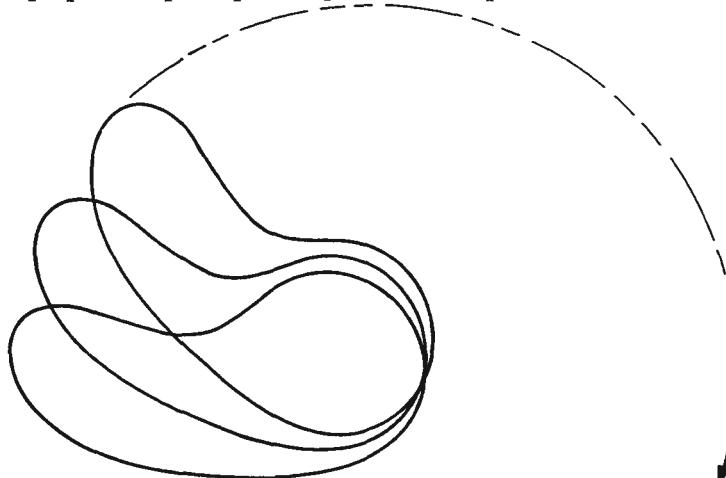


Fig. 1.

metry. In classical mechanics as well as in quantum mechanics, if a state is possible, the mirror image of that state is also possible. However, in classical theory no significant conclusion from this fact is possible. In quantum theory, original state and mirror image can be superimposed, with equal or oppositely equal coefficients. In the first case, the resulting state is symmetric with respect to reflections, in the second case antisymmetric. The great accomplishment of Lee and Yang, which was mentioned earlier¹, was just a very surprising reinterpretation of the physical nature of one of the reflection operations, that of space reflection, with the additional proof that the old interpretation cannot be valid. The consideration of «time inversion» requires rather special care because the corresponding operator is antiunitary. Theoretically, it does lead to a new quantum number and a classification of particles¹⁶ which, however, has not been applied in practice.

My discussion would be far from complete without some reference to approximate invariance relations. As all approximate relations, these may be very accurate under certain conditions but fail significantly in others. The critical conditions may apply to the state of the object, or may specify a type of phenomena. The most important example for the first case is that of low velocities. In this case, the magnetic fields are weak and the direction of the spins does not influence the behavior of the other coordinates. One is led to the Russel-Saunders coupling of spectroscopy¹⁷. Even more interesting should be the case of very high velocities in which the magnitude of the rest mass becomes unimportant. Unfortunately, this case has not been discussed in full detail even though there are promising beginnings¹⁸.

Perhaps the most important case of special phenomena in which there are more invariance transformations than enumerated before is rather general : it comprises all phenomena, such as collisions between atoms, molecules, and nuclei, in which the weak interaction, which is responsible for beta decay, does not play a role. In all these cases, the parity operation is a valid invariance operation. This applies also in ordinary spectroscopy.

In another interesting special type of phenomena the electromagnetic interaction also plays a subordinate role only. This renders the electric charge on the particles insignificant and the interchange of proton and neutron, or more generally of the members of an isotopic spin multiplet, becomes an invariance operation. These, and the other special cases of increased symmetry, lead to highly interesting questions which are, furthermore, at the center of interest at present. However, the subject has too many ramifications to be discussed in detail at this occasion.

1. Chen Ning Yang, The law of parity conservation and other symmetry laws of physics; Tsung Dao Lee, Weak interactions and nonconservation of parity, *Nobel Lectures, Physics*, 1942-1962, Elsevier, Amsterdam, 1964, pp. 393-403 and pp. 406-418.
2. See, for instance, A. C. Crombie, *Augustine to Galileo*, Falcon Press, London, 1952, pp. 316 ff. The growth of the understanding of the realm of the explainable, from the end of the 13th century on, can be traced through almost every chapter of this book.
3. C.F.V. Weizsäcker, *Z. Astrophys.*, 22(1944) 319; S. Chandrasekhar, *Rev. Mod. Phys.*, 18 (1946) 94.
4. An interesting and well understood case is that of « focussing collisions » in which neutrons, having velocities which are rather high but with random orientation, are converted into lower velocity neutrons but with preferential directions of motion. See R.H. Silsbee, *J. Appl. Phys.*, 28 (1957) 1246; Chr. Lehmann and G. Leibfried, *Z. Physik*, 172 (1963) 465.
5. See, for instance, the section *What is the State Vector* in the writer's article, *Am. J. Phys.*, 31(1963) 6.
6. The possibility of an invariance principle in which velocities are replaced by positions, and conversely, was studied by M. Born, *Nature*, 141 (1938) 327; *Proc. Roy. Soc. (London)*, Ser. A, 165 (1938) 291, 166 (1938) 552.
7. The crossing relations were established by M.L. Goldberger, *Phys. Rev.*, 99 (1955) 979; M. Gell-Mann and M.L. Goldberger, *Phys. Rev.*, 96 (1954) 1433. For further literature, see, for instance M.L. Goldberger and K. M. Watson, *Collision Theory*, Wiley, New York, 1964, chapter 10. The relations of the various types of symmetry principles were considered in two recent articles: *Nuovo Cimento, Suppl.*, in the press, and *Phys. Today*, 17 (1964) 34. See also E. Wigner, *Progr. Theoret. Phys. (Kyoto)*, 11 (1954) 437.
8. V. A. Fock, *The Theory of Space, Time and Gravitation*, Pergamon, Oxford, 1959. The character of the postulate of invariance with respect to general coordinate transformations as a geometrical invariance was questioned already by E. Kretschman, *Ann. Physik*, 53 (1917) 575.
9. M. A. Melvin, *Rev. Mod. Phys.*, 32 (1960) 477.
10. A. Einstein, Zur Elektrodynamik bewegter Körper, *Ann. Physik*, 17 (1905) 891.
11. A. Einstein and S.B. Preuss, *Akad. Wiss.*, (1915) 778, 779, 844; *Ann. Physik*, 49 (1916) 769. Similar results were obtained almost simultaneously by D. Hilbert, *Nachr. kgl. Ges. Wiss. (Göttingen)*, (1915) 395.
12. J. Schwinger, *Phys. Rev.*, 76 (1949) 790. See also S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory*, Row, Peterson and Co., New York, 1961, Section 15, where further references can also be found.
13. See A. S. Wightman, *Quelques Problèmes Mathématiques de la Théorie Quantique Relativiste*, and numerous other articles in *Les Problèmes Mathématiques de la Théorie Quantique des Champs*, Centre National de la Recherche Scientifique, Paris, 1959.
14. G. Hamel, *Z. Math. Physik*, 50 (1904) 1; G. Herzlotz, *Ann. Physik*, 36 (1911) 493; F. Engel, *Nachr. Kgl. Ges. Wiss. (Göttingen)*, (1918) 171; E. Noether, *Nachr. Kgl. Ges. Wiss. (Göttingen)*, (1918) 235; E. Bessel-Hagen, *Math. Ann.*, 84 (1921) 258; the quantum theoretical derivation given by E. Wigner, *Nachr. Kgl. Ges. Wiss. (Göttingen)*,

(1927) 375, contains also the parity conservation law which was shown, in ref.1, to be only approximately valid. See also the article of ref.16.

15. I heard this remark, for the first time, from C. N. Yang, at the centennial celebration of Bryn Mawr College.
16. See the writer's article Unitary Representations of the Inhomogeneous Lorentz Group including Reflections, in *Elementary Particle Physics*, Gordon and Breach, New York, 1964, for a systematic discussion of the reflection operations.
17. See the writer's book, *Gruppentheorie und ihre Anwendung auf die Quantenmechanik der Atomspektren*, Vieweg, Braunschweig, 1931, or the English translation by J. Griffin, Academic Press, New York, 1959.
18. H.A.Kastrup, *Physics Letters*, 3 (1962) 78. The additional invariance operations probably form the conformal group. This was discovered by E. Cunningham (*Proc. London Math. Soc.*, 8 (1909) 77) and by H.Bateman (*Proc. London Math. Soc.*, 8 (1910) 223) to leave Maxwell's equations for the vacuum invariant, i.e., the equations which describe light, always propagating at light velocity. For more recent considerations, see T. Fulton, F. Rohrlich and L. Witten, *Rev. Mod. Phys.*, 34 (1962) 442 and Y. Murai, *Progr. Theoret. Phys. (Kyoto)*, (1954) 441. The latter articles contain also more extensive references to the subject.

RICHARD WILSTÄTTER

On plant pigments

Nobel Lecture, June 3, 1920

I accept with pleasure the invitation with which the Royal Swedish Academy of Sciences has honoured me to speak about my investigations on plant pigments. My task is lightened by the fact that several years have elapsed since the completion of this work. Consequently, it is as if I were able today to conduct this chosen audience to a height from which a survey can be made of the main lines of research without having to weary you by wandering through the twisted and tortuous paths which I have trod. The intention behind my work was to establish the constitutional characteristics of the most widely distributed plant pigments, of chlorophyll in particular, and to gain some criteria with regard to its chemical function.

Although a considerable literature was already in existence on the pigments of green leaves, my work is actually linked with that of Berzelius, who in 1837 and 1838 conducted the first chemical investigation of leaf pigments. The method used by Berzelius was to treat the chlorophyll with acids and alkalis, and in this method lies also the basis of my work. From the innumerable investigations by Berzelius there is actually only one result which is outstanding, which Hoppe-Seyler found and which Schunck and Marchlevski have stressed, viz. the recognition of a certain relationship between the pigments of the blood and of leaves, or more correctly, between the structural materials of their molecules. But the first questions with regard to chlorophyll were still unanswered: what features in the composition and structure distinguish chlorophyll from haemin analogous to the basically different functions of the leaf pigment and the blood pigment? Is there only one chlorophyll, or are there several or many? Gautier had laid down the principle that chlorophyll was different in monocotyledons and dicotyledons, and in a much admired work which appeared in 1906, Etard claimed to show that in a single plant there exists a large series of quite different chlorophylls and in the entire plant world an unlimited number of the most different leaf pigments. In this work every green-coloured wax or fat was regarded as a chlorophyll, whereas Berzelius had already clearly distinguished the leaf green from the resins, waxes, and fats.

In one point, however, my first experiments have already led me to take up a different view from that of Berzelius. According to him, chlorophyll should not be decomposable either by strong acids or by alkalis. This was an error, caused by the chlorophyll-green colour which is precisely special to the salts of the decomposition products formed with strong acids and alkalis. But chlorophyll itself does not produce a salt; it is inert in its intact state towards both acids and bases. It is extremely easily transformable, and it can be decomposed hydrolytically by acids and bases alike. Even in its neutral solutions, e.g. in alcohol or ether, it undergoes extraordinarily easily a substantial transformation - I designated it allomerization - which is not, it is true, revealed in the colour.

Chemical indifference, ease of change, and ease of solubility of the leaf pigment, which is mixed with yellow and a huge number of colourless accompanying substances - all these were obstacles to its isolation. Our method of work was, first of all, without isolating and investigating the chlorophyll itself, to deduce the peculiarities of its constitution by two methods of decomposition, i.e. from the examination of the two series of derivatives which are produced from the reactions with acid and with alkali.

Even a test-tube experiment shows that as a result of the action of alkali hydroxide on the neutral chlorophyll an acid is produced which forms salts soluble in water; the colour of these salts is still chlorophyll-green. This means that by this reaction, without a significant optical change, a component which was bound to an acid group must have been hydrolytically liberated.

It is evidently another part of the molecule onto which is directed the action of the acid, and the mildest action at that. The chlorophyll colour changes during this process into olive and at the same time the fluorescence is diminished. Whilst this is taking place there is not as yet a salt-forming group formed; this means that saponification is avoided here. Consequently it is possible when splitting by acid to spare that component of the chlorophyll, and to trace it in the decomposition product, which is split off by alkalis and which is lost in the mother liquors, that is, in the substances accompanying the extracts. Conversely, the alkali derivatives of the pigment must still present a characteristic atom group, which is destroyed so easily by acid and with such striking alteration of colour. This was the guiding thought of our work, from which it was possible, before the chlorophyll itself was known, to combine its features from the analysis and from the reactions of the derivatives formed by acid and alkali, and to do this so completely that

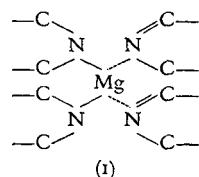
the analysis produced nothing new at all when it was finally possible to prepare the natural pigment in pure state.

With the mild action of oxalic acid on the alcoholic extract of leaves, a chlorophyll derivative - phaeophytin - difficultly soluble in alcohol, is almost quantitatively precipitated, free from colourless and yellow accompanying substances. It is a waxlike nitrogenous substance which leaves no ash residue, shows a very weak alkaline reaction and has no acid properties. This substance can be saponified like an ordinary wax; the result of this proves it to be an ester, it yields a mixture of high-molecular nitrogenous carbonic acids which have a pigment character, and a colourless and nitrogen-free alcohol, phytol, which corresponds to the formula $C_{20}H_{39}OH$. This alcohol, which has the appearance of glycerine, but is of course insoluble in water, is primary, unsaturated and aliphatic; its carbon chain has several branches, it is oxidized by ozone, with the loss of three carbon atoms, to a ketone.

This phytol occurs constantly as a component of chlorophyll and forms a third of its molecule. One of the first changes produced by the alkalies on the chlorophyll molecule is, therefore, the saponification of the phytol ester group. The phaeophytin, and therefore the chlorophyll itself, contains also a $COOCH_3$ group, which is subsequently hydrolysed. Furthermore, there is still a characteristic change resulting from the action of the alkalis, which is revealed by a striking change of colour, by the so-called brown phase. When an alcoholic solution of potassium hydroxide is added, the green solution immediately turns into an intense brown, then a few minutes later the original colour of the fluid returns. This behaviour, which the chlorophyll loses by allomerization when its solutions are merely left to stand, is to be attributed to the solution of an easily hydrolysable atom group and the formation of a similar group ; it can probably be explained as a re-lactamization, i.e. as the opening of an existing lactam ring and the closing of a new ring, similar indeed but resistant to alkalies.

When chlorophyll is saponified by potassium hydroxide, very easily decomposable chlorophyll-green carboxylic acids are produced-the chlorophyllins. They can be separated from the accompanying substances by the solubility in water of their salts and by transfer into ether upon acidification, and can be purified by careful transfer from the ether into disodium phosphate as an alkali, and liberated with monosodium phosphate as an acid. Analysis showed that the chlorophyllins are magnesium compounds. They contain the metal in an electrically non-dissociable state, as haemoglobin contains iron, it is bound in them to nitrogen with a complex bond. This

magnesium-containing group is very sensitive to acids, but is stable in an alkaline medium. Consequently, it remains intact during fundamental changes of the molecule, by which even the carboxyls of the chlorophyllins are split off, one after the other. The assumed type of bonding of the magnesium was confirmed during the prolonged action of the alkalies, i.e. by the decomposition of the chlorophyll with concentrated alcoholic alkalies at temperatures up to 250°. The series so produced consists of decomposition products which crystallize well, have splendid colours and fluoresce intensely-acids with three, two and finally with only one carboxyl group, are the so-called phyllins, e.g. glaucophyllin, rhodophyllin, pyrrophyllin. All contain magnesium and are moreover free acids; they have one atom of magnesium to four atoms of nitrogen. The atom group (I) with the basic metal

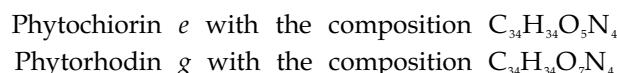


is the essential feature of chlorophyll and is to be regarded as the actual carrier of the chlorophyll function in the synthesizing life of the green plant.

The magnesium content of the chlorophyll is constant, close to 3%. It has been found to be similar in land and water plants of the most varied classes.

To the complete comparison of the leaf pigments of different origin there belongs also the analysis of the intricately composed nitrogenous carboxylic acids which appear upon decomposition, whether upon the hydrolysis of the phaeophytin by alkalis or upon the splitting of the chlorophyllins and other phyllins by acids. This was at first always a very complicated mixture of pigments of an acid nature. Its disentanglement and then the simplification of the results of the decomposition were only possible by a method of analysis, the development of which was initiated by my investigations. The method of determining and separating chlorophyll derivatives is based upon the alkaline properties of these amino acids, which are extremely varied, and upon the different distribution between ether and diluted hydrochloric acid. The more cautious and uniform were the conditions during hydrolysis, but also during the preparation of the extracts containing the chlorophyll, the more simple in its composition became the mixture of the alkaline cleavage products. Finally it was possible to conduct the production of the phaeophytin

and to arrange its splitting in such a way that there always appeared only two magnesium-free carboxylic acids,



the former being olive-green and the latter a splendid red in neutral solutions.

The determining factor for the identity of the chlorophyll when comparing preparations from over 200 plants of numerous classes of cryptogamia and phanerogamia, was therefore the content of magnesium and of phytol and the elucidation of the two characteristic cleavage products, phytochlorin and phytorhodin. With regard to the obstacles which had to be overcome in this comparison, it is noteworthy that the phytol content of the phaeophytin preparations at first showed great fluctuations; it often, indeed, reached one third of the molecule, but sank considerably in other cases, even to a small percentages and still smaller figures. For these differences, however, only one transformation was to blame which the pigment underwent during the extraction from the fresh or dried leaves with alcohol under the action of an enzyme. This enzyme, chlorophyllase, which belongs to the esterases, accompanies the chlorophyll in the green parts of the plant, often occurring very sparsely, but often very richly. Under its effect, which once recognized could be easily avoided, the phytol is replaced by ethyl or methyl alcohol or in an aqueous medium is removed by saponification. The action of this chlorophyllase also explained the appearance of the so-called crystallized chlorophyll, the remarkable and magnificent green crystals which Borodin had first observed in 1881 in microscopic leaf sections. The crystallized chlorophyll is ethylchlorophyllide; it originates through ethanolysis from the phytol compound. The formation of the crystallized chlorophyll was now no longer accidental; on the contrary, the effect of the chlorophyllase was utilized to an extensive degree for the purposes of preparation, also for partial synthesis of the chlorophyll from chlorophyllide and phytol.

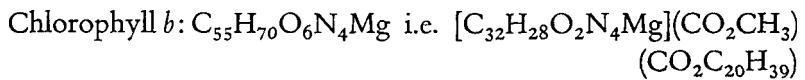
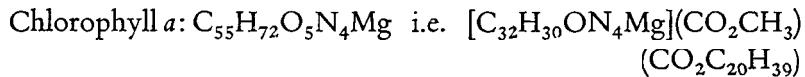
The same characteristics which serve for comparing the leaf pigments from whichever starting material allow us also to decide whether, in the course of the preparatory work, the chlorophyll remains unchanged in the finer details of its molecule. Hence this characterization was the basis for preparing the chlorophyll itself in an uninjured and pure state and for confirming, by analysis, the conclusions which had first been reached from the

investigation of its derivatives. The isolation of the chlorophyll relied upon calorimetric determination of the degree of purity of its solutions, and depended upon the systematic increase of this purity by methods of distribution between several immiscible solvents, such as petroleum ether and aqueous alcohol. By this method, colourless admixtures and the yellow pigments of the leaves were separated. The simultaneous occurrence of these yellow pigments, the carotenoids, with the green pigments, which seemed to indicate a special physiological role of the substances because of their great affinity to oxygen, gave rise to the preparation of the yellow substances in the pure state and their analysis. Two well-crystallized and nitrogen-free pigments occur in every green part of the plant and in many yellow parts. One of these, identical with the long-known carotene of carrots, is an unsaturated hydrocarbon of the formula $C_{40}H_{56}$. Its partner, xanthophyll, was still unknown in structure, although it predominates in leaves; according to composition and properties it is a carotene oxide ($C_{40}H_{56}O_2$). Only in the Phaeophyceae are carotene and xanthophyll accompanied and repressed in quantity by a third carotenoid, richer in oxygen, fucoxanthin, which can be isolated in crystal form and has the composition $C_{40}H_{56}O_6$.

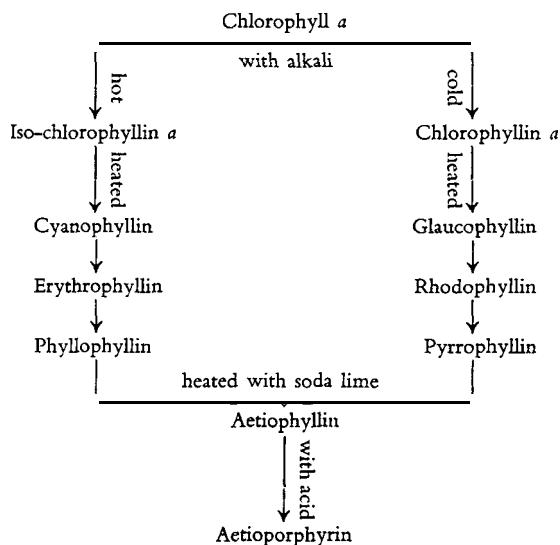
Whilst chlorophyll in the plant extracts is accompanied by other substances, as much as fifteen to eight times in quantity, solutions of about 70% pure chlorophyll can be obtained by separation methods. From this point on, a surprising observation helped to solve the problem. When the substance has reached a certain degree of purity, it reveals its true solubility characteristics, which were distorted before by the admixtures. Pure chlorophyll is not soluble in petroleum ether and is precipitated from the alcoholic solvent when the alcohol is washed away. The procedure permits, with good yields, the isolation of pure chlorophyll from dried or fresh leaves just as easily as that of any other plant substance, alkaloid or sugar.

However, this pure chlorophyll is not as yet a simple substance. The occurrence of phytochlorin *e* and phytorhodin *g* in the mixture upon its decomposition could also have been attributed to a cleavage of the phaeophytin into the two products or to successive decomposition to these. But in the course of the demixing operations which were used for the isolation it appeared - as had already been observed by the English physicist Stokes in 1864 in the course of small-scale experiments - that the chlorophyll consists of two differently coloured and differently soluble components, which are distributed unequally between methyl alcohol and petroleum ether. By means of systematic fractionation with these solvents the two pure chlorophyll com-

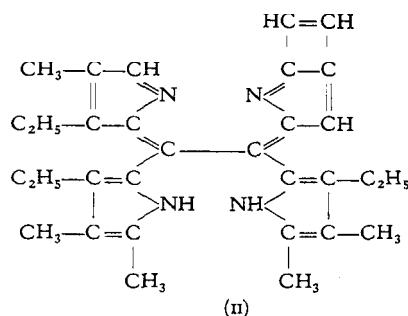
ponents were finally obtained from the mixture. One of them, chlorophyll a, is blue-green; and the other, chlorophyll b, is yellow-green; component a predominates; to almost three molecules of chlorophyll a in the leaves there is only one molecule of chlorophyll b. Despite the optical difference, the composition is similar. The difference is due to a different degree of oxidation. What the formulae of phytochlorin and phytorhodin allowed us to foresee is confirmed by the chlorophyll: the compounds of series b are derived probably from a by the fact that two atoms of hydrogen are replaced by one atom of oxygen, corresponding to the formulae:



With a specific reducing agent, magnesium alkylhaloid, the compound which is richer in oxygen can be converted to derivatives of the u-series. The decomposition of the chlorophyll by an alkali has also led, by way of different intermediate products, to the same end-products, and finally to one and the same carboxyl-free and altogether oxygen-free parent substance, aetio-phyllin, of the formula $C_{31}H_{34}N_4Mg$, from which is derived the magnesium-free aetioporphyrin of the formula $C_{31}H_{56}N_4$ - just as the carboxylic acids of the series, when acted upon by acid, are converted into analogous magnesium-free compounds, the porphyrins.



This parent substance became even more important because we were also able to decompose the blood pigment to the very same aetioporphyrin; thus for the first time a common transformation product has been obtained from haemin and chlorophyll, with its molecule still in close relationship to the pigments. From oxidation and reduction experiments it is clear that the molecule of the simplest porphyrin is built up from four multi-substituted pyrrole nuclei. Although there are as yet no definite details about the way in which these pyrrole nuclei are bound, nor any exact statements with regard to certain details of their substitution, we may be allowed nevertheless to draw a provisional picture, Formula (II), of the structure of aetioporphyrin, based on considerations of probability.



As with the proteins and the nucleic acids, so with haemoglobin and with chlorophyll there are still many difficult problems to solve regarding the detailed determination of their constitution. Nevertheless, the broad features of chemical knowledge of the leaf pigments already gained open new paths for the treatment of biological questions and for the investigation of the function of chlorophyll in the process of assimilation.

The English botanist Wager had submitted the hypothesis that chlorophyll is consumed in the assimilation process, that it builds up its molecule with carbon dioxide and forms its reduction product by its complete disintegration. The quantitative determination of the chlorophyll content allows us to make a decision on the correctness of this concept. It has been shown, however, that the chlorophyll content of a leaf remains absolutely constant throughout a long-continued and maximal assimilatory performance. We could therefore relate the assimilatory performance of the plant at a favourable temperature and moistening and with an excess supply of carbon dioxide and light to the amount of chlorophyll and check the proportion of the assimilated carbon dioxide to the amount of chlorophyll. This is subject to

great fluctuations according to the chlorophyll concentration in the leaves, and further to the growth and to the season of the year. From an examination of the cases in which this quotient differs very widely from the norm, we must conclude that in the assimilation process the chlorophyll works in collaboration with another internal factor, i.e. one of an enzymatic nature, an enzyme which is probably engaged with the decomposition of an intermediate product formed from chlorophyll and carbon dioxide. Chlorophyll, that is to say, each of the two components, indeed combines in a colloid state with carbon dioxide to form an addition product, which can be dissociated. This observation can form the basis of a theory of assimilation which assumes that the light which is absorbed does its chemical work in the chlorophyll molecule itself, of which carbon dioxide has become a part by attachment to the magnesium complex, in that by a regrouping of the valencies the carbon dioxide molecule is rearranged into a form which is suitable for spontaneous decomposition, which occurs in such a manner that the whole of the oxygen of the carbon dioxide is liberated.

It is therefore the knowledge of the complex nitrogen-magnesium group of the chlorophyll which contributes to the determination of its physiological function. In investigating the lesser or non-vital natural pigments, such as indigo blue and madder red, the analytical work was rewarded by the suggestions which were allotted to the organic synthesis by combinations which the imagination of the scientist would certainly not easily have invented. The easily crystallizing and stable plant pigments were the ones which the earlier generation of chemists succeeded in elucidating. What remained enigmatical were the wonderful but easily disappearing decorative and enticing colours of the vegetable kingdom which surround us in the flowers with their gay colours, in fruits and roots, in the barks and the red leaves. The first experiments led us to expect that in those cases we might be dealing with a class of pigments with many members. But here, as in the case of the assimilatory pigments, the laboratory of the plant cell works sparingly with chemical combinations and to some extent disappoints the chemist by the simplicity which is the basis of the natural wealth of colour. A fascinating problem of constitution was presented to me by the flower pigment in the test-tube when, during a holiday, I took a rose from my Zurich garden into the laboratory. The aqueous solution of the pigment prepared from this rose changed with sodium carbonate from red to emerald green; if, however, the anthocyanin solution was first acidified with a drop of mineral acid, then the colour changed with alkali suddenly from red to deep blue. Between the red

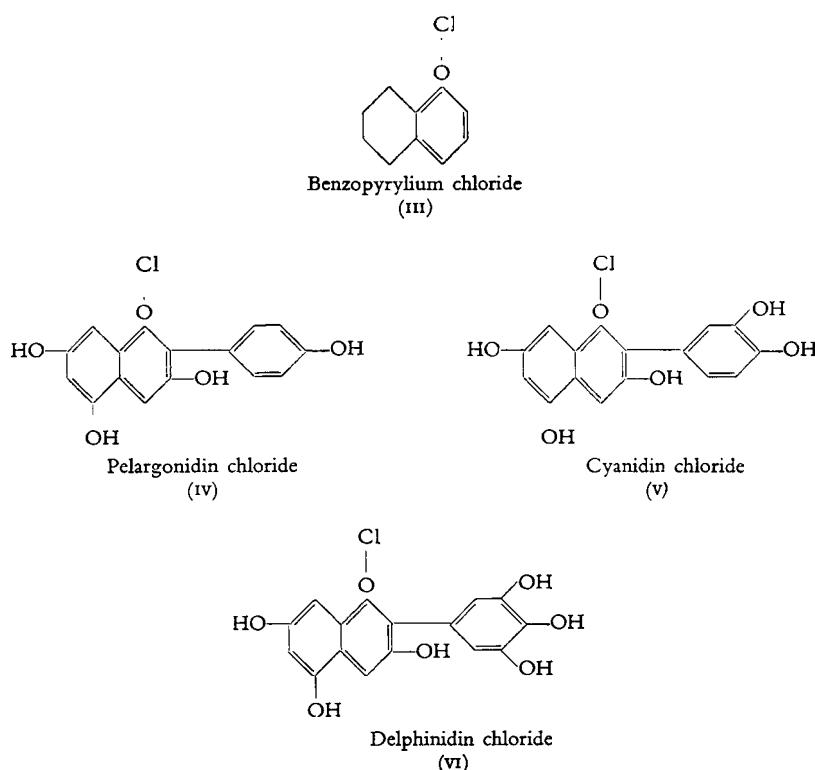
of the acid solution and the blue of the alkali solution there appeared a violet neutral colour, a puzzling phenomenon with a nitrogen-free pigment, but soon explained in the course of our work by the observation that the anthocyanins are quinoid compounds with tetravalent oxygen, that is to say, in Nature widely distributed representatives of the class of oxonium salts, the first examples of which were recognized by Collie in synthetically obtainable compounds of the pyrone series. The isolation of the anthocyanins in a pure state and their analysis are therefore founded upon their basic nature and through the preparation of well-crystallized salts with hydrochloric or picric acid. These acidic compounds are red, the alkali salts are blue, and the violet neutral forms are to be regarded as inner salts, as phenol betaines. Many variations of flower colours are caused solely by the occurrence of these three forms of compound. The neutral and alkaline solutions fade rapidly; this fading is caused by the conversion of the oxonium bases to carbinols, whose alkali salts are yellow and mix with the blue of the colour salt to form green.

The anthocyanins proved to be glucosides, in which the actual pigments, compounds with phenolic hydroxyl groups, are paired with one or two (or even more) molecules of sugars, with glucose, galactose, and rhamnose. The constitution of the sugar-free substances, the anthocyanidins, is elucidated by decomposition through fusion with alkali; the molecule is split into two pieces, into phloroglucin or one of its methyl ethers and an aromatic hydroxy-acid, *p*-hydroxybenzoic acid, protocatechuic acid or bile acid or a methyl ether of these phenolcarboxylic acids. If the variants are disregarded which are caused by the occurrence of the different methyl ethers, by the variety of the sugar components and their different ways of bonding, most of the anthocyanins can be traced back to only three closely related anthocyanidins, viz. pelargonidin, cyanidin and delphinidin, which are hydroxylated phenyl compounds of the benzopyrilium synthesized by Decker and von Fellenberg—cf. Formulae (III), (IV), (V), and (VI).

The assumptions of these constitutions were confirmed in several ways—by synthesis of cyanidin and pelargonidin from phloroglucinaldehyde as the starting-material, and by the reduction of quercetin with magnesium amalgam to the cyanidin, thus bridging a gap between the flavone pigments occurring throughout Nature and the anthocyanins.

The cyanidin appears in combination with two molecules of glucose, e.g. as pigment of the rose, the cornflower and the poppy; with one molecule of glucose in the anthocyanin of the aster and the chrysanthemum; and with galactose as the pigment of the cranberry. Pelargonidin is the base of the

anthocyanins of the pelargonium, the red sages and of certain dahlias; delphinidin gives the deep colours of the larkspur, the woodmallow, the petunia, the bilberry and the grape.



My students were actually planning to extend and complete the investigations of the series of the basic pigments when the World War broke out and destroyed the leisure time of the scientific workshops. The cultures in the flower-beds at Dahlem were neglected, and soon we were carrying baskets full of purple-red asters to the hospitals for the wounded soldiers.

The combination of the synthetic methods with analytical checks was also of decisive importance for the isolation of the anthocyanins. It was of two-fold dependency; at every step from the vegetable starting material to the chemical compound it consisted in quantitative determination of yield and quantitative pursuit of the degree of purity and concentration of the substance. This method of working, and in other respects also the described work in this field will, I hope, contribute still more to the solution of further problems of biochemistry. It seems to me that the most important future

problem of physiological chemistry is that of enzyme research, which has been successfully pioneered in this Laboratory of the Stockholm University by von Euler and his students. Will efforts be successful in making accessible as chemical individuals these most remarkable and active agencies of the animal and plant cell, whom we know only from their effects, and the nature of whose substance is not known at all? May the Swedish Academy of Sciences which has shown such goodwill towards the investigation of vegetable pigments devote its stimulating interest also to the chemical investigation of the enzymes, and may a successful and fortunate professional colleague soon be able on this spot to unveil the secret of the chemical nature of the enzymes!

JULIUS AXELROD

Noradrenaline: fate and control of its biosynthesis

Nobel Lecture, December 12, 1970

When I joined the National Institute of Mental Health in 1955, I began to think of an appropriate problem on which to work. In reading the literature I was surprised to learn that very little was known about the metabolism of noradrenaline and adrenaline. In 1946 Von Euler¹ isolated and identified noradrenaline in the sympathetic nervous system and was later to develop sensitive methods for measuring this catecholamine in tissues². In 1954 I had been working on the *in vivo*³ and *in vitro*⁴ metabolism of amphetamines and compounds related in structure to catecholamines. Because of this background, I decided to work on the metabolism of noradrenaline and adrenaline.

Just as this work was begun, Armstrong *et al.*⁵ identified 3-methoxy-4-hydroxymandelic acid in the urine of subjects with adrenaline-forming tumors. This observation immediately suggested that catecholamines might undergo an *O*-methylation reaction. Cantoni had shown that S-adenosylmethionine formed enzymatically from ATP and methionine can donate its methyl group to the nitrogen of nicotinamides and it appeared possible that S-adenosylmethionine could donate its methyl group to one of the hydroxy groups of catecholamines. In the initial experiment, a rat-liver fraction was incubated with ATP, methionine and Mg²⁺ and adrenaline, and the disappearance of the catecholamine was measured⁷. When these cofactors were added there was a marked disappearance of the catecholamine. When either cofactor was omitted no metabolism took place. The requirement for both ATP and methionine suggested that the liver extract was making S-adenosylmethionine. With S-adenosylmethionine instead of ATP and methionine, even greater metabolism of adrenaline occurred (Table 1). The *O*-methylated metabolite was isolated by solvent extraction and identified as 3-methoxy-4-hydroxyphenyl-2-methylamino ethanol (metanephrine). Metanephrine and normetanephrine were synthesized within two days after isolation by Senoh and Witkop at the NIH.

Rat urine and tissues were then examined by solvent extraction and paper chromatography for the normal occurrence of normetanephrine, meta-

Table 1
Enzymatic O-methylation of catecholamines

<i>Substrate</i>	<i>O-Methoxy product formed (mμmoles)</i>
<i>l</i> -Adrenaline	59
<i>l</i> -Adrenaline (AMe omitted)	0
<i>l</i> -Adrenaline ($MgCl_2$ omitted)	4
<i>dl</i> - <i>N</i> -Methyl-4-hydroxyphenylethanolamine	0
<i>d</i> -Adrenaline	62
<i>l</i> -Noradrenaline	63
<i>dl</i> -Octopamine	0
Dopamine	60
Tyramine	0
Dopa	63

Catecholamines or other amines (0.3 μ mole) were incubated at 37°C with partially purified catechol-O-methyltransferase from rat liver, 50 μ moles pH 7.8 phosphate buffer, 150 μ moles S-adenosylmethionine (AMe), 10 μ moles magnesium chloride in a final volume of 1 ml for 30 minutes. When adrenaline, noradrenaline or dopamine were used as substrates the O-methoxy derivatives were measured. With other substrates their disappearance was measured⁹.

nephrine and 3-methoxytyramine. All these compounds were present in brain, spleen and adrenal gland⁸. Later, another O-methylated metabolite, 3-methoxy-4-hydroxyphenylglycol, was identified. The administration of noradrenaline, adrenaline or dopamine resulted in an elevated excretion of O-methylated amines, acid and alcohol metabolites. As a result of these experiments the scheme shown in Fig. 1 was proposed for the metabolism of noradrenaline and adrenaline. Dopamine undergoes an analogous pathway.

Catechol-O-methyltransferase (COMT)

The enzyme that O-methylates catecholamines was partially purified from rat liver and its properties studied^a. It requires Mg^{2+} (Table 1), but other divalent ions such as Mn^{2+} , Co^{2+} , Zn^{2+} , Cd^{2+} and Ni^{2+} could be substituted. S-Adenosylmethionine is necessary as the methyl donor. All catechols examined were O-methylated by the enzyme, including adrenaline, noradrenaline, dopamine, dopa (Table 1), 3,4-dihydroxymandelic acid, 3,4-dihydroxy -

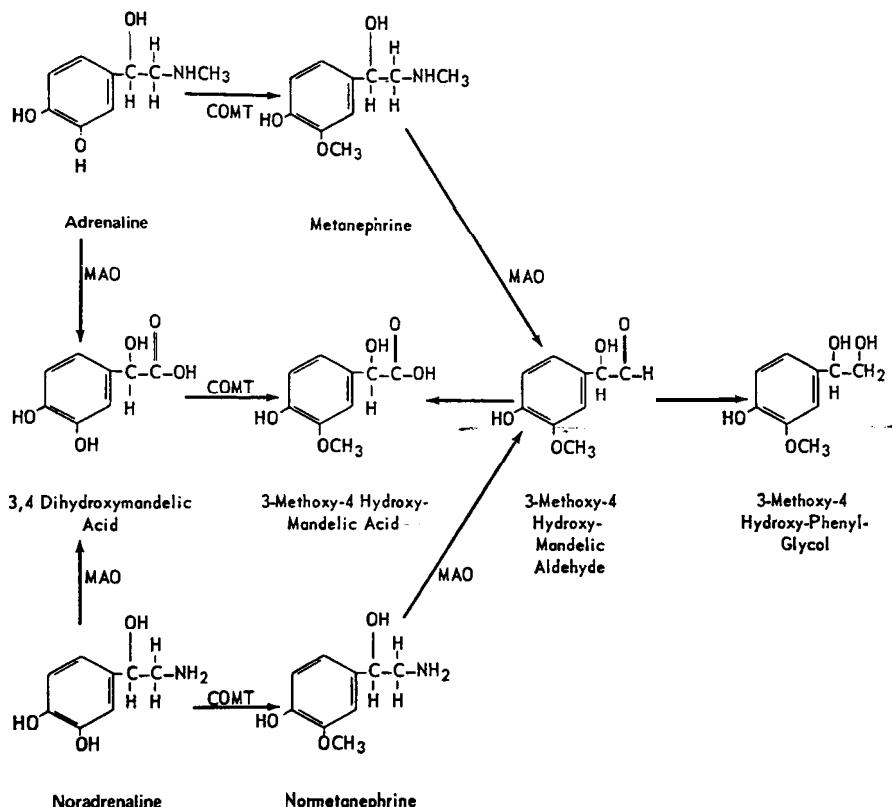


Fig. 1. Metabolism of noradrenaline and adrenaline. COMT is catechol-O-methyltransferase; MAO is monoamine oxidase.

phenylacetic acid, 3-hydroxyestradiol and ascorbic acid. Foreign catechols such as 3,4-dihydroxyephedrine, 3,4-dihydroxyamphetamine and many substituted catechols and polyphenols can serve as substrates for COMT. Monophenols are not O-methylated (Table 1). O-Methylation occurs mainly on the *meta* position. However, O-methylation *in vitro* occurs on both *meta* and *para* positions depending on the pH of the reaction mixture and the nature of the aromatic substrate¹⁰.

The purified enzyme has a molecular weight¹¹ of approximately 24000. At least two separate forms of the enzyme have been identified on starch block electrophoresis¹². The enzyme can be inhibited by polyphenols¹³, 3-hydroxyestradiol¹⁴ and tropolone¹⁵. The administration *in vivo* of COMT inhibitors results in a small, but definite, prolongation of the physiological effects of noradrenaline¹⁶.

COMT is present in all mammalian species examined⁹ and exists also in some plants¹⁷. Of all animal tissues, the liver and kidney exhibit highest activity. Unequally distributed in different regions of the brain, the enzyme's highest activity is present in the area postrema, and lowest activity is in the cerebellar cortex¹⁸. Catechol-O-methyltransferase occurs mainly in the soluble fraction of the cell, but small amounts are present in fat cell membranes¹⁹ and in microsome²⁰. COMT acts on catecholamines mainly outside the neurone, whereas monoamine oxidase, the other major enzyme for catecholamine metabolism, is localized mainly within the neurone. However, small amounts of COMT are present in the sympathetic nerves of the nictitating membrane and the vas deferens²². COMT is involved mainly in the metabolism of catecholamines released into the circulation²³ and in the inactivation of noradrenaline in tissues with sparse adrenergic innervation²⁴. It also appears to be associated with an extraneuronal uptake mechanismas. Recently we have observed that COMT is present within mammalian erythrocytes. This provided an easily available tissue to examine this enzyme in man. The activity of COMT in erytkocytes is reduced in women with primary affective disorders²⁷.

The discovery of COMT led to the description of other methyltransferases involved in biogenic amine metabolism: histamine-N-methyltransferase²⁸, hydroxyindole-O-methyltransferase²⁹, phenylethanolamine-N-methyltransferase³⁰, and a nonspecific methyltransferase³¹.

Uptake of noradrenaline by sympathetic nerves

Soon after the work on O-methylation was begun, the distribution of [³H]-adrenalin in animal tissues was investigated. Fortunately, Seymour Kety arranged for the synthesis of tritiated noradrenaline and adrenaline of high specific activity labeled on a 7-position. This made possible the administration of physiological amounts of the neurotransmitter and a study of the localization and metabolism of the circulating catecholamine. In collaboration with Weil-Malherbe and Whitby, specific methods for the measurement of adrenaline, noradrenaline and its O-methylated metabolites in tissues were developed. After the intravenous injection of [³H]adrenaline³² or [³H]noradrenaline³³ in cats, these catecholamines were rapidly and unequally distributed in tissues. The amines were selectively taken up in tissues heavily innervated with sympathetic nerves (heart, spleen). Since negligible amounts of

[³H]catecholamines were present in the brain, a blood-brain barrier to these compounds was indicated. O-methylated metabolites, [³H]metanephrine and [³H]normetanephrine, also occurred in tissues. When tissues were examined two hours following the administration of the catecholamines, long after the physiological effects had disappeared, they were found to have almost the same levels of [³H]adrenaline and [³H]noradrenaline as those found after two minutes. These experiments suggested that noradrenaline and adrenaline were taken up and retained in tissues in a physiologically inactive form. The selective binding of the catecholamines by tissues with a high adrenergic innervation pointed to the sympathetic nerves as the sites of retention. To examine this possibility the superior cervical ganglia of cats were removed unilaterally and sufficient time (7 days) was allowed to elapse for complete degeneration of the sympathetic nerves fibers. [³H]Noradrenaline was then given intravenously and the animals were killed one hour later and the [³H]catecholamine content was examined in structures innervated by the sympathetic cervical ganglia³⁴. There was a sharp reduction in the uptake of [³H]noradrenaline in the chronically denervated structures (Table 2). These results made it apparent that sympathetic nerve endings take up and retain the circulating catecholamine.

Table 2

Lack of uptake of [³H]noradrenaline after chronic denervation of the sympathetic nerves

	Chronic denervation		Acute denervation	
	Denervated	Innervated	Denervated	Innervated
Salivary gland	5	42	76	89
Lachrymal gland	3	45	—	—
Retractor muscle	2	11	13	13
Ocular muscle	6	48	25	26

Right superior cervical ganglia were removed from 6 cats. After 7 days cats were given 25 $\mu\text{g}/\text{kg}$ [³H]noradrenaline and the [³H]catecholamine assayed in innervated and denervated structures one hour later. In the acute denervation experiments right superior ganglia were removed 15 minutes before the administration of [³H]noradrenaline. Results are expressed as $\text{m}\mu\text{g}$ [³H]noradrenaline per g tissue³⁴.

To localize the intranenronal site of the noradrenaline retention, combined electron microscopy and autoradiography were carried out. [³H]Noradrenaline was injected; 30 minutes later the pineal was prepared for autoradiography and electron microscopy³⁵. The pineal gland was chosen because of

its rich sympathetic innervation. Electron microscopy showed a striking localization of photographic grains overlying non-myelinated axons which contained granulated vesicles of about 500 Å.

With Potter attempts were made to isolate the dense core vesicles associated with the [³H] noradrenaline^{36,37}. Previously Von Euler and Hillarp had isolated a high-speed noradrenaline-containing granular fraction from bovine splenic nerves³⁸. Again, [³H] noradrenaline was injected in rats and subcellular fractions of the heart and other tissues were separated in a continuous sucrose gradient³⁶. The predominant peak of the [³H]noradrenaline together with the endogenous catecholamine coincided with the "microsomal band" (Fig. 2). The noradrenaline containing particles had no pressor action unless they were lysed in dilute acid, suggesting that the catecholamine was bound. In addition to [³H] noradrenaline, the microsomal peak also contained large amounts of dopamine- β -oxidase³⁷, the enzyme that converts dopamine to noradrenaline. Further attempts to purify noradrenaline containing vesicles were unsuccessful.

The ability to take up and store [³H] noradrenaline enabled Hertting and me to label the neurotransmitter in the nerve endings of tissues and to study its fate on liberation from sympathetic nerves³⁹. Cats were given [³H]noradren-

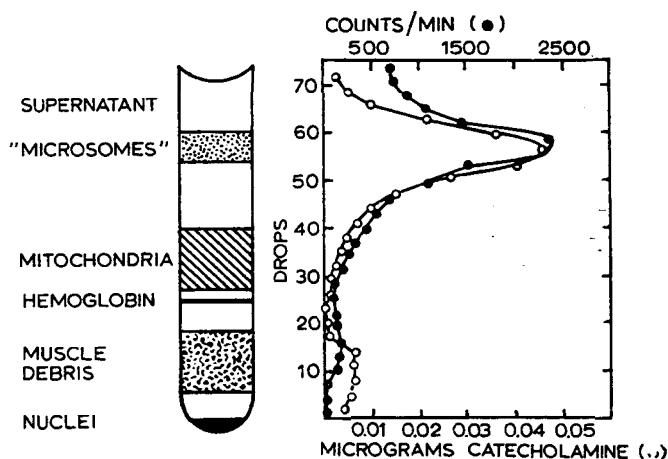


Fig. 2. Subcellular distribution of noradrenaline in the rat heart. Sprague-Dawley male rats were given 50 μ Ci [³H]noradrenaline and were killed 30 minutes later. The hearts were rapidly removed and homogenized in isotonic sucrose. A portion was layered on an exponential sucrose gradient and centrifuged in a Spinco preparative centrifuge using an SW39 rotor for 30 minutes³⁶. Drops were collected with a needle through the bottom of the tube and assayed for ³H and endogenous noradrenaline.

aline; the spleen, containing nerve endings labelled with [³H]noradrenaline, was perfused; the splenic nerve was stimulated, as described by Brown and Gillespie⁴⁰; and the radioactive catecholamine and its metabolites were measured in the venous outflow. After each series of stimulations a marked increase occurred in the concentration of [³H]noradrenaline in the venous outflow. There was also a small but measurable elevation of the O-methylated metabolite, normetanephrine, but no increase in deaminated metabolites. From these experiments we concluded that noradrenaline liberated from the nerve terminals was inactivated by several mechanisms. Part is discharged into the bloodstream; part is O-methylated by COMT, and part is taken up by the nerve terminals. Reuptake of noradrenaline by sympathetic neurones was examined in experiments performed with Rosell and Kopin⁴¹ using the vascular bed of the dog gracilis muscle *in situ*. The sympathetic nerves of the gracilis muscle were labeled by an infusion of [³H]noradrenaline, and the discharge of [³H]noradrenaline measured after nerve stimulation. When the vasomotor nerves were stimulated, an initial reduction in the outflow of [³H]noradrenaline was followed by a rise in outflow of the radioactive catecholamine (Fig. 3). The lag in the outflow was due to an increase in vascular resistance. This observation indicates a reduced capacity of the vascular bed to carry away the released noradrenaline. After the stimulus was ended, decline in [³H]noradrenaline outflow and return of the peripheral resistance were parallel. To block the constriction of the vascular bed, dogs were pretreated

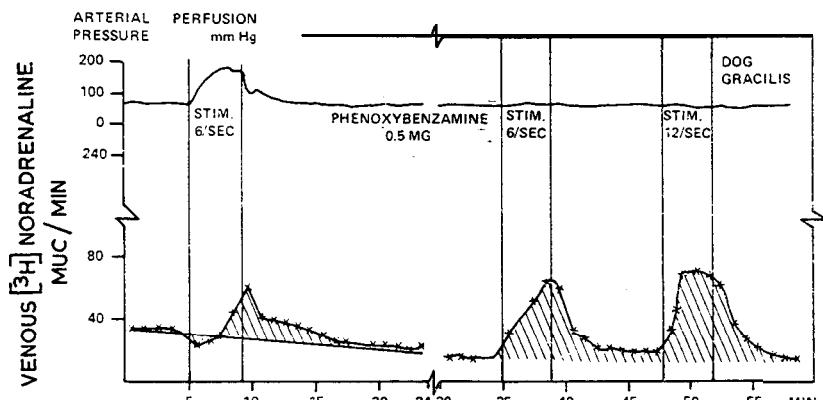


Fig. 3. Uptake and release of [³H]noradrenaline in dog gracilis muscle. Dog gracilis muscle was perfused with [³H]noradrenaline as described by Rosell Kopin and Axelrod⁴¹. Peripheral resistance and venous outflow of [³H]noradrenaline was measured during sympathetic nerve stimulation, before and after treatment with phenoxybenzamine.

with phenoxybenzamine, an adrenergic blocking agent shown to inhibit reuptake of noradrenaline. Vasomotor stimulation resulted in an immediate and larger increase in noradrenaline outflow. The larger and immediate outflow of noradrenaline was due to a blockade of noradrenaline reuptake by phenoxybenzamine. It was concluded from this and other experiments that reuptake by sympathetic nerves was a major mechanism for terminating the actions of the neurotransmitter noradrenaline. Subsequent work by several investigators, particularly Iversen^{42, 43} described the properties of the neuronal uptake mechanism. It obeys saturation kinetics of the Michaelis-Menten type: it is stereospecific for the Z-isomer of noradrenaline and requires Na⁺. Many other amines structurally related to noradrenaline can be taken up and stored in sympathetic nerves by a neuronal uptake process. The fate of noradrenaline at the sympathetic nerve terminal and circulation is shown in Fig. 4.

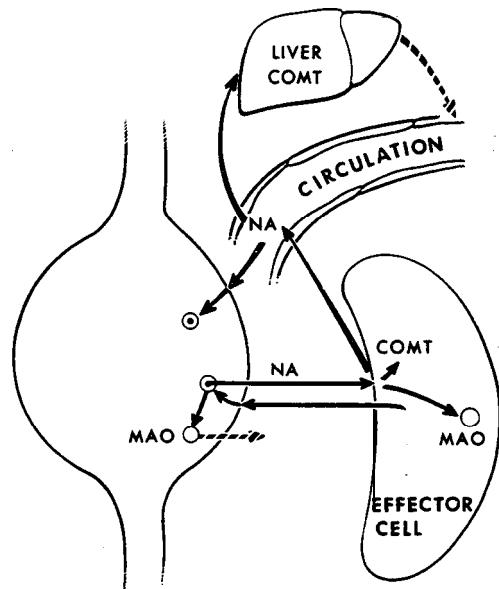


Fig. 4. Fate of noradrenaline (NA) at a varicosity of the sympathetic nerve terminal. COMT is catechol-O-methyltransferase; MAO is monoamine oxidase.

Noradrenaline can also be taken up by an extraneuronal process^{44,45} which has been shown⁴⁶ to be similar to Iversen's Uptake 2. This uptake is inhibited by adrenergic blocking agents and normetanephrine²⁵. Compounds such as isoproterenol which have a low affinity for intraneuronal uptake and a high affinity for extraneuronal uptake may be inactivated by the later process. Extraneuronal uptake operates at all concentrations of catecholamines⁴⁷ and serves

to transport amines into non-neuronal tissues in which they are subsequently metabolized.

Effect of drugs on neuronal uptake

The ability of the sympathetic nerves to take up [³H]catecholamines provided a relatively simple technique for studying the effect of a variety of drugs acting on the sympathetic nervous system. In early experiments of this kind, mice were treated with a variety of drugs and the rate of disappearance of [³H]-adrenaline was measured⁴⁸. A wide variety of drugs (imipramine, chlorpromazine, cocaine, reserpine, amphetamine, tyramine) increased the rate of disappearance of the catecholamine. Such experiments suggested that these drugs might increase the rate of metabolism by interfering with the binding and/or uptake of the catecholamines, thus exposing them to enzymatic attack by COMT or monoamine oxidase. This suggestion was supported by the observation that the catechol quercitrin markedly slowed catecholamine metabolism *in vivo*, presumably by inhibiting COMT.

The experiments that followed were more direct. Cats were treated with cocaine, and then [³H]noradrenaline was injected intravenously. One hour later, heart, spleen and adrenal gland were examined for [³H]noradrenaline⁴⁹. Cats pretreated with cocaine showed a dramatic decrease in tissue [³H]noradrenaline. In addition, there was a sharp elevation in plasma levels of [³H] - noradrenaline in cocaine-treated animals. This experiment revealed that cocaine markedly reduces the uptake of noradrenaline in tissues, presumably the sympathetic neurone. The inhibition of uptake by cocaine thus raised the extraneuronal concentration of noradrenaline (as reflected by the elevated level in plasma catecholamine). By blocking uptake into the nerves, cocaine caused an elevated concentration of noradrenaline to reach the receptor (Fig. 5), and this explains the effect of the drug and denervation of sympathetic nerves in producing supersensitivity.

Experiments similar to those described with cocaine were carried out with other drugs^{50,51}. The following compounds lowered the concentration of [³H]noradrenaline in tissues: imipramine, chlorpromazine, tyramine, amphetamine, guanethedine, reserpine and phenoxybenzamine. All of these drugs also elevated the initial blood level of the [³H]catecholamine. Such observations indicate that these drugs also interfere with the uptake of noradrenaline into the adrenergic neurone.

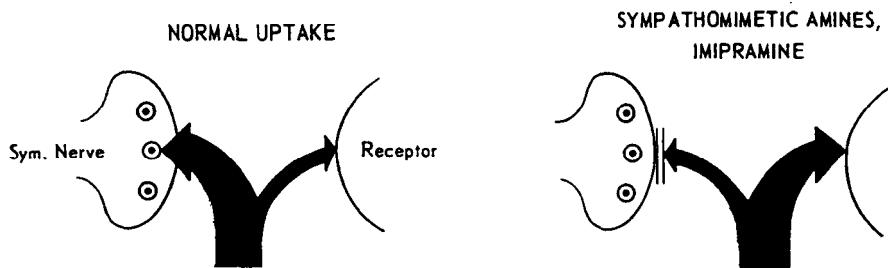


Fig. 5. Effect of drugs on uptake of noradrenaline at the sympathetic nerve terminal.

In addition to blocking uptake, these drugs could also prevent the storage or release of the bound [³H]noradrenaline. If a drug prevents noradrenaline uptake, it should lower the tissue levels of [³H]noradrenaline only when given before the [³H]catecholamine. If it reduces the concentration when given after [³H]noradrenaline, when the neurotransmitter is bound to tissue, then it releases the catecholamine. To distinguish between these two possibilities rats were given drugs before or after the intravenous injection of [³H]noradrenaline, and the amount of the [³H]catecholamine in the heart was measured⁵². Reserpine, amphetamine and tyramine reduced the [³H]catecholamine after it was bound. Pretreatment with imipramine (Fig. 6) or chlorpromazine lowered the concentration of cardiac [³H]noradrenaline only when given before [³H]noradrenaline, indicating that these drugs blocked uptake but did not

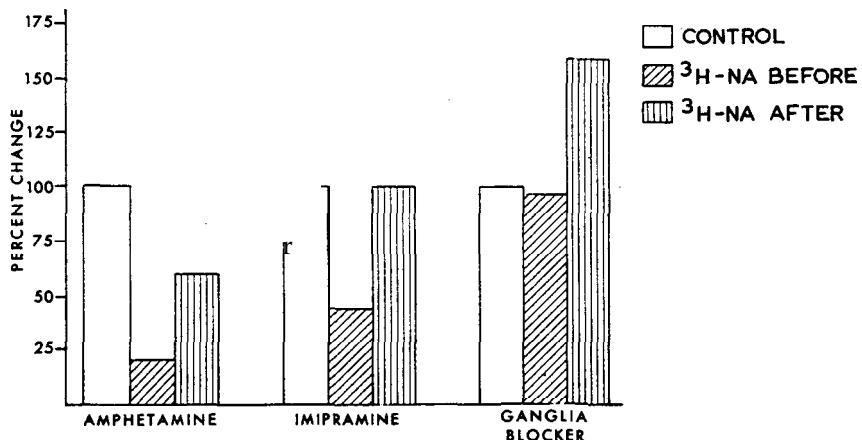


Fig. 6. Effect of drugs in uptake and release of [³H]noradrenaline in the rat heart. Rats were given 15 μ Ci [³H]noradrenaline 30 min before or after the administration of drugs and killed 24 h later. The hearts were examined for [³H]noradrenaline remaining^{52,54}. The ganglia blocker was chlorisondamine.

release the amine. Amphetamine caused a greater reduction when given before [³H]noradrenaline than after (Fig. 6), indicating that it not only blocks uptake but also releases the catecholamine. Many of these observations were confirmed and extended through direct visualization of the sympathetic neurone by histofluorescent techniques⁵³.

[³H]Noradrenaline was also used to measure the effect of drugs in blocking the spontaneous release of the neurotransmitter. Long-lasting ganglionic blocking agents (chlorisondamine; Fig. 6) and bretylium inhibited the spontaneous release of [³H] noradrenaline from the rat heart^{54,55}. Decentralization of the superior cervical ganglion also slowed spontaneous release of [³H]noradrenaline, again demonstrating that nerve impulses cause a release of the [³H]noradrenaline⁵⁴.

*Uptake, storage, release and metabolism of [³H]noradrenaline
in the rat brain*

In 1954 Vogt demonstrated the presence of noradrenaline in the brain and showed that it was unequally distributed^{5,6}. Drugs such as amphetamine and reserpine⁵⁶⁻⁵⁸ lowered the tissue concentration of endogenous noradrenaline, whereas monoaminoxidase inhibitors elevated the level of the catecholamine⁵⁸. In our earlier work we were unable to study the disposition of noradrenaline in the brain because of its inability to cross the blood-brain barrier³³. In 1964 Jacques Glowinski devised a technique for introducing [³H] - noradrenaline into the rat brain via the lateral ventricle⁵⁹. This provided a means of labeling the brain stores of noradrenaline and enabled us to study the fate of this compound in the brain and examine the effect of drugs. The initial concern was whether this exogenously administered [³H]noradrenaline mixed with the endogenous pool of brain catecholamines. We first examined the distribution of the [³H]noradrenaline in various brain areas. After an intraventricular injection, [³H]noradrenaline was selectively distributed in areas which contained high concentrations of catecholamines, the highest levels occurred in the hypothalamus and the lowest in the cerebral cortex and cerebellum⁶⁰. However, considerable amounts of [³H]noradrenaline were present in the corpus striatum, which normally contains high levels of dopamine and little endogenous noradrenaline. In autographic studies intense labeling was also found in the periventricular and ventromedial nuclei of the hypothalamus, medial forebrain bundle, in specific tracts of the spinal cord and in the apical

dendritic layer of the hippocampus. Subcellular distribution studies, using continuous sucrose gradients, showed the [³H] noradrenaline, after its intraventricular administration in the brain, was present in the synaptosomal layer (pinched off nerve endings) together with endogenous noradrenaline. These observations indicated that [³H]noradrenaline mixed to a considerable degree with the endogenous brain stores of the catecholamine. The [³H]noradrenaline persisted in the brain for long periods of time, indicating that it was stored and protected from metabolism. The radioactive metabolites formed were normetanephrine and O-methylated deaminated metabolites. The major product in the brain was [³H]3-methoxy-4-hydroxyphenylglycol.

Labeling of the brain stores of noradrenaline provided an opportunity to study the effects of drugs on the uptake, storage, release and metabolism of noradrenaline in the brain⁶². It was previously shown that imipramine and chlorpromazine blocked the uptake of [³H] noradrenaline in intact peripheral tissues⁵¹ and brain slices⁴². In the intact rat brain, imipramine reduced the accumulation of [³H] noradrenaline after its intraventricular injection⁶³, while chlorpromazine did not (Table 3). Other antidepressant drugs such as des-

Table 3

Antidepressant drugs and the inhibition of uptake of [³H]noradrenaline in the rat brain

<i>Treatment</i>	<i>Clinical antidepressant action</i>	[³ H]Noradrenaline g/brain cpm × 1000
None		30 ± 2.0
Imipramine	Yes	19 ± 1.0 ^a
Desmethylimipramine	Yes	19 ± 1.1 ^a
Amitryptyline	Yes	23 ± 2.1 ^b
Compound 2	No	30 ± 1.6
Compound 3	No	28 ± 1.2
Chlorpromazine	No	32 ± 3.1

<0.001

<0.05

Groups of 6 rats were given drugs (20 mg/kg) intraperitoneally 1 h before the administration of 0.07 µg of [³H]noradrenaline into the lateral ventricle the brain. Rats were killed 2 h later and assayed for [³H]noradrenaline. Compound 2 had the same structure as imipramine except that a dimethyl isopropyl side-chain was substituted for a dimethylaminopropyl side chain. Compound 3 had the same structure as chlorpromazine except that a dimethylaminoethyl ether side-chain was substituted for a dimethylaminopropyl side-chain⁶³.

methylimipramine and amitriptyline reduced the accumulation of [³H]noradrenaline in the brain, but structurally related derivatives of imipramine which are clinically inactive as antidepressants had no effect. Both monoamine oxidase inhibitors and imipramine are antidepressant drugs and cause an increased amount of physiologically active noradrenaline to react with the adrenergic receptors in the brain. Each of these compounds makes more noradrenaline available in the brain by different mechanisms. Imipramine and other tricyclic antidepressant drugs slow inactivation by reuptake into the neurone, and monoamine oxidase inhibitors prevent metabolism of the catecholamine. Amphetamine has multiple actions on the disposition of the catecholamine in the brain⁶². Like tricyclic antidepressants, it blocks uptake into the neurone, causes release of the catecholamine from its storage site, and inhibits monoamine oxidase. Amphetamine administration results in an increased formation of [³H]normetanephrine in brain, whereas reserpine causes an increase in deaminated metabolites. These metabolic changes reflect a release from the neurone of physiologically active noradrenaline by amphetamine and a release of inactive metabolites by reserpine.

Glowinski and Iversen performed a study on metabolism of noradrenaline in different brain regions. They found that all areas of the brain except the skiatum can convert dopamine to noradrenaline⁶⁴. Amphetamine blocked the reuptake of noradrenaline in all brain areas, whereas desmethylimipramine inhibited uptake in cerebellum, medulla oblongata and hypothalamus, but not in the corpus striatum⁶². Rates of turnover of brain noradrenaline were also examined by such different experimental approaches as measuring rates of disappearance of endogenous noradrenaline after inhibiting catecholamine biosynthesis, estimating rates of disappearance of [³H]noradrenaline formed from [³H]dopamine, and determining rates-of disappearance of [³H]noradrenaline after its intraventricular injection. These methods produced results in close agreement with one another. Cerebellum had the fastest turnover and the medulla oblongata and hypothalamus had the slowest turnover⁶⁵. With these techniques subsequent work has established that turnover of brain noradrenaline is altered by a variety of stresses, temperature changes and sleep.

Regulation of the biosynthesis of catecholamines

The catecholamines are synthesized as shown in Fig. 7. This biosynthetic pathway was first proposed by Blaschko⁶⁶ in 1939 and finally established by Uden-

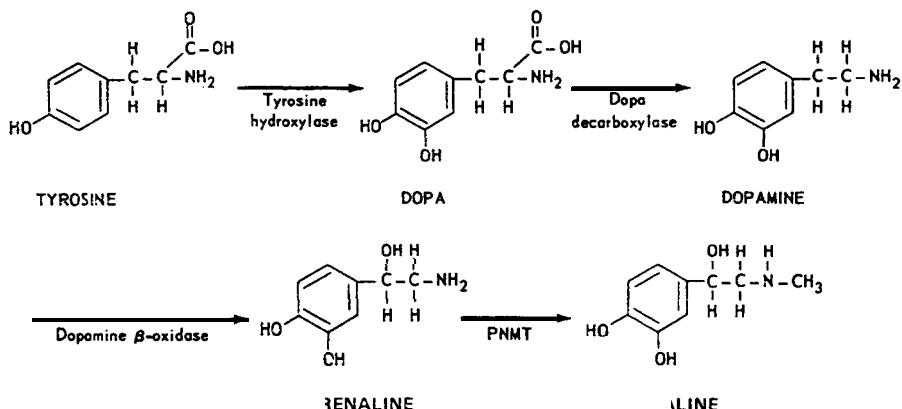


Fig. 7. The biosynthesis of catecholamines. PNMT is phenylethanolamine-N-methyltransferase.

friend and his coworkers⁶⁷. The first step is catalyzed by the enzyme tyrosine hydroxylase⁶⁷, the second by dopa decarboxylase⁶⁸, and the third by dopamine-B-oxidase⁶⁹. These reactions occur within the sympathetic nerve terminal. The final step is catalyzed by phenylethanolamine-N-methyltransferase (PNMT) and occurs almost exclusively in the adrenal medulla⁷⁰. In the adrenal gland the biosynthetic enzymes tyrosine hydroxylase, dopamine- β -oxidase and phenylethanolamine-N-methyltransferase are confined almost entirely to the adrenal medulla.

Noradrenaline in sympathetic nerves and catecholamines (noradrenaline and adrenaline) in the adrenal medulla are in constant flux. They are continuously being released, metabolized, and synthesized, yet they maintain a remarkably constant level in tissues. Recent work in our laboratory and those of others revealed several mechanisms that regulate the biosynthesis of catecholamines, involving long-term hormonal controls as well as short- and long-term neural regulation.

Hormonal control

In species such as dogfish, where the chromaffin tissue is located outside the adrenal gland, little or no adrenaline occurs⁷⁰. In species where the medulla is completely contiguous with the cortex (human and rat) almost all of the catecholamine content is adrenaline. This suggested to Wurtman and me⁷¹ that the adrenal cortex might affect the activity of the adrenaline forming

enzyme phenylethanolamine-N-methyltransferase. I had been working on the properties of this enzyme and developed a sensitive and specific assay for its measurement³⁰. In the initial experiment we measured the effect of hypophysectomy on the phenylethanolamine-N-methyltransferase in the adrenal gland⁷¹. The hypophysectomized rats showed steady fall of the adrenaline-forming enzyme until about 20 percent of the initial concentration remained (Fig. 8). The daily administration of either ACTH (Fig. 8) or dexamethasone for 21 days restored enzyme activity to normal levels in hypophysectomized

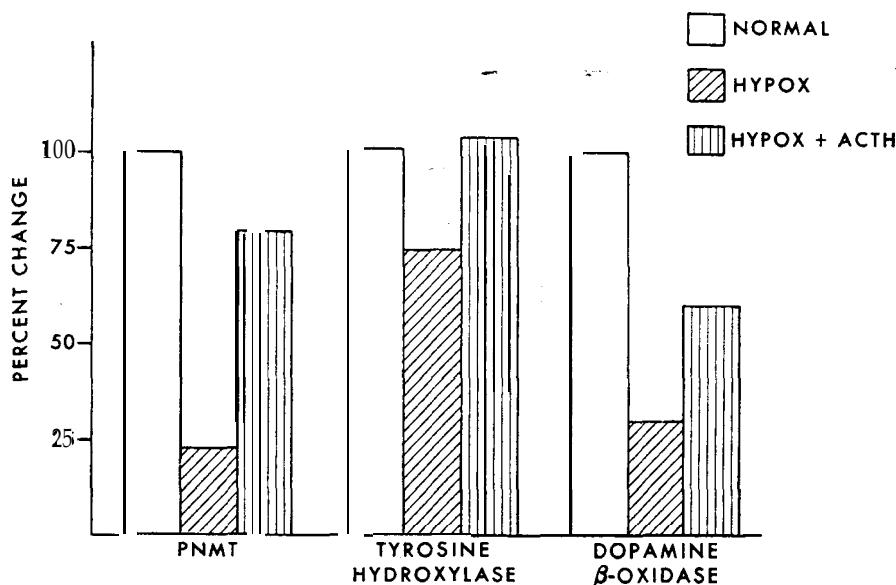


Fig. 8. Control of enzymatic synthesis of adrenaline in the adrenal medulla by ACTH. Phenylethanolamine-N-methyltransferase (PNMT)⁷¹ and dopamine- β -oxidase⁷³ were measured 21 days after hypophysectomy, and tyrosine hydroxylase 5 days after hypophysectomy⁷². ACTH was given, after hypophysectomy, daily for 6 days.

rats. To examine whether the corticoid-induced rise in PNMT was due to increased synthesis of new enzyme protein, dexamethasone was given to rats whose RNA-dependent protein synthesis had been inhibited by puromycin or actinomycin D. Both inhibitors of protein synthesis prevented the rise of enzyme activity caused by dexamethasone. However, repeated administration of ACTH or dexamethasone to intact rats failed to elevate adrenal PNMT activity above normal levels.

In view of the effect of hypophysectomy on the adrenal PNMT activity, the effect on other catecholamine biosynthetic enzymes was examined. After hypophysectomy there was a fall of adrenal-gland tyrosine hydroxylase⁷². (Fig. 8). Enzyme activity was reduced 25 percent in 5 days (Fig. 8) and to about half in 10 days. Repeated administration of ACTH restored tyrosine hydroxylase activity to normal values in hypophysectomized rats (Fig. 8). In contrast to PNMT, dexamethasone did not elevate tyrosine hydroxylase activity in hypophysectomized rats. Again repeated doses of large amounts of ACTH did not increase adrenal tyrosine hydroxylase in normal rats.

Dopamine- β -oxidase (the enzyme that converts dopamine to noradrenaline) activity was also examined in hypophysectomized rats⁷³. This enzyme decreases to about 30 percent of normal values after 21 days (Fig. 3). Administration of ACTH for 5 days caused **dopamine- β -oxidase** activity to increase, but full activity was not reached in this period of time. These observations indicate that the normal maintenance of the catecholamine biosynthetic enzymes in the adrenal glands requires ACTH.

Neural regulation

The biosynthesis of catecholamines in the sympathetic nerves and the adrenal gland is under precise control by nervous mechanisms, one of which is rapid and the other slower. After prolonged stimulation of the splanchnic nerve the sum of the amount of catecholamines released together with the amount remaining in the gland is greater than that initially present in the gland⁷⁴. This indicated that nerve impulses increases the biosynthesis of catecholamines. Weiner and his coworkers using an isolated preparation of the hypogastric nerve of the vas deferens showed that stimulation resulted in an increased synthesis of [¹⁴C]noradrenaline from [¹⁴C]tyrosine, but not from [¹⁴C]dopa⁷⁵. They also found that addition of noradrenaline prevented an increase in [¹⁴C]-catecholamine formation from [¹⁴C]tyrosine. However, stimulation of the vas deferens did not change the total amount of noradrenaline or tyrosine hydroxylase *in vitro*. The fact that noradrenaline is capable of inhibiting the conversion of [¹⁴C]tyrosine to noradrenaline indicated a rapid feedback inhibition at the tyrosine hydroxylase step.

Another type of regulation of catecholamine biosynthesis was uncovered in an unexpected manner. Tranzer and Thoenen⁷⁶ reported that 6-hydroxydopamine selectively destroyed sympathetic nerve terminals. Thoenen decided

to spend a sabbatical year in my laboratory, and together with Mueller we examined the effect of chemical destruction of sympathetic nerve terminals by 6-hydroxydopamine on the biosynthetic enzyme tyrosine hydroxylase. As expected, the enzymes completely disappeared within two days after the administration of 6-hydroxydopamine⁷⁷. However, when the adrenal gland was examined a marked increase in tyrosine hydroxylase was observed. Since 6-hydroxydopamine lowers blood pressure, the increase in enzyme activity caused by this compound might be due to a reflex increase in sympathetic adrenal activity. Consequently we examined the effect of reserpine, which is known to reduce blood pressure and increase preganglionic neuronal activity. Reserpine produced a marked increase in tyrosine hydroxylase activity over several days in the adrenal gland of the rat and several other species, in the superior cervical ganglion (Fig. 9) and in the brainstem of the rabbit^{78,79}. The adrenergic blocking agent phenoxybenzamine also caused a reflex increase in sympathetic adrenal activity. And again the administration of this compound resulted in an elevation in tyrosine hydroxylase activity in the adrenal gland. To examine whether the increased enzyme activity is due to the formation of new enzyme molecules, protein synthesis was inhibited prior to the adminis-

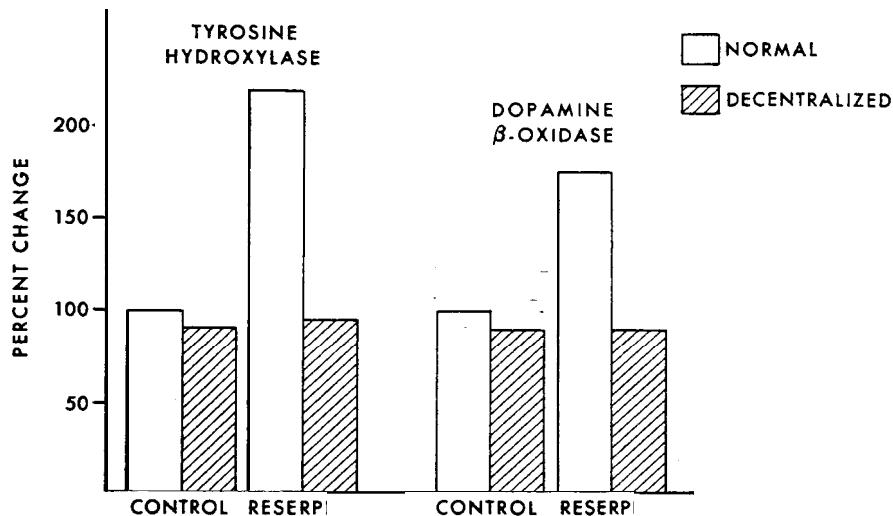


Fig. 9. Transsynaptic induction of noradrenaline biosynthetic enzymes. Right or left superior cervical ganglion was decentralized by transection of the preganglionic trunk from 2 to 6 days before reserpine treatment. Reserpine (5 mg/kg) was given 24 h before tyrosine hydroxylase was assayed in innervated and decentralized ganglia⁷⁸. In the case of dopamine- β -oxidase, reserpine (2.5 mg/kg) was given on 3 alternated days and enzyme examined on the 7th day after decentralization⁸³.

tration of the drugs. Inhibition of protein synthesis with either cycloheximide or actinomycin D prevented the drug-induced increase of tyrosine hydroxylase in the adrenal gland and ganglia⁸⁰. The most likely mechanisms for the increase in enzyme activity might be a blood-borne factor, as in the induction of PNMT by ACTH, or an increase in the activity of the preganglionic neurones. To examine the latter possibility, we cut unilaterally the splanchnic nerve supplying the adrenal gland⁸¹ and preganglionic fibers to the superior cervical ganglion and then administered reserpine⁷⁸. This drug caused the expected rise in tyrosine hydroxylase in the innervated side but the increase in tyrosine hydroxylase on the denervated side was completely prevented (Fig. 9). These results indicate that the increase in tyrosine hydroxylase is due to a transsynaptic induction of the enzyme. Studies on the molecular mechanisms that cause this induction across nerves have thus far proved unsuccessful. The neuronally-mediated induction of tyrosine hydroxylase after reserpine is also observed in the nerve terminals as well as the cell body. However, the increase in tyrosine hydroxylase in the nerve terminals lags behind the ganglia by two or three days⁸². Experiments with inhibitors of protein synthesis point to a local formation of induced tyrosine hydroxylase in the nerve terminals rather than the peripheral movement of the completed enzyme.

Similar studies on the induction of **dopamine- β -oxidase**, an enzyme present in the noradrenaline storage granule, were undertaken with the collaboration of Molinoff, Weinshilboum and Brimijoin⁸³. These experiments were made possible because a very sensitive assay for **dopamine- β -oxidase** was developed. This enzyme could be measured where it never has been found before. Repeated administration of reserpine caused a marked rise in **dopamine- β -oxidase** in rat stellate and superior cervical ganglia (Fig. 9) in the nerve terminals as well as in the adrenal medulla. The elevation of dopamine- β -oxidase in sympathetic ganglia was blocked by protein synthesis inhibitors or by surgical decentralization (Fig. 9). Recently we have found that **dopamine- β -oxidase** is present in the plasma of man and other mammalian species⁸⁴. Preliminary experiments indicate that the circulating **dopamine- β -oxidase** comes from sympathetic nerve terminals. The activity of PNMT in the adrenal gland also increased after reserpine and this elevation in enzyme was blocked by interrupting the splanchnic nerve⁸⁵.

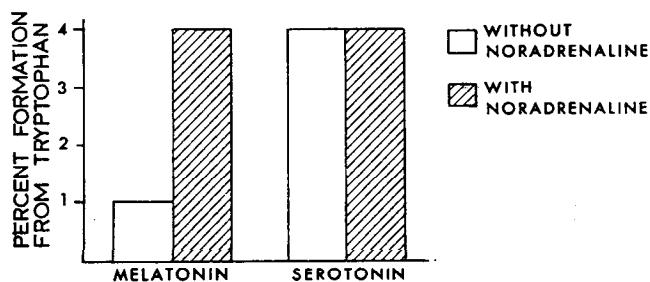
Because of the increasing implications of catecholamines in behavioral changes we examined the effects of psychosocial deprivation and stimulation on the biosynthetic enzymes tyrosine hydroxylase, and PNMT⁸⁶. One group of mice was isolated to prevent visual contact and another was exposed to in-

creased social stimulation by a specially designed cage system. After six months a marked decrease in adrenal tyrosine hydroxylase and PNMT activity occurred in the deprived mice and an increase in both these enzymes was found in the stimulated mice. In related experiments in Kopin's laboratory it was also observed that prolonged forced immobilization in rats also produced a rise in adrenal tyrosine hydroxylase and PNMT activities, and this elevation was abolished by interrupting the splanchnic nerve to the adrenal⁸⁷. All these results suggest that the increase in the catecholamine-forming enzymes in sustained stress may be neuronally mediated and that this response is not immediate, as in the case of a sudden discharge of noradrenaline and adrenaline in states of anger, fear or aggression.

Noradrenaline as a neurochemical transducer in the pineal gland

The pineal gland is exceedingly rich in sympathetic nerve fibers which originate in the superior cervical ganglia⁸⁸. This organ has the unique capacity to synthesize the hormone melatonin (5 -methoxy-N-acetyltryptamine)as follows: tryptophan → 5 - hydroxytryptophan → serotonin + N-acetylserotonin → melatonin⁸⁹. A year before the discovery of melatonin by Lerner⁹⁰ we found an O-methylating enzyme, COMT⁷. This stimulated a search for the enzyme that O-methylates indoles to form melatonin. Such an enzyme was found in the pineal gland and named hydroxyindole-O-methyltransferase²⁹. The enzyme O-methylates N-acetylserotonin to form melatonin, S-adenosylmethionine serving as the methyl donor. At about the same time the enzyme (N-acetyltransferase) that acetylates serotonin to N-acetylserotonin was described⁹¹. The latter enzyme subsequently proved to be critical in the control of melatonin by the adrenergic nervous system. That environmental lighting had something to do with the pineal was suggested by Fiske⁹² in 1961 when she found that continuous light changed the weight of the organ. Consequently rats were placed in continuous darkness or light and activity of the melatonin forming enzyme hydroxyindole-O-methyltransferase in the pineal was measured⁹³. In constant darkness the hydroxyindole-O-methyltransferase activity in the pineal was more than twice as great as that in constant light. Removal of the superior cervical ganglia abolished this difference in enzyme activity⁹⁴. Since noradrenaline is the neurotransmitter of the sympathetic nerves, it might be the agent involved in controlling melatonin synthesis. This possibility was reinforced by the demonstration that levels of nor-

adrenaline in the pineal are markedly influenced by environmental lightin gas. A possible approach to an examination of the mechanism whereby the neurotransmitter could influence the synthesis of melatonin (which occurs outside the neurone) was to use pineals in organ culture. Mainly through the efforts of Shein, we succeeded in growing pineal gland in organ culture. The pineal in organ culture was capable of carrying out all the steps in the formation of melatonin from tryptophan⁹⁶. Inhibition of protein synthesis completely stopped the conversion of tryptophan to melatonin, indicating that new enzyme protein was being formed. Addition of noradrenaline to the culture medium resulted in a sharp increase of melatonin, but not serotonin formation⁹⁷ from tryptophan over a period of 24 h (Fig. 10). However, noradren-



Stimulation of melatonin synthesis in pineal gland by noradrenaline. Culture tubes of pineal gland of rats were incubated with [¹⁴C]tryptophan in the absence or presence of noradrenaline ($3 \cdot 10^{-4} M$). After 24 h the pineal cultures were assayed for [¹⁴C]serotonin and [¹⁴C]melatonin⁹⁷.

aline had only a marginal effect on the hydroxyindole-O-methyltransferase activity. Klein et al. examined the enzyme that converts serotonin into N-acetylserotonin in pineal organ culture. He found that noradrenaline causes remarkable increase in the activity of this enzyme⁹⁸. When protein synthesis was blocked, noradrenaline no longer stimulated the N-acetyltransferase activity. These results show that noradrenaline released from sympathetic nerves stimulates the formation of the pineal hormone melatonin by specifically increasing the synthesis of new N-acetyltransferase molecules.

Concluding remarks

Since the demonstration by Otto Loewi⁹⁹ that sympathetic nerves exert their effects by the release of a chemical substance, numerous advances have oc-

curred. The neurotransmitter has been identified as noradrenaline and its biosynthesis, metabolism and inactivation elucidated. Although the complexities of the storage, release and regulation of noradrenaline and adrenaline have been partially unravelled, much remains to be done. Our understanding of central adrenergic mechanisms is still at the early stages but shows great promise for rapid development. Drugs therapeutically effective in the treatment of affective disorders and neurological and cardiovascular diseases have also been shown to influence the uptake, storage, release, formation and metabolism of catecholamines. These findings implicating the peripheral and central sympathetic nervous system have provided insight into the causes and treatment of mental depression¹⁰⁰, Parkinson's disease¹⁰¹ and hypertension¹⁰².

Acknowledgements

For their incisive help in several of these investigations, I thank the many research associates and visiting scientists who worked with me. I am grateful to the National Institute of Mental Health and the National Institutes of Health, Bethesda, Maryland, for providing generous support, superb resources and a stimulating environment.

1. U. S.von Euler, *Acta Physiol. Scand.*, 12 (1946) 73.
2. U.S.von Euler and I.Flooding, *Acta Physiol. Scand.*, 33, Suppl. 118 (1955) 57.
3. J.Axelrod, *J.Pharmacol.Exptl. Therap.*, 110 (1954) 315.
4. J.Axelrod, *J.BioZ.Chem.*, 214 (1955) 753.
5. M.D. Armstrong, A.McMillan and K.N.F.Shaw, *Biochim.Biophys. Acta*, 25 (1957) 422.
6. G. L. Cantoni, *J. Biol.Chem.*, 189 (1951) 203.
7. J.Axelrod, *Science*, 126 (1957) 400.
8. J.Axelrod, S.Senoh and B.Witkop, *J.Biol.Chem.*, 233 (1958) 697.
9. J.Axelrod and R.Tomchick, *J.Biol.Chem.*, 233 (1958) 702.
10. S. Senoh, J.Daly, J. Axelrod and B. Witkop, *J. Am.Chem. Soc.*, 81 (1959) 6240.
11. M. Assicot and C. Bohuon, *Europ. J. Biochem.*, 12 (1970) 490.
12. J.Axelrod and E. S.Vesell, *Mol.Pharmacol.*, 6 (1970) 78.
13. J.Axelrod and M. J.LaRoche, *Science*, 130 (1959) 800.
14. V.R.Knuppen, M. Holler, D.Tilmann and H.Breuer, *Z.Physiol.Chem.*, 350 (1969) 1301.
15. B.Belleau and J.Burba, *J.Med.Chem.*, 6 (1963) 755.
16. D. W. Wylie, S.Archer and A.Arnold, *J.Pharmacol.Exptl.Therap.*, 130 (1961) 239.

17. J.D.Mann, H.M.Fales and S.H.Mudd, *J.Biol.Chem.*, 238 (1963) 3820.
18. J. Axelrod, W. Albers and C. D. Clemente, *J. Neurochem.*, 5 (1959) 68.
19. G. J. Traiger and D.N.Calvert, *Biochem. Pharmacol.*, 18 (1969) 109.
20. J.K.Inscoe, J.Daly and J. Axelrod, *Biochem.Pharmacol.*, 14 (1965) 1257.
21. J.Axelrod, *Pharmacol.Rev.*, 18 (1966) 95.
22. B. Jarrot, *J.Neurochem.*, 18 (1971) 17.
23. I. J.Kopin, *Pharmacol. Rev.*, 16 (1964) 179.
24. J. A. Levin and R. F. Furchtgott, *J.Pharmacol.Exptl. Therap.*, 172 (1970) 310.
25. A. J.Eisenfeld, L.Landsberg and J.Axelrod, *J.Pharmacol.Exptl. Therap.*, 158 (1967) 378.
26. J.Axelrod and C.K.Cohn, *J.Pharmacol.Exptl. Therap.*, 176 (1971) 650.
27. C.K.Cohn, D.Dunner and J.Axelrod, *Science*, 170 (1970) 1323.
28. D.D.Brown, R.Tomchick and J.Axelrod, *J.Biol.Chem.*, 234 (1959) 2948.
29. J.Axelrod and H. Weissbach, *J.Biol.Chem.*, 236 (1961) 211.
30. J.Axelrod, *J.Biol.Chem.*, 237 (1962) 1657.
31. J. Axelrod, *J.Pharmacol.Bxptl. Therap.*, 138 (1962) 28.
32. J.Axelrod, H. Weil-Malherbe and R.Tomchick, *J.Pharmacol.Exptl. Therap.*, 127 (1959) 251.
33. L. G. whitby, J. Axelrod and H. Weil-Malherbe, *J. PharmacolExptl. Therap.*, 132 (1961) 193.
34. G.Hertting, J. Axelrod, I. J.Kopin and L.G. Whitby, *Nature*, 189 (1961) 66.
35. D.E. Wolfe, L.T.Potter, K. C.Richardson and J. Axelrod, *Science*, 138 (1962) 440.
36. L.T.Potter and J.Axelrod, *J.PharmacolExptl. Therap.*, 140 (1963) 199.
37. L.T.Potter and J.Axelrod, *J.Pharmacol.Exptl. Therap.*, 142 (1963) 291.
38. U. S.von Euler and N.-A.Hillarp, *Nature*, 177 (1956) 44.
39. G.Hertting and J.Axelrod, *Nature*, 192 (1961) 172.
40. G.L.Brown and J.S.Gillespie, *J.Physiol. (London)*, 138 (1957) 81.
41. S.Rosell, I.J.Kopin and J. Axelrod, *Am. J.Physiol.*, 205 (1963) 317.
42. H. J. Dengler, I. A. Michaelson, H. E. Spiegel and E. Titus, *Intern. J. Neuropharmacol.*, 1 (1962) 23.
43. L.L.Iversen, *The Uptake and Storage of Noradrenaline in Sympathetic Nerves*, Cambridge University Press, London, 1967.
44. A. J. Eisenfeld, J. Axelrod and L.R. Krakoff, *J. Pharmacol.Exptl. Therap.*, 156 (1967) 107.
45. J. S. Gillespie, D. N.H. Hamilton and R. J. A.Hosie, *J. Physiol. (London)*, 206 (1970) 563.
46. L. L. Iversen, *Brit. J.Pharmacol.*, 25 (1965) 18.
47. S. L. Lightman and L. L. Iversen, *Brit. J.Pharmacol.*, 37 (1969) 638.
48. J. Axelrod and R.Tomchick, *Nature*, 184 (1959) 2027.
49. L. G. Whitby, G. Hertting and J. Axelrod, *Nature*, 187 (1960) 604.
50. J. Axelrod, L. G. Whitby and G. Hertting, *Science*, 133 (1961) 383.
51. G.Hertting, J.Axelrod and L.G.Whitby, *J.Pharmacol.Exptl.Therap.*, 134 (1961) 146.
52. J. Axelrod, G. Hertting and L.Potter, *Nature*, 194 (1962) 297.
53. T.Malmfors, *Acta Physiol. Scand.*, 64, Suppl. 248 (1965) 1.

54. G.Hertting, L.T. Potter and J. Axelrod, *J.Pharmacol.Exptl. Therap.*, 136 (1962) 289.
55. G.Hertting, J.Axelrod and R.W.Patrick, *Brit.J.Pharmacol.*, 18 (1962) 161.
56. M.Vogt, *J.Physiol. (London)*, 123 (1954) 451.
57. A. Carlsson, E.Rosengren, A. Bertler and J. Nilsson, in S. Garattini and V. Ghetti (Eds.), *PsychotropicDrugs*, Elsevier, Amsterdam, 1957, p. 363.
58. B.B.Brodie, S. Spector and P.A. Shore, *Ann.N.Y.Acad.Sci.*, 80 (1959) 609.
59. J.Glowinski, I.J.Kopin and J.Axelrod, *J.Neurochem.*, 12 (1965) 25.
60. J. Glowinski and J. Axelrod, *Pharmacol. Rev.*, 18 (1966) 775.
61. J.Glowinski, S.H. Snyder and J. Axelrod, *J.Pharmacol. Exptl. Therap.*, 152 (1966) 282.
62. J. Glowinski, J. Axelrod and L. L. Iversen, *J.Pharmacol.Exptl. Therap.*, 153 (1966) 30.
63. J. Glowinski and J. Axelrod, *Nature*, 204 (1964) 1318.
64. J. Glowinski and L.L. Iversen, *J.Neurochem.*, 13 (1966) 655.
65. L.L.Iversen and J.Glowinski, *J.Neurochem.*, 13 (1966) 671.
66. H.Blaschko, *J.Physiol. (London)*, 96 (1939) 50P.
67. T. Nagatsu, M.Levitt and S. Udenfriend, *J.Biol.Chem.*, 239 (1964) 2910.
68. P. Holtz, R. Heise and K. Ludtke, *Arch.Exptl.Pathol.Pharmakol.*, 191 (1938) 87.
69. S.Kaufman and S.Friedman, *Pharmacol. Rev.*, 17 (1965) 71.
70. R.E. Coupland, *J.Endocrinol.*, 9 (1953) 194.
71. R.J.Wurtman and J.Axelrod, *J.Biol.Chem.*, 241 (1966) 2301.
72. R. A. Mueller, H. Thoenen and J. Axelrod, *Endocrinology*, 86 (1970) 751.
73. R. Weinshilboum and J. Axelrod, *Endocrinology*, 87 (1970) 894.
74. S.Bygdeman and U.S.vonEuler, *ActaPhysiol.Scand.*, 44 (1958) 375.
75. N. Weiner and M.Rabadjija, *J.Pharmacol.Exptl. Therap.*, 160 (1968) 61.
76. J. P. Tranzer and H. Thoenen, *Experientia*, 24 (1968) 115.
77. R. A. Mueller, H. Thoenen and J. Axelrod, *Science*, 163 (1969) 468.
78. H. Thoenen, R. A. Mueller and J. Axelrod, *Nature*, 221 (1969) 1264.
79. R. A. Mueller, H. Thoenen and J. Axelrod, *J.Pharmacol. Exptl. Therap.*, 169 (1969) 74.
80. R. A.Mueller, H. Thoenen and J. Axelrod, *Mol. Pharmacol.*, 5 (1969) 463.
81. H.Thoenen, R.A.Mueller and J. Axelrod, *J.Pharmacol.Exptl. Therap.*, 169 (1969) 249.
82. H. Thoenen, R. A.Mueller and J. Axelrod, *Proc.Natl. Acad. Sci. (U.S.)*, 65 (1970) 58.
83. P.B.Molinoff, W.S.Brimijoin, R.M.Weinshilboum and J.Axelrod, *Proc.Natl. Acad.Sci.(U.S.)*, 66 (1970) 453.
84. R. M. Weinshilboum and J. Axelrod, *Pharmacologist*, 12 (1970) 214.
85. H.Thoenen, R. A.Mueller and J.Axelrod, *Biochem.Pharmacol.*, 19 (1970) 669.
86. J. Axelrod, R. A. Mueller, J.P.Henry and P.M. Stephens, *Nature*, 22 (1970) 1059.
87. R. Kvemansky, V.K. Weise and I. J. Kopin, *Endocrinology*, 87 (1970) 744.
88. J.A. Kappers, *Z.Zellforsch.Mikroskop. Anat.*, 52 (1960) 163.
89. R.J.Wurtman, J.Axelrod and D.-E. Kelly, *The Pineal*, Academic Press, New York, 1968.
90. A.B.Lerner, J.D.Case and R.V.Heinzelman, *J.Am.Chem.Soc.*, 81 (1959) 6084.
91. H. Weissbach, B.G.Redfeld and J.Axelrod, *Biochim.Biophys.Acta*, 54 (1961) 190.
92. V. M. Fiske, K. Bryant and J. Putnam, *Endocrinology*, 66 (1960) 489.

93. R. J. Wurtman, J.Axelrod and L. S.Phillips, *Science*, 142 (1963) 1071.
94. R. J. Wurtman, J. Axelrod, E. W. Chu and J. E. Fischer, *Endocrinology*, 75 (1964) 266.
95. R. J. Wurtman, J. Axelrod, G. Sedvall and R.Y. Moore, *J.Pharmacol.Exptl.Therap.*, 157 (1967) 487.
96. R. J. Wurtman, F.Larin, J.Axelrod and H.M. Shein, *Nature*, 217 (1968) 953.
97. J. Axelrod, H. M. Shein and R. J. Wurtman, *Proc. Natl. Acad. Sci. (U.S.)*, 62 (1969) 544.
98. D.C.Klein and J. Weller, *Federation Proc.*, 29 (1970) 615.
99. O.Loewi, *Arch.Ges.Physiol.*, 189 (1921) 239.
100. J.J.Schildkraut and S.S.Kety, *Science*, 156 (1967) 21.
101. O.Hornykiewicz, *Pharmacol. Rev.*, 18 (1966) 929.
102. J. deChamplain, L.R.Krakoff and J. Axelrod, *Circulation Res.*, 24, Suppl. 1 (1969) 75.

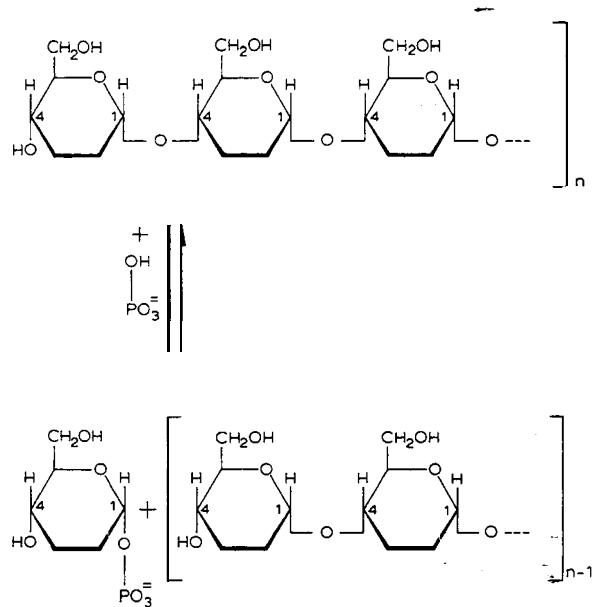
CARLF. CORI and GERTY T. CORI

Polysaccharide phosphorylase

Nobel Lectures, December 11, 1947

Part I - by Carl F. Cori

Polysaccharide phosphorylase is characterized as an enzyme which can break or make an α -1-4-glucosidic bond at the termination (non-reducing end) of a glycogen or starch chain. The process is illustrated below:



The interaction of phosphate with the terminal glucosidic bond results in the formation of glucose-1-phosphate and the loss of a chain unit; in the reverse reaction the glucose part of glucose-1-phosphate is added as a new chain unit and phosphate is set free. This reversible enzymatic polymerization occurs with little change in free energy, as may be calculated from the equilibrium constant. The reaction which involves expenditure of energy in the conversion of glucose to glycogen is the hexokinase reaction, the formation of glucose-6-phosphate from glucose and adenosine triphosphate.

For the discussion which follows it is important to note that the phosphorylated hexoses can enter into the following enzymatic equilibria (Table 1)

Table 1. Enzymatic equilibria at pH 7 at 30°.

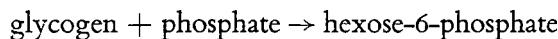
	glycogen + PO ₄ ⁻	Concentration in resting muscle (mole per kg)
Phosphorylase	23% ↓ ↑ 77% glucose-1-phosphate	1 × 10 ⁻⁴
Mutase	95% ↓ ↑ 5% glucose-6-phosphate	
Isomerase	30% ↓ ↑ 70% fructose-6-phosphate	3 × 10 ⁻³

Resting mammalian muscle contains about 0.003 mole of hexose-6-phosphate ($\frac{2}{3}$ glucose-6- and $\frac{1}{3}$ fructose-6-phosphate) per kilo. Assuming that the mutase reaction is also close to equilibrium in a resting muscle, we can calculate the concentration of glucose-1-phosphate to be expected as 0.0001 mole per kilo or less than 0.2 per cent of the total acid soluble phosphate content of muscle. It seems clear that the detection and isolation of this intermediate could not have been accomplished without a separation of phosphorylase activity from that of the other enzymes.

Formation of hexose-6-phosphate - The discovery of polysaccharide phosphorylase and glucose-1-phosphate can be traced to systematic work on the formation of hexosed-phosphate in muscle. Of particular importance was the fact that the method used for the determination of hexosed-phosphate consisted of two independent measurements, one based on the reducing power of the compound and the other on its phosphate content and that there was generally good agreement between these two measurements. In this manner it was found that a number of procedures led to an increase in the hexosed-phosphate content of muscle, among which may be listed, anaerobiosis, injection of epinephrine in intact animals, incubation of isolated frog muscle in Ringer's solution containing epinephrine, and electric stimulation of mammalian or frog muscle.

Balance experiments during aerobic recovery of previously stimulated and isolated frog muscle indicated that the hexose-6-phosphate which disappeared was in large part reconverted to glycogen; hence it was made probable that the reaction, glycogen → glucose-6-phosphate, was reversible. The next step was the finding that the increase in hexosed-phosphate in isolated frog muscle incubated anaerobically with epinephrine was accompanied by

a corresponding decrease in inorganic phosphate (average of 6 experiments per 100 g muscle, + 14 mg ester P, - 16 mg inorganic P). Phosphocreatine and adenosine triphosphate (ATP) remained unchanged, suggesting that they were not involved in the formation of hexose-6-phosphate, but since their regeneration through lactic acid formation was not excluded, the experiments were repeated with muscles poisoned with iodoacetate. The results were the same as with unpoisoned muscle and it was therefore concluded that hexosed-phosphate was formed from glycogen by esterification with inorganic phosphate:



These findings were presented in 1935 at the 15th International Physiological Congress and were discussed at that time with Professor Pamas who then stated that he had under consideration experiments with muscle extract. Prior to that time it has been assumed that glycogen reacted with ATP to form hexose diphosphate. Parnas and Baranowski found that a disappearance of inorganic phosphate could be demonstrated in a cell-free extract of muscle which did not contain phosphocreatine or ATP. This was of importance because it established beyond doubt the participation of inorganic phosphate in the splitting of glycogen, a process which has been aptly called "phosphorolysis" by Parnas. However, the mechanism of phosphorolysis remained unknown until glucose- α -phosphate had been isolated.

Formation of glucose-1-phosphate - The following experiments led to the detection and isolation of glucose-1-phosphate. Minced frog muscle was extracted three times with 20 volumes of cold distilled water, a procedure which removed most of the acid-soluble phosphates normally present in muscle, but did not remove glycogen. When the washed residue was incubated anaerobically at 20° in isotonic phosphate buffer at pH 7.2, some hexose monophosphate was formed. On addition of a catalytic amount of muscle adenylic acid, the formation of hexose monophosphate was very markedly increased. When phosphate was replaced by isotonic KCl, no ester formation occurred. The glucose part of the ester could have come only from glycogen, and the phosphate part only from the added inorganic phosphate, thus confirming the reaction postulated for intact muscle.

After short periods of incubation there was much more organic phosphate present in the hexose monophosphate fraction than corresponded to the reducing power of hexose-6-phosphate. Such a discrepancy had not been

encountered before in analyses of the hexose monophosphate fraction, and since the discrepancy became smaller or disappeared completely after longer periods of incubation, the formation of a precursor of glucose-6-phosphate was suspected. Short hydrolysis in NH_2SO_4 at 100° (conditions under which hexose-6-phosphate is not hydrolyzed) revealed the presence of a compound which yielded equivalent amounts of fermentable sugar and inorganic phosphate.

A representative experiment is shown in Table 2. Comparison of the second, third and the last columns shows that the additional organic phosphate present before hydrolysis is accounted for by this new compound. Furthermore, the disappearance of this compound in the third hour of incubation (-0.74 millimoles) is accounted for by a corresponding gain in reducing power ("hexose") before hydrolysis (+0.67 millimoles).

Table 2. Formation of glucose-1-phosphate in minced and washed frog muscle incubated in phosphate buffer plus adenylic acid.

The water-soluble, alcohol-insoluble barium salts (hexose monophosphate fraction) were isolated and analyzed for phosphate and reducing power before and after hydrolysis in H_2SO_4 for 10 minutes at 100°.

AU-values are given in millimoles per 100 g muscle.

Hours of incuba- tion	Before hydrolysis		After hydrolysis		Difference (Org. P— Inorg. P)
	Hexose	Organic P	Ferment- able sugar*	Inorganic P	
0	0.03	0.03			
1	0.44	1.68	1.22	1.26	0.42
2	0.73	2.22	1.42	1.45	0.77
3	1.40	2.16	0.68	0.65	1.51

* The sugar formed after hydrolysis was completely fermentable, while hexose-6 phosphate under the conditions chosen, was not fermented by the live yeast.

The new phosphate ester was isolated as the crystalline brucine salt in a large-scale experiment similar to that shown in Table 2 and identified as glucose-1-phosphate.

When glucose-1-phosphate was added to a cell-free frog or rabbit muscle extract, it was converted rapidly to glucose-6-phosphate by an enzyme which was named phosphoglucomutase. It was due to the leaking out of the mutase that glucose-1-phosphate accumulated in washed and minced

frog muscle. Mutase is greatly enhanced in its activity by magnesium ions. In order to demonstrate the formation of glucose-1-phosphate from glycogen and inorganic phosphate in muscle extract, it was necessary to remove magnesium ions by dialysis.

An experiment which shows the effect of magnesium ions as well as of adenylic acid is given in Table 3. Addition of magnesium ions to the dialyzed extract had no effect on the total amount of ester formed, but it prevented the accumulation of glucose-1-phosphate.

Experiments similar to those shown in Table 3 were performed with

Table 3. Formation of glucose-1-phosphate in dialyzed (17 hours) rabbit muscle extract.

Incubated for 60 minutes at 24° after addition of glycogen and inorganic phosphate.
All values are given in micromoles per 10 cc. of extract:

<i>Additions</i>				
<i>Aden-</i> <i>ylic</i> <i>acid</i>	<i>MgCl₂</i>	<i>1-ester</i>	<i>6-ester</i>	<i>Total</i> <i>ester</i>
0	0	7.8	15.0	22.8
0	80	0.0	21.7	21.7
5	0	71.2	22.2	93.4
5	80	5.6	86.2	91.8

dialyzed extracts of other mammalian tissues (brain, heart, liver, kidney) and of yeast. In all of these the formation of glucose-1-phosphate could be demonstrated, pointing to a wide distribution of the enzyme phosphorylase. Hanes has described the occurrence of this enzyme in higher plants, particularly in tubers and seeds. In general, the enzyme is present in tissues and cells which contain glycogen or starch.

Properties and synthesis of glucose-1-phosphate - The ester, having no free reducing group, does not react with alkaline copper solutions or with hypoiodite and is resistant to the action of strong alkali. Complete hydrolysis occurs in 10 minutes at 100° in 0.1 N HCl or H₂SO₄ and equivalent amounts of free glucose and inorganic phosphate are formed. The quantitative determination of the ester is based on this property. The neutral barium and potassium salts of the ester are sparingly soluble in 66 per cent alcohol. A crystalline dipotassium salt, containing 2 H₂O, has been described by Kiessling.

The ester has been synthesized by a condensation of α -tetraacetyl glucose-

I-bromide with trisilver phosphate. An intermediate product, tri- (tetraacetyl glucose-I)-phosphate, is formed which yields glucose-I-phosphate on hydrolysis in 0.2 N HCl in methyl alcohol. The synthetic, like the natural product, is the α -isomer. The β -isomer is obtained by substituting dibenzyl phosphate or "monosilver" phosphate as the phosphorylating agent; it is not acted upon by phosphorylase. Neither are the synthetically obtained α -isomers of mannose-I- or galactose-I-phosphate.

Reversibility - The first clue for a possible reversibility of the reaction, glycogen + phosphate \rightarrow glucose-I-phosphate, came from the observation that addition of glucose-I-phosphate to a reaction mixture containing enzyme, glycogen and phosphate was strongly inhibitory, while glucose- α -phosphate had only a weak inhibitory effect on the formation of glucose-I-phosphate. Further investigation showed that conditions for reversibility were unfavorable because the concentration of glucose-I-phosphate could not be maintained, owing to the activity of phosphoglucomutase even in electrodialyzed extracts and at pH 6.5 (which is less favorable for its action than for the action of phosphorylase). It became clear that a separation of the two enzymes was necessary in order to investigate reversibility. A partial separation was first achieved by adsorption of phosphorylase on aluminium hydroxide, followed by elution with disodium phosphate and dialysis to remove inorganic phosphate. When glucose-I-phosphate was added to this enzyme preparation, inorganic phosphate was set free and a polysaccharide was formed in equivalent amounts, showing the reversibility of the reaction. Independently Kiessling had prepared a protein fraction from yeast juice by fractionation with 0.3 saturated ammonium sulfate which also catalyzed the reaction in a reversible manner, while some qualitative observations with yeast extract based on iodine colors had been made earlier by Schäffner and Specht.

An original experiment with a partially purified preparation of muscle phosphorylase is reproduced here (Fig. 1) because it is instructive in relation to later developments. The curve in Fig. 1 shows a definite lag period; the polysaccharide which was formed gave a blue color with iodine and the reaction did not attain a true equilibrium owing to incomplete separation of phosphorylase from phosphoglucomutase. Reversibility could also be demonstrated with phosphorylase preparations of heart and brain. The iodine color of the newly formed polysaccharide was brown to reddish purple rather than blue as with muscle phosphorylase. Preparations of liver phosphorylase formed a polysaccharide which could not be distinguished from glycogen in iodine color and other properties.

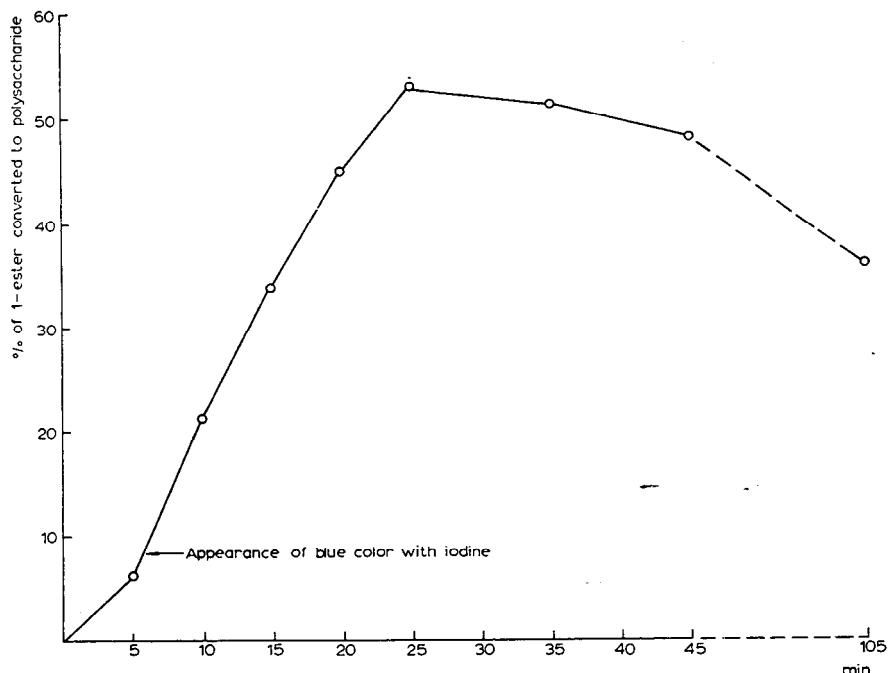


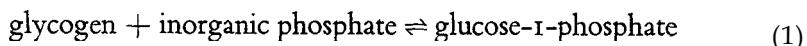
Fig. 1. Synthesis of polysaccharide with a partially purified phosphorylase preparation of muscle. (Experiment of March 11, 1939.)

After these observations had been made, it became clear that further progress depended on the isolation of phosphorylase. This is described in Part 2 which is inserted here in order to avoid an interruption in the sequence of exposition.

Part 2 – by Gerty T. Cori

For a detailed study of the action of phosphorylase and for an understanding of its mechanism, it appeared necessary to work with highly purified enzyme preparations. Muscle was chosen as starting material since in it the concentration of the enzyme is much higher than that found in other tissues. The rapid rate of glycogen breakdown connected with muscular contraction may explain the relatively high concentration in skeletal muscle of phosphorylase as well as of the other enzymes which are concerned with lactic acid formation.

The primer action of glycogen - The protein fraction of a muscle extract, precipitated by less than 0.5 saturation with $(\text{NH}_4)_2\text{SO}_4$, showed a marked rise in phosphorylase activity per unit of protein over the unfractionated starting material. This was however the case only when the enzyme was catalyzing the reaction toward the right:



When enzyme activity was tested in the opposite direction a puzzling difficulty was encountered. Activity set in only after a lag period; refractionation of the enzyme increased this lag period from minutes to hours and in some preparations completely abolished the activity toward polysaccharide formation. A similar observation was made by Kiessling with yeast phosphorylase and led him to conclude that he had separated two enzymes, one concerned with glycogen synthesis, one with its breakdown.

Liver phosphorylase, upon salt fractionation, was found to retain activity toward polysaccharide synthesis. Such preparations always contained traces of glycogen, while the purified muscle enzyme was free of glycogen. This observation offered a clue. Addition of glycogen to the reaction mixture in as low a concentration as 10 mg per cent led to immediate activity of muscle phosphorylase preparations, seemingly inactive when tested without gly-

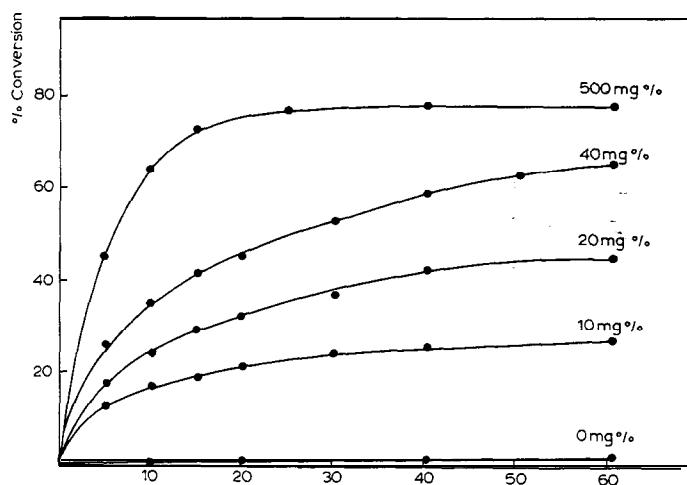


Fig. 2. Rate of conversion of glucose-1-phosphate to polysaccharide in the presence of crystalline muscle phosphorylase and increasing amounts of glycogen.

cogen addition. Fig. 2 illustrates these findings as well as the effect of increasing concentrations of glycogen on the rate of the reaction.

From these observations it followed that glycogen was needed for the activity of the enzyme in both directions. Since the equilibrium of the reaction is in favor of polysaccharide formation, it was of considerable advantage that the enzyme could now be tested when acting in that direction.

Enzyme test - Under certain conditions, Reaction (1) is of the unimolecular type; toward the left, it can be followed conveniently by determination of the amount of inorganic phosphate which is set free. Polysaccharide is formed in equivalent amounts. Since the reverse reaction occurs simultaneously, an equation for a reversible first-order reaction is applicable: $k = I/t \log x_e/(x_e - x)$; x_e represents the per cent of glucose-I-phosphate converted to inorganic phosphate (and polysaccharide) at equilibrium, 79 per cent at pH 6.7, while x represents the per cent converted at time t (in minutes). The position of the equilibrium changes with pH, hence a different value for x_e must be used at other pH values. Hanes has shown, however, that at equilibrium the ratio of the ionic forms of orthophosphoric and glucose-I-phosphoric acid remains constant over the pH range investigated. The mean ratio of the divalent ions of the two acids was found to be 2.2 for potato and 2.0 for muscle phosphorylase between pH 5 and 7.5.

In the standard test the enzyme is diluted with 0.03 M cysteine at pH 6.7. Unless a reducing agent is present the enzyme has low activity and the reaction is not of the first order. To a solution of glycogen, glucose-r-phosphate and adenylic acid (5-phosphoriboside), adjusted to pH 6.7, an equal volume of dilute enzyme solution is added to start the reaction. The order of addition of the different components of the system makes an appreciable difference in the initial rate of the reaction. When enzyme or glucose-1-phosphate is added last, the initial rate is the same, but when glycogen is added last, the initial rate is less and the shape of the rate curve is different.

The final composition of the reaction mixture in micromoles per cc. is : 15 cysteine, 16 glucose-I-phosphate, 0.5 adenylic acid and glycogen (10 mg per cc.). Inorganic phosphate is determined in an aliquot after 5 and 10 minutes of incubation at 30°, and the data are used for a calculation of the first-order velocity constant.

The number of enzyme units present are expressed as k multiplied by 1,000, for convenience. In order to compare different preparations, the activity of the enzyme is expressed as units per mg protein and is calculated for the amount of protein which is present in 1 cc. of reaction mixture.

The pH optimum of the reaction is between 6.5 and 6.8 and the temperature optimum is close to 38°. The conversion of glucose-r-phosphate (pK_2 6.1) to orthophosphate (pK_2 6.8) results in a shift in pH. This change is small and has a negligible effect on the rate during the early course of the reaction. The pH can be kept constant by the addition of various buffers but they all decrease the activity of the enzyme.

Crystalline muscle phosphorylase has an activity of about 3,000 units per mg protein at pH 6.7 and 30°. Calculated for the initial rate of the reaction, this corresponds to a turnover number of 40,000 molecules of glucose-I phosphate per molecule of enzyme (mol. wt. 400,000) per minute.

Phosphorylase a and b - It had been observed in experiments with dialyzed and aged muscle extracts that the rate of phosphorylase activity was increased 10 times or more by the addition of adenylic acid in low concentrations. Later it was found that phosphorylase preparations obtained by precipitation of a fresh-water extract of rabbit muscle at 0.41 saturation with $(NH_4)_2SO_4$, when freshly dissolved, had in the absence of added adenylic acid as high as 66 per cent of the full activity (i.e. activity in the presence of added adenylic acid). When the enzyme solution was kept for 1 hour at 25° its activity without adenylic acid dropped to zero; at this point addition of adenylic acid to the reaction mixture brought back the original activity.

These observations led to the conclusion that there exist two forms of phosphorylase, one active (phosphorylase *a*) and one inactive (phosphorylase *b*) without adenylic acid addition and that muscle tissue contains a factor, soon shown to be an enzyme, which converts the *a* into the *b* form. The assumption that this enzyme removes adenylic acid from the *a* form seemed justified at this point and the enzyme was designated as PR (prosthetic group removing enzyme). Later work, however, failed to establish the presence of adenylic acid in phosphorylase *a* and the nature of the prosthetic group which is removed by the PR enzyme remains obscure. There is evidence that an organic phosphate which is dialyzable and difficult to hydrolyse in hot acid is split off from phosphorylase *a* by PR.

In order to preserve the *a* form, it is necessary to prevent action of the PR enzyme. *In vivo* this is accomplished by avoiding as much as possible stimulation of the muscles before or during excision; *in vitro*, by the earliest separation of phosphorylase from the PR enzyme.

*Preparation of crystalline phosphorylase *a** - The procedure described below is based on experience gained in a large number of preparations. Information

concerning the various steps is given which has not been included in the original description of the method.

A rabbit is killed by injection of amytal and the muscles of the hind legs and back are rapidly excised and weighed. The following steps are carried out in a cold room at 4°. The muscles are passed through an ordinary meat grinder and extracted with one volume of water for about 10 minutes (extraction in a blender is to be avoided since it leads to conversion of the *a* to the *b* form). After the residue is separated by straining it through gauze it is re-extracted with another volume of water. The combined extracts are filtered through cotton and coarse filter paper; this should be accomplished in 1-2 hours. It is unnecessary to obtain a completely clear extract. The pH of the extract is measured with a glass electrode and if it is above 6.2, it is adjusted to that pH by adding 0.05 N HCl with stirring. The extract is then dialyzed in cellophane tubes (diameter 2.3 cm) against running tap water of 4-10° for 3-4 hours. The turbid extract is collected in a beaker and 0.05 N HCl is added to bring the pH to 5.7. A flocculent precipitate forms which contains most of the PR enzyme and only a small amount of phosphorylase. This is the case only in rabbit muscle. In other species this separation is less complete and so far the enzyme has only been crystallized from rabbit muscle. The isoelectric precipitate is removed by centrifugation followed by filtration. Table 4 shows that no purification is achieved in this step, which nevertheless, because of the separation from the PR enzyme, is essential for the success of the preparation.

To the filtrate which is red and must be perfectly clear, sodium glycero-phosphate or KHCO₃ is added in substance until the pH is 6.8. Then the extract is precipitated by bringing it to 0.41 saturation with a (NH₄)₂SO₄

Table 4. Preparation of crystalline muscle phosphorylase.

	Total protein (mg)	Phosphorylase protein (mg)	(per cent)
1. Muscle (210 g) extracted twice with 210 cc. H ₂ O	4,420	86	2
2. Dialyzed 3 hours, pH adjusted to 5.75, ppt. removed	3,910	71	2
3. Precipitated with 41% saturated (NH ₄) ₂ SO ₄	147	58	39
4. Dialyzed against cysteine-buffer mixture pH 6.8; crystals centrifuged off	58	52	90
5. Mother liquor of crystals	85	5	6

solution saturated at room temperature and neutralized to pH 6.8. Overnight the flocculent precipitate settles to the bottom and most of the supernatant fluid can be decanted. Finally the precipitate, which contains the enzyme is separated by centrifugation, preferably at high speed (10,000 r.p.m.). The well-drained precipitate is dissolved in water, about 3 to 4 ml per 100 g muscle used originally. Table 4 shows that with a small loss the enzyme has now been purified about 20 times and in this particular experiment was 39 per cent pure.

The solution, which is slightly turbid, is transferred to a cellophane tube (diameter 1.3 cm) and after a short dialysis (1/2 to 1 hour) against cold running tap water, the tube is transferred to a cysteine solution (0.005 M) brought to pH 6.6 to 6.8 with 0.03 M sodium glycerophosphate. Buffers other than glycerophosphate have been used successfully, while no crystals were obtained when glutathione was substituted for cysteine or when cysteine was omitted. Dialysis is continued at 0°, against several changes of the same cysteine-buffer solution, until most of the sulfate has been eliminated from the enzyme solution. Precipitation of the enzyme sets in before this is the case; the precipitate is sometimes, but not always, crystalline. It is separated by centrifugation, the supernatant fluid is carefully drained off and



Fig. 3. Crystalline phosphorylase prepared from rabbit muscle ($\times 130$).

the precipitate is stirred into a 0.01 to 0.03 M cysteine-glycerophosphate solution of pH 6.6 to 6.8 at 30 to 35°. The volume should be about the same as in the first crystallization. Solution of most of the protein should be rapidly achieved by vigorous stirring with a glass rod, however foaming should be avoided: the material is then centrifuged at room temperature (25°) at 10,000 r.p.m. for about 5 minutes. If a considerable amount of protein remains undissolved, it can be re-extracted. The small insoluble residue consists mostly of cystine crystals. The supernatant fluid should be perfectly clear and almost colorless. It is immediately transferred to an ice bath. Crystals appear rapidly, the rate depending, among other factors, on the concentration of enzyme in the solution. In order to obtain the large crystals shown in Fig. 3 it is necessary to use a dilute solution and slow cooling.

Table 4 shows that only 2 per cent of the protein in the crude extract was phosphorylase; 60 per cent of the enzyme was recovered in the first crystals. Recrystallization causes a slight rise in specific activity. Eventually the mother liquor has almost the same specific acitivity as the crystals.

During summer, when the temperature in St. Louis is well over 30° on many days, the phosphorylase content of the muscles drops to such a low level that crystallization becomes impossible. When the rabbits are kept for about 1 week at 13° the level rises sufficiently to obtain crystals.

Table 5. Comparison of properties of phosphorylase *a* and *b*.

	<i>Phosphorylase a</i>	<i>Phosphorylase b</i>
Molecular weight calculated from diffusion and ultracentrifugation	4×10^5	
Diffusion constant, $D_{20}(w \times 10^7)$	3.3	3.3
Electrophoretic mobility (sq. cm per volt per sec $\times 10^5$, phosphate buffer pH 7.15, μ o.i.; temperature 2°)	-3.25	-2.75
Isoelectric point	5.5-5.6	5.8
Solubility (pH 7.2, water)	insoluble	soluble
0.1 M KCl (24°)	poorly soluble	»
0.08 M KCl + 0.02 M cysteine (24°)	soluble	»
0.08 M KCl + 0.02 M cysteine (0°)	crystallizes	»
Crystal form	long needles	rhomboid plates
Activity, without adenylic acid (%)	65	none
With adenylic acid (%)	100	80
Dissociation constant for combination with adenylic acid (pH 6.7, 25°)	1.5×10^{-6}	5×10^{-5}
Phosphorus content (%)	0.08	0.02

Phosphorylase *b* was prepared by letting purified PR enzyme act on twice crystallized phosphorylase *a*. The solution was then brought slowly to 0.28 saturation with $(\text{NH}_4)_2\text{SO}_4$. The enzyme crystallized in the form of rhomboid plates. Table 5 gives a comparison of the properties of the two forms of the enzyme.

Determination of phosphorylase *a* and *b* in mixtures is based on two parallel activity determinations, one without and one with the addition of adenylic acid in 10^{-3} M concentration. In the former case the *a* form has 66 per cent of its full activity, while the *b* form is inactive; in the latter case both forms are fully active.

Conversion of phosphorylase a to b, in vitro - The PR enzyme originally obtained by isoelectric precipitation of dialyzed muscle extract, was considerably purified by fractionation with $(\text{NH}_4)_2\text{SO}_4$. It was shown that its activity was greatly increased in the presence of cysteine and that Mn^{++} ions had an activating effect. The conversion of the *a* to the *b* form follows the first-order reaction rate equation. PR enzyme units could thus be expressed in terms of the first-order velocity constant.

A conversion of phosphorylase *a* to *b* could also be effected by crystalline trypsin at pH 6 to 6.2, at which pH the proteolytic acitivity of trypsin is kept at a minimum. This conversion is not a first-order reaction and is not accelerated by Mn^{++} ions. Work by E. Krebs (unpublished) indicates that phosphorylases *b* obtained by PR and trypsin may not be identical.

Conversion of phosphorylase a to b, in vivo - Extracts of resting muscles contain predominantly phosphorylase *a*. During strong muscular contraction produced by strychnine or electric stimulation a large part of the enzyme is converted to the *b* form, presumably through the action of the PR enzyme *in vivo*. The sum of phosphorylase *a* and *b* does not differ significantly in resting and stimulated muscles. Determinations were carried out in crude, and crude dialyzed extracts and in the precipitate obtained with $(\text{NH}_4)_2\text{SO}_4$ at 0.41 saturation (Table 6). No crystals of phosphorylase *a* were obtained when its concentration fell below 25 per cent of the total phosphorylase.

Experiments on frogs gave results very similar to those obtained on rabbits. When electric stimulation was followed by a rest period of 10-15 minutes before the muscles were excised, the extract contained again preponderantly phosphorylase *a*. Conversion of the *b* to the *a* form has so far only been observed *in vivo* and nothing is known about its mechanism.

The progressive conversion of phosphorylase *a* to *b* in muscle as contraction continues might represent a regulatory mechanism, preventing exhaust-

Table 6. Effect of stimulation and recovery on phosphorylase-*a* content of muscle.

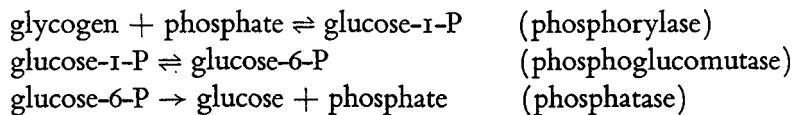
	<i>Phosphorylase a + b per g muscle</i> (units)	<i>Phosphorylase a in crude extract</i> (per cent)	<i>Phosphorylase a in dialyzed (NH₄)₂SO₄ precipitate</i> (per cent)
Rabbits			
12 resting	1,340	92	92
6 strychnine	1,420	10	8
4 electric stimulation	1,140	11	10
Frogs			
14 resting	1,290	86	91
7 electric stimulation	1,310	37	40
4 stimulated, 5' recovery	1,330	43 ←	64
7 stimulated, 10-15' recovery	1,270	82	91

tion of the glycogen stores. The limiting reaction in the conversion of glycogen to lactic acid is the phosphorylation of fructosed-phosphate to fructose-I, 6-diphosphate. Consequently the equilibrium ester (glucose-6- and fructose-6-phosphate) accumulates during muscular contraction. The piling up of even more of this intermediate substance is avoided by a temporary inactivation of phosphorylase. The adenylic acid content of muscle is too low to permit full activation of phosphorylase *b*.

Liver phosphorylase and blood sugar formation - Until recently it was assumed that *a*-amylase plus maltase convert glycogen to glucose in the liver; the absence of dextrins and maltose from liver tissue and blood could not be easily reconciled with this assumption.

Glycogen added to dialyzed extracts of perfused liver tissue disappeared very slowly, indicating weak amylase activity. When inorganic phosphate and traces of adenylic acid were added to such extracts, glycogen disappeared rapidly and its disappearance was balanced by the sum of hexose phosphate and glucose which accumulated.

Fractionation of the liver extracts led to separation of the enzymes involved and it was shown that glucose was formed by the following reactions:



The phosphatase seems to be specific for glucose-6-phosphate; its pH optimum is within the physiological range. The glucose formed diffuses into the blood stream and the phosphate can react again with glycogen.

It is the absence of phosphatase from skeletal muscle tissue which explains the fact that muscle does not contribute glucose to the blood. In the kidney, both the glycogen and phosphorylase content are very low and its contribution of glucose to the blood through the above system seems to be insignificant.

Part 3 - by Carl F. Cori

Plant phosphorylases - When Hanes first described the occurrence of phosphorylase in higher plants, he pointed out that there existed a close parallelism between the action of an enzyme system prepared from peas and potatoes and what was then known about the action of the corresponding enzyme system from animal tissues and from yeast. Hanes fractionated potato extract with ammonium sulfate and noted an initial lag period in the formation of starch from glucose-I-phosphate. In conformity with the results obtained with the animal enzyme he found that this lag period was abolished by the addition of a small amount of starch. Green and Stumpf found their purified potato phosphorylase preparations completely inactive in the direction of synthesis, unless a small amount of starch was added. Weibull and Tiselius showed that under certain conditions the reaction catalyzed by

Table 7. Concentration of substrates at which phosphorylases show half maximal activity.

Enzyme	Substrate	Concentration (M per liter)	Author
Muscle	glucose-1-phosphate	5.7×10^{-3}	Cori, Cori, and Green
Potato	"	2.6×10^{-3}	Weibull and Tiselius
Muscle	inorganic phosphate	6.8×10^{-3}	Cori
Potato	"	6.2×10^{-3}	Weibull and Tiselius
Muscle	glycogen	$1.2 \times 10^{-4}^*$	Cori, Cori, and Green
Potato	starch	$2.4 \times 10^{-4}^*$	Weibull and Tiselius

* Calculated per mole end-group, assuming that glycogen contains 10 per cent and starch 4.5 per cent end-groups.

potato phosphorylase was first order in either direction. The Michaelis-Menten constants of potato and muscle phosphorylase for substrates are very similar (Table 7).

Table 8. Differences in the action of muscle and potato phosphorylase.

Substance	Effect on phosphorylase	
	Muscle	Potato
Glucose (0.1 M)	inhibits	no effect
Cu ⁺⁺ (0.001 M)	inhibits	no effect
Reducing agents	activate	no effect
Adenylic acid	activates	no effect
Polysaccharides as primers	optimally primed by glycogen; not primed by short amylose chains	primed by starch or amylopectin more effectively than by glycogen; primed by short amylose chains

In spite of these and other similarities (nature of polysaccharide formed, effect of pH on equilibrium), there are certain important differences in the action of potato and muscle phosphorylase; these are summarized in Table 8.

Nature of polysaccharide formed - A summary of the properties of synthetic polysaccharides is given in Table 9. It is based on the work of Hanes, Hassid,

Table 9. Properties of natural and synthetic polysaccharides.

	X-ray diagram	Rel. intensity of iodine color at 660 m μ	Hydrolysis by β -amylase	Average chain length	Rel. ability to activate muscle phosphorylase
Starch (corn)	ring pattern	100	60	24-30	45
Amylopectin (corn)	diffuse	50	55	20-25	65
Amylose (corn)	ring pattern	310	100	250	10
Synthetic amylose					
potato phosphorylase	ring pattern	305	98	100	0
muscle phosphorylase	ring pattern	290	97	200	0
Glycogen (liver)	diffuse			12-18	100
Synthetic glycogen					
liver phosphorylase	diffuse				100

Bear, Cori and others. It seems clear that muscle and potato phosphorylase form a linear polysaccharide which closely resembles the amylose component of natural starch in all its properties.

In order to form a branched polysaccharide of the type of amylopectin or glycogen, a second enzyme is needed which forms α -1-6 glucosidic linkages at the points of branching. Such an enzyme has been found in both animal and plant tissues, but the mechanism of its action is not clearly understood. When crystalline muscle phosphorylase plus a second enzyme (called the branching factor and obtained from liver or heart) were allowed to act on glucose-1-phosphate, an autocatalytic type of curve was obtained (Fig.4). Traces of glycogen, introduced by the liver preparation, were present, but not enough to prime muscle phosphorylase when acting alone.

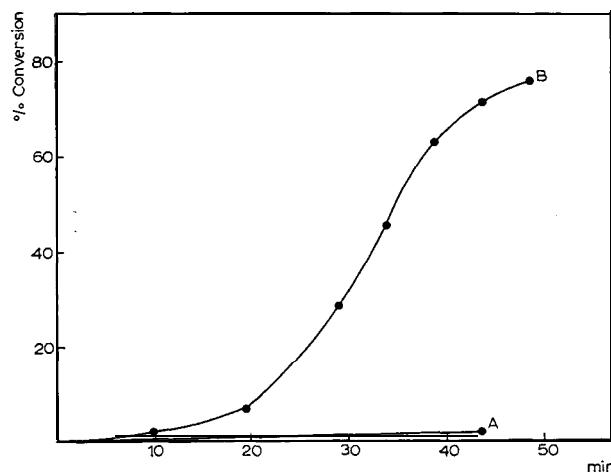


Fig. 4. Enzymatic synthesis of a branched polysaccharide from glucose-1-phosphate. In Curve B crystalline muscle phosphorylase was acting simultaneously with a supplementary enzyme prepared from liver. In Curve A the liver enzyme preparation was inactivated by heating before being added to muscle phosphorylase.

As shown in the last column of Table 9, branched and linear polysaccharides differ markedly in their activating effect on muscle phosphorylase. The autocatalytic type of curve would result from the increasing number of activating end-groups formed during the synthesis of a branched polysaccharide. It was suggested that a combined action of both enzymes was necessary for the formation of glycogen or amylopectin. Haworth and collaborators have, however, reported the conversion of amylose to amylopectin

by the action of an enzyme prepared from potato, a sort of cross-linking enzyme which would break the long amylose chains into smaller fragments and rearrange them in a laminated pattern.

Connected with the problem of formation of α -1-6 linkages is that of their disruption. Neither α - nor β -amylase can split this linkage. Recently an enzyme present in muscle has been investigated in this laboratory which in combination with phosphorylase is able to cause an almost complete degradation of glycogen. Phosphorylase alone cannot degrade branched polysaccharides beyond the branch points. In fact, the limit dextrin formed from amylopectin or glycogen by either potato or muscle phosphorylase is larger than that formed by β -amylase. The question whether the same enzyme is involved in the formation and disruption of the 1-6 linkage has not been settled.

Theory of the action of phosphorylase - In the phosphorolysis of glycogen or starch the reactants are (1) inorganic phosphate, and (2) the terminal glucose unit of the polysaccharide chains. The chain molecule is thereby shortened, one glucose unit at a time, but the concentration of terminal glucose units remains the same until the limit of degradation is reached.

The theory of the action of phosphorylase is based on the assumption that the terminal glucose unit is also a reactant in the reverse direction; in this case the polysaccharide chains would be lengthened by successive additions of new glucose units from glucose-1-phosphate, again without any change in the concentration of the terminal glucose units. The reaction catalyzed by phosphorylase in the direction of increasing chain length might therefore be expressed as follows.

Terminal glucose (of n chain units) + glucose- α -phosphate + terminal glucose (of $n+1$ chain units) + phosphate, etc.

The following observations are in accord with this formulation. (1) Muscle as well as potato phosphorylase remain inactive- when glucose-1-phosphate alone is added; they require the further addition of polysaccharide as an essential reactant before new polysaccharide can be formed. (2) When increasing amounts of glycogen or starch are added, the rate of polysaccharide formation from glucose-1-phosphate increases in a manner which is characteristic for a reacting molecule (see Fig. 2.) (3) Branched polysaccharides are more strongly activating than linear polysaccharides, the effect being roughly proportional to the number of end-groups present. (4) The equilibrium of the reaction is independent of the concentration of polysaccharide. This would follow from the fact that the concentration of one of the two

reactants in either direction, the terminal glucose units, remains constant. Hence this term cancels out when the usual mass-law expression is written and the equilibrium constant, $K = (\text{phosphate})/(\text{glucose-1-phosphate})$. (5) For the same reason, the reaction in either direction, although involving two reactants, has been found to be kinetically of the first order.

Several experiments have been devised to test this theory. The reaction, glucose- α -phosphate \rightleftharpoons polysaccharide + phosphate, in the presence of ^{32}P , should lead to an incorporation of ^{32}P in glucose-1-phosphate. Such an exchange occurred in the presence but not in the absence of polysaccharide primer, showing that the latter was an essential reactant for polysaccharide synthesis from glucose-1-phosphate.

A further consequence of the theory presented above is that the length of the newly formed polysaccharide chains should depend on the ratio of molar concentrations, glucose-1-phosphate which reacted/terminal glucose units of added primer. When this ratio is large, long chains, when it is small, short chains should be formed. This has been shown to be the case by a number of investigators, who used iodine color as a measure of chain length. A comparison has also been made of the chain length calculated from the above ratio and that actually found by end-group assay, with fair agreement.

A special case arises when the limit dextrin of glycogen (formed by exhaustive treatment with phosphorylase) is used as primer and an amount of glucose-1-phosphate is added which would allow no more than one chain unit to be added for each primer end-group. At the start the reaction can proceed only in the direction of synthesis, since there are no end-groups present in the limit dextrin which can be removed by phosphorolysis. As the reaction proceeds, cleavable end-groups are added, but the rate of phosphorolysis close to the limit of degradation is very slow. If polysaccharide synthesis consists in the addition of glucose units to primer end-groups, one would expect under these special conditions a shift in the equilibrium to the side of polysaccharide formation and a dependence of the equilibrium on the concentration of added primer. Hestrin who carried out this experiment (unpublished) found both these predictions fulfilled.

There is thus experimental support for the idea that polysaccharides act as primers in the direction of synthesis because they enter stoichiometrically into the reaction catalyzed by phosphorylase.

References to the literature quoted will be found in the following articles:

1. C. F. Cori and G. T. Cori, *J. Biol. Chem.*, 116 (1936) 169; *Proc. Soc. Exptl. Biol. Med.*, 34 (1936) 702; 39 (1938) 327.
2. C. F. Cori, S. P. Colowick, and G. T. Cori, *J. Biol. Chem.*, 121 (1937) 465; 123 (1938) 375; 124 (1938) 543.
3. C. F. Cori, G. Schmidt, and G. T. Cori, *Science*, 89 (1939) 464.
4. A. A. Green, C. F. Cori, and G. T. Cori, *J. Biol. Chem.*, 151 (1943) 21; 158 (1945) 315.
5. M. A. Swanson and C. F. Cori, *J. Biol. Chem.*, 172 (1948) 797.
6. M. Cohn and G. T. Cori, *J. Biol. Chem.*, 175 (1948) 89.

GEORGE DE HEVESY

Some applications of isotopic indicators

Nobel Lecture, December 12, 1944

The method of isotopic indicators had its ultimate origin in the Institute of Physics at the University of Manchester, which then was under the inspiring leadership of that great physicist the late Lord (then Professor) Ernest Rutherford.

The cradle of radium is the Czecho-Slovakian town Joachimstal; it was from Joachimstal pitchblende ore that Professor and Madame Curie isolated that element. The Austrian Government, the owners of these mines, generously supplied Professor Rutherford not only with radium, but also with the by-products of radium production, equally important for the worker in the field of radioactivity. One of the most significant by-products is radium D, which has a half-life period of 20 years and is found associated with the very substantial amounts of lead present in pitchblende. The Austrian Government presented to Professor Rutherford several hundred kilograms of such "radio-lead". In view of its association with very large amounts of lead, which absorb the radiation emitted by radium D, this precious radioactive material nevertheless proved to be almost useless. When I met Professor Rutherford one day in 1911 in the basement of the laboratory where the radio-lead was stored, he addressed me in his friendly and informal way, saying: "My boy if you are worth your salt, you try to separate radium D from all that lead". In those days, I was an enthusiastic young man and, on immediately starting to attack the problem suggested to me, I felt quite convinced that I would succeed. However, although I made numerous attempts to separate radium D from lead and worked for almost two years at this task, I failed completely. In order to make the best of this depressing situation, I decided to use radium D as an indicator of lead, thus profiting from the inseparability of radium D from lead. Suppose that we dissolve 1 g of lead in the form of nitrate in water, add radium D of negligible weight showing a radioactivity of one million relative units (an electroscope being used to measure the activity), and proceed to carry out the most intricate operations with this "labelled" lead. If we then ascertain the presence of one radioactive unit in a fraction obtained in the course of these operations, we

must conclude that 1/1000 mg of the lead atoms present in the lead nitrate we started from, are now present in the fraction.

Radium D cannot be separated from lead, but it can easily be obtained in the pure form from lead-free radium salt samples or from radium emanation, since radium D is formed in the course of the disintegration of these radioactive bodies and can readily be separated from them. At that time, the Vienna Institute of Radium Research had more radium and radium emanation at its disposal than any other institution. This fact induced me, late in 1912, to start work at the Vienna Institute in collaboration with Dr. Paneth, assistant at that Institute, who himself had made very extensive and abortive trials to separate radium D from lead. The first application of labelled lead¹ was the determination of the solubility of some very slightly soluble lead compounds such as lead chromate. In these experiments, not radium D but another isotope of lead, thorium B, was applied as an indicator. Labelled lead chromate was obtained by adding a solution of 100,000 relative units of thorium B to lead nitrate containing 10 mg of lead and converting the nitrate thus labelled into chromate.

After the saturated solution of this compound had been held at the desired temperature in a thermostat for a sufficient time, its composition was ascertained by evaporating a few cubic centimetres to dryness and measuring the activity of the almost invisible residue in the electroscope. From the number of units of thorium B found, the amount of lead was calculated, one unit corresponding to 10^6 grams of lead; finally, the solubility of the lead chromate in moles per litre ($2 \cdot 10^{-7}$) was computed.

The radioactive method is extremely simple, having the advantage that the presence of foreign ions in no way interferes with the measurements. The method may be applied without difficulty, for example, in determining the solubility of lead sulphate in the presence of calcium sulphate.

Simultaneously with the said experiments, we used labelled lead and labelled bismuth (the radioactive bismuth isotope radium E can easily be obtained from radium, radium emanation, or radio-lead) in an investigation of the manner in which unweighable amounts of metals are precipitated during electrolysis². The application of the well-known Nernst formula

$$e = \frac{RT}{n} \log_n \frac{c}{C}$$

was extended to concentrations of $10^{-8} N$ and even lower.

On the basis of Nernst's theory, we should expect an interchange to take

place between the metal of the electrodes and the ions in solution. The existence of such an interchange was demonstrated³. While, in the case of lead peroxide, the interchange was found to take place between the outermost layer of molecules of the geometrically calculated electrode surface, in the case of the lead electrodes numerous layers of molecules were found to participate in the interchange process. This finding is to be interpreted as a result of local currents due to variations in the structure of the metallic surface. An interchange between atoms of a lead foil and the lead ions present in a solution was found to occur very rapidly, while the lead ions adsorbed on colloidal lead particles were found to interchange at a slow rate only⁴.

In contrast to metallic surfaces, Paneth⁵ found that in the case of salt crystal surfaces the interchange was restricted to the uppermost molecular layer of the crystal. On this observation he based an important method for the determination of the surface areas of crystalline powders⁶.

When lead sulphate is shaken with its saturated solutions, a constant kinetic exchange occurs between the molecules of lead sulphate in the solution and those on the surface of the solid. If the solution contains marked molecules, after equilibrium has been attained the numerical ratio of marked molecules on the surface to those in solution is identical to that of total molecules on the surface to total molecules in the solution. As the distribution of the labelled molecules is determined by means of radioactive measurements, and as the lead sulphate content of the saturated solution is evaluated by the usual methods of analytical chemistry, the amount of lead sulphate present in the uppermost molecular layer can be computed. When the weight is known of a unimolecular layer of lead sulphate of 1 cm² area, the surface of the crystal powder can be calculated from the above data.

Among the numerous applications of radioactive indicators by Paneth I wish to emphasize the importance of his discovery of the existence of bismuth hydride⁷ and lead hydride⁸. After he gained experience regarding the best method of preparation and the stability of radioactive bismuth hydride and lead hydride, he succeeded in preparing these compounds from inactive bismuth and inactive lead, respectively.

Self-diffusion

The conception of the diffusion of a substance into itself, self-diffusion, was introduced by Maxwell. No further use was made of this concept until fifty

years later, when the method of radioactive (isotopic) indicators was introduced. The possibility of measuring self-diffusion by following the rate of penetration of the lead isotopes ThB or RaD into lead soon suggested the measurement of the self-diffusion in liquid and solid lead, using ThB or RaD as indicators. The measurements of the self-diffusion coefficient in liquid lead⁹ gave the result anticipated from the known diffusion rates of lead in mercury and other related elements. The diffusion rate in liquids is primarily determined by the radius of the diffusing particle and the viscosity of the liquid: thus, the replacement of a diffusing metal present in small concentration by another related metal will not appreciably influence the rate of diffusion. A very different behaviour was revealed, however, when the self-diffusion in solid lead¹⁰ was measured, using RaD as indicator. In the first experiments carried out in collaboration with Gróh, we soldered a piece of radio-lead to the bottom of a rod of ordinary lead, whereafter the system was kept at 280° for 140 days. After the lapse of that time, we cut the system into four equal parts, rolled the four lead pieces into thin plates, and placed them in an electroscope. No diffusion of the radio-lead into the ordinary lead could be ascertained, showing that the self-diffusion rate in lead must be at least several hundred times smaller than that of gold in lead, as determined by Roberts-Austin.

This result necessitated the introduction of special methods of great sensitivity for measuring diffusion. Since the rate of diffusion is inversely proportional to the square of the thickness of the layer, we worked out methods for the measurement of the penetration of radioactive lead into ordinary lead layers to a depth of only a few microns. The α -particles emitted by ThB (more correctly, by its disintegration products, ThC and ThC', which, however, attain radioactive equilibrium with the former within a few hours) produce scintillations on a zinc sulphide screen, the number of which is ascertained. The infinitesimal layer of ThB, which is in intimate contact with a lead foil placed below it, is then heated for a few hours to, say, 200°. If a diffusion of the ThB atoms into the lead foil takes place, the count of the scintillations will give a smaller value after the experiment than before. The range of the α -particles in lead being only about $30\ \mu$, a shift of a small percentage of the ThB atoms to depths less than $30\ \mu$ will suffice to reduce the counts of scintillations to a noticeable extent. This method, worked out in collaboration with Mrs. Obrusheva¹¹ was later replaced by a more sensitive and exact procedure applied in diffusion measurements in solid metals, and also salts, in collaboration with Seith¹². ThB was condensed on a foil

or a single crystal of the metal, and the ionisation produced by radiation emitted by the radioactive body was measured. A slight diffusion of the ThB into the lead after heating sufficed to diminish the ionising effect registered by an electrometer. Instead of the ionisation produced by the α -rays, the ionisation produced by recoil particles accompanying the emission of α -rays was measured in some cases. The range of the recoil particles in lead being only about 10^5 cm, these measurements made possible the determination of diffusion coefficients as small as 10^{12} cm 2 per day.

Our measurements led to the result that, while the diffusion coefficient of gold in lead is found to be $5 \cdot 10^3$ cm 2 day $^{-1}$ at 165° , the coefficient of self-diffusion in lead at the same temperature is only 10^6 cm 2 day $^{-1}$, the difference rapidly increasing with decreasing temperature. The change of the value of the coefficient of self-diffusion, D , in lead foils and single crystals is represented by the equation

$$D = 5.76 \cdot 10^5 e^{-27900/RT} \text{ cm}^2 \text{ day}^{-1}$$

Making use of this formula, we can show that, at room temperature, the atoms will change their places in a piece of lead on the average only once a day.

From the change of the coefficient of self-diffusion with temperature, the heat of activation of the diffusion process, the heat of loosening of the lead lattice, can be calculated. The value obtained and, for purposes of comparison, other thermal data are given in Table 1.

Table I. Thermal data for solid lead

	<i>kcal per g atom</i>
Heat of melting	1.1
Energy content at the melting point	3.5
Heat of lattice-loosening	27.9
Heat of evaporation	36.2

Roberts-Austin measured the diffusion rate of gold in solid lead. His measurements gave the first quantitative determination of diffusion rates in solids. The high values he obtained, shown in Fig. 1¹³, led his contemporaries to consider diffusion in solid metals a comparatively rapid process. The introduction of the conception of self-diffusion and the subsequent development led to a very different view and also to the elucidation of the remark-

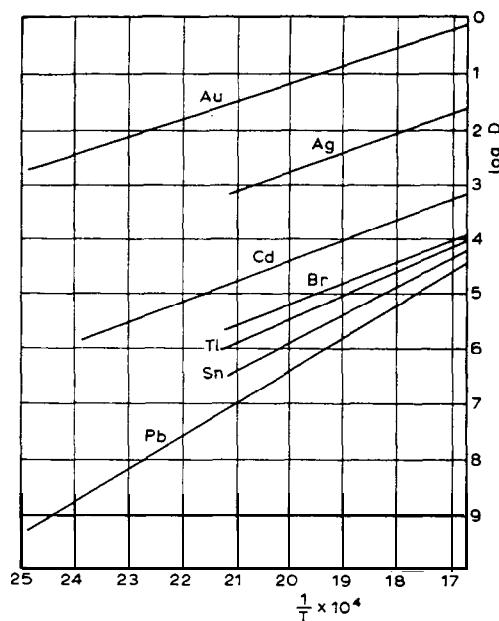


Fig. 1. Diffusion rates of metals in solid lead.

able nature of the gold-lead system investigated by that pioneer metallurgist.

The methods outlined above were also applied to determine the self-diffusion rate of Pb^{++} in solid lead chloride and lead iodide¹². The variation of the self-diffusion rates with temperature can be expressed by the equations

$$D = 1.06 \cdot 10^7 e^{-38120/RT} \text{ and } D = 3.43 \cdot 10^4 e^{-30000/RT}$$

respectively. As first shown by Nernst, the ionic mobilities in an electrolyte solution, and hence the conductivity of the solution, can be calculated if the diffusion rates of the ions are known. We can apply the same ideas to solid electrolytes¹⁴ and calculate, for example, the diffusion rate of Pb^{++} from the conductivity of lead chloride, on the assumption that the electrolytic conductivity of the salt is due solely to the transference of charges by the Pb^{++} . The diffusion rate of Pb^{++} thus calculated is, however, many thousand times larger than the value found experimentally, showing that the chloride ions are almost exclusively responsible for the conduction of electricity in solid lead chloride. The above data permit the calculation of the transport number of Pb^{++} in solid lead chloride. Not far from the melting

point, at 484° , the transport number of Pb^{++} is found¹⁵ to be 10^{-3} , at 273° 10^{-5} , and at 90° only 10^{-10} . By no other method can such small transport numbers be determined in any electrolyte.

In the case of lead iodide, the diffusion rate of Pb^{++} calculated from conductivity data, under the assumption that the whole conduction is due to Pb^{++} , is in good agreement with the measurement of the self-diffusion rate of Pb^{++} . This shows that, in contrast to the case of PbCl_2 , the conductivity in PbI_2 at high temperatures is due almost exclusively to the transference of charges by Pb^{++} . With decreasing temperature the role of Pb^{++} decreases and the transport number of I^- increases accordingly. At 260° , only 40% of the conductivity is due to Pb^{++} and 60 percent to the I^- ; at 155° the share of the former is only 0.4 percent. In the case of lead iodide, Tubandt's beautiful method for the determination of transport numbers could also be applied. The values obtained agreed well with those found by the measurement of self-diffusion.

Formerly, the self-diffusion rates of only lead, bismuth and a few other elements could be determined. These elements have natural radioactive isotopes. The discovery of artificial radioactivity greatly enlarged the possibilities for the determination of self-diffusion rates. By making use of the radioactive bromine isotope, we can determine the self-diffusion rate of Br^- in AgBr just as we determined that of Pb^{++} in PbCl_2 . Working at the Institute for Theoretical Physics at the University of Copenhagen, H. A. C. McKay determined the self-diffusion in gold. By the action of neutrons on gold a radioactive gold isotope, having the atomic number 198, can be produced. Neutrons having an energy of about 4 V are strongly absorbed in gold. A thin gold sheet exposed to the action of such neutrons will be more strongly activated on the side first struck by the neutron beam than on the opposite side. When the activated film is heated, the difference in the concentrations of the active gold atoms will decrease and, from the decrease of the activity difference shown by the two faces of the foil, the rate of self-diffusion in gold can be calculated¹⁶.

Svante Arrhenius' theory of electrolytic dissociation

If we dissolve sodium chloride and the equivalent amount of sodium bromide in water and then separate the two salts by crystallisation, it would have been expected in the time prior to Arrhenius that the chloride ions

would retain their original partners, the same applying to the bromide ions. According to Arrhenius, however, each chloride ion has the same chance of associating with a sodium atom originally bound to chlorine as with one initially associated with bromine. The correctness of the much debated views of Arrhenius was shown in different ways; the most direct proof, however, was provided through the application of isotopic indicators¹⁷. When equivalent amounts of $PbCl_2$ and labelled $Pb(NO_3)_2$ (or *vice versa*) were dissolved and the two compounds were separated by crystallisation, the labelled lead ions were found to be equally distributed between chloride and nitrate ions.

Very different results were obtained in all cases in which the lead atom was joined to carbon. Between lead chloride and lead tetraphenyl in pyridine, between lead acetate and lead tetraphenyl in amyl alcohol, and between lead nitrate and diphenyl lead nitrate in aqueous ethyl alcohol, no change in the places of lead atoms could be detected, although in every combination investigated one of the molecular types was capable of electrolytic dissociation.

The lack of interchange of atoms present in organic binding (hydrogen atoms bound to oxygen or nitrogen being an exception, as shown by Bonhoeffer⁸¹), such as that of carbon atoms in glycogen or phosphorus atoms in lecithin with other carbon and phosphorus atoms respectively, was found to be of great significance for the application of isotopic indicators in biochemical research. Owing to the absence of such an interchange, the presence of labelled carbon atoms in glycogen molecules, or of labelled phosphorus atoms in lecithin molecules, extracted from the organs, proved that a synthesis of these molecules took place after the labelled atoms were administered. This principle enables us to distinguish between "old" and "new" molecules and to determine the rates at which molecules of the different compounds are built up and carried to the different organs.

A prompt interchange of the electrical charges between Pb^{++} and Pb^{++++} ions was found to take place in experiments where plumbous acetate and labelled plumbic acetate (or *vice versa*) were dissolved in glacial acetic acid and then separated by crystallisation¹⁸. The same holds for Tl^+ and Tl^{+++} ion¹⁹. An interchange of lead atoms takes place between fused lead and fused lead chloride, lead oxide or lead sulphide²⁰.

After artificially radioactive isotopes became available as indicators, interchange processes were studied in numerous cases. A rapid interchange of charges was found to take place between Fe^{++} and Fe^{+++} , Cu^+ and Cu^{++} , etc.²¹.

Analytical chemistry

Analytical chemistry proved to be a fruitful field for the application of isotopic indicators. A knowledge of the total lead content of the earth's crust, for example, is of considerable chemical interest. In view of the small lead content of the average rock sample, the quantitative determination of its lead content involves some difficulties. These have been eliminated by making use of an isotopic indicator²². An amount of radium D, known in relative radioactive units, is added to the solution of the rock sample; the radium D is then recovered by electrolysis as peroxide. If 100 per cent of the added radium D is recovered, we may expect 100 percent of the lead present in the sample to have been recovered as well. If only 50 percent is recovered, for example, we have to multiply the amount of lead recovered by 2 in order to arrive at a correct analytical figure. The indicator method thus allows a correction for the shortcomings of the analysis. Such corrected analytical figures are seen in Table 2.

Instead of adding radium D to the solution to be analysed, we may add lead labelled by the presence of some radium D, for example 100 mg of lead having an activity of 1,000 units. If we subsequently isolate 10 mg of lead from the solution, this lead should show an activity of 100 units, under the assumption that the original sample does not contain lead. If the activity of the isolated 10 mg of lead is, for example, found to be 83 only, we have to conclude that the sample contains lead amounting to 20 mg.

In recent years, isotopic indicators have found an extended application in

Table 2. Lead content of igneous rocks

<i>Rock types</i>	<i>g lead per g rock precipitated by electrolysis</i>	<i>Percentage RaD recovered by electrolysis</i>	<i>Corrected value of g lead per g rock</i>
Gabbros and related types (composite of 67 samples)	$4 \cdot 10^{-6}$	80	$5 \cdot 10^{-6}$
Essexites and related types (composite of 40 samples)	$7 \cdot 10^{-6}$	80	$10 \cdot 10^{-6}$
Soda-granites and soda-syenites (composite of 26 samples)	$9 \cdot 10^{-6}$	73	$11 \cdot 10^{-6}$
Granite rocks (composite of 58 samples)	$18 \cdot 10^{-6}$	53	$30 \cdot 10^{-6}$
Basalt (Giant Causeway)	$4 \cdot 10^{-6}$	100	$4 \cdot 10^{-6}$

biochemical analyses. Schoenheimer and his colleagues⁸² determined the leucine content in the protein of the rat by adding to the hydrolysate a known amount of leucine containing heavy nitrogen. This tracer was also used⁸³ in the investigation of the occurrence of the amino acids of the *d* series in cancer proteins; while Chargaff, Ziff and Rittenberg²⁵ used bases containing a known amount of ¹⁵N in the analysis of the nitrogenous constituents of tissue phosphatides. Amino acids containing deuterium as indicator were used by Ussing²³ and the same tracer was applied by Rittenberg and Foster²⁴ in their determination of the palmitic acid content of rats' fat.

Early biological application

In contradistinction to the animal body, the uptake of mineral constituents by the plant is not followed by a loss of such constituents, and it was formerly considered that the ions taken up by the roots of the plant did not migrate in the opposite direction at all. The application of isotopic indicators, however, has shown that this is not the case. Ions taken up by the plant can be removed by an exchange process under the action of other ions present in the soil or in the nutrient solution. It was already found in 1923 that minute amounts of lead, labelled by the admixture of the lead isotope thorium B, when taken up by the roots of *Vicia faba*, could to a large extent be removed by an excess of non-labelled lead added to the nutrient solution²⁶. Most other ions were found to be much less effective in removing the labelled lead ions from the plant.

In recent years, the behaviour of essential constituents of plants has been investigated, making use of artificial, radioactive ions as indicators; similar results were obtained. Mullins and Brooks²⁷ placed cells of *Nitella coronata* first in a solution containing radioactive potassium and later in solutions of different chlorides. Sodium and lithium were found to be much less effective in removing labelled potassium than potassium itself, whereas rubidium was more effective. Jenny and Overstreet²⁸ and Broyer and Overstreet²⁹ found that ionic exchange could take place during periods of, and under conditions favourable for, active solute uptake. It was also observed³⁰ that for each six phosphate ions taken up by the roots of growing wheat seedlings, one phosphate ion migrated from the roots into the nutritive solution.

Early in the twentieth century, the application of bismuth compounds in syphilis therapy came to the fore. This induced Christiansen, Lomholt and

de Hevesy³¹ to investigate the absorption, circulation and excretion of labelled bismuth preparations. Lomholt³² succeeded in showing that, of all the preparations investigated, bismuth hydroxide suspended in oil was most suitable for therapeutic application.

Successes achieved by Blair-Bell in cancer therapy, using lead compounds, induced the investigation of the partition of labelled lead compounds between normal and tumorous tissue³³. Though this work gave a negative result, it nevertheless proved to be of great importance in the future development of isotopic indicators. It was in the course of these investigations that Schoenheimer became familiar with the method of isotopic indicators, which he applied several years later with such great success in the study of fat and protein metabolism and of numerous related problems. Never were more beautiful investigations carried out with isotopic indicators than those of the late Professor Schoenheimer, whose untimely and tragic death is much to be deplored. The discussion of the numerous important results obtained by Schoenheimer and Rittenberg and their collaborator³⁴ lies, however, beyond the scope of this lecture.

Heavy water

In 1931, Urey discovered deuterium, an isotope of hydrogen³⁵. This important discovery made possible the labelling of hydrogen. Deuterium is not an ideal indicator, its properties differing appreciably from those of hydrogen. The latter has a unique position: it is the sole element met with, though only transitorily, as a naked nucleus in chemical reactions. Chemical forces do not suffice to remove all electrons from any other element. Differences in the structure of the nucleus will therefore make themselves more noticeable in the chemical behaviour of hydrogen isotopes than in the case of any other element. Furthermore, the difference between the mass of the hydrogen atom and that of the deuterium atom amounts to as much as 100 percent, while, for example, the corresponding difference between a ³¹P and a ³²P atom is only 3 percent. The very small difference in the chemical properties of ³¹P and ³²P remains at the present time within the errors of our experiments, whereas between hydrogen and deuterium the difference is quite appreciable. The same applies to H₂O and D₂O. Dilute "heavy" water, however, contains mostly DOH molecules which exhibit in their chemical behaviour a very great resemblance to HOH.

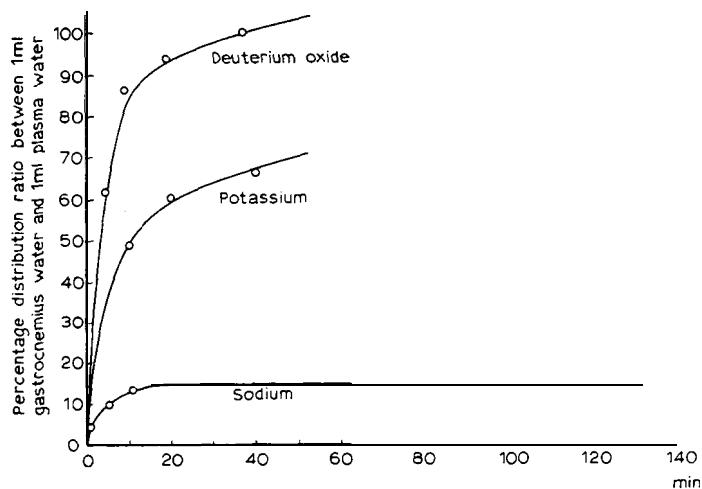


Fig. 2. Percentage distribution ratio of labelled sodium, potassium and deuterium oxide between plasma water and muscle water of equal weight.

In the study of the circulation of water in the organism, dilute heavy water can therefore safely be used as an indicator. In the determination of the life period of water molecules in the human organism, water containing $\frac{1}{2}$ per cent heavy water was used³⁶. While a small percentage (0.1) of the water was found to be excreted in so short a time as 26 minutes, the average life of the water molecules in the organism was found to be 13.5 days. In the excreted water, molecules were thus found which were taken in both a few minutes and several months before the investigation. By extrapolation, however, we arrive at the result that, though the number of water molecules present in an adult organism amounts to as much as ca. 10^{27} , the adult organism no longer contains a single water molecule taken up at birth.

The rate of admixture of administered water with the water present in the body was investigated in experiments on rabbits³⁷ and guinea pigs³⁸. While the water reaching the circulation was found to enter into exchange equilibrium with the extracellular water (about $\frac{1}{4}$ of the weight of the rabbit) in the course of a few minutes, the penetration of the water molecules into the cells took some time. As is seen in Fig. 2, about 30 min passes before the exchange equilibrium is reached between the water administered and extracellular and intracellular water present in the muscles of the rabbit. In the guinea pig, 73 per cent of the water in the blood is exchanged for extracellular water every minute.

Time does not permit me to discuss the extended application of heavy hydrogen, heavy carbon and heavy nitrogen, and to treat the numerous important results obtained by the use of these isotopes as indicators.

Application of artificially radioactive isotopes

During the lengthy operations preceding the early experiments of self-diffusion in lead, we often discussed the great progress which might be expected if radioactive indicators of the common elements were made available to chemical and biological research. This wish, which seemed utopian in those remote days, was fulfilled by Frederic Joliot and Irène Joliot-Curie's³⁹ fundamental discovery of artificial radioactivity, followed by Fermi's³⁹ work leading to the discovery of many more artificially radioactive isotopes. Soon after the announcement of these discoveries, we prepared the radioactive phosphorus isotope ³²P by neutron bombardment of carbon disulphide and used this isotope in collaboration with Chiewitz⁴⁰ in the study of phosphorus metabolism. In these experiments, 10 litres of carbon disulphide were used to absorb most of the neutrons emitted by a mixture of radium and beryllium kindly put at our disposal by Professor Niels Bohr. The ³²P formed was extracted by treatment with diluted nitric acid or with water, the carbon disulphide being immediately available after the extraction for further neutron-irradiation.

A few other radioactive isotopes, such as the radio-halogens can also be prepared by similar simple and convenient procedures. This is, however, not the case with the majority of radioactive isotopes. These were prepared in amounts sufficient to be utilized in indicator work only after the invention of the cyclotron⁴¹. Lawrence's highly significant invention also made available radio-phosphorus preparations of very much greater activity than could be obtained from neutron-sources containing as much as several grams of radium. The number of neutrons produced by the Berkeley cyclotron was stated by Birge⁴¹ in 1939 to correspond to the ionisation produced by 100 kg of radium; since that date, a still more powerful cyclotron has been brought into use. In our later investigations, radio-phosphorus generously put at our disposal by Professors Niels Bohr, Lawrence and Siegbahn was used.

The preparations of radioactive isotopes of numerous elements prepared in the Radiation Laboratory at Berkeley and in other laboratories found an extended application as indicators. Radioactive iron prepared at Berkeley, for

example, was used by Hahn, Whipple and their colleagues⁴² in extended studies of iron metabolism.

The application of cyclotron-prepared radio-carbon⁴³ revolutionized our views of the fundamental process of photosynthesis.

Radio-iodine⁴⁴ found an extended application in the study of the formation of thyroxine and diiodotyrosine; it led, *inter alia*, to the important finding that some thyroxine is formed in the organism even after total extirpation of the thyroidea⁴⁵.

Radio-phosphorus found, however⁴⁶, the most extensive application. This was due not only to the convenient mode of production and period of decay of this material, together with the low absorbability of the rays emitted by it, but mainly to the important part which phosphorus plays in a very great number of metabolic processes. These include skeleton formation, metabolism of carbohydrates and fats, cell division, and many other processes. The discussion of the role of phosphorus in metabolic processes is therefore well suited to demonstrate different applications of isotopic indicators in biological research. We shall therefore now describe some applications of radio-phosphorus. These examples represent only a small percentage of the investigations in which radio-phosphorus has been used as an indicator; many of the results to be discussed were obtained in Copenhagen.

Radio-phosphorus

Owing to the great sensitivity of the Geiger-Müller counter, which registers ^{32}P with an activity of only 10^{-6} microcurie, some of the radio-phosphorus administered can soon be located in all organs. Table 3 shows the distribution of ^{32}P in the organs of the rat 4 hours after subcutaneous injection of labelled sodium phosphate⁴⁷.

While 4 hours after the administration most ^{32}P is found in the skeleton, muscles, liver and the digestive tract, with increasing time more and more ^{32}P becomes incorporated with the skeleton; 98 days after the start of the experiment, 92 percent of all ^{32}P present in the rat, which corresponds to about one half of the total amount administered, is found in the skeleton. This result may be seen in Table 4. Most phosphorus taken up with the food, in so far as it is not excreted, ultimately finds its way into the skeleton, where it replaces "old" phosphorus which interchanges with the phosphorus present in other organs or is excreted.

Table 3. Distribution of ^{32}P between different organs in a rat, 4 hours after subcutaneous injection of labelled phosphate (Weight of the rat: 188 g)

<i>Organ</i>	<i>Percent ^{32}P present</i>	<i>Specific activity</i>
Bones	22.6	0.020
Muscles	18.7	0.191
Liver	17.6	0.475
Digestive tract	15.9	0.365
Skin	11.1	0.192
Lungs and heart	6.3	0.317
Blood	2.5	0.558
Kidneys	2.4	0.370
Spleen	1.3	0.256
Brain	0.02	0.032

Table 4. Percentage total ^{32}P found in some organs of rats

<i>Organ</i>	<i>Time after distribution of ^{32}P</i>						
	<i>Hours</i>		<i>Days</i>				
	<i>1/2</i>	<i>4</i>	<i>10</i>	<i>20</i>	<i>30</i>	<i>50</i>	<i>98</i>
Muscles	18.3	19.4	25.8	28.8	25.2	12.1	3.6
Total skeleton	19.1	23.4	43.1	43.1	51.8	76.5	92.0

From these results, however, no conclusions can be drawn concerning the extent of renewal of the skeleton, as the labelled phosphorus, i.e. phosphorus administered throughout the experiment, may be incorporated wholly or principally in the upper molecular layers of the apatite-like crystals which form the mineral constituents of the skeleton. We can determine the extent of renewal of the bone mineral phosphorus by comparing the ^{32}P content, i.e. the radioactivity of 1 mg bone mineral P, with the radioactivity of 1 mg free plasma P. Were the bone phosphorus entirely renewed in the course of the experiment, the ^{32}P would be distributed equally between the free P atoms of the apatite-like bone crystallites and the free P atoms of the plasma, the latter being the direct or indirect source of the bone phosphorus. If only 1 per cent of the bone apatite P were renewed in the course of the experiment, the specific activity of the bone apatite P would be only 1/100 of that of the free plasma P.

The determination of the degree of renewal based on the said administration is made difficult by the fact that the specific activity of the free plasma

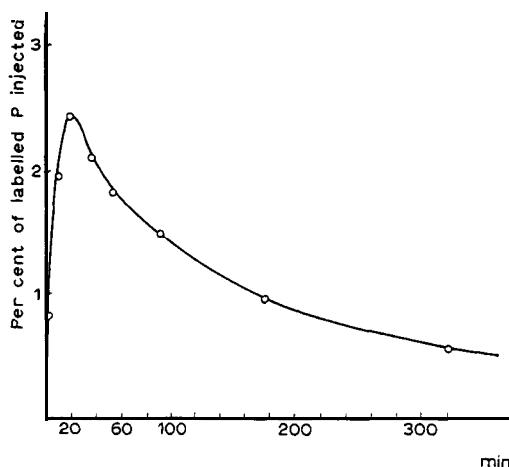


Fig. 3. Change in the specific activity of the plasma inorganic P after subcutaneous injection of labelled phosphate into a rabbit.

phosphorus varies throughout the experiment. After administration by the mouth or by subcutaneous injection, the specific activity first increases and subsequently decreases with time, while after administration by intravenous injection it first decreases very rapidly, and later at a moderate rate, as seen in Figs. 3 and 4⁴⁸. The determination of the extent of renewal of the mineral P in the skeleton is much facilitated by keeping the free plasma P activity at a constant or almost constant level. This can be attained by repeated injections of varying amounts of labelled phosphate throughout the experiment. Making use of this technique, the data given in Table 5 were obtained for

Table 5. Extent of renewal of the mineral constituents of the bone in the course of 50 days

	Percentage renewal
Femur epiphysis mineral P	29.7
Femur diaphysis mineral P	6.7
Tibia epiphysis mineral P	28.6
Tibia diaphysis mineral P	7.6
Costa mineral P	27.5
Femur phosphatide P	100

the extent of renewal of the different parts of the skeleton of the adult rabbit in the course of 50 days⁴⁹. It appeared that 72 per cent of the epiphysis and 93 percent of the diaphysis remained unchanged after this period, while 29

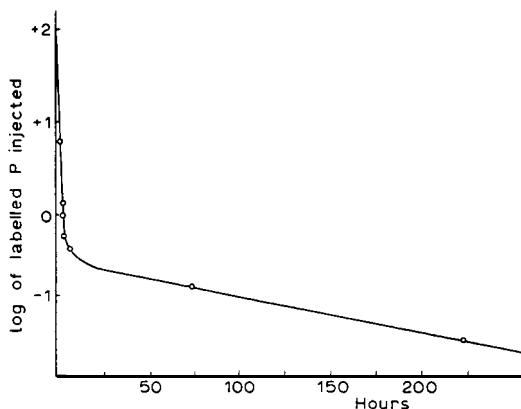


Fig. 4. Change with time in the logarithm of the labelled P content of the plasma after intravenous injection of labelled phosphate into a rabbit.

and 7 percent respectively were renewed not once, but, at least to some extent, repeatedly.

The restricted extent of renewal of the skeleton is due to the fact that while the P atoms of the uppermost molecular layer of the bone apatite crystals can promptly interchange with the free P atoms of the plasma (actually not the P atoms, but the phosphate ions interchange), a renewal of the main part of the apatite P can take place only when the crystal is dissolved and when crystals are formed from the plasma; from a labelled plasma, labelled crystals are formed. This "biological" recrystallization of the skeleton crystallites is a slow process. Moreover, we have to consider that if only the outer part of the crystal is renewed, this process can often be repeated without affecting deeper molecular layers of the crystal.

No data are available concerning the extent of renewal of the human skeleton; the relative rates of renewal of different parts of the skeleton were, however, determined by Erf⁵⁰.

The problem as to whether and to what extent the P atoms of the dental enamel are renewed has been a subject of extensive investigations⁵¹ which led to the result that, though some ³²P is found to be present in the enamel after administration of labelled phosphate, the extent of renewal of the enamel phosphate is almost negligible. Regarding the extent of replacement of the phosphorus present in the constituents of the dentine, about one millionth part of the food phosphorus was found to be located in the mineral constituents of the dentine of each tooth.

The bone tissue growing in a labelled organism is bound to become labelled. Of the labelled phosphate administered by mouth, after the lapse of 3 days, 2 percent was found to be present in the rapidly growing incisors of the rat⁵². As seen in Fig. 5, these phosphate ions are mostly found in the

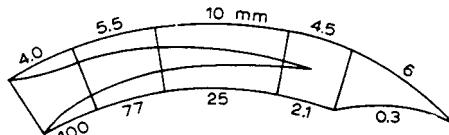


Fig. 5. Distribution of labelled phosphorus in the incisor of a rat killed 3 days after the administration of the phosphorus. The figures below give the relative amounts of labelled phosphorus present in 1 mg of fresh tissue in the section in question. The figures above give the length of the section in mm.

incisal part of the incisor, though a minor part are located at the apical end remote from the pulpa.

Permeability investigations

The above-mentioned rapid decrease in the plasma activity following intra-venous administration of ^{32}P is to a large extent due to the interchange of plasma phosphate with the phosphate of the extracellular fluid. From this fact it follows that the capillary wall is readily permeable to phosphate; similar results were obtained for the other labelled ions investigated. Sodium ions, which are mainly confined to the extracellular space, enter into exchange equilibrium with the plasma sodium within 20 minutes^{53,84,85}. This may clearly be seen from Fig. 6⁵³. For potassium and for phosphorus, elements mainly located in the tissue cells, a longer time is required for the attainment of such equilibrium⁵⁴. The low rate at which exchange equilibrium between the cellular and extracellular phosphorus is reached in the animal organism is mainly due to a very low rate of renewal of large parts of the skeleton.

As seen from the above examples, the method of isotopic indicators can be utilized with advantage in permeability investigations. It is with the aid of isotopic indicators that we best can measure the permeability of phase boundaries, since other methods do not indicate solely the resistance of the phase boundary to the penetration of ions, but a more complex phenomenon. Prior to the application of isotopic indicators, the high potassium content of the

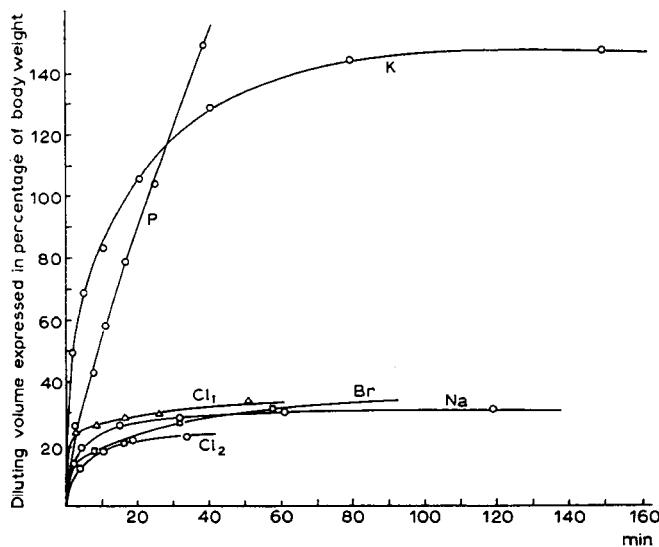


Fig. 6. Rate of disappearance of various labelled ions from the plasma.

erythrocytes of most species and their low sodium content were interpreted as being due to the impermeability of the erythrocyte membrane to potassium and sodium ions. The application of isotopic indicators, however, has disclosed the fact that potassium ions in the erythrocyte interchange quite easily⁵⁵ with those present in the plasma, and the same applies to the sodium ions. The high concentration of potassium and low concentration of sodium found in the erythrocytes of most animals can thus not be explained as being due to an impermeability of the corpuscle membrane to these ions.

Not only the resistance of phase boundaries to labelled ions, but also that to molecules of different kinds, can be measured with the aid of isotopic indicators. The rates of interchange of phosphatides present in the plasma and in different organs were determined in the following way⁵⁶. Labelled phosphate was administered to a rabbit. After the lapse of 2 days, when the plasma contained an appreciable amount of labelled phosphatides, part of the plasma of another rabbit (rabbit II) was replaced by the labelled plasma. On following the decrease with time in the activity of the phosphatides extracted from the plasma of rabbit II, it was found that half of the plasma phosphatide molecules had interchanged in the course of few hours with phosphatide molecules present in the organs of the rabbit. An investigation of the activity of the phosphatides isolated from the organs led to the result that a very

substantial part of the labelled phosphatide molecules injected was found in the liver.

Rate of formation

The rate of formation of labelled organic phosphorus compounds differs much for various compounds and varies greatly with the organ in which they are located. The labile P of adenosine triphosphate, for example, is renewed at a very remarkable rate^{57, 86}, the second P atom being renewed somewhat more slowly than the terminal atom⁵⁸. Hexose monophosphate, present in the red corpuscles, was found to be largely renewed within a few minutes⁵⁷. The formation of labelled phosphatides takes place in the liver and in the intestinal mucosa of the rabbit at a much more rapid rate than in the brain and more quickly than in any other organ⁵⁹. Desoxyribose nucleic acid, on the other hand, shows a behaviour opposite to that of the phosphatides. Its extent of renewal in the liver of adult rats is very low^{60, 61, 87}, amounting to only about 0.1 percent in the course of 2 hours. In the spleen and in the intestinal mucosa, the renewal of the desoxyribose nucleic acid is 20 and 30 times, respectively, more rapid than in the liver. High figures for the rate of formation were found in rapidly growing tissue.

X-rays were found to obstruct the formation of labelled desoxyribose nucleic acid molecules⁶¹. Following irradiation with 300 r or more both in the sarcoma and in the organs of the rat formation figures were obtained amounting only to about half of the value observed in the controls. In the study of the reduction in labelled nucleic acid formation under the action of X-rays, a new line of attack was opened in the study of the action of such radiation on cell division.

We calculate the extent of renewal of the compound in question, for example creatine phosphoric acid, by comparing the specific activity of the creatine phosphorus at the end of the experiment with the average specific activity of the free phosphorus present in the tissue cells during the experiment. This calculation is based on the assumption that the labelled free phosphate, or the phosphate of a donor whose P enters rapidly into exchange equilibrium with the free P present in the cells, is incorporated in the creatine phosphate molecule present in the cells of the organ investigated. In the case of phosphatides, the possibility cannot be entirely excluded^{45, 46} that a precursor of the phosphatides containing labelled phosphate is transferred from another organ into the organ investigated. In such a case, the calcula-

tion of the extent of the renewal of the phosphatide molecules would necessitate knowledge of the specific activity of the precursor P. The renewal figures obtained must therefore be interpreted with caution. Another more pertinent reason for the cautious interpretation of the results obtained is that the molecules of some organic compounds may possibly be built up within the phase boundary, where the specific activity of the free P may appreciably differ from that of the intracellular free P.

While we measure the rate of renewal of phosphatide molecules with respect to their phosphate content by employing ^{32}P as an indicator, the rate of renewal of the fatty acid constituents is determined by the use of deuterium⁶² as indicator, and the new-formation of the choline content by applying ^{15}N as a tracer⁶³. A molecule can clearly be renewed in various ways.

Site of formation of phosphorus compounds in the organism

Origin of yolk phosphatides

We shall first consider the site of formation of some constituents of the hen's egg. Where in the organism are the phosphatides found in the yolk synthesized? This question can be answered by comparing the specific activity of phosphatide P extracted from the yolk and from the different organs a few hours after administration of labelled sodium phosphate⁶⁴.

The results of an experiment in which the hen was killed 5 hours after subcutaneous injection of labelled P are seen in Table 6. The specific activities of the yolk phosphatide and the ovary phosphatide P were very low, showing that only a small part of the phosphatide molecules present in the said phosphatides had been formed within the last 5 hours. The plasma phosphatide P had a much higher specific activity than those extracted from the ovary and the yolk, while the liver phosphatide P had a higher specific activity than the plasma phosphatide P. The gradient indicating the presence of phosphatide molecules formed within the last 5 hours, thus falls off in the direction from the liver, through the plasma, into the ovary.

The conclusion that the formation of the phosphatide molecules of the plasma mainly occurs in the liver^{59,64,65} is strongly supported by the results obtained in the study of fat phosphorylation in the hepatectomized dog by Chaikoff and his colleagues⁶⁶. These authors injected labelled sodium phosphate intravenously immediately after removal of the liver. Practically no

Table 6. Specific activity of phosphatides extracted from the organs of a hen

<i>Organ</i>	<i>Relative specific activity (Activity of inorganic plasma P = 1)</i>
Liver	0.54
Plasma	0.43
Ovary	0.039
Yolk	0.0035
Intestine	0.11
Spleen	0.1

phosphatide ^{32}P was recovered in the plasma as late as 3-6 hours after extirpation of the liver; at these times 0.4 percent of the injected ^{32}P had been incorporated into phosphatides of both kidneys and about an equal amount into the whole small intestine.

Considerations similar to those applied to the origin of the phosphorus compounds of the yolk were used in an investigation of the phosphorus compounds of milk⁶⁷. As seen in Table 7, the milk phosphatides were found to have a much higher specific activity than the plasma phosphatides, indicating that the phosphatides must enter the milk from a source other than the plasma and must thus have been synthesized to a large extent in the mammary gland. The determination of the specific activity of the mammary gland phosphatides revealed a very high value, even higher than those found for the kidney and liver phosphatide P.

One often encounters the view that the milk fat originates from the phosphatides of the blood, which are decomposed into fatty acid and inorganic P in the mammary gland. The inorganic P present in the milk should, according

Table 7. Specific activity of the phosphatide P extracted from the organs of a goat
 $4\frac{1}{2}$ hours after administration of labelled sodium phosphate

<i>Fraction</i>	<i>Specific activity</i>
Milk	0.09
Plasma	0.02
Corpuscles	0.01
Mammary gland	0.13
Liver	0.09
Kidney	0.11
Plasma inorganic P	1.48

to this view, originate from phosphatide P. The fat content of goat's milk amounts to about 3 percent. Taking the ratio of fatty acid to phosphorus to be 20 : 1 in plasma phosphatides, the production of 3 percent fatty acid from phosphatides would set free 0.15 percent of inorganic phosphorus. This being about the inorganic P content of the milk, almost all inorganic P of the milk should originate from plasma phosphatide. A few hours after the administration of labelled phosphate, the milk phosphatides are only slightly active, while the milk inorganic P shows a strong activity. This is a decisive argument against the above view. The high activity already found for the milk inorganic P in the early stages of the experiment, is only compatible with the assumption that the milk inorganic P is derived from the plasma inorganic P. The latter acquires a high activity soon after subcutaneous injection of labelled sodium phosphate.

As a further example we may mention the origin of the phosphorus compounds in the chick embryo⁶⁸. Physiological sodium chloride solution (0.1 ml) containing traces of labelled sodium phosphate was injected into fertilized eggs. Several days after incubation, the phosphatides and other phosphorus compounds were isolated in turn from the embryo and the yolk, their activity and their phosphorus content being determined. As seen in Table 8,

Table 8. Specific activity of P extracted from the hen's egg incubated for 18 days

	<i>Specific activity</i>
Yolk residue	Acid-soluble P 1.56
	Phosphatide P 0.016
	Protein P 0.058
Embryo	Skeleton inorganic P 1.66
	Phosphatide P 1.59
	Protein P 1.44

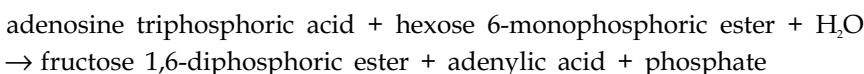
the specific activity of the embryo phosphatide P is very different from that of the yolk phosphatide P.

While the phosphatides of the yolk are scarcely active, the phosphatides extracted from the embryo are found to have a very strong activity. The phosphatide molecules present in the embryo must obviously have been newly synthesized. Similar considerations apply to the protein phosphorus present in the embryo.

Reaction path

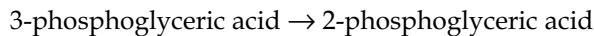
The path taken by organically bound phosphate radicals in glycolytic processes was investigated by using labelled compounds prepared under the action of enzymes present in muscle juice or yeast. When labelled adenylyl phosphate was added to fresh muscle pulp in which glycolysis occurred, no formation of active inorganic phosphate was found to take place, but active phosphate was detected in the Harden-Young ester formed during alcoholic fermentation.

In a study of the interaction of labelled adenosine triphosphoric acid with non-labelled hexose monophosphoric acid ester, which leads to the formation of fructose 1,6-diphosphoric acid ester, the labelled phosphate given off by the adenosine triphosphoric acid was found to be exclusively present in the fructose 1,6-diphosphoric ester. The fact that the free phosphate formed according to the equation:



was found to be inactive, indicates that the free phosphate originated exclusively from the hexose 6-monophosphoric ester⁶⁹.

When both hydrogen and labelled phosphate were transferred, neither of the two stable radicals of the cozymase molecule was found to be replaced by active phosphate. A similar negative result was obtained for the reaction



in the presence of active phosphate and also for the conversion of glucose monophosphoric acid into glucose hexaphosphoric acid in the presence of active phosphate. The ester fractions were found⁷⁰ not to have taken up ³²P.

Does a less pronounced formation of new molecules take place simultaneously with the autolysis observed in tissue slices? By the employment of isotopic indicators this question can be answered. On shaking liver, kidney or brain slices for a few hours with a Ringer solution containing ³²P at 37°C, Chaikoff and his colleagues⁷¹ found that the phosphatides isolated from the tissue slices contained ³²P; hence, side by side with an autolysis of the phosphatides in the tissue slices, an appreciable formation of labelled phosphatides also takes place.

The formation of labelled nucleic acid in slices of Jensen's rat sarcoma was likewise obtained⁷² in the investigation of the formation of desoxyribose nu-

cleic acid *in vitro*, when these slices were shaken with labelled blood or labelled Ringer solution. About 0.1 percent of the desoxyribose nucleic acid molecules present in the tissue slices were found to be labelled after the lapse of 4 hours; these molecules had accordingly been formed during the experiment. The presence of hydrogen sulphide, azide or carbon monoxide inhibits the formation of labelled phosphatides. Addition of cyanide, fluoride or monoiodoacetate to labelled blood or labelled Ringer solution is also found to inhibit the formation of labelled nucleic acid.

Dynamic state of body constituents

The most remarkable result obtained in the study of the application of isotopic indicators is perhaps the discovery of the dynamic state of the body constituents. The molecules building up the plant or animal organism are incessantly renewed. In the course of this renewal, not only the atoms and molecules taken up with the food participate, but atoms and molecules located in one organ or in one type of molecule will soon be found in another organ or in another type of molecule present in the same or in another organ. A phosphate radical taken up with the food may first participate in the phosphorylation of glucose in the intestinal mucosa, soon afterwards pass into the circulation as free phosphate, enter a red corpuscle, become incorporated with an adenosine triphosphoric acid molecule, participate in a glycolytic process going on in the corpuscle, return to the circulation, penetrate into the liver cells, participate in the formation of a phosphatide molecule, after a short interval enter the circulation in this form, penetrate into the spleen, and leave this organ after some time as a constituent of a lymphocyte. We may meet the phosphate radical again as a constituent of the plasma, from which it may find its way into the skeleton. Being incorporated in the uppermost molecular layer of the skeleton, it will have a good chance of being replaced by other phosphate radicals of the plasma or the lymph, but it may also have the good fortune to find a more or less lasting abode in the skeleton. This will be the case when it becomes embedded in a newly formed apatite-like bone crystallite.

There are indications that, in the growing organism, the rate of new formation of the molecules is still greater than in a fully grown organism. It was found, for example, by making use of heavy nitrogen, kindly put at our disposal by Professor Urey, as an indicator, that in "old" leaves of the sun-

flower, which did not develop further during the experiment, 12 percent of the protein molecules present were renewed within 12 days. In growing leaves, the replacement of old protein molecules was found to take place at a higher rate⁷³.

Schoenheimer and Rittenberg³⁴ have shown, by applying labelled nitrogen, that the peptide linkages in the proteins of the animal tissue are opened and reclosed with great ease. They found that the protein molecules in the living body continually change and renew their structures. This discovery is one of the most surprising and outstanding results arrived at with isotopic indicators.

Excretion studies

Chemical analyses of the food and of the excreta permit the determination of the extent to which the organism is in balance. Chemical methods, however, cannot determine to what degree the substances found in the faeces originate from undigested food and to what extent they have been carried into the digestive tract, coming from the body proper in the form of digestive juices. This problem can be solved under strictly physiological conditions with the aid of isotopic indicators.

The simplest procedure is the following⁷⁴. At a suitable time after administration of labelled sodium phosphate, we determine the specific activity of the urine P and that of the faeces P. Both originate from the blood plasma and, provided that we wait for a sufficient time, the specific activity of the P compounds carried into the digestive tract from the body will be about equal to that of the urine P. If the faeces P were entirely of endogenous origin, it should show a specific activity equal to that of the urine P. If we find the faeces P to be less active than the urine P, the active faeces P of endogenous origin must have been diluted by non-active P. Since the sole source of non-active P is the diet, the ratio of the specific activities of the faeces P and urine P tells us to what extent the endogenous faeces P has been diluted by food P.

The ratio $100 \times$ specific activity of faeces P / specific activity of urine P gives the percentage of P in the faeces which originates from the body proper. In the case of human subjects, $\frac{3}{4}$ to $\frac{4}{5}$ of the P present in the faeces was found to originate from non-absorbed P.

Labelled red corpuscles

As seen in Fig. 7⁷⁵, labelled phosphate penetrates at a fairly slow rate into the red corpuscles. On entering the corpuscles, however, the newly arrived phosphate ions participate rapidly in the formation of acid-soluble organic phosphorus compounds which occur in a comparatively high concentration in the corpuscles. The formation of new acid-soluble phosphorus compounds in the corpuscles is largely associated with glycolytic processes occurring there and is attended by the destruction of an equal or almost equal number of "old" molecules. As a result of these processes, the specific activity of the labile P atoms of adenosine triphosphoric acid and that of the P of some other compounds will soon acquire a specific activity almost as high as that shown by the free P of the corpuscles, but much lower than that of the free P of the plasma. This fact and the fairly slow rate of penetration of phosphate through the corpuscle wall explain the low rate of loss of ^{32}P by labelled corpuscles when brought into contact with unlabelled plasma, and make possible the application of such labelled corpuscles in the determination of the total circulating red corpuscle content of the organism^{76,77}. A detailed investigation of the erythrocyte content of human subjects, making use of labelled corpuscles, was carried out by Nylin^{78,88}. He estimated, furthermore, the rate at which injected blood and the circulating blood are homogeneously mixed. Fig. 8 shows the result obtained by Nylin in an experiment

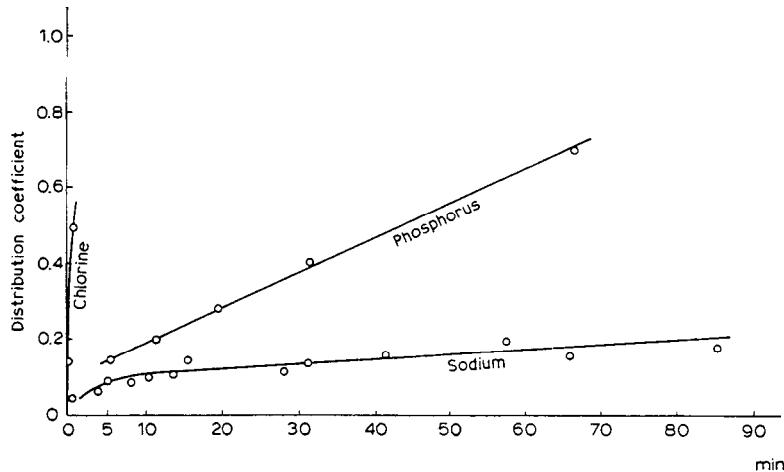


Fig. 7. Distribution of labelled ions between corpuscles and plasma of equal weight at 37°.

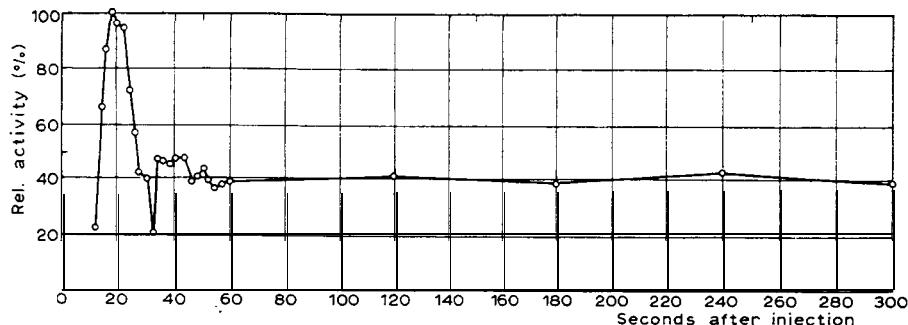


Fig. 8. Change of activity of the arterial corpuscles with time following intravenous injection of labelled corpuscles (G. Nylin).

where homogeneous distribution of the injected blood took only 60 seconds.

Corpuscles can also be labelled by introduction of radio-iron into the corpuscle haemoglobin. Such corpuscles were used by Hahn and his colleagues⁷⁹ in the determination of the red-corpuscle content of the dog. Radio-iron has a much longer half-life period than has radio-phosphorus; such "iron-labelled" corpuscles remain labelled for a much longer time than "phosphorus-labelled" corpuscles. While, however, the latter can easily be obtained by shaking blood with labelled phosphate, the former can be made only in the living organism, a fact which, together with the relatively great difficulty of obtaining radio-iron, restricts the applicability of iron-labelled corpuscles in the determination of the erythrocyte volume.

The determination of the total corpuscle volume of the organism demands only that the labelled corpuscles retain their labelling for some minutes; the determination of the life cycle of the corpuscles requires, however, the use of marked corpuscles which conserve their labelling for weeks. No worker has yet succeeded in achieving a labelling of mammalian red corpuscles that fulfils this condition. Iron-labelled corpuscles, although remaining labelled for a sufficient time, were found by Hahn and his colleagues⁸⁹ not to be suited to the purpose in hand. The life of the red corpuscles of the hen, however, was determined⁸⁰, making use of phosphorus-labelled corpuscles. In contradistinction to mammalian corpuscles, avian corpuscles contain large amounts of desoxyribose nucleic acid, and the nucleic acid molecules were found to remain unchanged throughout the life of the corpuscles. The newly formed corpuscles of a hen to which labelled phosphate is administered contain labelled desoxyribose nucleic acid.

By daily injection of labelled phosphate, the activity of the plasma phos-

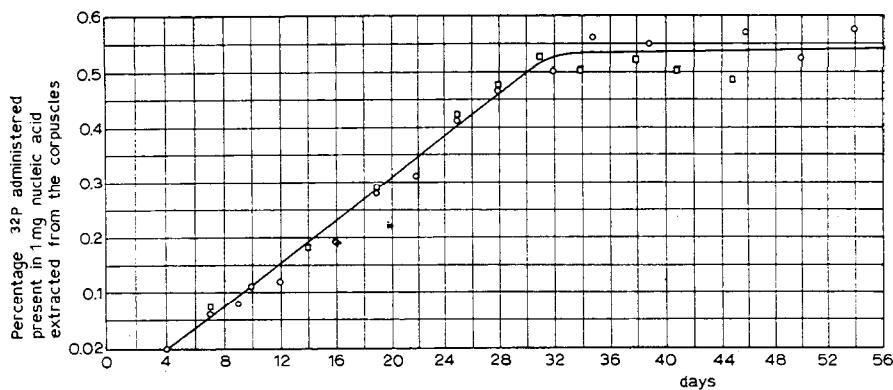


Fig. 9. Life cycle of the red corpuscles of the hen.

phate is kept at a constant level, and at suitable intervals the specific activity of the nucleic acid P extracted from the corpuscles is determined. Fig. 9 illustrates the results obtained, including the fact that, after the lapse of about 33 days, the specific activity of the nucleic acid P became constant. This indicates that all corpuscles present in the circulation of the hen were formed during the experiment. In the corpuscle samples taken in the course of the first four days, only minute amounts of labelled nucleic acid were found to be present. This may be interpreted by supposing that the formation of the corpuscles in the marrow, up to the point of their release into the circulation, requires four days; 3.5 percent of the corpuscle content of the hen is thus built up daily.

I have attempted to give a short review of the earliest applications of isotopic indicators and to discuss a few examples of their earlier and more recent employment. Their use may be much extended in the time to come.

1. G. de Hevesy and F. Paneth, *Z. Anorg. Chem.*, 82 (1913) 322.
G. de Hevesy and E. Róna, *Z. Physik. Chem. Leipzig*, 89 (1915) 294, 303.
2. F. Paneth and G. de Hevesy, *Monatsh.*, 36 (1915) 75.
3. G. de Hevesy, *Phys. Z.*, 16 (1915) 59. Comp. also O. Erbacher, *Z. Physik. Chem. Leipzig*, A 163 (1933) 196.
4. G. de Hevesy and M. Biltz, *Z. Physik. Chem. Leipzig*, B 3 (1929) 271.
5. F. Paneth, *Z. Elektrochem.*, 28 (1922) 113.
6. F. Paneth, *Radio Elements as Indicators*, New York, 1928.
7. F. Paneth and E. Winternitz, *Ber.*, 51 (1918) 1728.
8. F. Paneth and O. Nörring, *Ber.*, 53 (1920) 1693.

9. J. Gróh and G. de Hevesy, *Ann. Physik*, 63 (1920) 85.
10. *Ann. Physik*, 65 (1921) 216.
11. G. de Hevesy and A. Obrusheva, *Nature*, 115 (1925) 674.
12. G. de Hevesy and W. Seith, *Z. Physik*, 56 (1929) 791.
13. G. de Hevesy and W. Seith, *Metallwirtschaft*, 13 (1934) 479.
G. de Hevesy and W. Seith, *Z. Physik*, 57 (1929) 869.
G. de Hevesy, W. Seith and A. Keil, *Z. Physik*, 79 (1932) 197.
14. G. de Hevesy, *Sitzber. Akad. Wiss. Wien*, 129 (1920) I.
C. Wagner, *Z. Physik. Chem. Leipzig*, B 15 (1932) 147.
15. W. Seith, *Ber. Naturforsch. Ges. Freiburg*, 30 (1930) 1.
16. H. A. C. McKay, *Trans. Faraday Soc.*, 34 (1938) 845; O. Frisch, G. de Hevesy and H. A. C. McKay, *Nature*, 137 (1936) 149. Comp. also A. Sagrubskij, *Physik. Z. Sowjet-union*, 12 (1937) 118.
17. G. de Hevesy and L. Zechmeister, *Ber.*, 53 (1920) 410.
18. G. de Hevesy and L. Zechmeister, *Z. Elektrochem.*, 26 (1920) 151.
19. V. Majer, *Z. Physik. Chem. Leipzig*, A 179 (1937) 51.
20. G. de Hevesy, *Math.-Phys. Commun., Copenhagen Acad. Sci.*, 3 (1921) 12.
21. Comp. G. T. Seaborg, *Chem. Rev.*, 27 (1940) 199.
22. G. de Hevesy and R. Hobie, *Nature*, 128 (1931) 1038; *Z. Anal. Chem.*, 88 (1932) I.
23. H. Ussing, *Nature*, 144 (1939) 977.
24. D. Rittenberg and G. L. Foster, *J. Biol. Chem.*, 133 (1940) 737.
25. E. Chargaff, M. Ziff and D. Rittenberg, *J. Biol. Chem.*, 144 (1942) 343.
26. G. de Hevesy, *Biochem. J.*, 17 (1923) 439.
27. L. J. Mullins and S. C. Brooks, *Science*, 90 (1939) 256.
Comp. also S. C. Brooks, *Trans. Faraday Soc.*, 33 (1937) 1002; *Proc. Soc. Exptl. Biol. Med.*, 38 (1938) 856.
28. H. Jenny and R. Overstreet, *J. Phys. Chem.*, 43 (1939) 1185.
29. T. C. Broyer and R. Overstreet, *Am. J. Botany*, 27 (1940) 425.
30. G. de Hevesy, *Botan. Commun., Stockholm Acad. Sci., Arkiv Bot.*, 33 A, Nr. 2 (1946).
31. J. A. Christiansen, G. de Hevesy and Sv. Lomholt, *Compt. Rend.*, 178 (1924) 1324; 179 (1924) 241. Comp. also B. Behrens, *Arch. Exptl. Pathol. Pharmakol.*, 109 (1925) 332.
32. Sv. Lomholt, *Brit. J. Venereal Diseases*, Jan. 1925.
33. G. de Hevesy and O. H. Wagner, *Arch. Exptl. Pathol. Pharmakol.*, 149 (1930) 336.
34. R. Schoenheimer, *The Dynamic State of Body Constituents*, Cambridge, Mass., 1942.
35. H. Urey, *Les Prix Nobel*, 1934.
36. G. de Hevesy and E. Hofer, *Klin. Wochschr.*, 13 (1934) 1524; *Nature*, 134 (1934) 879.
37. G. de Hevesy and C. F. Jacobsen, *Acta Physiol. Scand.*, I (1940) II.
L. Hahn and G. de Hevesy, *ibid.*, I (1941) 347.
38. L. B. Flexner, A. Gellhorn and M. Merrell, *J. Biol. Chem.*, 144 (1942) 35.
39. F. Joliot, *Les Prix Nobel*, 1935; I. Joliot-Curie, *Les Prix Nobel*, 1935; E. Fermi, *Les Prix Nobel*, 1938.
40. O. Chiewitz and G. de Hevesy, *Nature*, 136 (1935) 754; *Biol. Commun., Copenhagen Acad. Sci.*, 13 (1937) 9.
41. E. O. Lawrence, *Les Prix Nobel*, 1939.

42. P. F. Hahn, W. F. Bale, E. O. Lawrence and G. H. Whipple, *J. Am. Med. Assoc.*, III (1938) 2285; *J. Exptl. Med.*, 69 (1939) 739; 71 (1940) 731; W.M. Balfour, P.F. Hahn, W. F. Bale, W. T. Pommerenke and G. H. Whipple, *J. Exptl. Med.*, 76 (1942) 15; P. F. Hahn, W. F. Bale, J. F. Ross, W. M. Balfour and G. H. Whipple, *J. Exptl. Med.*, 78 (1943) 169; W. B. Hawkins and P. F. Hahn, *J. Exptl. Med.*, 80 (1944) 31. Comp. also M. E. Antoni and D. M. Greenberg, *J. Biol. Chem.*, 134 (1940) 27.
43. S. Ruben, W. Z. Hassid and M. D. Kamen, *J. Am. Chem. Soc.*, 61 (1939) 661; 62 (1940) 3443. I. H. C. Smith and D. B. Cowie, *Plant Physiol.*, 16 (1941) 257. A. W. Frenkel, *Plant Physiol.*, 16 (1941) 654.
44. S. Hertz, A. Robert and E. D. Evans, *Proc. Soc. Exptl. Biol. Med.*, 38 (1938) 510. J. G. Hamilton and M. H. Soley, *Am. J. Physiol.*, 126 (1939) 521. C. P. Leblond, P. Sue and A. Chamorro, *Compt. Rend. Soc. Biol.*, 133 (1940) 540. C. P. Leblond and P. Sue, *Compt. Rend. Soc. Biol.*, 133 (1940) 540. I. Perlman, I. L. Chaikoff and M. E. Morton, *J. Biol. Chem.*, 139 (1941) 433. M. E. Morton, I. Perlman and I. L. Chaikoff, *ibid.*, 140 (1941) 603. S. Hertz and A. Roberts, *J. Clin. Invest.*, 21 (1942) 31. S. Hertz, A. Roberts and W. Salter, *J. Clin. Invest.*, 21 (1942) 25. W. Mann, Ch. P. Leblonnet and S. L. Warren, *J. Biol. Chem.*, 142 (1942) 905. A. S. Keston, R. P. Ball, V. K. Frantz and W. W. Palmer, *Science*, 95 (1942) 362. M. E. Morton and I. L. Chaikoff, *J. Biol. Chem.*, 147 (1943) 719. C. P. Leblond, J. Gross, W. Peacock and R. D. Evans, *Am. J. Physiol.*, 140 (1943/1944) 671. A. L. Franklin, I. L. Chaikoff and S. R. Levner, *J. Biol. Chem.*, 153 (1944) 151.
45. M. E. Morton, I. L. Chaikoff, W. O. Reinhardt and E. Anderson, *J. Biol. Chem.*, 147 (1943) 757.
46. The survey by G. de Hevesy in *Ann. Rev. Biochem.*, 9 (1940) 641 includes papers on the application of ^{32}P as an indicator published prior to November 1, 1939. A summary of the application of ^{32}P and other labelling agents to the study of phosphatide metabolism is given by I. L. Chaikoff, *Physiol. Rev.*, 22 (1942) 291. Comp. also J. G. Hamilton, *J. Appl. Phys.*, 12 (1941) 440.
47. G. de Hevesy, *J. Chem. Soc.*, (1939) 1213.
48. G. de Hevesy and L. Hahn, *Biol. Commun., Copenhagen Acad. Sci.*, 15 (1940) 5.
49. G. de Hevesy, H. Levi and O. Rebbe, *Biochem. J.*, 34 (1940) 532. Comp. also R. S. Manly, H. C. Hodge and M. L. Manly, *J. Biol. Chem.*, 134 (1940) 293.
50. L. A. Erf, *Proc. Soc. Exptl. Biol. Med.*, 47 (1941) 287.
51. A survey of these investigations is given by W. D. Armstrong, *Ann. Rev. Biochem.*, (1942). Comp. also P. O. Pedersen and B. Schmidt-Nielsen, *Schweiz. Monatsschr. Zahnheilk.*, 51 (1941) 647; *Acta Odontol. Scand.*, 4 (1942) I.
52. G. de Hevesy, J. J. Holst and A. Krogh, *Biol. Commun., Copenhagen Acad. Sci.*, 13 (1937) I.
53. L. Hahn and G. de Hevesy, *Acta Physiol. Scand.*, I(1941) 347.
54. J. Ariel, W. F. Bale, V. Downing, H. C. Hodge, S. van Voorhis, S. L. Warren and H. J. Wilson, *Am. J. Physiol.*, 132 (1941) 346. D. M. Greenberg, R. B. And, M. D. O. Boelter, W. Wesley Campbell, W. E. Cohn and M. M. Murayama, *Am. J. Physiol.*, 149 (1943/1944) 147.
54. M. Joseph, W. E. Cohn and D. M. Greenberg, *J. Biol. Chem.*, 128 (1939) 673. T. R.

- Noonan, W. O. Fenn and L. Haege, *Am. J. Physiol.*, 129(1940) 432. G. de Hevesy and L. Hahn, *Biol. Commun., Copenhagen Acad. Sci.*, 16 (1941) I. G. de Hevesy, *Acta Physiol. Scand.*, 3 (1941) 123. W. E. Cohn, *Am. J. Physiol.*, 133 (1941) 242. R. B. Dean, L. Haege and W. O. Fenn, *J. Gen. Physiol.*, 24 (1941) 3 13.
55. W. E. Cohn and T. E. Cohn *Proc. Soc. Exptl. Biol. Med.*, 41 (1939) 455. L. Hahn, G. de Hevesy and O. Rebbe, *Biochem. J.*, 33 (1939) 1540. L. I. Mullins, T. R. Noonan, L.F. Haege and W.O. Fenn, *Am. J. Physiol.*, 133 (1941) 394; 135 (1941/1942) 93. A. Krogh, *Acta Physiol. Scand.*, 6 (1944) 203. A. Krogh, A. L. Lindberg and B. Schmidt-Nielsen, *ibid.*, 7 (1944) 221.
56. G. de Hevesy and L. Hahn, *Biol. Commun., Copenhagen Acad. Sci.*, 15 (1940) 6. D. B. Zilversmit, C. Entenman, M. L. Montgomery and I. L. Chaikoff, *J. Gen. Physiol.*, 26 (1943) 3 3 3. Comp. also F. L. Haven and W. F. Bale, *J. Biol. Chem.*, 129 (1939) 23.
57. L. Hahn and G. de Hevesy, *Mem. Carlsberg*, 22 (1938) 188. O. Meyerhof, P. Ohlmeyer, W. Genmer and H. Maier-Leibnitz, *Biochem. Z.*, 298 (1938) 396. E. Lundsgaard, *Scand. Arch. Physiol.*, 80 (1938) 291. G. de Hevesy and A. H. W. Aten, *Biol. Commun., Copenhagen Acad. Sci.*, 14 (1939) 5. G. de Hevesy and L. Hahn, *Biol. Commun., Copenhagen Acad. Sci.*, 15 (1940) 7.
58. R. F. Furchtgott and E. Shore, *J. Biol. Chem.*, 151 (1943) 65. E. V. Flock and I. L. Bollman, *ibid.*, 152 (1944) 371. H. M. Kalckar, *J. Biol. Chem.*, 154 (1944) 267.
59. C. Artom, C. Perrier, M. Santangelo, G. Sarzana and E. Segré, *Nature*, 139 (1937) 836; *Arch. Intern. Physiol.*, 45 (1937) 43 and 47 (1938) 245. L. Hahn and G. de Hevesy, *Scand. Arch. Physiol.*, 77 (1937) 148. C. Entenman, S. Ruben, I. Perlman, F. W. Lorenz and I. L. Chaikoff, *J. Biol. Chem.*, 126 (1938) 493. B. A. Fries, I. Ruben, I. Perlman and I. L. Chaikoff, *ibid.*, 123 (1938) 587. G. W. Changus, I. L. Chaikoff and S. Ruben, *ibid.*, 126 (1938) 493. E. Chargaff, *ibid.*, 128 (1939) 587. E. Chargaff, K. B. Olson and P. F. Partington, *ibid.*, 134 (1940) 505. G. de Hevesy and L. Hahn, *Biol. Commun., Copenhagen Acad. Sci.*, 15 (1940) 5. B. A. Fries, G. W. Changus and I. L. Chaikoff, *ibid.*, 132 (1940) 23. B. A. Fries and I. L. Chaikoff, *ibid.*, 141 (1941) 479. B. A. Fries, H. Schachner and I. L. Chaikoff, *ibid.*, 144 (1942) 59.
60. G. de Hevesy and J. Ottesen, *Acta Physiol. Scand.*, 5 (1943) 237. E. Andreasen and J. Ottesen, *Acta Microbiol. Scand., Suppl.* LIV, (1944) 26.
61. H. v. Euler and G. de Hevesy, *Biol. Commun., Copenhagen Acad. Sci.*, 17 (1942) 8. *Chem. Commun., Stockholm Acad. Sci.*, 17 A, Nr. 30 (1944). L. Ahlström, H. v. Euler and G. de Hevesy, *Chem. Comm., Stockholm Acad. Sci.*, 18 B, Nr. 13 (1944); 19 A, Nr. 9 (1944); 19A, Nr. 13 (1945).
62. H. M. Barrett, C. H. Best and J. H. Ridout, *J. Physiol.*, 93 (1938) 367. B. Cavenagh and H. S. Raper, *Biochem. J.*, 33 (1939) 17. W. M. Sperry, H. Waelsch and V. A. Stryanoff, *J. Biol. Chem.*, 135 (1940) 28. A. Waelsch, W. M. Sperry and V. A. Stryanoff, *ibid.*, 135 (1940) 291; 140 (1940) 885.
63. D. Stetten jr., *J. Biol. Chem.*, 138 (1941) 437; 140 (1941) 143; 142 (1942) 629.
64. G. de Hevesy and L. Hahn, *Biol. Commun., Copenhagen Acad. Sci.*, 14 (1938) 2. A. H. W. Aten, *Diss. Utrecht*, 1939. F. W. Lorenz, I. Perlman and I. L. Chaikoff, *Am. J. Physiol.*, 138 (1943) 318.
65. G. de Hevesy and E. Lundsgaard, *Nature*, 140 (1937) 275. L. Hahn and G. de Hevesy, *Biochem. J.*, 32 (1938) 342.

66. D. B. Zilversmit, C. Entenman, M. C. Fishler and I. L. Chaikoff, *J. Gen. Physiol.*, 26 (1943) 333. M. C. Fishler, C. Entenman, M. L. Montgomery and I. L. Chaikoff, *J. Biol. Chem.*, 150 (1943) 47.
67. A. H. W. Aten and G. de Hevesy, *Nature*, 142 (1938) III. A. H. W. Aten, *Diss. Utrecht*, 1939.
68. G. de Hevesy, H. Levi and O. Rebbe, *Biochem. J.*, 32 (1938) 2147.
69. J. K. Parnas, *Enzymologia*, 5 (1938-1939) 166. G. de Hevesy, T. Baranowski, A. J. Gutke, P. Ostern and J. K. Parnas, *Acta Biol. Exptl. Warsaw*, 12 (1938) 34. J. K. Parnas, *Bull. Soc. Chim. Biol.*, 21 (1939) 1059. T. Korzybski and J. K. Parnas, *ibid.*, 21 (1939) 713.
70. O. Meyerhof, P. Ohlmeyer, W. Gentner and H. Meyer-Leibnitz, *Biochem. Z.*, 298 (1938) 396.
71. A. Robinson, I. Perlman, S. Ruben and I. L. Chaikoff, *Nature*, 144 (1938) 119. H. Bulliard, J. Grundland and A. Moussa, *Compt. Rend.*, 207 (1938) 745; 208 (1939) 843. S. A. Fries, H. Schachner and I. L. Chaikoff, *J. Biol. Chem.*, 144 (1942) 59. A. Taurog, I. L. Chaikoff and I. Perlman, *ibid.*, 145 (1942) 281.
72. L. Ahlström, H. v. Euler and G. de Hevesy, *Chem. Commun., Stockholm Acad. Sci.*, 21 A, Nr. 6 (1945).
73. G. de Hevesy, K. Linderström-Lang, A. S. Keston and C. Olsen, *Mem. Carlsberg*, 23 (1940) 213.
74. G. de Hevesy, L. Hahn and O. Rebbe, *Biol. Commun., Copenhagen Acad. Sci.*, 14 (1939) 3. K. Kjerulf-Jensen, *Acta Physiol. Scand.*, 3 (1942) 193.
75. L. Hahn and G. de Hevesy, *Acta Physiol. Scand.*, 3 (1942) 193.
76. L. Hahn and G. de Hevesy, *Acta Physiol. Scand.*, 1 (1940) 1. G. de Hevesy and K. Zerahn, *ibid.*, 4 (1942) 376.
77. G. de Hevesy, K. H. Köster, G. Sorensen, E. Warburg and K. Zerahn, *Acta Med. Scand.*, 116 (1944) 561.
78. G. Nylin and M. Malm, *Cardiologia*, 7 (1943) 153.
79. P. F. Hahn, W. M. Balfour, J. F. Ross, W. F. Bale and G. H. Whipple, *Science*, 93 (1940) 87.
80. G. de Hevesy and J. Ottesen, *Nature*, 156 (1945) 534.
81. K. Bonhoeffer, *Z. Elektrochem.*, 40(1934) 469.
82. R. Schoenheimer, S. Rattner and D. Rittenberg, *J. Biol. Chem.*, 130 (1939) 703.
83. S. Graff, D. Rittenberg and G. L. Foster, *J. Biol. Chem.*, 133 (1940) 745.
84. J. F. Manery and W. F. Bale, *Am. J. Physiol.*, 126 (1939) 578; 132 (1941) 215.
85. J. H. E. Griffiths and B. G. Maegraith, *Nature*, 143 (1939) 159.
86. G. de Hevesy and O. Rebbe, *Nature*, 141 (1938) 1907.
87. A. M. Brues, M. M. Tracy and W. E. Cohn, *J. Biol. Chem.*, 155 (1944) 619.
88. G. Nylin, *Chem. Commun., Stockholm Acad. Sci.*, 20 A, Nr. 17 (1945).
89. P. F. Hahn, W. F. Bale and W. M. Balfour, *Am. J. Physiol.*, 135 (1941-1942) 800.

THE NERVE GROWTH FACTOR: THIRTY-FIVE YEARS LATER

Nobel lecture, December 8, 1986

by

RITA LEVI-MONTALCINI

Istituto di Biologia Cellulare, via G. Romagnesi 18/A, ROMA, Italy

- 1) Neurogenesis and its early experimental approach
- 2) Experimental neuroembryology in the forties
- 3) The unexpected break: a gift from malignant tissues
- 4) NGF at its early in-vitro and in-vivo debuts
- 5) The vital role of NGF in the life of its target cells
- 6) NGF as a retrograde trophic messenger and tropic factor
- 7) Neuronal and non-neuronal target cells
- 8) The I.D. card of NGF
- 9) NGF, growth factors and protooncogenes
- 10) NGF in exocrine glands: a fortuitous presence or a biological function?
- 11) Foreseeable approaches and predictions of the unpredictable

Neurogenesis and its early experimental approach

"Embryogenesis is in some way a model system. It has always been distinguished by the exactitude, even punctilio, of its anatomical descriptions. An experiment by one of the great masters of embryology could be made the text of a discourse on scientific method. But something is wrong, or has been wrong. There is no *theory* of development in the sense in which Mendelism is a theory that accounts for the results of breeding experiments. There has therefore been little sense of progression or timeliness about embryological research. Of many papers delivered at embryological meetings, however good they may be in themselves, one too often feels that they might have been delivered five years beforehand without making anyone much the wiser, or deferred for five years without making anyone conscious of a great loss" [1].

This feeling of frustration, so incisively conveyed by these considerations by P. Medawar, pervaded in the forties the field of experimental embryology which had been enthusiastically acclaimed in the mid-thirties, when the upper lip of the amphibian blastopore brought this area of research to the forefront of the biological stage. The side branch of experimental neuroembryology, which had stemmed out from the common tree and was entirely devoted to the study of the trophic interrelations between neuronal cell populations and between these and the innervated organs and tissues, was then in its initial vigorous

growth phase. It in turn suffered from a sharp decrease in the enthusiasm that had inflamed the pioneers in this field, ever since R. G. 'Harrison delivered his celebrated lecture on this topic at the Royal Society in London in 1935 [2].

Although the alternate 'wax and wane' cycles are the rule rather than the exception in all fields of human endeavor, in that of biological sciences the 'wane' is all too often indicative of a justified loss of faith in the rational and methodical approach that had at first raised so much hope.

A brief account of the state-of-the-art of experimental neuroembryology in the forties, when interest in this approach to the study of the developing nervous system was waning, is a prerequisite for understanding the sudden unforeseeable turn of events which resulted in the discovery of the Nerve Growth Factor.

Experimental neuroembryology in the early forties

The replacement, in 1934 by Viktor Hamburger, of the chick embryo with that of the amphibian larva as object of choice for the analysis of the effects of limb bud extirpation on spinal motor neurons and sensory nerve cells innervating the limbs [3], signed the beginning of a long series of investigations centered on the analysis of this and related experimental systems in avian embryos. Here I shall only list the major advantages offered by the chick embryo over amphibian larvae as object of neurological investigations.

The avian nervous system is built according to a more elaborate design than that of amphibians, and it lends itself to a more rigorous analysis of its nerve centers than that of lower vertebrates. Extensive fundamental studies on the nervous system of the chick embryo, with use of the invaluable silver specific techniques by Ramon y Cajal and coworkers, extended recently by myself and other investigators [4, 5], provided a very accurate blueprint of most nerve centers and their developmental history during neurogenesis. This allowed the detection of even small infractions to normal developmental rules in experimentally manipulated embryos. At variance with ontogenetic processes in amphibians, the same processes in chick embryos unfold according to a rigidly scheduled time sequence which never departs from the anticipated. It is therefore possible to compare the central and peripheral nerve centers of experimental and control specimens in embryos incubated at the same temperature and environmental conditions. The analysis, in amphibian larvae, was extended to the brain, spinal cord and peripheral nervous system under various experimental situations. In the chick embryo, it was mainly confined to the study of the effects called forth by extirpation of limb primordia or implantation of additional wing or leg buds on their innervating motor and sensory nerve centers. In 1934, Viktor Hamburger published an article [3] on the effects of wing bud extirpation on the development of the brachial spinal motor segment and sensory dorsal root ganglia innervating the wing. He came to the conclusion that the hypoplasia of motor nerve cells of the ventral horn and of other nerve cells of the same hemisection of the spinal cord resulted from lack of stimuli centripetally transmitted by nerve fibers of the first differentiated neurons. These normally exert a regulatory effect on proliferation and differen-

tiation of neighboring nerve cells. A reinvestigation of the effects produced by limb bud extirpation prospected a different control mechanism of the developing nerve centers by peripheral tissues. Through serial studies of silver stained embryos, the conclusion was reached that the severe hypoplasia of nerve centers deprived of their fields of innervation, resulted from death of differentiated neurons and not from failure of recruitment of neurons from a pool of still uncommitted nerve cell precursors [6, 7]. In 1947, Hamburger invited me to join him for the purpose of reinvestigating this problem. This invitation marked the beginning of a thirty year period that I spent at Washington University and of my life-long friendship with Viktor. Our 1949 article [7] confirmed the hypothesis previously submitted by G. Levi and myself [6]. The satisfaction of this confirmation of an important theoretical issue, and the successful analysis of other neuroembryological problems [8,9] was, however, perturbed by the awareness of the low resolution power of the techniques in our possession for in depth exploration of the tremendously complex neurogenetic processes. The temptation to abandon the experimental analysis of the developing nervous system and move into the phage field, in full blossom in the forties, did not take hold, however, thanks to unpredictable and most fortunate events which occurred at the same time and opened a new era in developmental neurobiology.

The unexpected break: a gift from malignant tissues

In a 1948 article, a former student of Viktor Hamburger, Elmer Bueker, reported the results of a bold and imaginative experiment consisting in grafting fragments of mouse sarcoma 180 in to the body wall of three-day chick embryos. The histological study of the embryos fixed 3-5 days later, showed that sensory nerve fibers emerging from adjacent dorsal root ganglia had gained access into the neoplastic tissue while no motor nerve fibers entered into the tumor [10]. The author concluded that histochemical properties of the fast growing mouse sarcoma offered a favorable field for growth of sensory fibers. This condition, in turn, resulted in a slight but consistent volume increase of these ganglia as compared to that of homologous ganglia innervating the wing of the contralateral side. Viktor and I reinvestigated this remarkable phenomenon adopting the method I had developed during my first neuroembryological studies, i. e., that of a daily inspection of control and experimental embryos serially sectioned and impregnated with a specific silver technique. Our results confirmed those reported by Bueker, but at the same time uncovered other effects elicited by the grafts of the mouse tumor, which hardly fit in with the hypothesis that they were in the same range and of the same nature as those called forth by transplants of normal embryonic tissues. They differed from the latter in the following, most significant, respects: sympathetic and not only sensory fibers gained access into the neoplastic tissues where they built a network of extraordinarily high density; nerve fibers branched at random between tumor cells without, however, establishing synaptic connections with them; sensory and sympathetic ganglia innervating the tumor underwent a progressive increase in volume, attaining, in the case of sympathetic ganglia,

a size about six times larger than that of same control ganglia [11]. Subsequent experiments uncovered another astonishing deviation from the norm in embryos bearing transplants of mouse sarcoma 180 or of another tumor of identical origin, known as sarcoma 37. It was found that embryonic viscera which in normal specimens are devoid of innervation, such as the mesonephroi, or which become scarcely innervated only in late developmental stages, such as the sex glands, the thyroid, the parathyroid and the spleen, were loaded with sympathetic nerve fibers during early embryonic stages (12). A patent infraction of all developmental rules came to light with the finding of thick sympathetic fiber bundles inside the veins of the host where they protruded in the form of large neuromas obstructing blood circulation (Fig. 1). All sympathetic chain ganglia, and not only ganglia adjacent to or in direct connection with neoplastic tissues, were enormously enlarged. The hypothesis that these anomalous effects could result from the release by neoplastic cells of a soluble, diffusible agent which altered the differentiative and growth properties of its target cells, received full

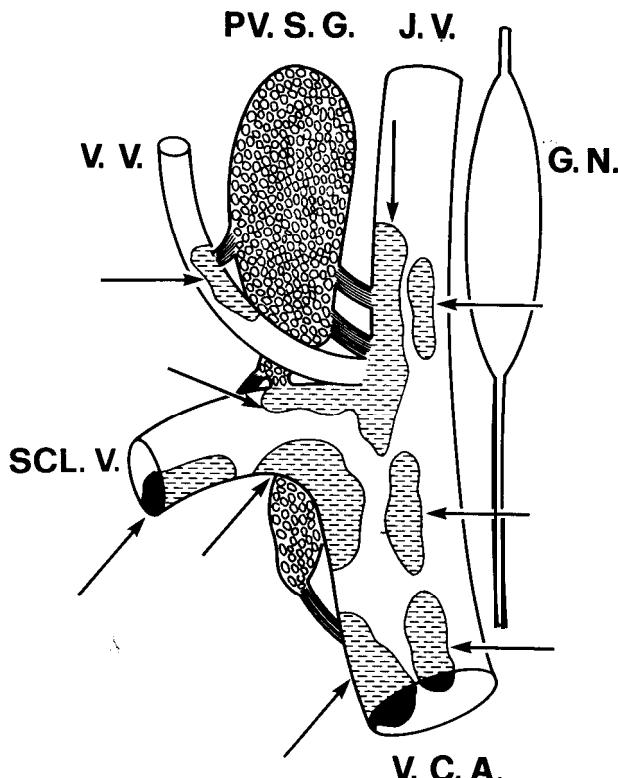


Fig 1. Sixteen-day chick embryo with intra-embryonic tumor (sarcoma 180). Ingrowth of sympathetic nerve fibers into the Jugular, Vertebral, Subclavian Anterior Caval Veins. GN, Ganglion Nodosum; JV, Jugular Vein; Pv.SG, Paravertebral Sympathetic Ganglion; SCL.V, Subclavian Vein; VCA, Anterior Caval Vein; VV, Vertebral Vein. Arrows point to nerve agglomerations. (from Ref. 12)

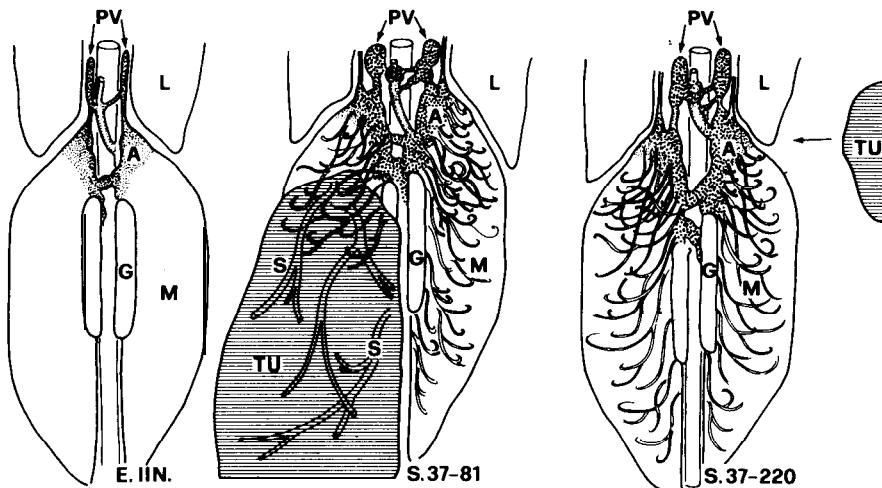


Fig. 2. Semi-diagrammatic reconstruction of a normal 11-day chick embryo (E.I of an 11-day embryo carrying an intra-embryonic transplant of mouse sarcoma (S 37-81) and of an 11-day embryo with transplant of sarcoma 37 on the chorio-allantoic membrane (S 37-220). Note the hyperplastic growth of the prevertebral chain ganglia in embryos carrying tumor transplants. Visceral nerve fibres from these ganglia invade the nearby mesonephroi. A, adrenal; G, gonad; L, lung; M, mesonephros; PV, prevertebral ganglia; S, sensory nerves; Tu, tumor. (from Ref. 12)

confirmation from experiments transplanting one or the other mouse sarcoma onto the chorio-allantoic membrane of 4 to 6-day chick embryos, in such a position as to prevent direct contact between embryonic and neoplastic tissues (Fig. 2). Embryonic and tumor tissues were, however, in reciprocal connection through the circulatory system. The finding that these extra-embryonic transplants elicited the same effects as intraembryonic grafts gave definite evidence for the diffusible nature of the tumoral nerve growth promoting factor [12, 13].

Attempts to replicate these effects by implanting dried tumor pellets or by injecting extract of either sarcoma were unsuccessful. I then thought of resorting to the tissue culture technique, which I had practiced with G. Levi at the University of Turin. Lack of facilities in this field in the Department of Zoology at Washington University, prompted me to ask hospitality from Professor Carlos Chagas, Director of the Biophysics Institute of the University of Brasil in Rio de Janeiro. There, a friend of mine, Hertha Meyer, had built and was director of a most efficient tissue culture unit. Upon approval and invitation by Professor Chagas, I boarded a plane for Rio de Janeiro, carrying in my handbag two mice bearing transplants of mouse sarcomas 180 and 37.

The Nerve Growth Factor at its early in-vitro and in-vivo debut

'The tumor had given a first hint of its existence in St. Louis, but it was in Rio de Janeiro that it revealed itself, and it did so in a theatrical and grand way, as if spurred by the bright atmosphere of that explosive and exuberant manifestation of life that is the Carnival in Rio' [14].

The discovery of the growth response elicited by a soluble tumoral agent revealed the receptivity of developing nerve cells to hitherto unknown humoral factors and in this way opened a new area of investigation. The *in vitro* bioassay offered a practical and invaluable tool for uncovering the identity card of this factor and paved the way for the study of its mechanism of action. Ink drawings, which I enclosed in several letters mailed from Rio to Viktor, give an eloquent account of the spectacular way in which this still unknown agent revealed itself. Sensory and sympathetic ganglia explanted from 8-day chick embryos in a semi-solid medium in proximity to, but not in contact with, fragments of mouse sarcoma 180 or 37 produced, in a 24 hour period, a halo of nerve fibers of maximal density on the side facing the tumor [15] (Fig. 3). The euphoric state elicited by this discovery was, however, soon damped by the discovery that normal mouse tissues, at variance with those of chick embryos, induce a milder, but not substantially different effect from that of mouse sarcomas. In retrospect, this should have alerted us to a novel and even more significant aspect of these *in vitro* experiments; namely, the widespread presence of the factor endowed with nerve growth promoting activity in normal and neoplastic tissues. The failure to realize the significance of this 'mouse effect' was beneficial rather than detrimental, since for the next two years our attention was entirely focussed on the study of the chemical nature of the factor released by the two mouse sarcomas, in much larger quantities than from normal mouse tissues.

A young biochemist, Stanley Cohen, who joined our Group shortly before my return from Rio, isolated from the two tumors a nucleoprotein fraction endowed with the *in vitro* nerve growth promoting activity [16]. Chance, rather than calculated search, signed a new, most fortunate turn of events. In order to degrade the nucleic acids present in this active fraction, Stan made use of snake venom which contains, among other enzymes, also the nucleic acid degrading enzyme, phosphodiesterase. Its addition in minute amounts to the nucleoprotein tumor fraction, was expected to suppress the formation of the fibrillar halo if nucleic acids rather than the protein were responsible for the nerve growth promoting effect elicited by this fraction. The startling result was a marked increase in the density of the fibrillar halo around the ganglia incubated in the presence of the tumoral fraction treated with snake venom. Since a dense fibrillar halo was produced also around ganglia cultured in the presence of minute amounts of snake venom alone, it became apparent that the venom itself was a most potent source of nerve growth promoting activity. On the basis of biochemical studies, Cohen was in fact able to show that equivalent growth stimulation effects were obtained by 15,000 µg of a sarcoma 180 homogenate and 6 µg of the moccasin snake venom. From the latter he isolated, after several purification steps, a non-dialyzable, heat-labile substance endowed with nerve growth promoting activity, identified as a protein molecule with a molecular weight in the order of 20,000 [17,18]. Microgram quantities of the purified snake venom fraction endowed with nerve growth promoting property, injected daily into the yolk of 6 to 8-day chick embryos for a 3 to 5-day period, resulted in the overgrowth of sensory and sympathetic ganglia and excessive

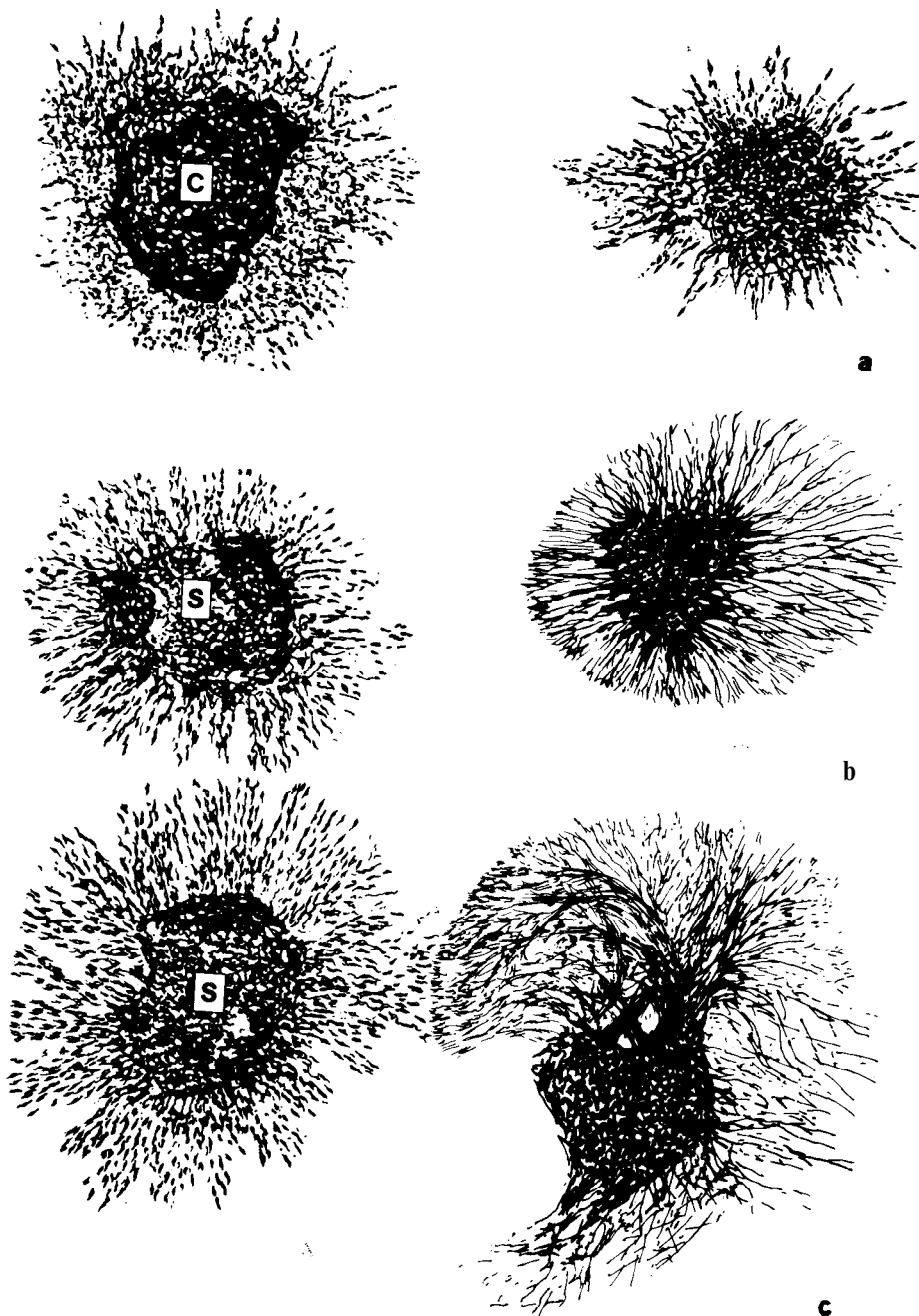


Figure 3. Drawings illustrating the in vitro "halo" effect on 8-day chick embryo sensory ganglia cultured in the presence of fragments of mouse sarcoma 180 for 24 hours (b) or 48 hours (c). In (a), the ganglion, which faces a fragment of chick embryonic tissue, shows fibroblasts but few nerve fibers. In (b) and (c), the ganglia, facing fragments of sarcoma 180, show the typical "halo" effect elicited by the growth factor released from the sarcoma. Note in (c) the first evidence of a neurotropic effect of the growth factor.

production of their fibers. Sympathetic nerve bundles branched profusely into the viscera and protruded into the cavity of the veins, mimicking in all details the effects elicited by grafts of mouse sarcomas [19].

If chance brought to our attention the unforeseeable presence of two nerve growth promoting sources, mouse sarcomas and snake venom, the sub

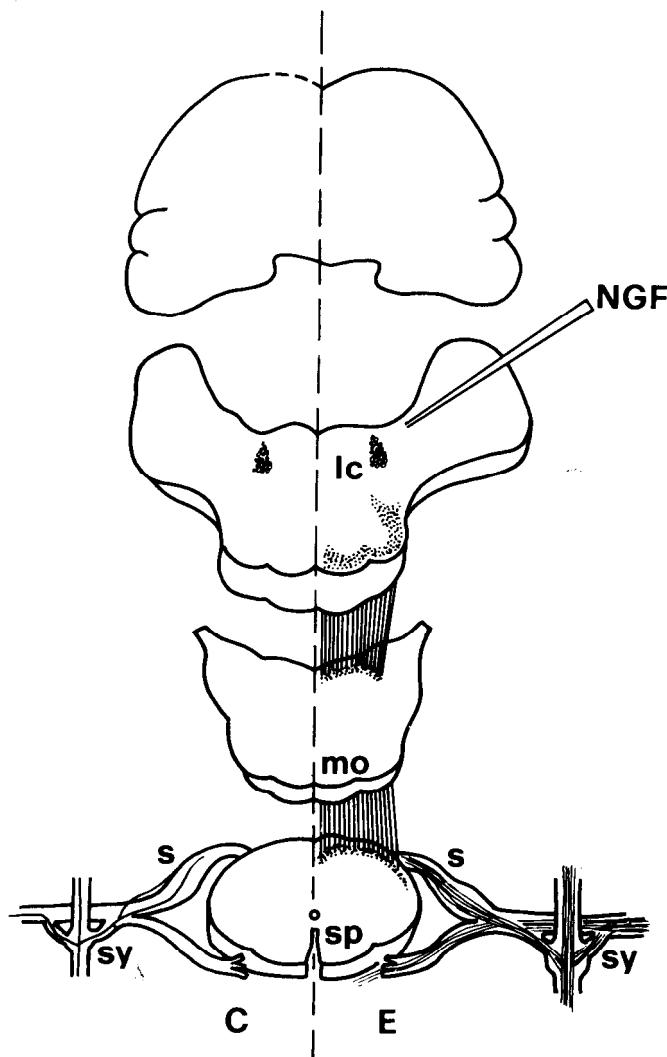


Figure 4. Diagrammatic representation of sympathetic fiber bundles which enter the spinal cord and medulla oblongata from adjacent sympathetic ganglia in intracerebrally NGF injected neonatal rats. Left half: control (C). Right half: experimental (E) embryo. NGF, site of injection of NGF into the floor of the fourth ventricle; lc, locus coeruleus; mo, medulla oblongata; sp, spinal cord; s, sensory ganglia; sy, sympathetic ganglia. Sympathetic fibers run across the sensory ganglion and enter into the neural tube with the dorsal roots. (from Ref. 40)

sequent finding that mouse submandibular salivary gland extract added in a minute quantity to the culture medium elicits an even denser and more compact fibrillar halo, was the result of a calculated search. These glands, as the homologue of the snake venom glands, were chosen by Stanley Cohen [20] as more likely than other organs screened with the in vitro bioassay, to store the nerve growth factor (NGF). These results were soon followed by purification and identification by Cohen of the salivary factor as a protein molecule with a molecular weight of 44,000 [20]. Its availability in larger quantities than the venom NGF, and its moderate toxicity when injected in a highly purified form, made possible the exploration of its biological activity in neonatal, young and adult mammals [21]. The results of these investigations signed the beginning of an ever more extensive and systematic in vivo and in vitro analysis of the salivary NGF, its chemical structure, as well as its mechanism and spectrum of action. Only the most significant findings reported from several laboratories in original and review articles will be considered in the following pages.

The vital role of NGF in the life of its target cells

In spite of, or perhaps because of its most unusual and almost extravagant deeds in living organisms and in-vitro systems, NGF did not at first find enthusiastic reception by the scientific community, as also indicated by the reluctance of other investigators to engage in this line of research. The finding that a protein molecule from such diverse and unrelated sources as mouse sarcomas, snake venom and mouse salivary glands, elicited such a potent and disrupting action on normal neurogenetic processes, did not fit into any conceptual preexisting schemes, nor did it seem to bear any relationship to normal control mechanisms at work during ontogenesis. It was in this skeptical atmosphere that NGF asserted, in a most forceful way, its vital role in the life of its target cells. Previous in vitro experiments had shown that incubation of snake venom with its antiserum inhibited the fiber outgrowth induced by the venom NGF. A specific antiserum to salivary NGF likewise abolished the formation of the in vitro fibrillar halo. These results suggested testing the effect of daily injections of small amounts of this antiserum (AS-NGF) in neonatal rodents. The inspection of treated mice, performed at the end of the first month with stereo and optic microscopes, revealed the near total disappearance of sympathetic para- and prevertebral chain ganglia [22-24]. This dramatic effect, which deprives newborn rodents and other neonatal mammals injected with antiserum to salivary NGF of the sympathetic system, without interfering with their normal development and vitality, became known as immunosympathectomy [25, 26]. The same treatment produces much less damaging effects in young and adult animals.

Two alternative hypotheses were submitted to explain the mechanism underlying the destructive effects of the antiserum: 1) a complement-mediated cytotoxic effect, or 2) inactivation of NGF or of an NGF-like protein essential for differentiation and survival of sympathetic nerve cells. Although the first hypothesis was favored in early articles, the second progressively gained support and is now generally accepted on the basis of this and an in-

vitro experimental approach which provided additional, unequivocal evidence of the essential role NGF plays in the early differentiation stages of its target cells. The in vitro experiments consisted of the dissociation of sensory and sympathetic nerve cells from ganglia of 8-11 day chick embryos and their incubation in minimum essential media. Nerve cells failed to survive unless nanogram quantities of NGF were added daily to the culture medium [27]. The in vitro evidence for the role of NGF in the early phases of development of sensory nerve cells, received confirmation from subsequent experiments which proved that administration of NGF antiserum to rodent fetuses [28, 29] and autoimmunization of pregnant rodents against endogenous NGF, [30] result in failure of sensory ganglia to undergo normal development.

NGF as a retrograde trophic messenger and tropic factor

The evidence in favor of the hypothesis that immunosympathectomy results from removal, by AS-NGF [31], of circulating endogenous NGF, raised the questions of how NGF reached its target cells and what were its sources of production. Subsequent experimental pharmacological and surgical approaches provided satisfactory answers to both questions, and in view of the interest in these problems, techniques and main findings will be briefly reported.

Administration to neonatal rodents of drugs such as 6-hydroxydopamine, which destroys adrenergic nerve endings [32], or of vinblastine, which blocks axonal transport [33], results in death of the large majority of sympathetic nerve cells in their most active phase of differentiation and growth. The degenerative effects produced by these drugs are of the same magnitude as those produced by administration of AS-NGF and result in the destruction of para- and prevertebral sympathetic ganglia through a process which became known as chemical sympathectomy [32,33]. A third experimental manipulation, consisting of the surgical transection of postganglionic axons of superior cervical ganglion performed in neonatal rodents, results in death of about 90% of immature sympathetic cells in this ganglion [34]. The experimental evidence that in all instances nerve cell death is prevented by an exogenous supply of NGF [30,35-37] demonstrates the vital role played by this molecule in the life and differentiation of these cells. The subsequent demonstration that labelled NGF is taken up by the nerve endings of sympathetic [38] or sensory fibers [39] and is retrogradely transported to the cell perikarya, lent strong support to the concept of NGF as a strophic messenger, conveyed through nerve fibers from peripheral cells to the innervating neurons. Disconnection of the partners by chemical or surgical axotomy results in death of differentiating nerve cells deprived of this essential molecule.

At the same time as the vital role of NGF in developing sympathetic and sensory nerve cells was assessed and its retrograde transport from peripheral tissues was well documented, another important property of NGF - its ability to direct growing or regenerating axons of sensory and sympathetic fibers along its concentration gradient (neurotropism) - was definitely established through different in vivo and in vitro experimental approaches [40-45].

The first strong evidence for a NGF neurotropic effect was obtained from experiments of daily micro-injections of NGF into the 'floor of the fourth ventricle. A 7-day treatment resulted in the penetration of fiber bundles originating from sympathetic ganglia inside the neural tube and in their ending at the level of experimentally produced NGF pools [40-41] (Fig. 4). In vitro experimental approaches gave more rigorous proof that neurites of NGF target cells grow along a NGF concentration gradient and deflect their route according to the changed position of the NGF releasing pipette [42]. While these studies unequivocably establish the NGF neurotropic effect as independent from its trophic action, they leave unanswered the question of whether this effect is exerted via a local control of growth cone motility [43], altered adhesion of this locomotor organelle to the substratum [44-45], or other mechanism(s) [46-47].

Neuronal and non-neuronal target cells

As indicated in Table I, targets of NGF action that have been well characterized up to now, can be classified under three main categories: 1) neural-crest derivatives, 2) central nervous system (CNS) neurons; and 3) cells of non-neuronal origin. For a thorough analysis of the many and diversified effects exerted by NGF on each one of these cells, the reader is referred to review articles on this specific matter [47-52]. In this context I only wish to make some general considerations.

A generally valid rule is that all cells are maximally responsive to NGF action during their early differentiation; the response undergoes progressive restriction in the adult without, however, being totally effaced. Long sympathetic neurons and sensory neurons, with particular reference to those of the dorso-medial quadrant of spinal ganglia in chick embryo [12], provided a most valuable system for demonstrating the three main activities of NGF, i.e., 1) its vital trophic role during the early developmental stages, 2) its property of enhancing differentiative processes such as neurite outgrowth, and 3) of guiding the growing or regenerating neurites along its own concentration gradient [43-44]. These same cells offered an *in vivo* model system to study the induction of enzymes involved in neurotransmitter synthesis [53] and were also instrumental in providing the first demonstration of the retrograde transport of NGF [54] and its role as a trophic messenger [55]. If sensory and sympathetic cells played a key role in revealing these properties of NGF, chromaffin cells and their neoplastic counterpart, the clonal cell line PC12, became the model of choice for studying the capacity of NGF to modulate the phenotypic expression and molecular mechanism subserving this process [56]. The phenotypic shift induced by NGF both in chromaffin [57-58] and PC12 cells [56,59], resulting in their neuronal differentiation accompanied by a plethora of chemical, ultrastructural and morphological changes characteristic of the neuronal rather than glandular phenotype, is too well known to warrant a detailed description [50]. These cells, moreover, uncovered the startling capacity of NGF to act both as a mitogenic [60] and a antimitotic agent [56], even within the context of the same clonal cell line PC12 and of a mutated version of it [61].

This, in turn, clearly pointed to the 'versatility' of NGF receptors and of their transduction machinery, whose message is evidently read and interpreted in different ways according to the cell type and previous cell history. The 'priming model' prospected to give a molecular account for the very fast and very slow onset of neurite outgrowth occurring, respectively, in sensory and sympathetic cells [24] on the one hand and PC12 cells on the other [50], is an excellent example of the contribution of these latter cells to studies on the mode of action of NGF.

Other examples of the wide and at the same time diversified NGF effects are illustrated by other sympatho-adrenal derivatives such as paraganglia, small intensely fluorescent cells (SIF) and carotid body cells [62-64]. A particularly impressive evidence of the capacity of NGF to modulate phenotypic expression is the case of SIF cells which have been hypothesized as immediate precursors of both sympathetic and chromaffin cells. When these cells are cultured under appropriate conditions, they can be channelled towards the first or the second phenotype in media supplied with NGF or with dexamethasone [63,64]. Such an interplay, even in fully differentiated cells, between NGF and steroid hormones, is also indirectly suggested by *in vivo* studies on the short adrenergic neurons which innervate the genito-urinary system in both sexes [65].

In more recent years, two new populations came to the forefront of studies on NGF target cells: CNS neurons and cells originating from the hematopoietic system.

Small and large neuronal populations located in different brain areas have been shown to exhibit all properties and responses typical of sensory and sympathetic cells, such as: 1) the presence of specific receptors [66], 2) retrograde transport of NGF [67], 3) increased neurotransmitter synthesis with special reference to acetylcholine [68-70], and 4) trophic response manifested as protection by exogenous NGF administration to selective noxious treatments or surgical transection otherwise leading to cell death [71,72]. A role for NGF in the development of as yet unidentified hypothalamic brain centers has been suggested by the finding that injections of affinity purified polyclonal antibodies against NGF in rat fetuses induce a severe postnatal neuroendocrine syndrome [29]. The loop of an unquestionable NGF role in brain is completed by the demonstration that other nerve cells, especially those located in the hippocampus and cortical areas, manufacture large quantities of NGF mRNA and NGF protein, thus closing the functional link between NGF-producing and NGF-responding cells [73-75]. As prospected in the last section of this article, although the range of NGF action in the CNS is qualitatively comparable to that previously observed in peripheral neurons, the actual extent of the NGF role in brain is far from foreseeable due to the vast repertoire of possible responses from nerve cells in the CNS.

An analogous general consideration holds for the effect exerted by NGF on mast cells and possibly on other cells of the immune system. The increased *in-vivo* and *in vitro* number of mast cells following NGF treatment [76,77], as well as the effect of this growth factor on histamine release [78-80], point to an unquestionable role in the physiology of these cells. It is not yet clear, however,

whether such an effect is exerted through a generalized action on all mast cell precursors or through a sort of clonal selection mechanisti. The more recent report of an NGF effect on other spleen cells, such as mononuclear cells [81], and the existence of NGF receptors on thymocytes [82], clearly suggests that the NGF action extends also to cells belonging to a network of enormous functional significance. The role played by histamine as an immunomodulator and the obvious involvement of spleen cells in the immune response of the organism prospect new scenarios in which NGF may gain access, not through a back door, but through the main entrance.

The I.D. card of NGF

Sequencing of mouse submandibular gland NGF, achieved in 1971 [83], provided invaluable information not only on its primary structure but was recently instrumental in the preparation of synthetic oligonucleotides which resulted in the identification of NGF cDNA [84]. The cloning which followed in rapid succession of mouse [84], human [85], bovine [86] and chick [87] genes, demonstrated their high degree of homology. The NGF gene, located in the human species on the proximal short arm of chromosome 1 [88], codes for a large polypeptide of 307 amino acid residues which, upon cleavage(s), gives rise to the 118 amino acid mature NGF subunit protein and, possibly, to other peptides of unknown function and with no sequence homology with presently identified proteins [84]. NGF is a dimer composed of two identical subunits held together by non-covalent bonds. The dimer can be isolated as such [89] or under the form of a complex also consisting of two other proteins, one with an esteropeptidase activity, probably involved in the processing of an NGF precursor, and the other with an as yet unknown function [90-92]. While it remains to be established whether each NGF subunit is biologically active, it has been demonstrated that a covalently cross-linked form of the dimer maintains full activity [91,92]. Between the two well-indentified molecular entities of NGF and of its coding gene, which can be visualized as the summit and the base of an iceberg, are several other possible intermediate forms of unknown nature and biological properties. Their identification would answer important questions such as: Are other biologically active peptides coded for by the NGF gene? What is the significance of different splicing in different cells of NGF mRNA [93]? Is the processing of pre-pro-NGF identical in all neuronal and non-neuronal cells or, as in other peptides [94], do alternate processing pathways result in the production of peptides endowed with different biological functions? Since the same peptides may undergo post-transcriptional or post-translational modification, the submerged areas of the NGF iceberg loom very large.

Studies on the immunological and biological relatedness of NGFs purified from different species strongly support the hypothesis that the site(s) of interaction with their receptors has remained structurally more constant than is the case for other epitopes, probably free to mutate in view of their less fundamental biological functions [95].

NGF, growth factor and oncogenes

The discovery of NGF, soon followed by that of Epidermal Growth Factor (EGF), led to the biological identification of an ever-growing list of polypeptide growth factors [48]. In the seventies, another apparently unrelated area of biology came to the forefront of research with the discovery of single gene products (oncogenes) causing transformation. Polypeptide Growth Factors (PGF) and oncogene research, pursued at first independently of each other, converged when homology between some oncogenes and growth factors or their receptors was shown by sequence analysis. Evidence is steadily increasing that excessive synthesis, or an altered version of PGFs or of their receptors, may result in transformation of recipient cells [95-98]. More recently, the demonstration that the opposite is also true, namely, that certain oncogene products may induce differentiation of recipient cells, called attention to another facet of this intricate interplay between differentiative and transforming processes. The case of H-ras and that of v-src, whose expression into PC12 cells [99,100] result in mitotic arrest and neuronal differentiation comparable to those elicited by NGF, provide instances of a list most likely to grow. The obvious conclusion is that a given polypeptide growth factor, or intracellular proteins playing essential roles in the cell cycle or in differentiation of some cells, may exert markedly different actions in distinct cell types. In the case of NGF, one wonders if and how other actions are elicited by this versatile molecule. For instance, is an altered version of NGF or of its receptors capable of causing transformation of some recipient cells, as has been shown for other PGFs? If this is the case, could NGF in a modified version or its receptors be implicated in neoplasia in the central and peripheral nervous systems?

NGF in exocrine glands: a fortuitous presence or a biological function?

The early discovery that mouse submandibular glands synthesize and release large quantities of NGF into the saliva, that the synthesis of this protein molecule is under the control of testosterone and of thyroxine [101,102] and that the NGF protein content is about ten-fold higher in male than in female mice, remained for about three decades a puzzling and unexplained finding. The conflicting but altogether negative attempts to reveal the presence of this molecule in the circulating blood [49,51], and the lack of any adverse effects on sympathetic and sensory cells by removal of these glands, which deprived these rodents of such a large NGF source, militated against the hypothesis that salivary NGF gains access to their target cells. An alternative biological function for salivary NGF was first hypothesized by our group [103], and recently proved by us [104] and another investigator [105]. We demonstrated that intraspecific fighting, experimentally induced in adult male mice by 6-9 weeks of social isolation, results in massive NGF release into the blood stream, an event prevented by previous sialoadenectomy. Since injections of NGF induce weight and size increase of the adrenal glands [105] and stimulate the synthesis of the catecholamine key enzyme, TH [106], we suggested that such a massive discharge into the blood circulation of endogenous salivary NGF may be instrumental in the defence and/or offense mechanisms of vital significance for

male mice that engage in intraspecific lighting among individuals of the same sex. In favor of this hypothesis, is a recent report that aggressive behavior results in the release into the blood of another biologically active protein, renin, synthesized in the same tubular portions of these glands [107]. The mechanism triggering this NGF release is not yet understood, nor is it known whether other stations are activated and play a role in this specific stress syndrome.

As for the presence of large NGF sources in snake venom [18] and male genital organs [108,109], they may be conceived as instances of bizarre evolutionary gene expression. Alternatively, in these cases NGF may subserve other functions which may somehow be linked with the poisonous action of snake venom, or the reproductive activity of the genital apparatus. In the case of snake venom, one can envisage the possibility that a highly specific neurotropit molecule such as NGF is utilized by reptiles as a carrier of other neurotoxins devoid of specific receptors in the central and peripheral nervous systems. For instance, enzymes such as phospholipases, phosphodiesterases and proteases of various nature, which by themselves may lack specific recognition sites in target cells, may exploit NGF as a carrier to gain access inside cells wherever there are specific NGF receptors. The widespread distribution of these specific molecules also in several non-neuronal cells could offer some toxins or enzymes a better access to their target organs.

In the reproductive tract, NGF could participate in fertilization mechanisms by cytoskeletal mediated activation of spermatozoa locomotion much in the same way as in neurite outgrowth, or by favoring egg implantation, via inhibition of rejection through the immune system. This latter hypothesis is presently under investigation (Geraci, Cocchiara and Calissano) by assessing the effect of NGF on uterine mast cells which, through histamine release, are postulated to prevent the local immune reaction [110].

Foreseeable approaches and predictions of the unpredictable

The most obvious among the foreseeable approaches is the search for other NGF target cells, using the ever more sophisticated in vivo and in vitro techniques which became available in these last decades. It was this multimodal approach which in recent years led to the discovery of NGF target cells in the CNS of lower and higher vertebrates and in cell lineages playing a role in the immune system. This list is likely to increase, as the search extends to other neuronal and non-neuronal cell populations. Furthermore, one should take into account the fact that some of these populations are receptive to NGF mainly during developmental stages in prenatal life. This was already demonstrated in sensory cells of avian and mammalian species [49,51,52], and in cells lining the third ventricle in amphibian tadpoles [111] and prenatal and neonatal rodents (Aloe and Levi-Montalcini, unpublished observations). Likewise, the systematic screening of neuroendocrine and hematopoietic cell lines in in vitro and in vivo systems may reveal other as yet uncovered roles of this growth factor.

Another approach now in progress in many laboratories is the search NGF-like factors active on other neuronal populations. These factors may be subdivi-

ded into two major classes: 1) those coded by the NGF gene itself but processed through alternate post-transcriptional or post-translational pathways leading to PGFs with a somewhat different structure and function; 2) other proteins or peptides having the trophic, chemotactic and/or differentiative activity of NGF, but coded by other genes.

The search for factors belonging to the first group and their identification will take advantage of the techniques of molecular biology and immunology. These should provide valuable information on some of the still unexplored, submerged areas of the NGF iceberg, dealing with the processes of the NGF gene transcription or translation. Of particular importance would be the identification of the NGF sequence responsible for its binding to receptors which may presumably trigger a given cellular response [47]. As previously surmised, [95], this portion has possibly been better conserved than other parts of the molecule. Once identified, it will be feasible to introduce, in its synthetic counterpart, amino acid substitutions and/or chemical modifications and explore the biological potency of the newly manufactured peptide. This approach should not only provide invaluable information on the nature and properties of the NGF active center, but, hopefully, will result in the synthesis of peptides endowed with an activity even higher than that of NGF itself, so brilliantly achieved in the field of other biologically active peptides [112, 113].

Within this category of studies on NGF and its coding gene, one can conceive a strategy aimed at exploiting the property of non-neuronal cells in peripheral tissues and of neurons and satellites in the CNS to manufacture and release NGF by resorting to pharmacological agents that modify NGF gene expression or processing. The well-established findings that NGF synthesis is increased following transection of nerve fibers connecting NGF receptive nerve cells to their targets [114] or via hormonal action [101,102], are an additional indication of the remarkable plasticity of the mechanisms controlling the expression of the NGF gene. It is conceivable that this property might be modulated by pharmacological agents acting on the same path as those involved in the regulation of the synthesis and release of NGF.

The search for neurotrophic factors coded by genes other than the NGF gene could take advantage, at least in its main lines, of the classical approach so successfully applied in the isolation and identification of NGF. Two main causes may explain why extensive work invested in this attempt has not been so successful in providing evidence for the existence of other PGFs activating different neuronal cell lines: 1) the lack of fast and reliable bioassays such as those developed for NGF and 2) the failure to detect large sources of these factors comparable to those fortuitously discovered in early NGF studies. The availability of rapid, highly reliable bioassays can, however, now be achieved by resorting to the use of most stringent, chemically defined media, permitting survival and differentiation of only given cell types, upon addition to the medium of putative growth factors extracted from different sources and screened with the *in vitro* bioassay for their potential specific growth enhancing activity. The problem of finding by sheer chance large sources of NGF-like peptides, such as those which played a key role in the discovery of NGF, can be

solved by resorting to techniques of protein chemistry and recombinant DNA technology. A few micrograms of purified protein are sufficient to decipher the sequence, prepare the corresponding cDNAs, identify the gene of the PGF in question, and express it in bacteria, thus replacing a search once guided by unpredictable strokes of luck, with a rational and systematic strategy.

Predictions of the unpredictable are encouraged by the same history of NGF which may be defined as a long sequence of unanticipated events which each time resulted in a new turn in the NGF unchartered route, and opened new vistas on an ever-changing panorama. This trend, which became manifest from the very beginning and in fact alerted me to the existence of NGF, is perhaps the most attractive, even though elusive trait of this thirty-live year long adventure. One can at present only predict where future developments are most likely to occur. The main causes of unpredictability of the findings, reside in the intricacy of the new surroundings where NGF is moving - the CNS and the immune system-rather than in NGF itself. The enormous complexity of these two networks, which on the basis of recent findings are closely interrelated and influence each other through bidirectional signals [115,116], opens endless possibilities for NGF activation of distinct repertoires of cells belonging to one or the other system. How many indirect effects can be elicited by direct NGF action on cholinergic, adrenergic and peptidergic neurons interlinked via fiber pathways and humoral channels or through short-distance diffusion? Likewise, how many effects could follow the simple histamine release by NGF activated mast cells, considering the well-established role of this amine as an immunomodulator or an immunosuppressor? These considerations hold also for the potential utilization of NGF in brain and immunosystem disorders. For instance, whenever cell death of specific neuronal populations may be linked to a decreased local availability of neurotrophic factors, such as NGF, its exogenous supply or stimulation of its endogenous production via pharmacological agents may offer a promising approach to presently incurable diseases.

I shall end this account of the unfolding of the NGF story with a remark made more than a decade ago by Viktor Hamburger: " - - - the fact that this discovery, which grew out of a seemingly peripheral problem (peripheral in every sense of the word), has blazed so many new trails is its greatest contribution in neuroembryology" [117]. Studies in this last decade have not only provided new strong evidence of the most important contributions of NGF in the field of neuroembryology, but brought to the fore its significance in the more general field of neuroscience and also prospect its role in that of the immune system.

I dedicate this article to Viktor Hamburger, who promoted and took part in this search, and to whom I am forever indebted for invaluable suggestions and generosity. Without him, the Nerve Growth Factor would never have come to our attention.

To my dear friends, Pietro Calissano and Luigi Aloe, I wish to express my deepest gratitude for their fundamental contributions. In this thirty-five year long investigation, a large number of colleagues, research associates and graduate students took part in this scientific adventure. I am particularly indebted

and I very gratefully acknowledge the most important work performed by two of them: Drs. Piero Angeletti and Vincenzo Bocchini. To Professor Carlos Chagas, for his generous hospitality in the Biophysic Institute of the University of Brasil, and to Dr. Hertha Meyer who helped me in devising the tissue culture bioassay of NGF, my warmest thanks.

Table 1. NGF TARGET CELLS

NEURAL CREST DERIVATIVES	
<i>Long sympathetic neurons</i>	
<i>Short sympathetic neurons</i>	
<i>Cells of parangglia (carotid & abdominal parangglia)</i>	
<i>Sympathoadrenal</i>	
<i>SIF</i> (small, intensely fluorescent) cells	
	<i>Chromaffin cells</i> {
	normal
	neoplastic (PC12)
<i>Sensory neurons</i>	
	<i>Cholinergic neurons:</i>
	{
	<i>Adrenergic, indoleaminergic,</i>
	nucleus diagonal band of Broca
	<i>Peptidergic neurons</i>
	Xenopus laevis tadpoles
	{
	<i>Mast cells</i>
<i>CENTRAL NERVOUS SYSTEM</i>	
<i>NON NEURONAL ORIGIN</i>	

REFERENCES

1. Medawar P. B., in *The Art of the Soluble*, (Methuen & Co., LTD, 1967) pp. 106-107.
2. Harrison R. G., *Proc. Roy. Soc. London Series B* 118, 155- 196 (1935).
3. Hamburger V., *J. Exp. Zool.*, 68,449-494 (1934).
4. Tello J. F., *Trabajos Lab. Invest. Biol. Univ. Madrid* 21, 1-93 (1922).
5. Levi-Montalcini R., *Prog. Brain Res.* 4, 1-29 (1964).
6. Levi-Montalcini R., Levi G., *Arch. Biol. Liege* 54, 183-206 (1943).
7. Hamburger V., Levi-Montalcini R., *J. Exp. Zool.* 111, 457-502 (1949).
8. Levi-Montalcini R., *J. Comp. Neurol.* 91, 209-242 (1949).
9. Levi-Montalcini R., *J. Morphol.* 86, 256-283 (1950).
10. Bueker E. D., *Anat. Rec.* 102, 369-390 (1948).
11. Levi-Montalcini R., Hamburger V., *J. Exp. Zool.* 116, 321-362 (1951).
12. Levi-Montalcini R., *Ann. N.Y. Acad. Sci.* 55, 330-343 (1952).
13. Levi-Montalcini R., Hamburger V., *J. Exp. Zool.* 123, 233-288 (1953).
14. Levi-Montalcini R., in *The Neurosciences: Paths of Discovery*, F. G. Worden, J. P. Swayzey, G. Adelman, Eds. (MIT Press, 1975), pp. 245-265.
15. Levi-Montalcini R., Meyer H., Hamburger V., *Cancer Res.* 14, 49-57 (1954).
16. Cohen S., Levi-Montalcini R., Hamburger V., *Proc. Natl. Acad. Sci. USA* 40, 1014-1018 (1954).
17. Cohen S., Levi-Montalcini R., *Proc. Natl. Acad. Sci. USA* 42, 571-574 (1956).
18. Cohen S., *J. Biol. Chem.* 234, 1129-1137 (1959).
19. Levi-Montalcini R., Cohen S., *Proc. Natl. Acad. Sci. USA* 42, 695-699 (1956).
20. Cohen S., *Proc. Natl. Acad. Sci. USA* 46, 302-311 (1960).
21. Levi-Montalcini R., Booker B., *Proc. Natl. Acad. Sci. USA* 46, 373-384 (1960).
22. Levi-Montalcini R., Booker B., *Proc. Natl. Acad. Sci. USA* 46, 384-391 (1960).
23. Levi-Montalcini R., *Science* 143, 105-110 (1964).
24. Levi-Montalcini R., *The Harvey Lectures* 60, 217-219 (1966)
25. Levi-Montalcini R., Angeletti P. U., *Pharmacol. Rev.* 18, 819-828 (1966).
26. Steiner G., Schönbaum E. Eds., *Immunosympathectomy* (Elsevier Publishing Co., Amsterdam, 1972).
27. Levi-Montalcini R., Angeletti P. U., *Dev. Biol.* 7, 653-659 (1963).
28. Levi-Montalcini R., Aloe L., Calissano P., Cozzari C., in 1st Meeting of the Internatl. Society for Dev. Neuroscience, Strasbourg, vol. 1, pp. 5 (1980).
29. Aloe L., Cozzari C., Calissano P., Levi-Montalcini R., *Nature* 291, 413-415 (1981).
30. Johnson E.M., Gorin P.D., Brandeis L.D., Pearson J.. *Science* 210, 916-918 (1980).
31. Goedert M., Otten U., Schaefer T., Schwab M., Thoenen H., *Brain Res.* 201, 399-409 (1980).
32. Angeletti P. U., Levi-Montalcini R., *Proc., Natl. Acad. Sci. USA* 65, 114- 121 (1970).
33. Calissano P., Monaco G., Menesini-Chen M. G., Chen J. S., Levi-Montalcini R., in *Contractile Systems in Non-muscle Tissue*, S. W. Perry, A. Margret, R. S. Adelstein, Eds. (Elsevier, Amsterdam, 1976a), pp. 201-211.
34. Hendry I.A., *Brain Res.* 90, 235-244 (1975).
35. Levi-Montalcini R., Aloe L., Mugnaini E., Oesch F., Thoenen H., *Proc. Natl. Acad. Sci. USA* 72, 595-599 (1975).
36. Hendry I. A., and Campbell, J. J. of *Neurocytol.* 5, 351-360 (1976).
37. Aloe L., Mugnaini E., Levi-Montalcini R., *Arch. Ital. Biol.* 113, 326-353 (1975).
38. Stöckel K., Paravicini U., Thoenen H., *Brain Res.* 76, 413-421 (1974).
39. Hamburger V., Brunso-Bechtold V. J. K., Yip J. W., *J. Neurosci.* 1, 60-71 (1981).
40. Levi-Montalcini R., *Prog. Brain Res.* 45, 235-258 (1976).
41. Menesini-Chen M. L., Chen J. S., Levi-Montalcini R., *Arch. Ital. Biol.* 116, 53-84 (1978).

42. Gundersen R. W., Barrett J. N., *Science* 206, 1079-1080 (1979).
43. Gundersen R. W., Barrett J. N., *J. Cell Biol.* 87, 546-554 (1980).
44. Campenot R. B., *Dev. Biol.* 93, 1-12 (1982).
45. Campenot R.B., *Dev. Biol.* 93, 13-41 (1982).
46. Pfenninger K. H., Johnson M. P., *Proc. Natl. Acad. Sci. USA* 78, 7797-7800 (1981).
47. Calissano P., Cattaneo A., Aloe L., Levi-Montalcini R., in *Hormonal Proteins and Peptides*, C. H. Li, Ed. (Academic Press, 1984), Vol. XII, pp. 1-56.
48. Bradshaw R.A., *Annu. Rev. Biochem.* 47, 191-216 (1978).
49. Thoenen H., Barde Y. A., *Physiol. Rev.* 60, 1284-1335 (1980).
50. Greene L.A., Shooter E. M., *Annu. Rev. Neurosci.* 3, 353-402 (1980).
51. Calissano P., Cattaneo A., Biocca S., Aloe L., Mercanti D., Levi-Montalcini R., *Exp. Cell Res.* 154, 1-9 (1984).
52. Levi-Montalcini R., Calissano P., *Trends Neurosci.* 9, 473-476 (1986).
53. Thoenen H., Angeletti P. U., Levi-Montalcini R., Kettler R., *Proc. Natl. Acad. Sci. USA* 68, 1598-1602 (1971).
54. Paravicini U., Stoeckel K., Thoenen H., *Brain Res.* 84, 279-291 (1975).
55. Johnson E. M., Yip H. K., *Nature* 314, 751-753 (1985).
56. Greene L. A., Tischler A. S., *Proc. Natl. Acad. Sci. USA* 73, 2424-2428 (1976).
57. Unsicker K., Krisch B., Otten U., Thoenen H., *Proc. Natl. Acad. Sri. USA* 75, 3498-3502 (1978).
58. Aloe L., Levi-Montalcini R., *Proc. Natl. Acad. Sci. USA* 76, 1246-1250 (1979).
59. Greene L. A., Liem R. K. H., Shelanski M. L., *J. Cell Biol.* 96, 76-83 (1983).
60. Lillien L. E., Claude P., *Nature* 317, 632-634 (1985).
61. Burnstein, D. E., Greene L. A., *Dev. Biol.* 94, 477-482 (1982).
62. Levi-Montalcini R., Aloe L., *Adv. Biochem. Psychopharmacol.* 25, 3-16 (1980).
63. Doupe A. J., Patterson P. H., Landis S.C., *J. Neurosci.* 5, 2143-2160 (1985).
64. Doupe A. J., Landis S. C., Patterson P. H., *J. Neurosci.* 5, 2119-2142 (1985).
65. Owman C., Sjoberg N. O., in *Proceedings of the 5th International Congress of Endocrinology*, V. H. T. James, Ed. (Excerpta Medica, Amsterdam, 1977), vol. I, pp. 205-209.
66. Szutowicz A., Frazier W.A., Bradshaw R.A., *J. Biol. Chem.* 251, 1516-1523 (1976).
67. Seiler M., Schwab M., *Brain Res.* 30, 33-39 (1984).
68. Gnahn H., Hefti F., Heumann R., Schwab M. E., Thoenen H., *Dev. Brain Res.* 9, 45-52 (1983).
69. Hefti F., Dravid A., Hartikka J., *Brain Res.* 293, 305-311 (1984).
70. Mobley W. C., Rutkowski J. L., Tennekoon G. I., Buchanan K., Johnston M. V., *Science* 229, 284-287 (1985).
71. Williams L. P., Varon S., Peterson G. M., Wictorin K., Fischer W., Bjorklund A., Gage F. H., *Proc. Natl. Acad. Sci. USA* 83, 9231-9235 (1986).
72. Kromer L. F., *Science* 235, 214-216 (1987).
73. Korschning S., Auburger G., Heumann R., Scott J., Thoenen H., *EMBO J.* 4, 1389-1393 (1985).
74. Shelton D. L., Reichardt L. F., *Proc. Natl. Acad. Sci. USA* 83, 2714-2718 (1986).
75. Whittemore S. R., Ebendal T., Lärkfors L., Olson L., Seiger A., Stromberg I., Persson H., *Proc. Natl. Acad. Sci. USA* 83, 817-821 (1986).
76. Aloe L., Levi-Montalcini R., *Brain Res.* 133, 358-366 (1977).
77. Böhm A., Aloe L., *Acad. Naz. dei Lincei* 80, 1-6 (1986).
78. Bruni A., Bigon E., Boarato E., Leon A., Toffano G., *FEBS Lett.* 138, 190-192 (1982).
79. Sugiyama K., Suzuki Y., Furuta H., *Arch. Oral. Biol.* 30, 93-95 (1985).
80. Mazurek N., Weskamp G., Erne P., Otten U., *FEBS Lett.* 198, 315-320 (1986).
81. Thorpe L. W., Werrbach-Perez K., Perez-Polo J. R., *2° Internat. Workshop on Neuroimmunomodulation*, Dubrovnik, Abstract, pp. 151 (1986).

82. Cattaneo A., Secher D. S., *Exp. Cell Res.* (in press).
83. Hogue-Angeletti R., Bradshaw R. A., *Proc. Natl. Acad. Sci. USA* 68, 2417-2420 (1971).
84. Scott J., Selby M., Urdea M., Quiroga M., Bell G., Rutter W. J., *Nature* 302, 538-540 (1983).
85. Ullrich, A., Gray A., Berman C., Dull T. J., *Nature* 303, 821-823 (1983).
86. Meier R., Becker-Andre M., Gotz R., Heumann R., Shaw A., Thoenen H., *EMBO J.* 5, 1489-1493 (1986).
87. Ebendal T., Larhammar D., Persson H., *EMBO J.* 5, 1483-1487 (1986).
88. Francke V., De Martinville B., Coussens L., Ullrich A., *Science* 222, 1248-1250 (1983).
89. Bocchini V., Angeletti P. U., *Proc. Natl. Acad. Sci. USA* 64, 787-792 (1969).
90. Varon S., Nomura J., Shooter E. M., *Biochemistry* 6, 2202-2210 (1967).
91. Varon S., Nomura J., Shooter E. M., *Biochemistry* 7, 1296-1303 (1968).
92. Stach R. W., Shooter E. M., *J. Biol. Chem.* 249, 6668-6674 (1974).
93. Edwards R. H., Selby M. J., Rutter W. J., *Nature* 319, 784-787 (1986).
94. Eipper B.A., Mains R. E., Herbert E., *Trends Neurosci.* 100, 463-467 (1986).
95. Doolittle R., Hunkapiller M. W., Hood L. E., DeVare S.G., Robbins K. L., Aaronson S. A., Antoniades H. N., *Science* 221, 275-276 (1983).
96. Downward J., Yraden Y., Mayes E., Scrace G., Totty N., Stockwell P., Ullrich A., Schlessinger J., Waterfield M. D., *Nature* 307, 521-527 (1984).
97. Sherr C. J., Rettermier C. W., Sacca R., Roussel M. F., Look T. A., Stanley E. R., *Cell* 41, 665-676 (1985).
98. Weinberger C., Hollenberg S. M., Rosenfeld M. G., Evans R. M., *Nature* 318, 670-673 (1985).
99. Bar-Sagi D., Feramisco J. R., *Cell* 42, 841-848 (1985).
100. Alemà S., Casalbore P., Agostini E., Tatò, F., *Nature* 316, 557-559 (1985).
101. Levi-Montalcini R., Angeletti P. U., in *Salivary Glands and their Secretions*, L. M. Sreebny, J. Meyer, Eds. (Pergamon, Oxford, 1964), pp. 129-141.
102. Aloe L., Levi-Montalcini R., *Exp. Cell Res.* 125, 15-22 (1980a).
103. Aloe L., Cozzari C., Levi-Montalcini R., *Brain Res.* 332, 259-265 (1985).
104. Aloe L., Alleva E., Böhm A., Levi-Montalcini R., *Proc. Natl. Acad. Sci. USA* 83, 6184-6187 (1986).
105. Lakshmanan J., *Am. J. Physiol.* 250, E 386-391 (1986).
106. Otten U., Schwab M., Gagnon C., Thoenen H., *Brain Res.* 133, 291-303 (1977).
107. Bing J., Poulsen K., Hackenthal E., Rix E., Taugner R., *J. Histochem. Cytochem.* 28, 874-880 (1980).
108. Harper G. P., Barde Y. A., Burnstock G., Carstairs J. R., Dennison M. E., Suda K., Vernon C. A., *Nature* 279, 160-162 (1979).
109. Harper G. P., Thoenen H., *J. Neurochem.* 34, 893-903 (1980).
110. Beer D.J., Matloff S. M., Rocklin R.E., *Adv. Immunol.* 35, 209-215 (1984).
111. Levi-Montalcini R., Aloe L., *Proc. Natl. Acad. Sci. USA* 82, 7111-7115 (1985).
112. Kaiser E. T., Lawrence D. S., *Science* 226, 505-511 (1984).
113. Rajashekhar B., Kaiser E.T., *J. Biol. Chem.* 261, 13617-13623 (1986).
114. Ebendal T., Olson L., Seiger A., Hedlund K. O., *Nature* 286, 25-28 (1980).
115. Roszman T. L., Jackson J. C., Cross R. J., Titus M. J., Markesberry W. R., Brooks W. H., *J. Immunol.* 135, 769s-772 (1985).
116. Hall N. R., McGillis U. P., Spangelo B. L., Goldstein A. L., *J. Immunol.* 135, 806-811.
117. Hamburger V., *Perspectives in Biol. and Med.* 18, 162-178 (1975).

EMILIO G. SEGRÈ

Properties of antinucleons

Nobel Lecture, December 11, 1959

I must begin by thanking the Swedish Academy for the great honor they have bestowed on me. The names of the previous recipients of the Nobel Award, while lending great prestige to the Award, make me feel humble and dubious about my merits to join the company. However, I can only repeat my gratitude and think that my constant devotion to science may have something to do with the choice, apart from any success, in which there is perforce an element of luck. At the onset I must also mention the names of two people who have had, in different ways, a very great influence upon all my work. Of Enrico Fermi I would only say, quoting Dante as he himself might have done,

*Tu se' lo mio maestro e il mio autore;
Tu se' solo colui da cui io tolsi
Lo bello stilo che mi ha fatto onore.*

Thou art my master and my author;
thou alone art he from whom I took
the good style that hath done me honor.

I learned from him not only a good part of the physics I know, but above all an attitude towards science which has affected all my work. Ernest Orlando Lawrence created the instruments with which most of my work was done. Although I belong scientifically to a different tradition and outlook, it was only through the instruments developed at his instigation and under his leadership that most of my own researches became possible. This is especially true for the last one: the antiproton.

By 1954 the Bevatron had been developed and tested. It had been purposefully planned for an energy above the threshold for forming nucleon-antinucleon pairs, and many physicists including my colleagues and I naturally thought of means for hunting the elusive antiproton. Although its existence was very probable, a definite experimental proof was lacking and, being

aware of the crucial importance of the problem for the extension of Dirac's theory from the electron to the nucleon, we tried to design an experiment which would give a definite answer¹. The final apparatus has been described in the preceding lecture by Dr. Chamberlain².

Other experiments involving photographic detection were also planned at that time and came to fruition soon after the success of the first experiment³.

Dr. Chamberlain has described to you what an antiproton is and how it was found, and I have nothing to add to his lecture on these matters.

The properties used for the identification of the antiproton were predicted by Dirac long ago and were used as a guide in finding the particle. However, once it was found we faced a host of new problems, and it is to those that I will direct the rest of my speech.

I will be very brief concerning the experimental developments.

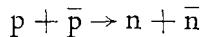
Here great emphasis has been put on the development of better antiprotons beams. By « better » I mean beams in which there are more antiprotons per unit time and in which the ratio of the number of antiprotons to unwanted particles is higher. It suffices to say that now it is possible to have at Berkeley, beams with about 10 antiprotons per minute instead of one every fifteen minutes as in 1955 and beams in which antiprotons are about one in ten particles instead of one in 50,000 as in 1955. The improved beams allow more difficult and complicated experiments, and the development of electronics and bubble chambers has kept pace with the increased possibilities. I may add that the complications in which we are entering now are by no means a cause of joy to the experimenters who have to cope with them, and that they are properly considered as the heavy price to be paid in order to obtain more detailed physical information.

Some of the problems raised by the identification of the antiproton have a predictable solution, although the prediction does not derive from anything as solid as Dirac's theory. We could, for instance, expect with complete confidence the existence of the antineutron and of all the antiparticles of the baryons, although it might require considerable skill to find them. In fact, antineutrons are certainly formed copiously at the Bevatron but the primary antineutrons are very difficult to identify. For this reason immediately after the discovery of the antiproton, it was suggested that the antineutron should be found by investigating the charge exchange reaction in which a proton and an antiproton give a neutron and an antineutron⁴. In a very ingenious and elegant counter experiment, Cork, Lambertson, Piccioni, and Wenzel



Fig. 1. An antiproton enters a propane bubble chamber, and at the point marked with an arrow undergoes charge exchange. The antineutron originates the annihilation star (directly below). Density of propane 0.42 g/cm^3 . Real distance between charge exchange and origin of star 9.5 cm. $T_{\bar{p}}$ at charge exchange $\sim 50 \text{ MeV}$. (From Agnew *et al.*⁶)

did demonstrate the existence of the antineutron some time ago⁵. Their method was based on a counter technique and uses the reaction



which is called charge exchange because we can interpret it as the passage of the electric charge from the proton to the antiproton. The product antineutron is recognizable by its annihilation properties. Namely, an antineutron on annihilation forms an annihilation star extremely similar to an antiproton star. Instead of reproducing their experimental arrangement, I will show in a slide (Fig. 1) a graphical picture of these phenomena as observed in a bubble chamber by the joint efforts of Professor Wilson Powell and his group, and my own group⁶.

Similarly, the antilambda was found by Baldo-Ceolin and Prowse⁷ in photographic emulsions exposed to a pion beam and was confirmed in the hydrogen bubble chamber. Also the antisigma has been recently seen in a hydrogen bubble chamber by the Alvarez group in Berkeley⁸.

It is also possible to predict with certainty some of the nucleonic properties of the antinucleons, specifically the spin, I-spin, 3rd component of the I-spin, and parity to be those shown in Table 1.

Table 1. Spin, parity, I-spin of nucleons, and antinucleons.

	Proton	Neutron	Anti-proton	Anti-neutron
Spin, S	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
I-spin, T	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
Third component of I-spin, T_3	$\frac{1}{2}$	-	-	$\frac{1}{2}$
Parity	+	+	-	-

But in addition to these interesting questions of systematics of particles, which can be summarized by the diagram shown in Fig. 2, there are problems for which we know much less what to expect because they involve more than general symmetry properties. They require a fairly detailed knowledge of interactions and subnuclear structure which at present we do not have. Indeed these are the most interesting and challenging problems.

For instance, we know that a nucleon and an antinucleon may annihilate each other, but what are the products of the annihilation? What is their energy? What are the collision cross sections? It is in this direction that we are

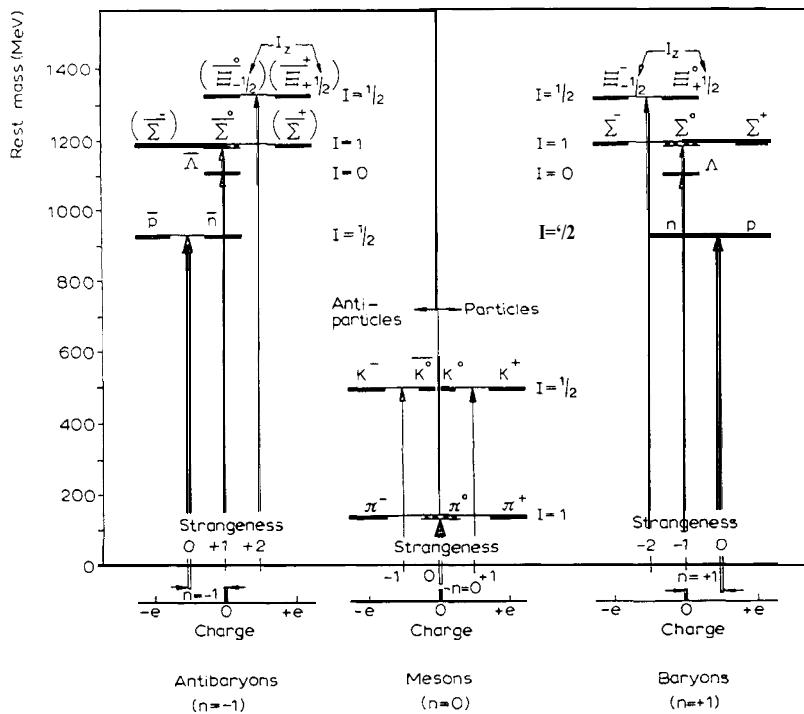


Fig. 2. A diagram showing all strongly interacting particles as known or predicted today. The particles still unobserved are in parenthesis. The weak interacting particles not reported in this diagram are the μ^\pm -meson, the electron and positron, the neutrino and antineutrino, and the light quanta. (From Gell-Mann and Rosenfeld, *Ann. Rev. Nucl. Sci.*, 7 (1957) 407.)

working now and here we must be guided mainly by experiment, at least for the time being, and also be prepared for surprises.

The first surprise came immediately after the discovery of the antiproton when we found that this particle has an unusually large collision cross section. This fact has now been studied intensively for some time. The simplest situation occurs in the case of proton-antiproton collisions. There, in addition to the charge exchange process mentioned above, there are two other possibilities, elastic scattering and annihilation, at least until we reach energies such that inelastic processes (pion production) also become possible. Thus we have three cross sections: for scattering, for annihilation, and for charge exchange. All three have been measured for a wide energy interval and the results are shown in Fig. 3.

The magnitude of these cross sections is striking when we compare them

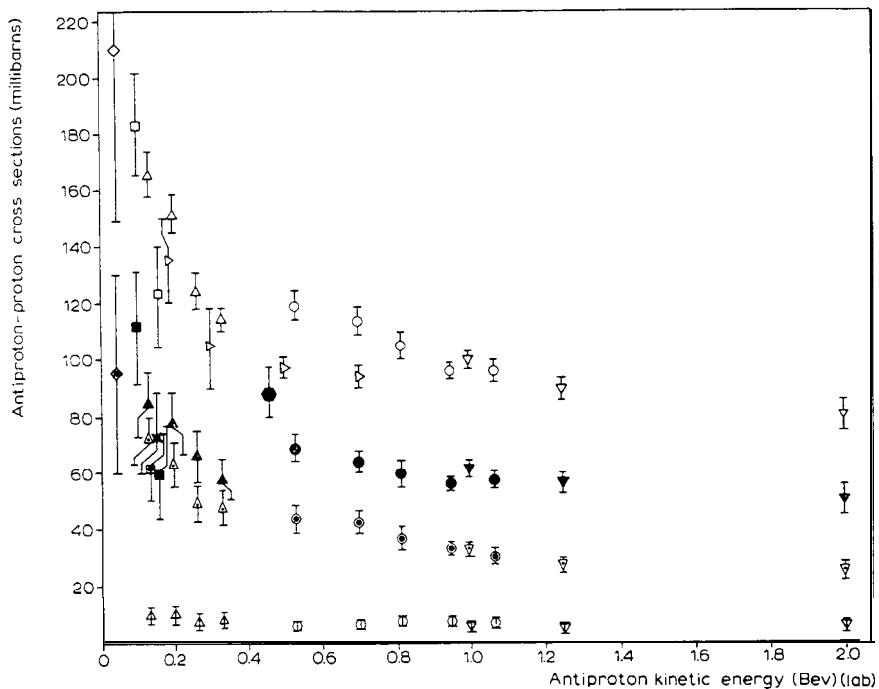


Fig. 3. All p-p cross sections published up to November, 1959. The *open symbols* are total cross sections; *closed symbols* are inelastic cross sections (which are due to annihilation only for $T_{\bar{p}} \lesssim 290$ MeV); *open symbols encircling a dot* are elastic cross sections; *open symbols crossed by a vertical line at the bottom of the figure* are charge exchange cross sections.

The various symbols are referenced as follows:

- \square Agnew, Elioff, Fowler, Gilly, Lander, Oswald, Powell, Segrè, Steiner, White, Wiegand, Ypsilantis, *Bull. Am. Phys. Soc.*, Ser. II, 4 (1959) 357.
- ∇ Armenteros, Coombes, Cork, Lambertson, Wenzel, *Bull. Am. Phys. Soc.*, Ser. II, 4 (1959) 356.
- \circlearrowleft Chamberlain, Keller, Mermod, Segrè, Steiner, Ypsilantis, *Phys. Rev.*, 108 (1957) 1553.
- \triangle Coombes, Cork, Galbraith, Lambertson, Wenzel, *Phys. Rev.*, 112 (1958) 1303.
- \circ Elioff, Agnew, Chamberlain, Steiner, Wiegand, Ypsilantis, *Phys. Rev. Letters*, 3 (1959) 285.
- \triangleright Cork, Lambertson, Piccioni, Wenzel, *Phys. Rev.*, 107 (1957) 248.
- \diamondsuit Horwitz, Miller, Murray, Tripp, *Phys. Rev.*, 115 (1959) 472.

* Emulsion results of many authors compiled and averaged by Baroni *et al.*, *Nuovo Cimento*, 12 (1959) 564.

with those obtained in proton-proton collisions. A tentative theory of this phenomenon has been put forward by Chew⁹ and his associates, and also by Koba and Takeda in Japan¹⁰.

The model is based on the Yukawa theory of nuclear interactions in such a way as to stress the analogy between the nucleon-nucleon and the nucleon-antinucleon system. For the nucleon-nucleon system a model consisting of a hard repulsive core of a radius of about $\frac{1}{3}$ of the Compton wavelength of the pion (0.45×10^{-13} cm) surrounded by a pion cloud has been reasonably successful in explaining the experimental results of the scattering and polarization experiments. The pion cloud which dominates the interactions at moderate distance can be treated from first principles of pion theory. The hard repulsive core on the other hand is unaccounted for from a pion theoretical point of view and must be introduced *ad hoc* as a phenomenological hypothesis, although the existence of heavier mesons such as the K mesons may have something to do with it. For a nucleon-antinucleon system the pion cloud of the antinucleon is substituted by its charge conjugate according to the expectations of meson theory and the medium range interactions are treated on the basis of this theory. The overlap of the cores, however, is now supposed to bring annihilation instead of strong repulsion. On the basis of this model it has been possible to account for most of the observations made thus far, which however do not extend to energies above 1 BeV where some critical tests of the theory await us.

In addition to the total cross sections for scattering, annihilation and charge exchange mentioned above, the angular distribution on scattering has been measured. Here a large diffraction peak in the forward direction has been found. It is directly related to the annihilation.

The extension of the cross section studies to complex nuclei has been started. The deuteron has been first investigated with the hope of finding information on the neutron-antiproton interaction. Here the data are still very rough, mainly because the subtraction techniques which we were forced to use introduce considerable errors. The qualitative feature seems to be that there is not much difference between proton-antiproton and neutron-antiproton collisions.

For heavier nuclei the data from the nucleon-antinucleon collision have been fed into an optical model treatment and the results agree with the experimental data as far as they are available. This gives a consistent picture connecting the more complicated case to the simpler one.

There are however still some crucial tests to be performed on the $p-\bar{p}$ case

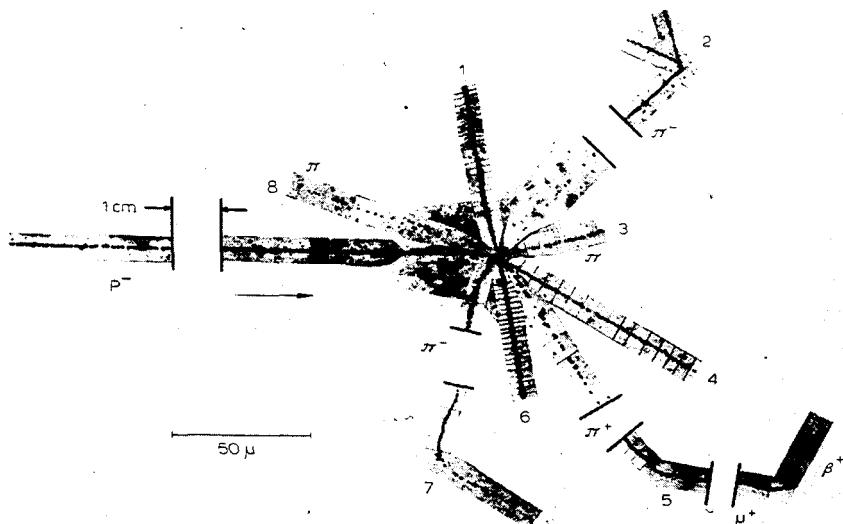


Fig. 4. An annihilation star showing the particles as numbered.

No.	1	2	3	4	5	6	7	8
Identity	p?	π^-	$\pi^?$	p	π^+	$H^3(?)$	π^-	π
T(MeV)	10	43	175	70	30	82	34	125

Total visible energy 1,300 MeV. Total energy release $> 1,400$ MeV.

in order to validate the Chew model. At high energy, say 2 BeV, the annihilation cross section should be essentially the cross section of the core, and hence considerably smaller than the one observed at lower energy: $3 \times 10^{-26} \text{ cm}^2$ would be a generous guess. If this expectation is not fulfilled it will be necessary to look for some other model. I will not go further into the numerous problems connected with cross-section studies, and will turn now to the annihilation.

The annihilation process itself has been fairly well investigated experimentally, but the theoretical situation leaves much to be desired. Initially the effort was mainly directed toward establishing the fact that the energy released was $2mc^2$ (m is the mass of the proton, c the velocity of light), thus furnishing a final proof of the annihilation. In the early investigations with photographic emulsions carried out in my group especially by Gerson Goldhaber and by a group in Rome led by Amaldi, we soon found stars showing

a visible energy larger than mc^2 , giving conclusive evidence of the annihilation in pairs of proton and antiproton¹¹. With great pleasure I recognized in the Nobel diploma the image of the first star of this type, found in Berkeley by Prof. Gösta Ekspong, now of Stockholm. It is shown in Fig. 4.

The observations on annihilation have been performed with many techniques. Initially, immediately after the identification of the antiproton, these particles were stopped in a block of heavy glass and the showers due to the gamma rays resulting from the decay of neutral pions were observed by Dr. Moyer and his co-workers¹². This method was not however very quantitative.

Photographic emulsions were also exposed to antiprotons at the earliest possible moment. Here we see only the charged annihilation products, although much detailed information is obtainable. The great observational effort needed here was shared in a large cooperative experiment in which many laboratories in the U.S.A. and in Europe participated¹³.

Bubble chambers have also been used, both of the propane and of the hydrogen type.

By now we know a good deal about annihilation. It gives rise prevalently to π mesons. These in a time of the order of 10^{-8} seconds decay into μ mesons and neutrinos. The π mesons in a time of the order of microseconds decay into electrons or positrons and neutrinos, and the electrons and positrons finally recombine to give gamma rays. In a few microseconds the total rest mass of the nucleon-antinucleon pair degrades to particles with rest mass zero, travelling away from the spot of the annihilation with the velocity of light.

Direct annihilation into photons may occur, but is expected to be rare and thus far has never been observed with certainty.

The reason for this difference between the behavior of electron-positron and nucleon-antinucleon pairs is, of course, that the latter can annihilate not only through the electromagnetic interaction giving rise to light quanta, but also through the specific nuclear interaction whose quanta are the pions. This last interaction is much stronger than the electromagnetic one and when both are simultaneously present its effects overwhelm those of the electromagnetic interaction, which is the only available to the electron-positron pair.

The most significant result of the annihilation studies is that the annihilation process gives rise to an average of 4.8 pions per annihilation, about equally divided among positive, negative, and neutral pions. These pions



Fig. 5. Annihilation of an antiproton in carbon giving rise to a K meson and a Λ hyperon.

escape with a continuous energy distribution, the average kinetic energy being about 200 MeV. In about 4 percent of the annihilation cases at rest strange particles, K mesons, are emitted. See Fig. 5.

The escaping pions give rise in complex nuclei to secondary processes and thus a number of nucleons or light nuclei is also found among the particles

emitted on annihilation. Sometimes the relatively rare K mesons interact producing a Λ hyperon and even more complicated hyperfragments have been observed (Ekspong).

In hydrogen the multiplicity of the prongs, referring of course only to charged particles, is given in the following little table, for annihilations at rest. Naturally only even numbers of charged prongs may appear because the total charge of the proton-antiproton system is zero. (See Table 2.)

Table 2.

<i>Charged multiplicity</i>	0	2	4	6	8
<i>Numbers of stars (total, 222)</i>	10	89	109	14	0

From the theoretical point of view, we do not yet have an entirely satisfactory picture of the annihilation process. It has been mostly analyzed on the basis of a statistical theory put forward many years ago by Fermi, which does not take into account any detailed mechanism, but only the obvious and necessary features determined by phase space. This theory contains only one free parameter, namely, the volume into which the energy released on annihilation is concentrated at the beginning of the phenomenon. Naturally this volume is supposed to be the one corresponding to a sphere of the radius equal to radius of action of nuclear forces. If one calculates what is to be expected on this basis, one finds a result which is in rather poor agreement with experiment, namely, the multiplicity of pions produced is larger than that predicted by the model. Clearly the average energy and the multiplicity are connected, and hence the average energy also disagrees with the naive statistical prediction. The model can be made to yield correct results by increasing beyond what seems plausible the volume in which the energy comes to equilibrium. Many attempts have been made to refine Fermi's theory and to bring it into agreement with facts. Some of these attempts are very ingenious and one would wish that there were more success than there is. The ratio between K mesons and pions is another element of the puzzle that has to be taken into account and seems rather intractable for the time being.

It is however hardly to be expected that a purely statistical theory should explain quantitatively the annihilation process, inasmuch as selection rules, strong interactions of the escaping particles and other important factors completely omitted in the theoretical picture, are at work. I think that the future study of the annihilation process with its bearing on the core of the nucleon,

a region of which we know so little, will give some important results. Antinucleons are especially suited for this study because they will exhibit more clearly than other particles the effects of the core.

And now let me say some words on the popular subject of the « antiworld ». Already Dirac in his Nobel Lecture of 1933 said:

« If we accept the view of complete symmetry between positive and negative electric charge so far as concerns the fundamental laws of nature, we must regard it rather as an accident that the earth (and presumably the whole solar system), contains a preponderance of negative electrons and positive protons. It is quite possible that for some of the stars it is the other way about, these stars being built up mainly of positrons and negative protons. In fact, there may be half the stars of each kind. The two kinds of stars would both show exactly the same spectra, and there would be no way of distinguishing them by present astronomical methods. »

We can now add that the proved existence of the antinucleons has very strongly corroborated this possibility, although we also know that the symmetry between electric charges breaks down for weak interactions. As far as astronomical means are concerned, a verification seems impossible in principle, because they depend on electromagnetic phenomena, which are invariant under charge conjugation. It is however interesting that the recent important discoveries about β -decay and the neutrino now give a method which, while still impossible in practice, is sound in principle, being based on weak interactions which are not invariant under charge conjugation. This method, if it could be executed, would solve unambiguously the question of the existence of antiworlds. If we observe a star and from its astronomical characteristics can decide that most of its energy comes from a known cycle, as for example the carbon cycle, which is dominated by β -decays, we can see whether the antineutrinos coming from it are or are not of the same kind as the antineutrinos coming from a pile or from our sun by performing an inverse β -decay experiment. If it should turn out that they are neutrinos, i.e. different from those coming from the sun, then the star is of antimatter.

Let me finish this lecture with a remark and some acknowledgements. As in many investigations in high energy physics in recent times, this experiment is the result of a large cooperative effort. The credit for the success is shared by many individuals and even by a machine, which was obviously necessary to produce particles above the threshold for nucleon pair production. Since it is impossible to mention all the numerous contributors, I shall limit myself to a few. Dr. O. Piccioni helped materially in the early plan-

ning of the experiment, especially by suggesting the use of magnetic quadrupole lenses. Dr. E. Lofgren most ably directed the operation of the Bevatron. Dr. H. Steiner supplied invaluable help during the whole experiment. Dr. T. J. Ypsilantis, our colleague and co-author, also worked with us all the time. Above all, however, our co-author and comrade of 20 years of work, Dr. Clyde Wiegand, was indispensable and deserves a major part of the credit for the success of our investigation.

1. See for instance: P. A. M. Dirac, *Les Prix Nobel en 1933*.
2. O. Chamberlain, E. Segrè, C. Wiegand, and T. Ypsilantis, *Phys. Rev.*, 100 (1955) 947.
3. O. Chamberlain, W. W. Chupp, G. Goldhaber, E. Segrè, C. Wiegand; and E. Amaldi, G. Baroni, C. Castagnoli, C. Franzinetti, A. Manfredini, *Phys. Rev.*, 101 (1956) 909.
4. O. Chamberlain, E. Segrè, C. Wiegand, and T. Ypsilantis, *Nature*, 177 (1956) II.
5. B. Cork, G. R. Lambertson, O. Piccioni, and W. A. Wenzel, *Phys. Rev.*, 104 (1956) 1193.
6. L. E. Agnew, T. Elioff, W. B. Fowler, L. Gilly, R. Lander, L. Oswald, W. Powell, E. Segrè, H. Steiner, H. White, C. Wiegand, and T. Ypsilantis, *Phys. Rev.*, 110 (1958) 994.
7. M. Baldo-Ceolin and D. J. Prowse, *Bull. Am. Phys. Soc.*, 3 (1958) 163.
8. Button, Eberhard, Kalbfleisch, Lannutti, Maglić, Stevenson, *Phys. Rev. Letters* (to be published).
9. J. S. Ball, and G. F. Chew, *Phys. Rev.*, 109 (1958) 1385.
10. Z. Koba, and G. Takeda, *Progr. Theoret. Phys. (Kyoto)*, 19 (1958) 269.
11. O. Chamberlain, W. W. Chupp, A. G. Ekspong, G. Goldhaber, S. Goldhaber, E. J. Lofgren, E. Segrè, C. Wiegand; and E. Amaldi, G. Baroni, C. Castagnoli, C. Franzinetti, A. Manfredini, *Phys. Rev.*, 102 (1956) 921.
12. J. M. Brabant, B. Cork, N. Horwitz, B. J. Moyer, J. J. Murray, R. Wallace, and W. A. Wenzel, *Phys. Rev.*, 101 (1956) 498.
13. W. H. Barkas, R. W. Birge, W. W. Chupp, A. G. Ekspong, G. Goldhaber, S. Goldhaber, H. H. Heckman, D. H. Perkins, J. Sandweiss, E. Segrè, F. M. Smith, D. H. Stork, L. van Rossum; and E. Amaldi, G. Baroni, C. Castagnoli, C. Franzinetti, A. Manfredini, *Phys. Rev.*, 105 (1957) 1037.

SELMAN A. WAKSMAN

Streptomycin: background, isolation, properties, and utilization

Nobel Lecture, December 12, 1952

*The Lord hath created medicines out of the earth;
and he that is wise will not abhor them.*

Ecclesiasticus, xxxviii, 4

The highest scientific award and honor presented to me the day before yesterday gives me the opportunity to summarize briefly the discovery and utilization of streptomycin for disease control, notably in the treatment of tuberculosis, the "Great White Plague" of man.

Historical background

Streptomycin belongs to a group of compounds, known as antibiotics, which are produced by microorganisms and which possess the property of inhibiting the growth and even of destroying other microorganisms. Antibiotics vary greatly in their chemical nature, mode of action upon different organisms, and effect upon the animal body. The selective action of antibiotics upon bacteria and other microorganisms is known as the antibiotic spectrum. Some antibiotics are characterized by a very narrow spectrum, whereas others possess a wide range of activity. Some are active only against certain bacteria and not upon others, whereas some are active against fungi, and some against viruses. There is not only considerable qualitative variation in the activity of different antibiotics, but also wide quantitative differences. Antibiotics are produced by bacteria, fungi, actinomycetes, and to a limited extent by other groups of microorganisms.

It has been known for more than six decades that certain fungi and bacteria are capable of producing chemical substances which have the capacity to inhibit the growth and even to destroy pathogenic organisms. Only within the last 12 or 13 years however, have antibiotics begun to find extensive application as chemotherapeutic agents. Among these, penicillin and streptomycin have occupied a prominent place. Penicillin is largely active upon gram-posi-

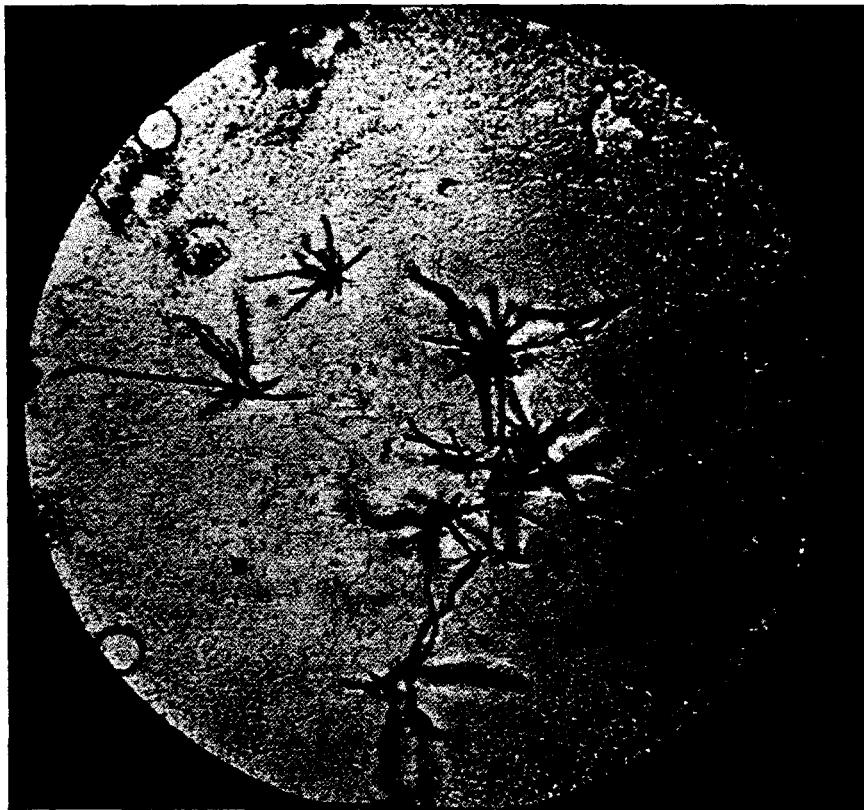


Fig. 1. *Streptomyces griseus*, streptomycin-producing strain. Vegetative and aerial mycelium.

tive bacteria, gram-negative cocci, anaerobic bacteria, spirochaetes, and actinomycetes; streptomycin is active against a variety of gram-negative and acid-fast bacteria, as well as upon gram-positive organisms which have become resistant to penicillin. Neither of these antibiotics is active upon rickettsiae, viruses, and fungi. They too differ in their physical and chemical properties and in their toxicity to animals.

Since the discovery of streptomycin, the production and clinical application of this antibiotic have had a phenomenal rise. *Streptomyces griseus*, the streptomycin-producing organism, was first isolated in September 1943, and the first public announcement of the antibiotic was made in January 1944. Before the end of that year, streptomycin was already being submitted to clinical trial. Within two years, the practical potentialities of streptomycin for disease control were definitely established.

The most spectacular of all the clinical applications followed recognition that streptomycin was highly effective against the tuberculosis organism, not only *in vitro* but also *in vivo*. Several clinical centers undertook to test the sensitivity of different freshly isolated strains of *Mycobacterium tuberculosis* to streptomycin, its practical evaluation in control of tuberculosis in experimental animals, and finally its use in the control of tuberculosis in the human body. Following the lead of Dr. William H. Feldman and Dr. H. Corwin Hinshaw at the Mayo Clinic, the American Trudeau Society and the National Tuberculosis Association took an active part in these investigations. The first conference arranged for the evaluation of the clinical results of the use of streptomycin was held in 1945, and others soon followed.

Within three years after the announcement of the isolation of streptomycin came the almost complete elucidation of its chemistry, its first derivative (dihydrostreptomycin) was obtained, and the first 1,000 clinical cases were reported. From a laboratory curiosity, streptomycin production soon grew into a large industry, with a monthly output of more than 25 thousand kilograms in the United States alone, and with many plants being established abroad for its manufacture.

This marked rise in the development of streptomycin was due partly to the fact that the spectacular increase, between 1941 and 1943, in the use of penicillin for disease control suggested the possibility that there were other antibiotics which could be utilized for the treatment of diseases not sensitive to penicillin. It was the knowledge of the great abundance and wide distribution of actinomycetes, which dated back nearly three decades, and the recognition of the marked activity of this group of organisms against other organisms that led me in 1939 to undertake a systematic study of their ability to produce antibiotics. Between 1940 and 1952 we isolated in our laboratories 10 different chemical substances from cultures of different actinomycetes, beginning with *actinomycin* and ending with *candididin*. Of these, *streptomycin* proved to be by far the most important chemotherapeutic agent.

Production of antibiotics by actinomycetes

The isolation of streptomycin was the culminating point of a painstaking search for antimicrobial agents produced by actinomycetes, a group of organisms closely related to the bacteria. This was preceded by long and continuous research, dating back to 1915, on actinomycetes, their occurrence

and abundance in nature, their systematic or taxonomic position, their role in soil processes, notably in the decomposition of plant and animal residues and in the formation of humus, and finally their associative and antagonistic effects upon bacteria and fungi. It was finally established that as many as 20 to 50 per cent of all the actinomycetes found in the soil and in other natural substrates had the capacity to inhibit the growth of other microorganisms.

Streptomyces griseus, the organism which comprised the streptomycin-producing strain, was known in our laboratories from the beginning of our work on actinomycetes, although it was not tested at that time for its antibiotic-producing properties. The ability of actinomycetes to exert injurious effects upon bacteria and fungi has been known for many years. Lieske

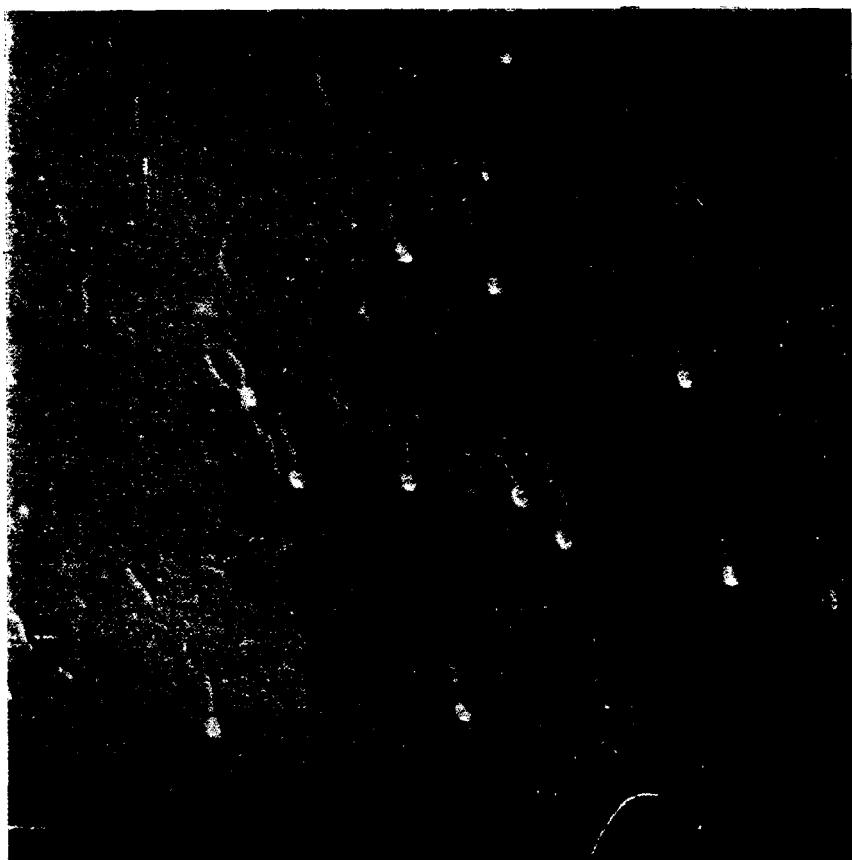


Fig. 2. Electron micrograph of actinophage, type I, of streptomycin-producing *Streptomyces griseus*, $\times 31,000$. (Courtesy of Squibb Institute for Medical Research.)

showed in 1921 that certain strains are able to bring about lysis of many bacteria and antagonize their growth. This process is selective in nature: some of the bacteria are affected, and others are not. Other investigators, notably Gratia and Dath and Rosenthal, demonstrated in 1925 that cultures of organisms designated as *Streptothrix*, and now known to be actinomycetes, are capable of dissolving living and dead bacterial cells. Nakhimovskaia in 1937 made the first survey of the occurrence of antagonistic actinomycetes in the soil: 80 cultures were isolated, of which 47 were able to repress bacterial growth, but only 27 liberated into the medium substances which had the capacity to inhibit the growth of gram-positive bacteria, but not of gram-negative bacteria or fungi.

When we began our investigation on the production of antibiotics by actinomycetes in 1939, only two preparations were known to possess antimicrobial properties. These were not true antibiotics, or at least were not recognized as such. One was obtained by Gratia and had the capacity to lyse dead typhoid cells and certain living bacteria; it was later designated by Welsch as actinomycetin. The other was believed to be a lysozyme, which had lytic principles, and was studied by Krassilnikov and Koreniako in 1939.

The first true antibiotic to be derived from a culture of an actinomycete was isolated in our department in 1940. The organism, *Actinomyces antibioticus*, yielded a substance which was designated as actinomycin. It was soon crystallized, and its chemical and biological properties were established. This antibiotic proved to be a quinone-like pigment with an approximate molecular formula of $C_{41}H_{56}N_sO_{11}$. It was highly active against various gram-positive bacteria but to a much lesser degree upon the gram-negative organisms. It proved to be extremely toxic to experimental animals.

This practical failure was followed by a comprehensive program of screening actinomycetes for their ability to produce different antibiotics. The fact was revealed that these antibiotics vary greatly in their chemical nature, toxicity to animals, and antimicrobial activities. Some were active against bacteria and actinomycetes, but not upon fungi; some were largely active against certain bacteria, especially the gram-positive types, but not, or to only a limited extent, against the gram-negative forms, whereas others were active alike upon various gram-positive and gram-negative bacteria. Some had a very narrow spectrum, being active only or largely upon one group of organisms, such as the mycobacteria, or only upon certain fungi, such as yeast-like forms, or upon certain viruses, such as the influenza-B types; others

had a much wider spectrum, being active upon both bacteria and fungi, or upon bacteria, rickettsiae, and the so-called larger viruses.

It was further found in our laboratory and in others that one antibiotic, such as actinomycin and streptothricin, could be produced by several different species. Some organisms give rise to different modifications of the same antibiotic, as actinomycin A, B, and C. This is true also of streptomycin, various modifications of which are produced by different species of *Streptomyces*, notably *S. griseus*, *S. bikiniensis*, and *S. griseocarneus* (streptomycin, mannosidostreptomycin, hydroxystreptomycin). Some organisms produce a mixture of different antibiotics, as in the case of different strains of *S. griseus*, which give rise not only to streptomycin, but also to the antifungal agents actidone and candicin.

Isolation of streptothricin and streptomycin

The isolation of actinomycin, the first true antibiotic from a culture of an actinomycetes, pointed a way to the formation and isolation from species belonging to this group of organisms of other chemical substances possessing antimicrobial properties. A new type of substance, designated as *streptothricin*, was soon isolated, in 1942. It showed distinct promise as a chemotherapeutic agent. It was active against both gram-positive and gram-negative bacteria and was not toxic to animals. The organism producing it was identical with *Actinomyces lavendulae*, a culture found in the soil 25 years previously. This new substance possessed highly desirable physical, chemical, and antimicrobial properties and gave promise of filling the gap left by penicillin in the treatment of infectious diseases due to gram-negative bacteria. It was a basic compound soluble in water, with an optimum activity at a slight alkalinity. It was active against a number of bacteria, not only *in vitro*; but also *in vivo*, as well as against various fungi. It was resistant to heat and to the action of different microorganisms and enzymes. A study of its pharmacology, however, brought out the fact that streptothricin exerted a residual toxic effect upon the animal body; its use in the treatment of infectious diseases was, therefore, limited.

The experience gained in the study of streptothricin made possible effective planning of a more comprehensive study of the production of a definite type of antibiotic by actinomycetes. Particular attention was paid to substances that possessed chemical and biological properties similar to those of

streptothricin but that would be less toxic to the animal body. The method of cultivation of the organism (notably the use of submerged cultures), the method of isolation of the active substance (notably its adsorption on charcoal and its removal with dilute acid), the method of evaluation of its antimicrobial properties (notably its activity against gram-negative bacteria), were now well understood. It was, of course, desirable that any new substance should possess a spectrum which would be broader than that of streptothricin, that it be particularly active against a greater variety of bacteria which were resistant to penicillin, and, if possible, that it be active against tuberculosis. (Streptothricin was later found also to possess this property.) Less than six months after many freshly isolated cultures of actinomycetes were screened, an organism was obtained which appeared to produce the long looked-for antibiotic.

Since actinomycin was named after the organism that gave the name to the group of actinomycetales (*Actinomyces* by Harz in 1877) and since the first promising antibiotic, streptothricin, was derived from the first name given to a member of this group (*Streptothrix* by F. Cohn in 1872), it was logical that the first chemotherapeutically active substance should be designated by the newly coined generic name of actinomycetes (*Streptomyces* by Waksman and Henrici in 1943). Thus *streptomycin* was born.

The organism producing streptomycin was identical with a culture isolated in our laboratory in 1916 and described at that time as *Actinomyces griseus* changed, according to the 1943 classification, to *Streptomyces griseus*). Although numerous additional cultures of streptomycin-producing strains of *S. griseus* have since been isolated, it is one of the 1943 cultures that has been utilized for the commercial production of streptomycin all over the world.

In addition to streptomycin, *S. griseus* was later found to produce mannosidostreptomycin, as well as certain other antibiotics, such as actidione and streptocin. Other strains of this organism produce no antibiotics at all or are able to form other substances, such as grisein and candididin.

The medium used for the production of streptomycin contained organic sources of nitrogen and carbon, mineral salts, and certain growth-promoting substances. The culture had to be well aerated and incubated at 25 to 30°C for 3 to 5 days. The activity of streptomycin reached 100 to 200 $\mu\text{g}/\text{ml}$; in commercial production, much higher yields are usually obtained.

Recovery of the streptomycin was accomplished in a series of operations, involving removal of the mycelium by filtration, adsorption of the streptomycin on activated carbon or on some other adsorbent, elution by di-

lute acid, neutralization of the eluate, concentration by evaporation and dehydration or by solvent precipitation, and filtration and drying. Various methods have been used for further purification and crystallization of the antibiotic.

Streptomycin was found to be soluble in water and insoluble in organic solvents, such as ether, chloroform, and acetone. The chloride was completely soluble in methanol, less soluble in ethanol, and virtually insoluble in butyl alcohol, acetic acid, and pyridine. The sulfate was only slightly soluble in methanol and virtually insoluble in the other solvents.

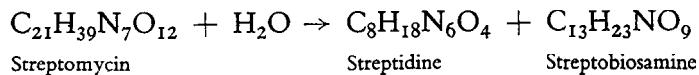
A unit of streptomycin was designated as that amount of material which was just sufficient to inhibit the growth of a standard strain of *E. coli* in 1 millilitre of culture medium. When streptomycin was later isolated in a crystalline state, it was found that 1 unit was comparable to 1 ecrogram of the pure base. The United States Food and Drug Administration, in consultation with the different manufacturers, decided to establish the potency of streptomycin on the basis of the weight of the active material.

The activity of streptomycin against sensitive organisms was found to be influenced by the presence of certain chemical compounds. Glucose reduced its activity appreciably; hydroxylamine, thioglycollate, and cysteine largely inactivated it. This phenomenon was utilized in testing for sterility of streptomycin preparations or for determining the admixture of other antibiotics. The activity of streptomycin was found to be influenced but little by body fluids, pus, or normal tissue juices. The optimum reaction was pH 8.0; increasing acidity resulted in a decreased bacteriostatic potency, the greatest reduction taking place between pH 6.6 and pH 5.9.

Chemical nature of streptomycin

Streptomycin belongs to the glucosides in which a diguanido-group is linked to a nitrogen-containing disaccharide-like compound. It is a strong base, with three basic functional groups. The molecular weight determination on the trihydrochloride in water gave about 800 for the free base after the necessary corrections for the chloride ion.

Upon hydrolysis, it splits into two compounds:

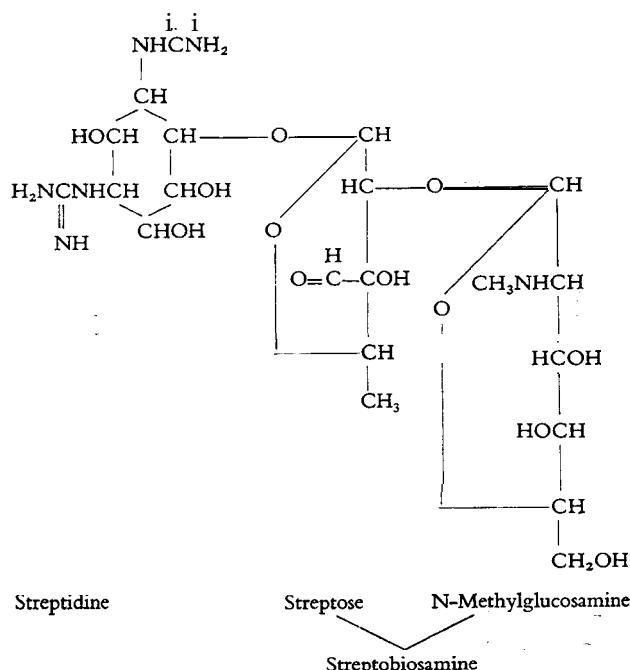


On further hydrolysis with strong mineral acids, the streptobiosamine gives a 6-carbon sugar and glucosamine:



The basic nitrogen atom in the streptobiosamine is not present as a primary amino group.

The structure of the streptomycin molecule is usually given as follows :



Antibacterial properties of streptomycin

Streptomycin is active against a large number of bacteria found among the gram-negative, gram-positive, and acid-fast groups and among the spirochaetes; it has relatively little activity against anaerobic bacteria, fungi, protozoa, and viruses. No absolute value can be given for the sensitivity of an organism to streptomycin; this depends not only upon the particular species, but also upon the strain of the organism and upon the composition of the

medium in which it is tested. Thus any given figure for a single organism is quite arbitrary, since such values can vary greatly.

Streptomycin is more active upon bacteria in young, actively growing cultures than in older cultures, although this difference does not appear to be so great as that for penicillin. When added to a 3-hour-old culture of *E. coli*, 2 μg streptomycin reduced the numbers by 75 per cent, and 5 μg by 95 per cent; the corresponding reduction for 24- to 48-hour-old cultures were 19 and 42 per cent. The size of the inoculum influences the bacteriostatic concentration of streptomycin: for instance, an inoculum of 30,000 cells of *Bacillus abortus* required 1 $\mu\text{g}/\text{ml}$ whereas an inoculum of 30,000,000 cells required 4 $\mu\text{g}/\text{ml}$ for inhibition of growth. Less streptomycin was needed to inhibit growth of various strains of *Br. abortus* and *Br. suis* under aerobic conditions than in the presence of 10 per cent CO₂.

Since streptomycin is not active against viruses, it can be utilized in their isolation and purification from bacterial infections. It is also used in the preservation of bull semen, in the purification of cultures of *Trichomonas vaginalis* and *T. foetus*, and in the development of media for the isolation of pathogenic fungi. Because of its marked antibacterial properties, streptomycin can be used in preventing wound infections and in surgical operations.

Among the antimicrobial properties of streptomycin, its bacteriostatic and bactericidal action upon different strains of *Mycobacterium tuberculosis* is particularly significant. In concentration of 0.05 $\mu\text{g}/\text{ml}$, it has been found capable of inhibiting growth of the human pathogen to a limited extent, 0.2 $\mu\text{g}/\text{ml}$ inhibits growth to a marked extent, and 0.4 $\mu\text{g}/\text{ml}$ stops growth completely, after an incubation time of 16 to 20 days. In other experiments, streptomycin gave complete inhibition of growth of *M. tuberculosis* in concentrations of 0.3 $\mu\text{g}/\text{ml}$, although occasional colonies occurred even with 0.5 $\mu\text{g}/\text{ml}$. The bactericidal action of streptomycin also varies with its concentration and the length of contact with the organism, 0.3 $\mu\text{g}/\text{ml}$ exerting a marked effect in 48 hours, and 20 $\mu\text{g}/\text{ml}$ in 6 hours.

Resistance of bacteria to streptomycin

The problem of variation in sensitivity of different strains of the same organism to streptomycin and the increasing resistance of the bacteria on prolonged contact with it are of considerable theoretical and practical importance. Variation in sensitivity may differ from day to day. This is believed to

be due, at least in part, to variations in the number of viable organisms, age and density of the culture, and the particular species involved.

Freshly isolated cultures of tubercle bacilli from patients with pulmonary tuberculosis are uniformly sensitive to streptomycin. When the cultures are exposed to relatively low concentrations of this antibiotic, growth of multiplying cells but not that of the non-multiplying cells is inhibited. The organism develops resistance to streptomycin *in vitro* at a rapid rate, and this resistance persists for a considerable time; it is not accompanied by any diminution in virulence. In a growing culture of tubercle bacilli, there was found to be, however, a decrease in the proportion of streptomycin-resistant cells with an increase in the age of the culture. The principal effects of streptomycin on the morphology of the organism were a loss of acid-fastness, an increase in granulation, and, in highly bacteriostatic concentration, a shortening of the rods.

Successive transfers of bacteria commonly occurring in infections of the urinary tract resulted in an increase in resistance to more than 1,000 $\mu\text{g}/\text{ml}$ of streptomycin. After 29 subsequent daily transfers on common media, the bacteria lost none of their resistance. It was suggested, therefore, that streptomycin-fastness may be a major factor in the failure of streptomycin therapy in certain infections. When the disease condition presents physical barriers to the penetration of the streptomycin, as in cases of abscesses and pyelonephritis, the organisms are exposed to sublethal concentrations, thus resulting in increased resistance. In infections with *Aerobacter cloacae*, an increase of more than 1,000 times was obtained after 36 days of treatment.

Streptomycin in experimental infections

Streptomycin proved to be highly effective in the treatment of a large number of infectious diseases brought about in experimental animals by various bacteria. This was true of *Salmonella schottmüller*, *Shigella gallinarum*, *Brucella abortus*, *Pseudomonas aeruginosa*, *Klebsiella* or Friedländer's bacillus, *Diplococcus pneumoniae*, and a number of organisms commonly found in urinary tract infections.

Subcutaneous doses of 100 to 200 μg of streptomycin were sufficient to protect mice against lethal infections with *S. schottmüller*. Oral doses of about 3 mg were required to protect mice infected by the intraperitoneal injection of 10 lethal doses. Following oral administration, streptomycin remained in

the intestinal tract and gave very low blood concentrations. Urine recoveries in the animals ranged from 70 to 80 per cent when streptomycin was given parenterally and less than 10 per cent when given orally.

The effectiveness of streptomycin against tularemia, caused by *Pasteurella tularensis*, has been most striking. Complete protection was also obtained in experimental infections due to *Proteus vulgaris*: in chick embryos, 1,000 μg of streptomycin were required to arrest 18-hour infections; to protect 50 per cent of the embryos, 250 μg in a single dose or 150 μg in three daily doses were required. Excellent protection was obtained by the use of 150 μg streptomycin against fowl typhoid caused by *Sh. gallinarum* in 11-day-old chick embryos.

Streptomycin was also found to have a marked effect upon experimental plague, *Pasteurella pestis* being inhibited by 1 $\mu\text{g}/\text{ml}$. A series of mice infected with the organism were treated with 400 μg streptomycin daily, beginning 2 days after inoculation and continuing for 6 days. Nine of the ten mice survived 14 days, as compared with the survival of four out of eleven for sulfadiazine-treated mice and one out of nine for the controls. Experimental infections caused by *D. pneumoniae* and *Staphylococcus aureus* were readily controlled when adequate doses of streptomycin were employed.

Of particular importance was pioneering work done by Feldman and Hinshaw on the effectiveness of streptomycin in experimental tuberculosis in guinea pigs. On the basis of an arbitrarily established index of infection, microscopically determined, 100 represented the maximum possible amount of tuberculosis. The control animals, sacrificed after 61 days, exhibited an index of 67 as contrasted to 5.8 for those which had received streptomycin. In another experiment, the corresponding values were 81.9 for the untreated and 2.8 for the treated animals. The daily administration of streptomycin per guinea pig varied from 1,387 to 6,000 pg. Two different strains of the human tubercle bacillus were equally sensitive of streptomycin. *M. tuberculosis* was recovered from the spleens of only one of the guinea pigs treated with streptomycin, the animal having received 1,387 μg streptomycin daily for 54 days.

The conclusion was reached that streptomycin is the most effective tuber-culo-chemotherapeutic agent so far studied. Its relatively low toxicity for guinea pigs, its high efficacy in resolving and suppressing what would otherwise be lethal tuberculosis, established streptomycin as a drug worthy of serious consideration for the treatment of tuberculosis.

Toxicity of streptomycin

To produce toxic manifestations of streptomycin in animals, it was necessary to administer either extraordinarily large doses of the pure material or smaller amounts of certain impure preparations. No correlation was obtained between the acute toxicity of different lots to experimental animals and clinical tolerance by human beings. The behavior of the material in man could not, therefore, be foretold if antibacterial potency alone was assumed to be the criterion of purity; highly concentrated material had on occasion been unsatisfactory clinically, whereas considerably less active preparations had often produced no undesirable effects.

Among the various toxic reactions resulting from the administration of streptomycin, the otic complications received particular attention. In one experiment, 81 patients treated with streptomycin were examined for evidence of aural toxicity. Two months after cessation of therapy, three of these patients showed an absence of vestibular response. During the third month after therapy, one of these three showed increasing recovery of response. All three complained spontaneously of dizziness, which decreased steadily during treatment. One patient showed a high tone loss, but the audiogram was normal in these three patients. Another patient showed a severe loss of hearing 2 weeks after treatment; 2 weeks later he had shown decided improvement in hearing; he had a normal vestibular response throughout. It was not certain that the low tone and high tone losses were due to streptomycin.

Diseases responding to streptomycin

A variety of human and animal diseases caused by various bacteria respond readily to streptomycin treatment. This was brought out in the first comprehensive study of a group of diseases which could be considered to be definitely controlled by streptomycin or to give promise that they would respond favorably. These included tularemia, urinary tract infections, especially those resistant to sulfa drugs and to penicillin, *Klebsiella* and *Hemophilus* infections, bacteremia due to penicillin-resistant organisms, various forms of meningitis, and whooping cough. Streptomycin was also found to be helpful in treatment of a variety of other diseases, such as leprosy, typhoid fever, brucellosis, certain forms of tuberculosis, and probably also bacillary dysentery, cholera, and bubonic plague.

The Committee on Chemotherapy, originally organized by the Committee on Medical Research of the OSRD, undertook the supervision and co-ordination of the first large-scale series of investigations on the use of streptomycin in the treatment of bacterial infections. The various infectious diseases have been divided, in their relation to streptomycin, as follows:

Diseases definitely indicated for streptomycin treatment

1. All cases of tularemia.
2. All cases of *H. influenza* infections:
 - Meningitis
 - Endocarditis
 - Laryngotracheitis
 - Urinary tract infections
 - Pulmonary infections
3. All cases of meningitis due to :
 - E. coli*
 - Pr. vulgaris*
 - . *K. pneumoniae*
 - B. lactis-aerogenes*
 - Ps. aeruginosa*
 - S. paratyphi*
4. All cases of bacteremia due to gram-negative organisms:
 - E. coli*
 - Pr. vulgaris*
 - A. aerogenes*
 - Ps. aeruginosa*
 - K. pneumoniae*
5. Urinary tract infections due to:
 - E. coli*
 - Pr. vulgaris*
 - K. pneumoniae*
 - B. lactis-aerogenes*
 - H. influenzae*
 - Ps. aeruginosa*

Streptomycin found to be a helpful agent but position not yet definitely defined

- (1) Peritonitis due to gram-negative bacteria.
- (2) B. Friedländer's pneumonia.
- (3) Liver abscesses due to gram-negative bacteria.
- (4) Cholangitis due to gram-negative bacteria.
- (5) Penicillin-resistant but streptomycin-sensitive organisms infecting heart valves.
- (6) Tuberculosis.
- (7) Chronic pulmonary infections due to mixed gram-negative flora.
- (8) Empyema due to gram-negative infections.

The Committee evaluated the first 1,000 cases treated with streptomycin. In 409 urinary tract infections, the general recovery rate was 42 per cent. Of 100 *H. influenzae* meningitis infections, recovery was obtained in 79 per cent; in 17 of the fatal cases, treatment was too late to be effective. In 14 cases of meningitis caused by other gram-negative organisms, only 4 died. In 91 bacteremia cases, 49 recovered, 12 improved, 4 showed no response and 26 died; in this group as well, the fatal cases were treated too late. Striking results were obtained in tularemia (63 recoveries out of 67 cases), and in otitis media (7 immediate recoveries out of 8 cases). In 44 cases of pulmonary infections, recovery or improvement was obtained in 61 per cent. Streptomycin was extremely effective against acute Friedlander bacillus infections.

In the evaluation of streptomycin in tuberculosis, 75 cases were used. These included miliary, meningeal, larynx, skin, and renal tract forms. It was recommended that a minimum period of treatment should be 3 to 6 months with daily doses of 1.5 to 3.0 g. Hinshaw, Feldman, and Pfuetze, on the basis of treatment of 100 tuberculosis cases, came to the rather optimistic conclusion, that streptomycin is an antibacterial agent which possesses the unique ability to inhibit the growth of *M. tuberculosis* both experimentally and clinically. It was recommended that treatment of tuberculosis consist of large doses of streptomycin for prolonged periods.

Various forms of tuberculosis were soon found to respond promptly to streptomycin. These included ulcerating tuberculous lesions of the larynx, hypopharynx, and the tracheobronchial tree, which received combined intramuscular and aerosol treatment. Some types of extrapulmonary tuberculosis also responded well, especially chronic, long-standing draining sinus tracts, which usually closed within a few weeks. These remained closed only

if treatment was continued for several weeks after superficial healing. Encouraging results were also obtained in treatment of tuberculosis of bones and joints. Some patients with early tuberculous meningitis responded promptly to adequate intramuscular and intrathecal administration, but frequently these gains were not permanently sustained; temporary clinical remission was frequently realized after streptomycin treatment of early but acute and severe tuberculous meningitis; consciousness could be regained, sometimes within a few days, fever would decline within a few weeks, the patient appearing normal for several months; subsequent exacerbations of the disease were likely to occur and were not likely to respond to treatment.

Disseminated hematogenous tuberculosis of the miliary type heretofore did not respond to treatment, spontaneous recoveries being extremely rare. By the use of streptomycin it was possible to bring about a complete clinical and roentgenologic remission. Tuberculosis of the alimentary tract and tuberculous peritonitis also showed striking symptomatic improvement.

The rapid progress made in the use of streptomycin in tuberculosis was thus well expressed by Walker, Hinshaw, and Barnwall: "The introduction of streptomycin into clinical medicine has been peculiarly exciting to phthisiologists. Not only have they never had an *antibiotic* which was helpful in the treatment of tuberculosis in man, they have never had a proved effective drug of any sort. So far as chemotherapy is concerned, not only was their situation comparable to that of the syphilologist prior to the discovery of penicillin, it was worse than his position before the appearance of Ehrlich and arsenic. Phthisiologists have been dependent on rest in bed, and, more recently, on collapse therapy. The relatively rapid evaluation of a new drug - rapid for tuberculosis if not for pneumococcal pneumonia - is partly the result of experience gained from the investigation of other drugs in other diseases. It is partly a result of the employment, here and in England, of co-operative methods of investigation which have yielded data, reasonably uniform, rapidly, and in large amount."

I need hardly survey the subsequent developments in the use of streptomycin for the different infectious diseases, especially tuberculosis. I need hardly dwell upon the various regimens for treatment, based on special dosages and modes of administration, and its combined use with para-amino-salicylic acid (PAS). I need not list the numerous antibiotics that have been isolated from cultures of actinomycetes, some in our own laboratories, such as neomycin, and some in numerous other laboratories throughout the world. Suffice to say that streptomycin pointed the way, both through the planned

screening programs and through its specific activity against the gram-negative bacteria and tuberculosis, to many of these antibiotics. The rest is history. Medical science and clinical practice have been revolutionized. One may look forward to further discoveries of agents that will combat diseases not now subject to therapy, to more active and less toxic agents than those now available, and to combined therapy of several antibiotics or of antibiotics and synthetic compounds which will prove to be more effective than the use of single substances.

Summary

The discovery of streptomycin as a product of a rather obscure group of microorganisms, the actinomycetes, led to the study of these organisms as potential producers of other chemotherapeutic substances. Following streptomycin, there came in rapid succession chloramphenicol, chlortetracycline, neomycin, oxytetracycline, and, more recently erythromycin and others. These antibiotics have found extensive application as chemotherapeutic agents in the treatment of numerous infectious diseases never before subjected to therapy.

In the treatment of tuberculosis, the more controlled dosage of streptomycin and the supplementary use of PAS tended to overcome some of the limitations of the antibiotic, notably its toxicity and the development of bacterial resistance. The recent introduction of isonicotinic acid hydrazide suggests the possibility that its combined use with streptomycin will tend further to control the disease and overcome undesirable complications. The conquest of the "Great White Plague", undreamt of less than 10 years ago, is now virtually within sight.

I wish to take this opportunity to thank all my collaborators, associates, and graduate students who have participated in the investigations necessary to the development of our broad antibiotic program. Special thanks are due to the following assistants and graduate students: Misses Elizabeth S. Horning, Elizabeth Bugie, Doris Jones, D. Hutchinson, H. C. Reilly and Dorothy G. Smith; Drs. M. Welsch, W. Geiger, W. Garson, E. A. Swart, and Messrs. H. B. Woodruff, Albert Schatz, H. Lechevalier, S. R. Green, D. A. Harris, W. P. Iverson, D. Reynolds, H. J. Robinson, and Otto Graessle; to the chemists, pharmacologists, bacteriologists, and engineers of Merck & Co.; to Drs.

W. H. Feldman and H. C. Hinshaw of the Mayo Clinic; to Dr. C. S. Keefer of the Committee on Medical Research; and to numerous clinical investigators who have helped to make streptomycin the important chemotherapeutic agent it is today.

1. S. A. Waksman and R. E. Curtis, The actinomyces of the soil, *Soil Sci.*, 1(1916) 99-134.
2. S. A. Waksman, Cultural studies of species of actinomyces, *Soil Sci.*, 8 (1919) 71-215.
3. S. A. Waksman and A. T. Henrici, The nomenclature and classification of the actinomycetes, *J. Bacteriol.*, 46 (1943) 337-341.
4. S. A. Waksman, Associative and antagonistic effects of microorganisms. I: Historical review of antagonistic relationships, *Soil Sci.*, 43 (1937) 51-68.
5. S. A. Waksman, Antagonistic relations of microorganisms, *Bacteriol. Revs.*, 5 (1941) 231-291.
6. S. A. Waksman, *Microbial Antagonisms and Antibiotic Substances*, The Commonwealth Fund, N.Y., 1st ed., 1945; 2nd ed., 1947.
7. S. A. Waksman, E. S. Horning, M. Welsch, and H. B. Woodruff, Distribution of antagonistic actinomycetes in nature, *Soil Sci.*, 54 (1942) 281-296.
8. S. A. Waksman and H. B. Woodruff, *Actinomyces antibioticus*, a new soil organism antagonistic to pathogenic and non-pathogenic bacteria, *J. Bacteriol.*, 42 (1941) 231-249.
9. S. A. Waksman and H. B. Woodruff, Streptothrin, a new selective bacteriostatic and bactericidal agent, particularly active against gram-negative bacteria, *Proc. Soc. Exptl. Biol. Med.*, 49 (1942) 207-210.
10. A. Schatz, E. Bugie, and S. A. Waksman, Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria, *Proc. Soc. Exptl. Biol. Med.*, 55 (1944) 66-69.
11. S. A. Waksman, E. Bugie, and A. Schatz, Isolation of antibiotic substances from soil microorganisms with special reference to streptothrin and streptomycin, *Proc. Staff Meetings Mayo Clinic*, 19 (1944) 537-548.
12. D. Jones, H. J. Metzger, A. Schatz, and S. A. Waksman, Control of gram-negative bacteria in experimental animals by streptomycin, *Science*, 100(1944) 103-105.
13. A. Schatz and S. A. Waksman, Effect of streptomycin and other antibiotic substances upon *Mycobacterium tuberculosis* and related organisms, *Proc. Soc. Exptl. Biol. Med.*, 57 (1944) 244-248.
14. H. J. Robinson, D. G. Smith, and O. E. Graessle, Chemotherapeutic properties of streptomycin, *Proc. Soc. Exptl. Biol. Med.*, 57 (1944) 226-231.
15. J. Fried and O. Wintersteiner, Crystalline reineckates of streptothrin and streptomycin, *Science*, 101 (1945) 613-615; J. Fried and H. E. Stavely, Streptomycin. X: The structure of mannosidostreptomycin, *J. Am. Chem. Soc.*, 74 (1952) 5461-5468.
16. F. A. Kuehl, R. L. Peck, A. Walti, and K. Folkers, Streptomyces antibiotics. I: Crystalline salts of streptomycin and streptothrin, *Science*, 102 (1945) 34-45.

17. H. E. Carter, R. K. Clark Jr., S. R. Dickman, Y. H. Loo, P. S. Skell, and W. A. Strong, Isolation and purification of streptomycin, *J. Biol. Chem.*, 160 (1945) 337-342.
18. F. R. Heilman, Streptomycin in the treatment of experimental tularemia, *Proc. Staff Meetings Mayo Chic.*, 19 (1944) 553-558.
19. W. H. Feldman and H. C. Hinshaw, Effects of streptomycin on experimental tuberculosis in guinea pigs: a preliminary report, *Proc. Staff Meetings Mayo Clinic*, 19 (1944) 593-599; 20 (1945) 313-318.
20. W. H. Feldman, H. C. Hinshaw, and F. C. Mann, Streptomycin in experimental tuberculosis, *Am. Rev. Tuberc.*, 52 (1945) 269-298.
21. H. C. Hinshaw, W. H. Feldman, and K. H. Pfuetze, Streptomycin in treatment of clinical tuberculosis, *Am. Rev. Tuberc.*, 54 (1946) 191-203.
22. C. S. Keefer, F. G. Blake, J. S. Lockwood, P. H. Long, E. K. Marshall Jr, and W. B. Wood Jr, Streptomycin in the treatment of infections. A report of one thousand cases, *J. Am. Med. Assoc.*, 132 (1946) 4-10, 70-77.
23. S. A. Waksman, *Streptomycin, its nature and practical application*, Williams & Wilkins Co., Baltimore, Md., 1949.

OTTO WALLACH

Alicyclic compounds

Nobel Lecture, December 12, 1910

The extraordinary honour of being allowed to address this illustrious assembly and to express my thanks for the superb recognition bestowed on my modest work by the Royal Swedish Academy of Sciences, gives me a feeling of happy pride at having been found worthy of distinction by such a distinguished body of men. But I am also deeply moved by another emotion. This place, where Jacob Berzelius once worked is, from the scientific point of view, a holy land for chemists, where, on entering, one still seems to sense the presence of this reverence-inspiring immortal genius. Berzelius sowed the seed, the fruit of which we are reaping now. It is no mere accident, but a logical historical evolution that Friedrich Wöhler concluded his studies here in *Stockholm*, which enabled him to carry out his pioneering work on the essential oil of bitter almonds in *Göttingen*, and that today Wöhler's pupil and successor to the professorial chair in Göttingen has been invited to report on the progress achieved during the recent decades in the field of "cyclic compounds", among which many components of essential oils must be counted.

The chapter of organic chemistry which I shall discuss first, really belongs to that field which Berzelius had included under the heading of "vegetable chemistry".

From a very early age onwards people's attention had been attracted to the volatile substances, characterized by strong smells or flavours, which are among the large variety of substances which form within plants; these were used partly for therapeutic purposes, but in particular for increasing pleasurable sensations by nerve stimulation. Food was flavoured with fragrant herbs; eminent persons whom one wanted to honour, were anointed with exquisite oils; the air in places where acts of worship were performed, was saturated with incense, gum benzoin and myrrh and other scented narcotic drugs; and even the dead were enveloped in fragrant substances before burial.

In consequence of their wide application, spices and seasonings became important trading articles early in history. Trading caravans and later on the

proud ships of the maritime nations dominating the seas, brought even in earliest times to the West and to the northern countries less favoured by Nature, where cinnamon, vanilla, cloves, cardamon and camphor cannot thrive, these precious *aromata*, from the blessed fields of the Orient, where even today there is a strong predilection for high-flavoured spices and heavily scented substances.

Initially only the drugs themselves were used, but later people learnt to extract from the fragrant plants their essential components, the essences, just as they learnt how to produce spirit of wine from the fruit of the vine, and, so to speak, present them in a more concentrated form. This is the start of the history of essential oils.

This is not the place to discuss the methods by which the volatile vegetable substances were obtained, whether by simple heating (so-called dry distillation) or by distillation with steam, or by extraction; neither is it possible to discuss how each of these methods was gradually improved until the present degree of perfection was reached, above all by fractionation of the substances in a vacuum or under reduced air pressure instead of under normal atmospheric pressure; these methods made distillation easier and prevented the formation of undesirable decomposition products during the process.

With time the number of known volatile vegetable substances has increased considerably. It has been possible to trace historically back to a very early age the taxes which were imposed on medicines, spices and similar substances in German towns. Thus, for instance, one finds that in the year 1500 thirteen, in 1540 thirty-eight and in 1708 already one hundred and twenty vegetable oils are mentioned. Nowadays their number is many times larger.

In vast modern factories engaged in extracting essential oils, of which the firm Schimmel & Co. in Miltitz near Leipzig is one of the most important, any kind of plant containing these oils is systematically processed, if obtainable in large enough quantities, and the oils are extracted for a detailed investigation of their components.

The scientific study of the components of the vegetable oils, however, has only been undertaken in modern times. Initially, nobody knew what to do with them from a chemical point of view. The investigation of the essential oil of bitter almonds, undertaken jointly, as already mentioned, by Wöhler and Liebig in 1832 was of epoch-making importance. These famous chemists proved that bitter almonds contain, besides liquid and non-volatile fatty oil, an inodorous solid substance, namely amygdalin, which, under the influence

of acids or ferments absorbs water and splits into glucose, prussic acid and the volatile, strong smelling bitter almond oil. From the series of ferments or enzymes, as called nowadays, which initiate the splitting, one is contained in the almond itself, namely the emulsin. The smell of bitter almonds becomes noticeable as the emulsin causes this splitting reaction. Such substances causing chemical reactions by their mere presence are called catalysts, following Berzelius' proposal. The process itself is called catalysis. In recent times the catalytic phenomena have again been in the foreground of theoretical and practical chemical interests. In this connection it is therefore important to mention briefly the observations made on amygdalin by Wöhler and Liebig, because these illustrate by which processes volatile, odorous substances can form from non-volatile, odourless matter in plants. Substances which have the ability of splitting into more simple volatile ones etc. in the presence of vegetable ferments, are those with a more complex composition. We shall refer to this again later on.

In the first instance we are interested in bitter almond oil itself, this fragrant main component of an important essential oil.

The investigation of this substance by Wöhler and Liebig is regarded as one of the most classic and memorable in chemical science. It showed for the first time that in organic as well as in inorganic compounds, certain compound radicals could play the part of simple substances. This was, in Berzelius' formulation, the sign of a new dawn breaking with regard to organic chemistry. And yet, almost another generation had to go by before the dawn gave way to the full light and the structure of "*benzoyl*", the chemical radical contained in bitter almond oil, became quite clear. Only in 1865 did the ingenious conception of Kekulé find the solution to the up till then mysterious nature of the "aromatic substances", to which bitter almond oil belongs. In bitter almond oil, like in a great number of other substances that previously had been counted among the "aromatic compounds" on behalf of their strong smell, a derivative of benzene is present. The special properties of benzene and its derivatives are caused by the typical arrangement of their carbon atoms. Kekulé demonstrated that the arrangement of the atoms in the normal organic compounds is chain-like, whereas in benzene and its derivatives it is in the form of rings. It is precisely the substances appertaining to the latter group which are distinguished by characteristic smells, particularly those belonging to the aldehydic type of compounds. Bitter almond oil proved to be nothing other than the simplest aldehyde of the benzene series, namely benzaldehyde.

Following this, it was not difficult to determine that substances, the structure of which is quite analogous, can cause the strong odour of vegetable matter; thus, e.g. cinnamic aldehyde causes the smell of cinnamon oil and cassia oil. Anisic aldehyde, vanillin, heliotropin are also aldehydes derived from benzene and occur in the well-known plants to which their names refer. Mention can also be made of substances appertaining to other types of compounds in the benzene series, which are vital components of strong-smelling plants. Among these are e.g. coumarin in *Asperula odorata* (woodruff), some of the phenols, such as thymol from the oil of thyme and in particular a large number of esters. For instance, it has been known for a long time that oil of winter green consists mainly of salicylic acid methyl ester. More recent research has shown that the ester of anthranilic acid, an acid containing nitrogen (*o*-aminobenzoic acid, $C_6H_4(NH_2)CO_2H$), plays a role in some strong-smelling oils.

But, however important the benzene compounds may have proved to be for the aromatic nature of many oils, it became apparent that such benzene derivatives could not be regarded as vital components of numerous other essential oils. In this connection mention must be made of the turpentine oils proper, the various oils extracted from the needles of Coniferae, orange-peel oil, caraway oil (carvone), peppermint oil, eucalyptus oil, fennel oil, thuja oil, camphor oil, and many others. Initially there was no explanation available on the nature of the substance occurring in this group of oils.

It was impossible to define these substances chemically, and the components which had been observed were at first divided into two groups according to their physical properties.

Compounds which are liquid like turpentine oil and remain in a liquid state even at low temperatures, were called "terpenes", and those that were solid, like camphor were called "camphors". Berzelius had already objected to the usage of designating all components of essential oils that precipitated in solid form as camphors, and proposed calling the liquid substances *elaeoptenes*, and the solid ones *stearoptenes*. In spite of the justified objections raised by Berzelius, who considered that the old designations were misleading, as many solid precipitations from oils have no connection whatsoever to camphor, this old distinction between terpenes and camphors has not disappeared. However, the latter designation was later only retained for oxygenous compounds with identical or similar composition as the ordinary camphor; the liquid compounds of such structure were also called camphors.

Various difficulties were encountered when a start was made on identifying

the chemical composition of the oils in question analytically, because in most cases one was dealing with mixtures of substances, difficult to separate from each other. However, very soon a characteristic fact became evident, namely that all important bodies appertaining to this group have 10 carbon atoms.

In organic chemistry we have learnt to derive from compounds containing only carbon and hydrogen, i.e. from the hydrocarbons, all other types of combinations, such as alcohols, aldehydes, ketones, acids, etc. Systematic scientific research therefore had to start with the terpene hydrocarbons. These, the terpenes proper, have 16 hydrogen atoms linked with 10 carbon atoms, therefore the chemical formula is $C_{10}H_{16}$.

Round about the middle of last century eminent French chemists (e.g. Berthelot and Ribon), in particular, had already investigated these hydrocarbons. Of great importance in these investigations was the fact discovered by Biot round about 1816 that turpentine oil, camphor and related compounds could deflect a polarized beam of light. The direction and intensity of the deflection appeared to be characteristic properties of optically active bodies. Since active terpenes which showed larger or smaller divergences in their optical behaviour were found in a great variety of plants, these substances of different origins were regarded as being different from each other. Accordingly, they were given names derived from the starting material, such as terebenthene, camphene, citrene, carvene, cynene, cajeputene, eucalyptene, hesperidene and so on. Gradually in this way, a great number of terpene hydrocarbons were listed in the literature, and the matter became more and more confused and obscure.

It was improbable from the start that there should be so many different hydrocarbons $C_{10}H_{16}$ occurring naturally, as had been assumed and it was obvious that in many cases the divergences observed in the properties had been caused by the presence of impurities. But there were no means available for ascertaining this, because not enough chemical characteristics were known about either the terpenes or their derivatives.

It is true that as early as 1803 Kindt has produced a characteristic solid compound containing chlorine (the present pinene hydrochloride $C_{10}H_{17}Cl$) when he treated turpentine oil with hydrochloric acid; this compound, on account of its smell and the already discussed usage of designating solid precipitates from such oils, was simply called "synthetic camphor". A little later Thenard produced in the same manner another solid compound (the present dipentene dihydrochloride $C_{10}H_{18}C_{12}$) and very much later (1877) Dr. Albert Atterberg in Uppsala made the important discovery that another, easily

crystallizing compound (sylvestrene dihydrochloride) could be obtained from certain Swedish oils, and at about the same time Tilden discovered that some terpenes combine with nitrosyl chloride to form substances which are difficultly soluble. But all knowledge gained by the middle of the 1880's was not adequate to clear up these matters, all the more as the investigations were up against another difficulty: namely the fact that the terpenes are extraordinarily unstable substances. They resinify not only easily by absorbing oxygen when exposed to air, but to a certain extent also when stored for some time without air. This is manifested by the change in physical properties, boiling point, density, and optical behaviour. In particular, these substances suffer displacements in the structure of the molecule under the influence of chemical reagents without, however, changing their composition; *isomerizations*, as they are called in accordance with Berzelius' proposal, take place and this phenomenon is more pronounced in the terpenes than in any other previously known substances.

The premise for any successful investigation therefore was the task of determining definite and reliable distinguishing characteristics for those terpenes which really differ from each other.

Several useful new observations, among others those on crystallized bromination products of the hydrocarbons, have enabled me, by taking into account and further developing the reactions already known, to achieve this aim to such a degree that it is now possible to name sufficient distinguishing characteristics for the various substances within this group. It now became evident that the number of the known terpene hydrocarbons that differ from each other from a chemical point of view, is in fact quite small; at the same time it was possible to separate, facilitated by previous investigations, the existing modification into clearly defined main groups, according to density, boiling point and refractive power. After the accurate diagnosis on the presence of certain compounds had been ensured—even if they only appeared as a part constituent as in essential oils—it was possible to tackle the next task, namely the determination of the mutual chemical relations of the oxygenous and oxygen-free constituents of essential oils.

By suitably converting hydrocarbons into oxygenous compounds and, reciprocally, oxygenated substances into hydrocarbons, and by effectuating the transitions, it was practicable to find the connection between a very large number of substances occurring in essential oils and appertaining to the terpene group and to trace the genetic relationship between the various substances.

It would be hopeless to try and attempt to discuss within the short time available the long series of experiments that were successfully carried out with this purpose in mind. I have given a short tabular survey of my experiments on pages 38-51 of my book "*Terpene und Campher*" (Terpenes and Camphors).

Listing only a few examples here, I want to mention that I succeeded in converting the main component of the ordinary turpentine oil, namely pinene, $C_{10}H_{16}$, into the odiferous constituent of caraway oil, namely *carvone* $C_{10}H_{14}O$, and then reconverting this into the isomers of pinene, i.e. limonene and terpinene. It was possible to convert this same pinene via the lilac-scented alcohol terpineol to eucalyptole, $C_{10}H_{18}O$, the main constituent of worm-seed oil (chenopodium oil) and of the oil from *Eucalyptus globulus*, and to convert this in turn to *l*-limonene, and so on.

As already mentioned, the main purpose of these investigations was the attempt to find links between the various known compounds in order to gain understanding of their mutual relationships; among other factors, reduction of an unsaturated compound to one with a higher degree of saturation played an important part. For instance, the following series is known: $C_{10}H_{14}O$, carvone; $C_{10}H_{16}O$ dihydrocarvone; $C_{10}H_{18}O$ tetrahydrocarvone. It was a simple matter to convert the active carvone $C_{10}H_{14}O$ into dihydrocarvone by direct addition of hydrogen. The last step, however, to $C_{10}H_{18}O$ could not be accomplished directly by the normal chemical reduction methods. Dihydrocarvone does not continue to absorb nascent hydrogen produced by a chemical process.

This aim, therefore, could only be reached in a roundabout way. Dihydrocarvone was rearranged into an isomeric compound, namely carvenone. This then absorbed hydrogen and resulted in $C_{10}H_{18}O$. However, carvenone is no longer optically active and therefore neither is its reduction product. So, up to that time it had not been possible to obtain active tetrahydrocarvone directly from carvone. I am very happy, however, to be able to tell you that within the last few weeks I was able to eliminate conclusively this difficulty. I found, quite unexpectedly, that carvone can be reduced directly to active tetrahydrocarvone by allowing free hydrogen to act on carvone at normal temperatures and in the presence of colloidal palladium. This process can also be excellently applied to other compounds of our series, which otherwise would have been completely unaffected by direct reduction. It has made Possible for certain cases the realization of transitions which had previously seemed impossible and represents a definite step forward. It is

particularly remarkable to note how much more effectively the molecular hydrogen reacts as compared with the nascent hydrogen, which so far has been regarded as being of far superior efficacy.

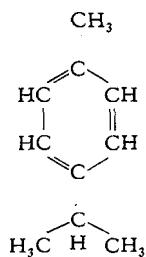
The investigations on the mutual relationships among the terpene compounds have been profitable in other directions as well.

As already mentioned, most of the terpenes are optically active substances. We know that the hypothesis of Le Bel and Van 't Hoff (1874) was the correct interpretation of the phenomenon of optical activity and that prior to this Pasteur had found when mixing equal quantities of laevo-rotatory and dextro-rotatory modifications of a substance in solution, a new substance with quite different properties may result. A mixture of laevo-tartaric acid and dextro-tartaric acid resulted in inactive meso-tartaric acid, the properties of which are entirely different from those of the active constituents. But up to the year 1888 this was the only experimentally established example of *racemization*, as this phenomenon was called (derived from the name acidum racemicum for tartaric acid).

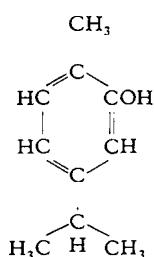
With the discovery of the occurrence of laevo-limonene in the needles of *Pinus sylvestris*, whereas so far only dextro-limonene had been known (as is found e.g. in the orange-peel oil) and with methods of producing well-crystallized derivatives of limonene at my disposal, I was now in a position to obtain whole series of racemic compounds from their active constituents; a secondary result was the discovery that the terpene hydrocarbon dipentene, which so far, on account of the peculiarity of its derivatives had been regarded as a special terpene, is nothing but inactive limonene.

The experiences which were gathered in investigating the mutual relationships between the terpene compounds also assisted in preparing the final solution of the question which is always the most important and most interesting one for theoretical chemists, namely what the internal structure of the molecule, or the constitution, as chemists call it, of this group of compounds would be.

All recent experiments had already confirmed the accuracy of an older assumption, namely that the hydrocarbons $C_{10}H_{16}$ occurring naturally, are related to a benzene hydrocarbon $C_{10}H_{14}$, the so-called *cymol* (isopropyl-*p*-methylbenzene):



Pinene, camphor and other compounds can be converted by appropriate chemical processes to cymol or cymol derivatives, e.g. *carvacrol*:



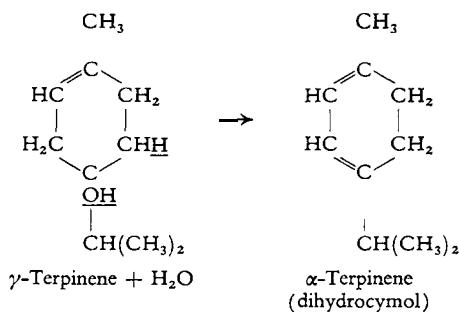
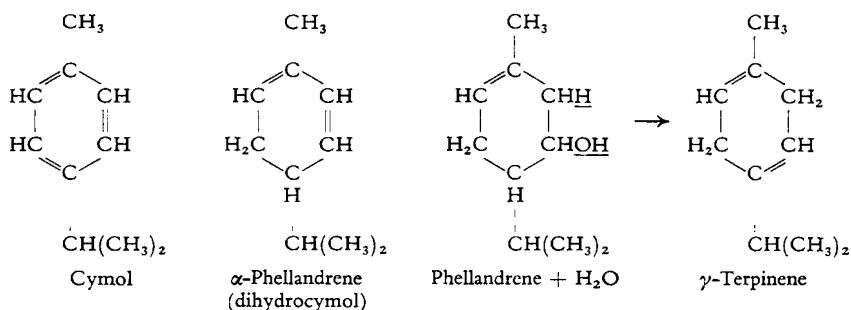
Kekulé, the brilliant originator of the benzene theory, had already expressed the opinion that terpenes were partially reduced cymol. He assumed in the benzene compounds - as shown by our formula - a nucleus of six carbon atoms, which carry, when cyclically linked, three so-called double bonds between the carbon atoms. If one then imagines, for example, one of these double bonds converted into a single bond by the addition of two hydrogen atoms, one arrives from $\text{C}_{10}\text{H}_{14}$ at $\text{C}_{10}\text{H}_{16}$, i.e. a substance with terpene composition. When Kekulé advanced this opinion, the practicability of dissolving bonds in benzene by absorption of hydrogen was only a hypothesis. The credit for proving by experiments that such an additive process is really feasible must essentially go to Ad. von Baeyer, who proved that by doing so, substances could be obtained which have lost their characteristics of benzene compounds, although they still have a ring-like or cyclic (as it is now usually called) arrangement of atoms, and rather resemble in their behaviour the ordinary or aliphatic combinations with chain-like pattern of atoms. Later the designation *alicyclic* was introduced for such compounds (also for other cyclic systems).

In accordance with this concept therefore the terpenes would appear to be

a special case of alicyclic compounds, which form by accident - or for reasons which we as yet do not understand - with particular ease in plants.

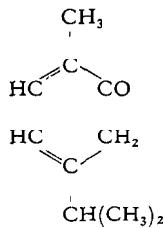
In the main this concept has proved to be correct, even if not quite adequate, as we shall now see. It explains, among other things, why terpene compounds isomerize with such ease (i.e. displace the bonds contained in them). Unsaturated substances (i.e. in our view those containing carbon double bonds) can be added to and can absorb e.g. water, acids, etc. by dissolving existing bonds. The added components can then, in certain circumstances, be split off in directions different from those in which they had been picked up. Thus we arrive at so-called bond displacements, i.e. isomerizations.

The following structural formulae give an example which makes the relation of *phellandrene* and *terpinenes* to *cymol* clear and explains the isomerizations which may occur when water is added and split off, respectively.



However, Kekulé's interpretation could not be used to explain the behaviour of some terpenes, in particular that of the most widely used one, which has been known longest, i. e. *pinene*, the main constituent of the French,

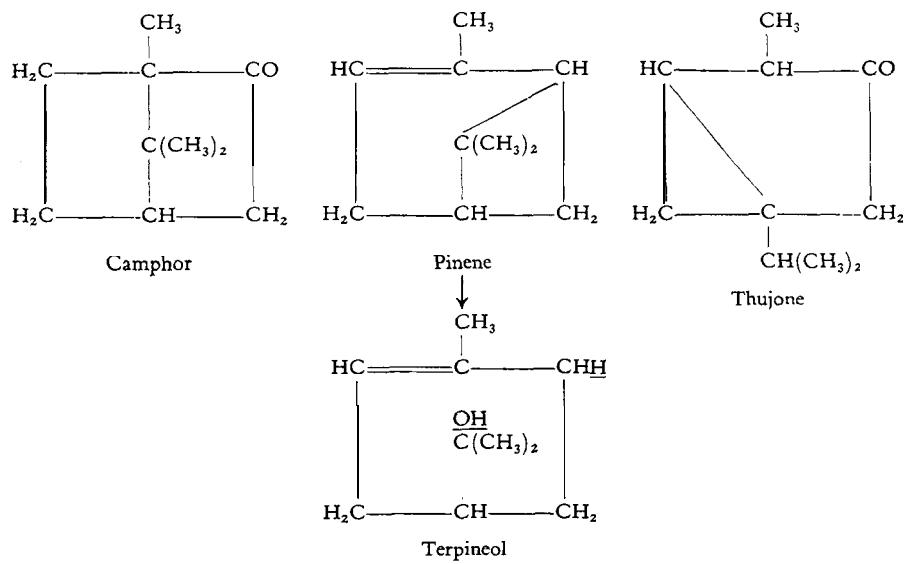
American and Swedish turpentine oil. The same applies to camphor, which Kekulé regarded as the ketone of dihydrogenated cymol, something like:



I must confine myself here to the brief statement that it has now been accepted with certainty that the substances mentioned, and also analogous ones, have their molecule constructed from two mutually interlaced carbon rings, in such a way that when one ring dissolves, the double-ring system again becomes a single one.

The simplest case of this kind occurs with the so-called diagonal bond in the molecule. Bredt proved by a series of excellent experimental investigations that camphor contains two intertwined five-ring compounds. In pinene a four-membered ring is combined with a six-membered one; thujone combines a three-membered ring with a five-membered one.

The following formulae will illustrate the prevailing concept and will also show how, e.g. the double-ring system of pinene can change into a single ring system by water absorption (conversion of pinene to terpineol) :



Credit is principally due to Semmler for having demonstrated that another quite different group of substances contained in essential oils, and distinguished by particular smells, does not belong to the cyclic compounds at all, but to the category of unsaturated aliphatic compounds. These, however, have a strong tendency towards ring formation, and therefore towards conversion to cyclic compounds. One of these substances for instance is citral, which imparts to the oil of lemons its lemon odour, and geraniol, which smells of roses.

As soon as chemists have a definite conception of the internal structure of the molecule of an organic compound, they are able to tackle the task of producing these substances by artificial methods, i.e. by synthesis, as we call it. As far as the terpene compounds occurring naturally were concerned, peculiar difficulties in connection with the synthesis had to be overcome even after their structure was known; these were mainly connected with the sensitivity of these substances to chemical reagents. However, even in this field synthetic preparation had made great progress in recent years. Numerous methods have been made known, first of all for producing alicyclic compounds of terpene-like character; here the chemist differentiates between lower or higher homologous terpene compounds, according to whether the carbon content is smaller or greater than in the natural representatives, which contain 10 carbon atoms. However, also true terpene compounds such as limonene, lately also sylvestrene, terpineol, menthol, camphor and others, have been synthesized from their elements. Among the chemists, to whom great credit is due in this connection, W. H. Perkin Jr. deserves special mention. In any case, one may well say today: there are no more insurmountable difficulties in this field. Even if many of the problems have not been solved as yet: the veil has been removed from the mysterious picture of the terpenes and camphors and there is now nothing which impedes further rapid progress!

Distinguished Audience! It is by no means easy to treat simply and briefly the special subject that I have to discuss, even when addressing specialists, on account of the difficult material. It is even more arduous to make these matters clear in the difficult language of organic chemistry to an audience, which, although so broadly educated, is none the less not expert. I must therefore beg your great indulgence for the topics I have just been expounding to you in terms as general as possible. The majority of the distinguished audience

will probably find the following easier to understand; I shall now discuss the reaction which this progress in our field has had on the development of the industry, which is engaged in the preparation and practical utilization of the essential oils and their components.

Until about 25 years ago, the manufacture of essential oils was purely a matter of trial and error. The plants which contained odiferous constituents were simply distilled and the distillate was put on the market. In view of the lack of knowledge regarding their chemical nature, these products were not always handled in the right way, and above all, the door was wide open for adulteration. There was no remedy for this, even if it was carried out with only moderate skill. A strongly scented oil, like for instance the precious rose absolute, could easily be greatly diluted by worthless additives and the consumer had no reliable means of checking at his disposal. Now all this has been changed greatly. Thanks to the clear characterization of the individual components of essential oils, we now have available adequate analytical methods, by which adulterations can be proved and protection given against it. Apart from the purely chemical methods, investigations relating to the physical properties of the substances, like boiling point, density, refractive index, rotatory power are of great importance in these tests; all these properties were determined with the greatest care in the course of the work on the terpene compounds.

A further progress is denoted by the fact that the importance of the individual components of the essential oils can be correctly assessed on the basis of the knowledge gained. Previously it was thought that the hydrocarbons were of special importance for the scent properties of an essential oil. Now the contrary is known, they are often regarded as useless ballast, which the plant produces together with the really valuable substances, and a start has been made on the production of "terpene-free essential oils", which naturally have a much higher commercial value and are more suitable for practical use.

The endeavours to determine which constituents really impart the specific and valuable properties to an essential oil have led to another result important for the estimation of perfumes. Most of the oils which are valued as scents are mixtures of substances; only the combined effect of these leads to the known result. And in this connection the unexpected fact became apparent that the especially pleasantly scented natural oils not seldom contained a very small quantity of extremely malodorous substances, which nevertheless influence the effect of the scent. These are decomposition products of vege-

table proteins and are closely related to those which are formed by the decomposition of animal proteins in the digestive process and which give such a repugnant smell to faeces. One of these substances is indole. This, for example, is a characteristic component of jasmine oil, as was proved by Dr. Albert Hesse in a brilliant investigation.

This and similar experiences which could only be achieved by the progress in scientific investigations, have been in particular applied to good use by that branch of industry which is engaged in the artificial composition of scents. The preparation of such artificial mixtures is of quite old standing, the best known example being Eau de Cologne, which was invented in 1725, and is a combination of neroli oil and other substances.

Nowadays, however, artificial flower oils - such as rose-blossom oil, jasmin-flower oil, orange-blossom oil - are produced in excellent qualities on a scientific basis; these are not at all, or only slightly, inferior to the natural oils, but are not so exorbitantly expensive; an idea of the price of the natural oils may be given by the calculation that a single kilogramme of essential violet-blossom oil would cost 80,000 Mark and would require 33,000 kilograms of blooms.

The clever utilization for practical purposes of scientific results is the only cause of a highly flourishing essential-oil industry in Germany, with the result that this northern country is now an important competitor in the production of perfumery to the climatically much more favoured Romanic countries (especially the South of France).

The present output of the German essential-oil and artificial-scent industry is estimated at 40-50 million Marks annually. According to their own information, the firm of Schimmel in Miltitz alone have a stock of material of an average value of 2 million Marks.

How closely linked this flourishing industry is with the scientific development is apparent from the fact that the total chemical industry in Germany, which, as is well known, has been constantly increasing, has only doubled its production value in the period from 1550-1895, whereas the essential-oil industry has quadrupled its output during the same period.

We may well expect a lot more practical successes in the utilization of the synthetic methods developed during the last few years. For, as we have seen, the natural terpene compounds are only a special case in the infinitely large group of alicyclic compounds. We are now able to prepare substances which are analogous to the ones that occur naturally and which are distinguished by their smell, and these will therefore have similar properties, because similarly

constructed molecules usually have a similar effect on our nervous system. We can now predict, with certain limitations, that synthetically producible compounds with a certain molecular structure will smell of peppermint, or camphor, or caraway seed or lilac, etc. As soon as we can wrest from Nature the secret of the internal structure of the compounds produced by her, chemical science can then even surpass Nature by producing compounds as variations of the natural ones, which the living cell is unable to construct. In such a manner the great progress was partly achieved by the dyestuff industry, who has now produced compounds of e.g. the indigo and alizarine groups, which cannot be found in Nature. In the same way the pharmaceutical industry is producing compounds which are related to atropine and cocaine, which possess the physiological effect of these remedies, but which have been usefully modified somewhat for certain purposes.

Thus, in our field, too, chemical variation will lead to practical results.

Complete success has already been achieved by the chemical synthesis with regard to the already mentioned artificial preparation of ordinary camphor; in the past we have had to rely almost entirely on the production of the island of Formosa and the goodwill of the Japanese. Even if for various reasons the artificial production of this important substance is not an economic proposition yet, the certainty of having developed methods for the production of synthetic camphor from turpentine oil will prevent the branches of industry which need camphor, like the celluloid industry, from having to face another emergency as was temporarily experienced during the Russo-Japanese War.

As soon as science has solved one problem, new ones arise. This is the essence of science and it applies, of course, also to the field of essential oils.

In principle we now know what the substances are which are produced by plants in the form of essential oils; we can also determine their presence by reliable reactions; we have an idea of their molecular structure, and we can even produce them artificially - an accomplishment which Berzelius still thought impossible for all times.

But behind all this there looms a vast new problem, in comparison to which the one already solved seems quite small: the problem, what kind of chemical processes in the plant organism cause the formation of essential oils. How can we explain their infinite variety?

One would think that plants belonging to the same genus would always produce identical or at least similar oils. But this is by no means so.

As an example let us look at the family of the *Eucalyptus* species, to which the "fever tree", *Eucalyptus globulus*, belongs which is planted so extensively in Italy. This plant contains in its leaves, fruit and other parts eucalyptole or cineole, which imparts the strong smell and which is an oxygenous compound $C_{10}H_{18}O$. On the other hand, another species, *Eucalyptus amygdalina* produces no eucalyptole, but the terpene $C_{10}H_{16}$, i.e. phellandrene, which is characterized particularly by its inconstancy. Yet another species, *Eucalyptus citriodora* contains citronellal, which smells of lemons. *Eucalyptus piperita*, however, supplies a peppermint-scented substance.

Thus, plants botanically closely related form quite different products, and vice versa, some not related at all, produce sometimes identical substances. For instance, eucalyptole is produced just as abundantly from a composita, namely *Artemisia cinsae*, as from *Eucalyptus globulus*. Where is the connection? This can only be found out by a detailed investigation on the manner in which the essential oils in plants are produced and stored. We already know that many plants deposit these substances in special oil cells. This applies, e.g. to the oil in orange peel and orange blossoms. Therefore, these substances can easily be extracted from the plant parts by distillation with vapour. But this by no means applies to all cases. The fresh blooms of jasmine and tuberoses contain only very little oil, and distillation with vapour supplies only very small quantities of unpleasantly smelling substances. Odiferous substances, however, continue to develop in these plants, when the blooms have been cut and the plant begins to die.

Therefore, in these cases the valuable aromata are obtained by a quite different process as used for the preparation of, say, the orange peel or orange blossom oil, namely by the process of the so-called *enfleurage*. The cut plants are enclosed in a so-called "chassis", i.e. cases equipped with sheets of glass which have been covered with grease. The volatile substances which continue to form during that time are absorbed by the grease and can then be extracted from these fragrant pomades.

The processes which take place during this slow formation of odiferous substances in plants such as tuberoses, which can be closely watched, remind us of the facts which we learnt about the formation of the bitter-almond oil from the non-odorous amygdalin. Under the influence of fermentation, complex compounds are gradually split by delivering up the volatile constituents. But to find out what these complex molecules are and under what conditions they are formed must be the subject of further investigations.

We know that the production of essential oil in the plant is connected on

the one hand with its vegetative state, and with the cultivation conditions and other factors on the other hand. In recent times the French research scientist Charabot, in particular, has set himself the task of investigating this subject and these studies will bring us nearer to the solution of the question which substances should really be regarded as the parent substances for the production of the terpenes and camphors in plants.

Distinguished Audience! Theoretically and practically we have made great strides forward within the last quarter of a century in the field of which I have had the honour of giving you a short survey. But we still see in front of us a large fertile field which is waiting to be cultivated. The magnanimous Founder, whose memory we are honouring these days, in accordance with his high-minded intentions to foster science, wished to give recognition to scientific work that had been accomplished, but at the same time he wanted to inspire us to undertake new work. May the work for the further development of chemical science, which has its strongest roots in this beautiful, strong and hard-working country of Sweden, continue to flourish in the future, for the promotion of culture and the benefit of mankind.