



nature physics

ISSN 1745-2473 VOL 8 NO 6

DECEMBER 2012

FERMION PHYSICS

From three to four

TOPOLOGICAL INSULATORS

Large bandgap family found

SUPERCONDUCTOR DEVICES

Coupler pairs couple conductors

The forces behind cell migration

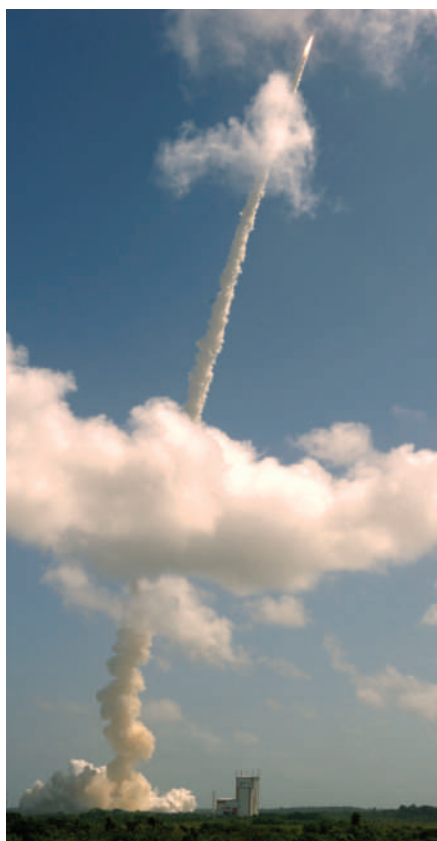
Something to look forward to

It's tempting to look back, but there's so much more to come.

At *Nature Physics*, we often find ourselves thinking about the past. Anniversaries pop up regularly — 20 years (in 2006) since the discovery of high-temperature superconductivity, 50 years in space (2007), and, earlier this year, the anniversaries of C. P. Snow's 'Two cultures' lecture and of Darwin's *On the Origin of Species* (the latter celebrated in this journal as an intellectual achievement that still has a bearing on physicists' thinking). In this issue, in two short articles, we're ruminating on the history of quantum mechanics (page 383) and the centenary of a remarkable development in physics: Geiger and Marsden's α -scattering experiment that led to Rutherford's nuclear atom (page 380).

It's fascinating to look back. Physics has a rich history. Following the line of thought through which a concept has been devised and developed — which may span centuries — is often a good route to understanding. It's a history also rich in personality, as many biographies of physicists attest. But it is the past. So, with the twenty-first century touted as 'the century of biology', is it all over for physics?

Hardly. For all those centuries of achievement, there is still so much to do. And the breadth of enquiry is astounding,



Lift-off: ESA's Planck and Herschel missions begin.

from the most intricate of studies on a lab bench, to profound questions about the nature of the Universe — answers to which could soon be beamed back from a satellite more than a million kilometres from Earth. ESA's Planck mission, and Herschel telescope (Planck, Herschel — again the backward glance, to honour Max Planck and eighteenth-century astronomer William Herschel), launched faultlessly on board an Ariane 5 rocket from French Guiana last month. Both will orbit the second Lagrangian point, from where Herschel will probe the evolution of stars and galaxies at infrared wavelengths and Planck will map the cosmic microwave background radiation in greater detail than ever before. Science operations begin in the next few months.

It's an over-worn phrase, but both missions are truly a 'new window' on the Universe, and on physics. And that's how it works — finding new windows through which to tackle the questions that have become so well formed through the efforts of physicists of preceding centuries. We seek new windows not only on the grand scale of the Universe, but on the scale of the atom, the electron, the electron spin. And we look forward to finding the answers. After all, every century is a century of physics. □

What did you do?

Nature Physics now requires a statement of authors' contributions to a paper.

In the past, at the acceptance of a manuscript for publication in *Nature Physics*, we have gently suggested to the corresponding author that a statement of 'author contributions' be included in the final version. It seemed good practice, and most authors obliged. Now, such a statement is mandatory for any paper published in *Nature* or a *Nature* journal.

It's nothing onerous: just a simple listing of the initials of all authors with a comment on what they contributed to the work reported — be it designing,

preparing, performing, collecting, analysing, modelling, writing or something else. There is no required format beyond that. These statements can vary greatly in their degree of detail, as is appropriate when the content of *Nature Physics* papers is so diverse and 'authors' may constitute anything from a lone pair of theorists to a substantial experimental collaboration. (For examples of statements, see <http://tinyurl.com/39mmyw>.)

It's not hard to imagine, either, why it is a good idea for all papers in the journal to carry an author-contributions statement.

It might serve to ensure that the author list includes only those who really did make a recognizable contribution to the paper. It might be useful for younger scientists, seeking grants or positions, to point to a record of their efforts in producing particular results. It will certainly be beneficial for all concerned, if ever a question of fraud is raised, to know exactly who was responsible for what. □

Details of our editorial policies on authorship are available at http://www.nature.com/authors/editorial_policies/authorship.html.

Fusion–fission hybrids revisited

Jeffrey P. Freidberg and Andrew C. Kadak

With the increasingly urgent need to find solutions to the impending energy crisis, there is growing interest within the fusion community in revisiting the concept of the fusion–fission hybrid reactor. But how soon could such reactors be realized, and could they meet the challenges of the coming century?

The worldwide demand for energy continues to grow at an unprecedented rate. This, combined with the dwindling supply and increasing cost of fossil fuels plus the realization of the threat they pose to the global climate, has led to a renaissance in the nuclear power industry. Yet despite improvements in the design of nuclear–fission reactors, there remain public concerns about their safety and the waste that they produce. Many have suggested that harnessing nuclear fusion could provide a better way to meet the world's energy demands, with greater safety and long-term sustainability, and much smaller quantities of long-lived waste than that produced by nuclear fission. But even the most optimistic of assessments predicts that this technology will not be able to produce electricity on a commercial scale for at least another three decades. Such issues have led to a resurgence of interest in a third nuclear option, which combines aspects of both technologies in the form of the fusion–fission hybrid reactor.

The fusion–fission hybrid

One of the basic properties of nuclear fission is the requirement of a constant flux of thermal neutrons to drive the splitting of heavy nuclei. In a conventional reactor these neutrons are supplied by the fission reaction itself, which requires a certain concentration of the correct fissile isotopes (typically ^{235}U or ^{239}Pu) to be present in the reactor's fuel rods. Each fission reaction releases a huge amount of energy (about 200 MeV) but requires a chain reaction for the reactor to remain self-sustaining.

By comparison, a fusion reactor generates an abundance of neutrons without the need for a chain reaction, but releases less energy per reaction. Each time a deuterium nucleus fuses with a tritium nucleus (the two hydrogen isotopes that are the most promising for use in a fusion reactor), it produces a 3.5-MeV alpha particle and a 14.1-MeV neutron, a total of only 17.6 MeV. For the fusion reaction to

be sustained, a constant supply of new fuel must be fed into the reactor.

The fusion–fission hybrid was conceived to capitalize on the advantages and minimize the disadvantages of both processes, neatly encapsulated by Lidsky in his 1975 review of the subject¹: “Fusion reactors are ‘neutron rich’ and ‘power poor’ while fission reactors are ‘neutron poor’ and ‘power rich’”. The idea, then, is to build a hybrid device, the core of which consists of a fusion reactor whose purpose is to supply a steady flux of neutrons to a surrounding blanket of fissile materials (see Fig. 1). Such a reactor could generate electricity, produce fuel for conventional fission reactors or provide a way to transmute the long-lived actinides of nuclear waste into shorter-lived and materials that are more safely disposable. These are not new ideas, with the earliest reference dating back to Sakharov in 1951 (ref. 2). Yet despite the numerous detailed studies of their potential that followed^{1,3–9}, the conclusions have been the same: further development could not be justified, as the fusion component of such a system was technologically complex, scientifically risky and significantly more expensive than alternatives. So what has happened to change these conclusions? To see why none have ever gone beyond the drawing board — and why some researchers feel that it is time that they should — it is important to understand the pros and cons of each application, and to relate them to their wider context.

Subcritical electricity production

An often cited but not technically significant advantage of a hybrid is that it allows a fission blanket to be operated subcritically. To ensure an adequate supply of neutrons to drive the fission process, a conventional nuclear reactor must operate at a level that is self-sustaining: at least as many neutrons are produced as are consumed in the fission process or non-productively absorbed in other

non-fissionable materials. Because the neutrons of a hybrid are supplied externally from the fissile fuel, there is no need to concentrate this fuel to maintain criticality. Thus a criticality accident in a hybrid is physically impossible. But such accidents are also extremely unlikely to occur in a properly designed light water reactor (LWR) with negative temperature coefficients. In such a design, as the temperature goes up, the neutron population goes down, thereby shutting down the reactor. This is in contrast to the Chernobyl reactor design, which did not have this feature. There, the reactor had a positive temperature coefficient, meaning that as the temperature went up so too did the neutron population.

In practice, however, the safety of a conventional nuclear power plant is not primarily determined by criticality accidents but by other accidents and transients resulting in a failure to deliver enough cooling water to the core after a reactor is shut down. Nevertheless, for the past 25 years the safety record of the nuclear power industry in the United States has been nothing less than remarkable. This situation has improved still further with the advent of advanced LWR designs that are substantially simplified in comparison with older reactors, therefore involving fewer components that might fail, and incorporating passive fail-safe features that cause them to shut down safely without manual intervention when something goes wrong. Most scientists and engineers consider the safety issues of conventional reactors to have been more than adequately resolved by these designs.

And so it is difficult to make a compelling case on the basis of safety for the development of more complex and costly hybrid designs, as both technologies will require emergency core-cooling systems to remove decay heat from the fission process should there be a ‘loss of coolant’ accident. In other words, if

producing electricity safely is the goal, then LWRs are the most economical nuclear solution.

Fissile fuel production

A future concern for LWRs is that the fuel supply will eventually run out or become too expensive to mine. However, with a sufficient flux of hybrid-produced neutrons, the desired fissile fuel isotopes can be produced by neutron capture from the much more abundant non-fissile isotopes ^{238}U or ^{232}Th . This generates ^{239}Pu in the case of the former and ^{233}U in the case of the latter, both of which can readily be used to power LWRs or used directly in sodium-cooled fast breeder reactors, which produce more fuel than they consume, owing to the production of ^{239}Pu in the blanket from ^{238}U .

Even so, a recent study¹⁰ has estimated that the accessible reserves of natural uranium are sufficient for the next 50 to 100 years, even assuming a 10-fold increase in the number of nuclear reactors from those that exist today. Consequently, the issue of fuel supply is a long-term issue, and not one that demands the large and immediate investment that would be necessary to develop a hybrid on a short timescale (25–30 years). For the foreseeable future, the most economical way to obtain fuel for LWRs is to dig it out of the ground.

Safe nuclear waste disposal

As has already been noted, demand for CO_2 -free electricity is growing. As yet, there is no demonstrated industrial-scale process to sequester the enormous amounts of CO_2 produced by coal. Wind and solar power, because of their intermittent nature, their high cost and the absence of inexpensive large-scale energy storage, are not well suited to replace baseload electricity. Simply put, nuclear fission is the only large-scale option available today for the next generation of CO_2 -free, baseload power plants.

There are currently about 100 LWRs in the United States, producing 20% of the nation's electricity. Although there have been no new reactors ordered since the Three Mile Island accident in 1979 — an event that turned majority public opinion in the United States against nuclear power — there have been 26 applications for Construction and Operating Licenses and four new orders for US plants during the past two years, spurred by the growing demand for CO_2 -free electricity plus the existence of a long-term stable fuel supply. Internationally, the growth in the number of nuclear power plants is even more rapid.

Despite this demand and the advances made in the safety and design of fission

reactors, the issue of waste remains a barrier to the public's willingness to accept (or at least be comfortable with) the re-emergence of nuclear power. This concern has recently focused on the proposed nuclear waste disposal facility at Yucca Mountain in Nevada, and the decision of the Obama administration to terminate consideration of this facility, a decision based on politics and not science. With no clear long-term political solution to the growing problem of what to do with the nuclear waste produced in the United States and throughout the world, this issue represents the most promising near- to mid-term application for a fusion–fission hybrid.

One possible solution is transmutation, which converts long-lived nuclear waste (the minor actinides plus, if desired, plutonium) into short-term non-fissile waste that can be disposed of more efficiently in a geological repository. Transmutation is accomplished by bombarding the long-lived waste with high-energy neutrons and is one way to ultimately increase the storage capacity of a geological repository like that proposed at Yucca Mountain by an order of magnitude. Most recent fusion–fission hybrid studies have recognized this fact and focused on the transmutation of long-lived actinides.

With half-lives of the order of thousands of years, the actinides represent the most important long-term radioactive toxicity hazard. It should, nevertheless, be noted that the total volume of actinides including plutonium produced by a LWR generating 1 GW electrical power for one year is only about a cubic foot — not a very large volume. However, it is not just the actinides that need to be considered. Several of the

lighter, long-lived byproducts of fission, such as ^{99}Tc and ^{129}I , are more soluble in moist soil and therefore can more easily be transported to underground water supplies, so they also represent a risk to public safety. Moreover, these lighter elements are inherently more difficult to transmute.

At present, there are two alternative non-fusion approaches that have been proposed for transmuting nuclear waste: non-breeding sodium-cooled fast-spectrum reactors or particle-accelerator-driven spallation hybrids. Both of these options are more developed than the fusion–fission hybrid. Which one will ultimately turn out to be the more desirable from the technological, proliferation, economic and environmental points of view remains to be seen. But from a purely economic point of view it seems that the best approach is not transmutation but disposal of waste in a permanent repository, storage in an interim repository or burial in deep bore holes. In fact the US Department of Energy has already demonstrated long-term disposal for transuranic military wastes quite effectively at the WIPP facility in New Mexico.

The disposal situation for commercial spent fuel is more complicated. Until policy makers decide whether to bury 'once-through' nuclear waste permanently, for instance in a repository such as Yucca Mountain, or instead to reprocess (that is, chemically separate out the actinides and plutonium), the best solution is likely to be an interim storage facility or the continued use of on-site storage at existing nuclear plants. This strategy makes both economic and technological sense. There is no crisis, because commercial waste has already been successfully stored on site in spent

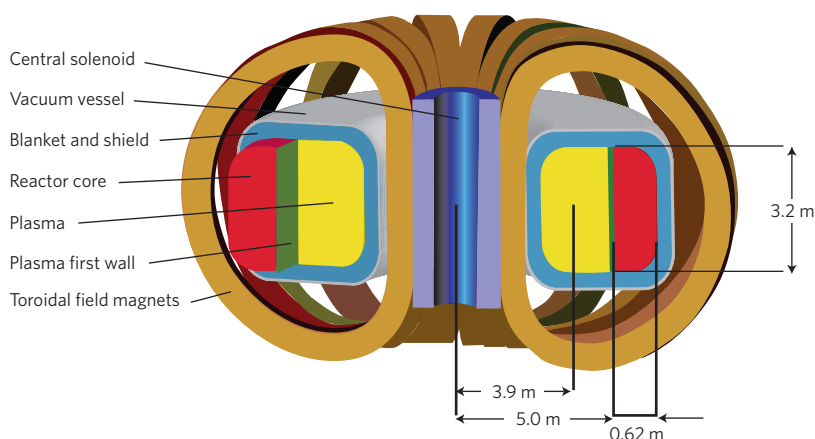


Figure 1 | The most common hybrid design consists of a fusion reactor core surrounded by a blanket of fissile material such as uranium or thorium. The generation of neutrons by the fusion of hydrogen isotopes in the core drives fission reactions in the blanket. These neutrons can be used to generate electricity, produce nuclear fuel for LWRs or transmute waste. Reproduced with permission from ref. 7; © 2008 ANS.

fuel storage pools of water and dry cask storage, as well as in government facilities around the country for almost 50 years. Furthermore, no matter how much waste is transmuted, there will always be some left over so that a long-term geological repository will still be needed, although the storage capacity of a given repository could increase by an order of magnitude compared with the once-through fuel cycle.

What is the bottom line with respect to the nuclear industry? The most useful application of the fusion–fission hybrid is in nuclear waste management. But this does not address the most important immediate problem facing the industry, which is the high capital cost of the plant. The cost of a hybrid further adds to the cost of electricity, as it would be more expensive than either interim or on-site storage because of the complexity and scale of the fusion machine required. The overall implication is that even if the hybrid were available tomorrow, it would not have a game-changing short-term impact on the fission industry because of economics. Stated differently, there may be industrial interest in the use of hybrids for waste management, but only on the mid- to long-term timescale.

The US Department of Energy, on the other hand, has a shorter, more immediate need. The government has the legal responsibility to dispose of nuclear waste and is way behind schedule in doing so. Here the hybrid could help. There might also be an indirect benefit to industry in that if an economically and environmentally credible solution to waste disposal was available that did not require multiple repositories of the type found at Yucca Mountain, the government might be more supportive of the nuclear renaissance. In short, the hybrid could represent a perceived technical solution to an immediate political problem.

At present, then, the transmutation of nuclear waste seems to be the nearest-term application of hybrids, with the government being the primary customer.

Hybrids in fusion research

Fusion research has now been going on for about half a century, with most of the attention focused on understanding the plasma physics required to confine and heat a plasma to thermonuclear temperatures at a high enough density for net power production. Great progress has been made, and it seems that much, although not all, of the scientific uncertainty about fusion associated with early hybrid studies has been greatly reduced.

Still, there are remaining plasma physics issues facing the leading magnetic

fusion concept, the tokamak. First, plasma confinement in the presence of a dominating amount of fusion produces alpha-particle heating; second, efficient non-inductive steady-state operation, which is not possible using the tokamak's transformer; and third, plasma disruptions, which are large-scale magnetohydrodynamic instabilities, can quite literally inflict material damage on the first wall and vacuum chamber of the reactor. These issues are expected to be addressed by the ITER project.

There are also engineering and technology problems associated with the hybrid. An important unsolved plasma engineering problem involves interactions between the plasma and the first wall (that is, the first material surface that makes contact with the plasma). The related issues involve heat load, neutron wall loading, refuelling and impurity removal. There are equally important basic fusion technology problems — materials, blanket design, tritium breeding, magnet development, gyrotron and neutral beam source development, remote handling, reliability and maintainability. All of these problems must be solved in an integrated fashion if the hybrid is to achieve an economically satisfactory capacity factor. Overall, the engineering and technology problems are of comparable difficulty to the plasma physics problems, yet there has been only a small and rapidly diminishing basic research effort in the US fusion programme devoted to engineering and technology problems.

From a plasma physics perspective, a key advantage of the hybrid is that the fusion power gain (the ratio of fusion power produced to power required to maintain the plasma) need only be a value of about 2, as compared with over 50 for a reactor that generates electricity from fusion alone. The reduced plasma physics requirements have led many members of the fusion community to conclude that the hybrid is an attractive intermediate goal on the path to pure fusion electricity. There is merit to this position, although not as much as one might think because of the unsolved engineering and technology problems. A hybrid could be developed in perhaps 25–30 years as compared with 35–50 years for pure fusion electricity.

Finally, there is a rarely discussed issue facing the fusion community that further motivates the development of the hybrid. Pure fusion electricity, when developed, will still have to compete with LWRs for market share, assuming that solutions are in hand to the LWR waste problem and, in the longer term, fuel supply problems. This is likely to be a difficult competition because of the inherent high capital cost associated

with the fusion core, the result of lower power density and increased technological complexity. Interestingly, the fusion–fission hybrid may offer attractive solutions to the LWR problems of waste and fuel supply, thereby postponing the time when pure fusion electricity will be needed.

The conclusion is that the hybrid could serve as an intermediate stepping stone to pure fusion electricity but it also may serve as an end goal in itself, making fission power a sustainable source of electricity for thousands of years.

Outlook

The hybrid may be attractive as a short- to mid-term fusion goal for the fusion community. But there is not a sufficiently compelling case to demand an urgent, Manhattan-like project for its development. The nearest-term (25–30 years) commercial application is to transmute nuclear waste from spent fuel. The hybrid may indeed be a very good way to process radioactive waste but there are other, more mature options (fast reactors and accelerator hybrids) that will be competitive technologically and economically, with the ultimate winner to be determined. A longer-term (50–100 years) commercial application of the hybrid is the production of fissile fuel, which has the important advantage of enabling the nuclear industry to remain focused on LWR technology rather than the breeder. The weakest link in the design of a hybrid, and other fusion systems in general, comes from the unaddressed engineering and technology problems associated with the fusion component of the reactor. Addressing these problems is an indispensable first step in any fusion application, whether it is for electricity, fuel production or waste management, and an important area for future research. □

Jeffrey P. Freidberg and Andrew C. Kadak are in the Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139-4307, USA.
e-mail: jpfreid@mit.edu; kadak@mit.edu

References

1. Lidsky, L. M. *Nucl. Fusion* **15**, 151–173 (1975).
2. Sakharov, A. *Memoirs* (transl. Lourie, R.) (Knopf, 1990).
3. Bethe, H. *Phys. Today* **32**, 44–51 (1979).
4. Moir, R. W. *Nucl. Eng. Design* **63**, 375–394 (1981).
5. Gohar, Y. *Fusion Eng. Design* **58–59**, 1097–1101 (2001).
6. Manheimer, W. J. *Fusion Energy* **25**, 121–139 (2006).
7. Stacey, W. M. et al. *Nucl. Technol.* **162**, 53–79 (2008).
8. https://lasers.llnl.gov/missions/energy_for_the_future/life/
9. Kotschenreuther, M., Valanju, P. M., Mahajan, S. M. & Schneider, E. A. *Fusion Eng. Design* **84**, 83–88 (2009).
10. Forsberg, C. et al. *The Future of the Nuclear Fuel Cycle*. MIT Energy Initiative Report (MIT, in the press).

Attack of the cyberspider

The quest for quantum computing seems a little like that for fusion energy — the necessary technology always seems another decade away, receding into the future almost as fast as we chase it. So far, no one has built a device that has more than about 10 qubits, or carried out any computation that would be impossible on a classical device. The potential for new information devices exploiting quantum physics is surely real enough — as quantum random-number generators and systems for quantum key distribution have already hit the market — but quantum computation, despite unambiguous progress, still seems a reality that we are approaching asymptotically, to be realized (according to current estimates) around 2020 or so.

Even so, it's clear that computation will be very different well before 2020, and even without quantum technology. The more immediate transforming technology is emerging from techniques for controlling and manipulating single molecules, especially information-rich biomolecules, and for constructing what is coming to be called 'molecular cybernetics'.

Double-stranded DNA may be the basis of life, but single-stranded DNA may turn out to be more promising for computing. In impressive work over the past few years, for example, Milan Stojanovich of Columbia University and colleagues have used it to design a number of simple logical gates based on chemical activity. They base their gates on a nucleic acid enzyme — a deoxyribozyme — which catalyses certain DNA reactions. By attaching to this enzyme so-called stem loops — short single-stranded oligonucleotides that bind and inhibit the enzyme activity — they can make this activity sensitive to the presence or absence of further strands of DNA that can disrupt the effects of the stem loops.

This is a little complicated, but the result is a chemical logic in which the presence or absence of specific DNA strands represents logical inputs — 1 or 0 — and the enzyme being active or inactive (reflected in its ability to cleave a certain test oligonucleotide) gives the output of a gate, also 1 or 0, acting on those inputs. With fluorescent markers to detect such activity, the readout can show up directly in colours.

Using various different stem loops, these researchers have managed to design a



Interesting physics is likely to emerge very rapidly from this new field of molecular computing and control.

variety of logic gates including NOT, AND, OR and so on, and, by combining them, to devise automata capable of playing simple games, such as Tic-Tac-Toe (or 'noughts and crosses', as it's also known). The most recent such implementation uses more than 100 basic logic gates distributed among a set of fluid wells making up the game board. In their set-up, the automaton always goes first, choosing the middle well, after which a human player can respond by choosing any of the other eight wells. The chemical system of gates then calculates subsequent moves, and fluorescent markers make the wells turn red (for the automata) and green (for the human). The chemical system has yet to be defeated.

This isn't meant, of course, to be anything other than a proof of principle, and Stojanovich and colleagues point out that this kind of computation will never compete with solid-state devices in terms of speed. Indeed, the Tic-Tac-Toe automaton takes about 30 minutes to make each of its moves, as the underlying DNA chemistry is quite slow. But computation isn't only about speed, even if this is what we usually emphasize in our thinking about it. The potentially revolutionary aspect of this technology is that the computation works in solution, and could, for example, be used to carry out information processing in biological fluids.

Hence, it is easy to imagine future devices carrying out tasks considerably more complex than Tic-Tac-Toe, and doing so autonomously within living cells. There is clearly potential for the engineering of molecular control systems able to detect specific DNA sequences, for example, and to release specific drugs or molecules in response. Just as computation has spread rapidly into every corner of engineering control, we can expect the same kind of transformation of biology and medicine into fields dominated by control based on flexible chemical computation, and an

intelligent and active chemistry that can gather molecular information and calculate delicate actions based on it.

In this regard, one of the most exciting recent developments is the creation, also by Stojanovich and colleagues, of molecular 'spiders' — biomolecular systems with 'legs' made of single-stranded DNA segments having lengths of the order of 10 nm. These spiders can move over a surface covered with single-stranded DNA segments that are complementary to its legs, as they repeatedly bind, dissociate and bind again. The movement of such spiders in large numbers can be controlled by engineering the properties and geometry of the surface, as well as by the physical conditions influencing the statistics of the binding process.

As a result, it's clear that interesting physics is likely to emerge very rapidly from this new field of molecular computing and control, and some physicists have already become actively involved. In recent work, Tibor Antal and Paul Krapivsky have noted that there's an interesting feedback between these spiders and their environment. When a spider leg binds to a site, it typically breaks a bond within the oligonucleotide at that site; hence, a spider's first visit to a site alters that location in an irreversible way, and on future visits, the legs don't bind so strongly. Hence, the terrain a spider sees and its physical influence on a spider varies with the spider's activity, and its diffusion is significantly more complicated than that of a passive particle.

Here one has diffusion with a memory effect, and the proper description of such motion, as Antal and Krapivsky point out, demands the solution of non-Markovian transport models, which they have only begun to explore.

The possibilities for controlling spiders and other novelties will only become richer with incredible objects like the DNA box reported last month (*Nature* **459**, 73–76; 2009). Using a technique known as DNA origami, Ebbe S. Anderson and colleagues were able to make a 3D box a few tens of nanometres on each side with a lid that can be opened by presenting certain DNA keys.

Our future, it seems, may well be written in DNA, just not of the same form as our biological past. □

MARK BUCHANAN

Isn't it demonic

FILM

I tried, really I did. I tried to go with it. It's not everyday this particle physicist formerly employed at CERN gets to see it all up there on the silver screen. *Angels and Demons*, based on the book of the same name by Dan Brown, tells of the theft of a canister of antimatter from CERN by a centuries-old secret society, who then threaten to use it to wipe out the Vatican City. Actually, that makes it sound quite straightforward — in fact, the plot is so convoluted, so weighed down with information that it sinks under the strain.

There's no denying, however, that it's executed in style. Director Ron Howard delivers a good-looking mix of action, hand-held camera work and panoramic shots across the Vatican (only in some of which could I detect the malign hand of CGI: Howard has revealed that, despite a Vatican ban on filming in the city, he was able to get his footage using cameramen disguised as tourists). Rome — the real one and the bits recreated on a Hollywood sound-stage — looks fantastic. I liked Tom Hanks in his role as Harvard academic Robert Langdon. I even couldn't take exception (straining against every prejudiced sinew in my body) to

Ayelet Zurer's pleasingly low-key portrayal of CERN physicist Vittoria Vetra. The problem is in what Hanks, Zurer and Rome are asked to pass off as the plot.

But first, the science bit. The appearance of CERN in *Angels and Demons* is actually rather brief, cut down considerably from the narrative of the book. Gone are the wheelchair-bound Director General and the supersonic plane that does Cambridge, Massachusetts, to Geneva in an hour (shucks). But in their place we have something better — real footage of the Large Hadron Collider and of the ATLAS detector, under construction at the time of filming. There's a pleasing amount of real physics, but it is quickly spattered with the erroneous. Some points are trivial — if only the passage through beam injection to collisions were quite as instantaneous as the film suggests (it's a delicate process that usually takes hours, but a bit of dramatic licence is fine). Plus the control room is peopled by the most unlikely looking physicists: not only are they wearing the requisite Hollywood white coats, they're also sporting ties. Ties!

But then we take a sharp left from the land of the improbable to the territory of the preposterous. Someone shrieks at Vittoria, our female physicist, "The collider was never intended to generate antimatter!" Quite. But Vittoria is siphoning it off, into a

canister-cum-Penning-trap with helpfully transparent walls, so that you can see the spookily glowing antimatter inside ("Does it really fizz like that?", whispered my friend Tim). Vittoria wants to investigate antimatter as a source of clean energy, but is stopped in her tracks by the theft of her canister and the grisly murder of her colleague.

Says Howard, "What goes on at CERN is exploration of the most awesome kind". Agreed. He continues, "What I find incredible is that Dan Brown wrote his novel, setting it here at CERN, about ten years ago — and now, a decade later, CERN is in the news, everybody is talking about the experiments they conduct here. It just shows how ahead of the curve he is." Pardon? CERN was making headlines in the *New York Times* ("Europe 3, US Not Even Z-Zero", in 1983 when the Z boson and the two W bosons were discovered at CERN's SPS collider) long before *Angels and Demons* was even a dollar-sign glint in Dan Brown's eye. The lab enjoyed an unprecedented level of publicity and public interest around the start-up of the Large Hadron Collider last year, but I have my doubts about Brown's prowess as a prophet of scientific progress.

So anyway, the canister containing an "extremely combustible substance called antimatter" (combustible?) is now hidden somewhere in the Vatican, aggravating an already tricky situation, as the Pope has died and the four favoured candidates to succeed him have been kidnapped by a shadowy organization known as the Illuminati. Unless the canister can be found and the antimatter kept in harmless suspension inside by replacing the canister battery, it's annihilation all round. It's over to Tom Hanks' Harvard symbolologist to decipher a coded trail laid centuries before, follow the Illuminati's 'path of illumination' across Rome and save the city.

Now that certainly sounds like the makings of a good thriller, even an intelligent good thriller. But that's the problem with *Angels and Demons*: the throwing in of evocative words, such as 'antimatter' and 'Illuminati', and not just blurring the facts behind them but actively distorting them. The Illuminati was an eighteenth-century German society, and hence did not count the likes of Galileo Galilei or sculptor Giovanni Bernini among their number, as this film (and a webful of



It glows, you know: antimatter stolen from CERN poses a threat to the Vatican City in *Angels and Demons*.

conspiracy-theorists) might have us believe. Antimatter and matter annihilate, but the quantity of antimatter allegedly contained in Vittoria's canister would take a facility like CERN billions of years to create. But, ok, it's a movie, let's pretend.

Well, as I said, I tried. Finally, the weight of pretending got too much to bear, under the battery of all the other words thrown in as cross-bracing to the precarious plot. There's a bandying-around of phrases such as "the new god, Science" and (of course) a rant about the audacity of physicists in devising "the God particle", but no engagement in a real dialogue (literally) between science and religion. Vittoria is not only a beautiful particle physicist, but is qualified too in "bioentanglement", which apparently makes her conveniently expert in toxicology and pathology at the precise

moment that a plot hole opens up. We even get an odd vignette in St Peter's Square about stem-cell research. You see, once you realize one of the 'facts' isn't what it seems, it all comes crashing down.

Then there are some oddities that really could have been ironed out of the screenplay. How weird is it that a Harvard 'symbolist' who has been campaigning for ten years to access Vatican-held original documents by Galileo Galilei speaks no Italian and doesn't understand Latin? Ewan McGregor, in the role of Camerlengo to the deceased Pope, gets saddled with an awkward biographic speech about losing his parents in a UVF bombing (ah, he's supposed to be Irish?) and then being trained as a helicopter pilot. Without wishing to spoil the plot, I think I can reveal that later in the film when his character does get into a helicopter, it

doesn't make it all one iota less preposterous knowing that he is qualified to pilot it.

At the film's premiere in Rome, Howard contended, "Despite all the supposed controversies, despite all that's been said, remember that it's just a film." I don't think that's good enough. You can't appropriate all this stuff — the Renaissance splendour of Rome, the marvel of CERN and its physics, the ongoing confrontation of science and religion — string it together in a manner so cavalier, and then say "it's just a film". Not if I'm supposed to enjoy it. □

REVIEWED BY ALISON WRIGHT

Angels and Demons is now on worldwide release from Columbia Pictures and Imagine Entertainment.

PUBLISHED ONLINE: 17 May 2009

Boldly going...where?

FILM

The most recent voyage of the Starship Enterprise is without doubt the most exciting and thrilling in many a ringed moon — massively enjoyable. With *Star Trek*, J. J. Abrams and crew have performed the seemingly impossible task of taking a franchise equally enriched and

encumbered by a considerable canon, and producing something new and vital, all while pleasing existing fans by acknowledging this heritage.

In the now popular tradition of the prequel, this film traces the backstory of the crew of NCC-1701 (The Original Series to the uninitiated), detailing how they all came to take their places aboard the Enterprise. Conflict is brought about by the "seriously disturbed Romulan", Captain Nero (Eric Bana), who blames Spock (Zachary Quinto) for the destruction of his home-world and swears transtemporal revenge. The Enterprise and its not-yet crew set about putting paid to his diabolical plot.

The action is compelling and the special effects exceed even the considerable expectations of fans. What is unexpected, however, is that the film is crackling with witty reference, both spoken and visual, to other science fiction, especially previous Star Treks. This device enables



© 2009 PARAMOUNT PICTURES CORPORATION

the film to please long-time fans while simultaneously reinvigorating the Star Trek project to entice a whole new generation of Trekkies.

Sadly, this scorching display of talent masks a serious failure. Central to science fiction — and Star Trek in particular — is the exploration of contemporary social and political problems, in a future setting. What has lent such endurance to Star Trek is its sustained ability to deliver moral and political ideas in a popular format, thus securing a place in the hearts of many fans. It is a disappointment that this movie disavows such responsibility. Captain Nero, rather than embodying any ethos, is merely an evildoer to be hunted down. Such an asinine central plot reaffirms, rather than opposes, the

infantile public discourse of our day. Indeed, some characters, such as Chekov (Anton Yelchin), seem to appear on screen only for us to laugh at their comedic foreign accents. Such apparent racism is unbecoming of a franchise founded on the idea of describing an idealized multicultural society.

We trust, however, that like the young crew of the Enterprise, the crew behind this fantastic film will grow into their role and set their prodigious abilities to not merely telling a story fabulously well, but, in the best tradition of Star Trek, telling a story worth telling. □

REVIEWED BY EDMUND JACKSON

Edmund Jackson is chief engineer aboard a quantitative hedge fund.

Spinning around

Phys. Rev. D **79**, 103001 (2009)

Some newly formed neutron stars are believed to be magnetars — named for their incredibly strong magnetic fields, which may be as high as 10^{15} gauss. So far, 16 potential magnetars have been identified by astronomers, but the origin of their magnetic fields is not certain: Kohta Murase and colleagues suggest that there could be valuable clues in neutrinos detected on Earth.

If there is a dynamo mechanism at work in generating the magnetar fields, then these neutron stars are likely to be spinning with a rotation period of as little as a millisecond. Such rotation is good for particle acceleration, and Murase *et al.* show that, under such conditions, interactions of the cold nucleons and thermal photons around the magnetar (the remnant of its supernova) with cosmic-ray protons could produce distinctive fluxes of neutrinos — which could be picked up by large-volume neutrino telescopes such as IceCube, a cubic-kilometre detector array buried in the Antarctic ice and due to achieve full operational capability in 2010.

A better connection

Phys. Rev. A **79**, 040304 (2009)

A number of different systems have been investigated as bits for quantum information. Each has their strengths and weaknesses. David Petrosyan and co-workers now propose a technique for combining two of these approaches — atomic ensembles and superconducting circuits — to take advantage of the best aspects of each.

Superconducting qubits enable fast efficient logic processing, but decoherence means that the information is lost very quickly. Atomic ensembles, on the other hand, in which the information is stored as the collective spin excitation of many atoms, are much better for qubit storage. To transfer the quantum information from one system to the other and back again, Petrosyan *et al.* suggest using a microwave transmission line. This structure would strengthen the coupling between an optical field and the qubits. Quick state-transfer could be achieved by a judicious choice of resonant frequency for each part of the system.

The authors admit that the idea is likely to pose a number of challenges to experimentalists, but it is feasible and could enable high-fidelity quantum computing.

Less than wafer thin

Science doi:10.1126/science.1170775 (2009)

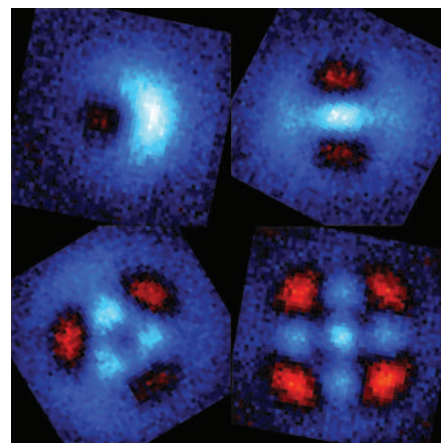
Superconductors are robust. You can pressurize them, freeze them, strain them, subject them to a magnetic field, add dirt to them — up to a limit, of course — but how thin can you make them? From tens of atomic layers downwards, quantum effects such as oscillations in the superconducting order parameter start to appear. Going further, Shengyong Qin and co-workers have made thin films of lead and observe superconductivity down to two atomic layers. Calculations show that one quantum level (or subband) of electronic states remains available for the formation of Cooper pairs.

The samples are grown on silicon substrates, which can affect the metal film. For instance, there are two lattice types: one with an underlying 1×1 atomic structure

(similar in lattice parameter to bulk lead) and another that is $\sqrt{3} \times \sqrt{3}$ (similar to the silicon substrate) and rotated by 30° . Clearly, the film–substrate interface is important and could be used as a tuning parameter for the superconductivity, which has a different transition temperature for each type (4.9 K and 3.65 K, respectively).

Quantum-state synthesizer

Nature **459**, 546–549 (2009)



© 2009 NPG

The superposition principle lies at the heart of quantum physics, and allows the preparation of a quantum system that is simultaneously in several distinct physical states. In practice, however, creating such non-classical states is challenging. But Max Hofheinz and colleagues now demonstrate that, in a superconducting resonator, they can synthesize ‘on demand’ a wide range of quantum-superposition states.

A resonant circuit possesses an infinite number of quantized energy levels, but accessing these levels is difficult, at least when the resonator is driven by a classical signal whose only adjustable parameters are amplitude and phase. Instead, Hofheinz *et al.* gain more control by using a superconducting qubit — a micrometre-sized circuit that acts as an effective two-level quantum system — which serves as a nonlinear element.

They have used a similar scheme before for creating states with a well-defined number of photons, but, with additional control over the amplitude and phase of the photons pumped into the resonator, they can now create superposition states in a deterministic manner — as shown here in these four Wigner tomograms for two- to five-photon states in superposition with the no-photon state.

Useful diversion

Phys. Plasmas **16**, 056110 (2009)

The key challenge in harnessing nuclear fusion for electricity generation is to confine a dense hydrogen plasma at temperatures of hundreds of millions of kelvin for long enough to extract useful amounts of energy. Progress with tokamaks has been made steadily for decades, and is measured by the product of the density, temperature and confinement time of a plasma. But increasing the power density of a tokamak-confined plasma may be limited by whether its exhaust — known as its divertor — could withstand the extreme temperatures and neutron fluxes generated.

Through redesigning the magnetic geometry of the divertor, Prashant Valanju and colleagues propose a solution. Simulations of their design, which they call a Super-X divertor, show that it increases the area of plasma in contact with the divertor plate two- or threefold, thereby reducing the flux and the temperature. If the allowable power density in the core of a plasma is thus increased, this could enable the construction of more compact fusion devices, to be used in the development of a hybrid fusion-fission reactor for processing nuclear waste (see Commentary on page 370).

BIOPHYSICS

Cells guided on their journey

The formation of complex organs, tissue repair and metastasis all require a coordinated regulation of the shape and movement of groups of cells. The mechanical means of communication between cells is crucial to understanding collective cell motions — so how can cells transmit physical forces within cell sheets?

Benoit Ladoux

How living cells are able to sense their environment and adequately respond remains one of the more puzzling issues in cell biology. This is particularly important in the context of embryonic development where a specific and complex architectural organization of biological tissues is defined. Embryonic cells adhere, migrate, segregate and differentiate in a selective and coordinated fashion. As histogenesis proceeds, specific cellular junctions are formed, which contribute to the mechanical cohesion of tissues and act as platforms that make communication between cells possible. Furthermore, dysfunctions in cell adhesion frequently lead to the loss of tissue homeostasis, which has serious physiopathological consequences such as tumour development and metastasis. It is thus important to understand how cells can establish and regulate precise contacts with adjacent cells, and how this depends on their physiological state and position in the embryo or tissue.

In 1917, D'Arcy Thomson published a treatise *On Growth and Form* in which he suggested that morphogenesis could be explained by forces and motion — in other words by mechanics¹. For a while this idea took a back seat in favour of genetics and chemical communication within cell assemblies, but it has recently been revisited and it has been suggested that mechanical forces are important in the organization, growth, maturation and function of living tissues.

Within living tissues, local tension changes can occur during the addition or removal of cells, cell movements linked to morphogenesis, tissue repair or tumour invasion. Therefore, contacts between cells, or between cell and extracellular matrix (ECM), are subjected to force fluctuations and adjust to changes in tension². Cell migration is commonly understood as the movement of individual cells, and this idea has led to a well-established model whereby cells move by the extension and adhesion of a leading edge pointed in the direction of migration, and the

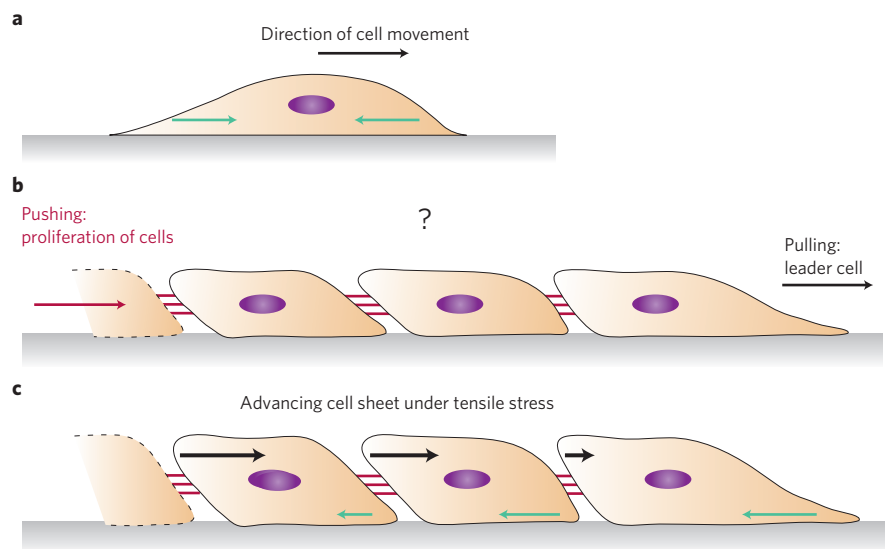


Figure 1 | Force distributions during cell migration. **a**, Schematic representation of the traction forces exerted by a single cell on its substrate. Large traction forces are localized at the leading edge and at the trailing edge, acting in opposite directions. **b**, Different mechanical processes can promote the growth of cell sheets. In particular, proliferation of cells inside the sheet far from the leading edge can induce the build-up of an internal pressure that pushes neighbouring cells outwards. In contrast, peripheral 'leader' cells can generate mechanical tension in such a way as to drive the movement of passive followers. **c**, The mechanical cooperation and the long-range force transmission (green arrows) within advancing epithelial cell sheets induces an increase of the accumulated stress on a plane perpendicular to the substrate and parallel to the cell edge (black arrows). Trepats *et al.*⁶ show that growing cell sheets are under mechanical tensile stress.

retraction and loss of adhesion of the trailing edge³. Here, the transmission of nanonewton-scale contractile forces required for the translocation of the cell body is generated at the points of contact with the surrounding substrate⁴ (Fig. 1a). Besides this well-established mode of cell migration, detailed knowledge obtained over the past 30 years suggests that at least one additional mechanism is important for cell translocation within tissues: the movement of cell groups, sheets or strands consisting of multiple cells connected by cell–cell junctions⁵. The regulation of such a migration mode, although ubiquitous in development, tissue repair and tumour invasion, has been largely

unexplored and has awaited experimental models to decipher important steps of force transmission during collective cell migration.

Xavier Trepats and colleagues⁶, reporting on page 426 of this issue, have used traction force microscopy to investigate how physical forces regulate the motion of epithelial cell sheets. These sheets represent a convenient *in vitro* model to describe many aspects of the migratory behaviour of cell groups. By culturing epithelial cells on flexible gels, researchers can analyse the traction forces exerted at the cell–substratum interface by looking at the deformation pattern of embedded particles that act as markers⁴. To determine the transmission of forces

within an advancing cell sheet, Trepap and colleagues have analysed the radial expansion of cell colonies as a function of time. Surprisingly, they find that large traction forces are observed many cell rows behind the leading edge, suggesting a mechanical cooperation from cell to cell over large distances within the cell sheet. As their results indicate long-range force transmission, this finding will undoubtedly fuel debates in the field, especially on the question of the push or pull mechanism that could drive collective migration (Fig. 1b).

When thinking about cells moving as a cohesive tissue, an important issue is the extent to which mechanical stress propagates within multicellular cohorts to control migration behaviour. Various *in vivo* and *in vitro* situations^{7,8} suggest that extrinsic cues can drive the movement of tissues, not by acting directly on all members of the group, but rather by instructing smaller numbers of peripheral leader cells that in turn seem to be responsible for the guidance of naive followers. But it remains an open question whether the global motion is coordinated by 'leader' cells pulling on cells behind or by internal pressure due to cell division and proliferation that would expand cell sheets outwards. Additionally, other mechanisms could involve submarginal cells that extend 'cryptic' lamellipodia several rows behind the wound margin of epithelial cell monolayers and thus could collectively drive cell-sheet movement⁹.

The direct measurement of physical forces within advancing epithelial cell monolayers provides some evidence to discriminate between these various plausible mechanisms. First, mapping the traction forces in directions normal to and parallel to the front edge at different locations

underlines that large traction forces exerted by cells on the substrate are observed far away from the leading edge. Moreover, both traction distributions show non-Gaussian behaviours with exponential tails, independent of the distance from the edge. It further suggests that cells within the sheet share a common mechanical behaviour with a long-range force transmission. The idea of leader cells moving outwards, normal to the free boundary, 'dragging' passive followers behind them, is not sufficient to explain this complex mechanical process.

The propagation of physical forces depends on cell interactions not only with the substrate but also with neighbouring cells. This implies a 'tug of war' between both types of adhesion¹⁰. Physical signals from the substrates tend to induce a migration of cells away from each other, whereas a stronger mechanical input from cell-cell interactions would drive them towards each other. Thus the importance of cell-cell junctions in the force transmission requires a cell sheet to transmit physical forces in a cooperative way. Consistent with these arguments, Trepap *et al.*⁶ show that the average traction stress exerted by cells on the substrate in the direction perpendicular to the edge is not concentrated at the leading edge but decays slowly with the distance from the edge over several cell diameters, keeping values larger than zero. By applying Newton's third law at various distances from the leading edge of the cell sheet, they are able to use the mechanical force balance to calculate the accumulated stress within the cell sheet. They show that this stress transmitted through cell-cell junctions increases as a function of the distance from the edge. These combined findings clearly demonstrate that guidance within tissues is due to a cohesive

and coordinated movement, and that the growth of the epithelial cell sheet is induced by a global state of tensile stress (Fig. 1c). Interestingly, such a tensile stress rules out the possibility of the build-up of an internal pressure due to cell proliferation that would push neighbouring cells outwards.

A remaining question is how such a tensile state is modulated by external cues, such as the stiffness of the environment. As collective cell migration and traction forces are affected by substrate rigidity^{10,11}, one would expect to observe changes in the value of the tensile stress with the stiffness, providing important information about the reciprocal modulation of tension induced by cell-cell and cell-ECM adhesions. The study by Trepap *et al.* opens a promising possibility of testing the impact of mechanical stress on tissue remodelling and repair by dissecting the respective contributions of local and global forces. □

Benoit Ladoux is at the Laboratoire Matière et Systèmes Complexes (MSC), Université Paris Diderot, and CNRS, UMR 7057, Paris, France.
e-mail: benoit.ladoux@univ-paris-diderot.fr

References

1. Thomson, D. A. W. *On Growth and Form* 2nd edn (Cambridge Univ. Press, 1942).
2. Vogel, V. & Sheetz, M. *Nature Rev. Mol. Cell Biol.* **7**, 265–275 (2006).
3. Lauffenburger, D. A. & Horwitz, A. F. *Cell* **84**, 359–369 (1996).
4. Munevar, S., Wang, Y.-L. & Dembo, M. *Biophys. J.* **80**, 1744–1757 (2001).
5. Friedl, P., Hegerfeldt, Y. & Tusch, M. *Int. J. Dev. Biol.* **48**, 441–449 (2004).
6. Trepap, X. *et al. Nature Phys.* **5**, 426–430 (2009).
7. Lecaudey, V. & Gilmour, D. *Curr. Opin. Cell Biol.* **18**, 102–107 (2006).
8. Poujade, M. *et al. Proc. Natl Acad. Sci. USA* **104**, 15988–15993 (2007).
9. Farooqui, R. & Fenteany, G. *J. Cell. Sci.* **118**, 51–63 (2005).
10. De Rooij, J. *et al. J. Cell Biol.* **171**, 153–164 (2005).
11. Saez, A. *et al. Proc. Natl Acad. Sci. USA* **104**, 8281–8286 (2007).

TOPOLOGICAL INSULATORS

The next generation

Spin-orbit coupling in some materials leads to the formation of surface states that are topologically protected from scattering. Theory and experiments have found an important new family of such materials.

Joel Moore

Topological insulators are materials with a bulk insulating gap, exhibiting quantum-Hall-like behaviour in the absence of a magnetic field. Such systems are thought to provide an avenue for the realization of fault-tolerant quantum computing because they contain surface

states that are topologically protected against scattering by time-reversal symmetry. However, topological phases in condensed matter generally behave like 'hothouse flowers'; they are beautiful but fragile and, until now, were thought to be impossible to create without extremes of temperature and

magnetic field. This conventional wisdom may be overturned by a pair of papers in this issue^{1,2}, which show that a certain class of three-dimensional topological insulator material can have protected surface states and display other topological behaviour potentially up to room temperature without

magnetic fields. On page 398, Zahid Hasan and colleagues¹ report the observation of characteristic signatures of a topological insulator in the band structure of Bi_2Se_3 studied using angle-resolved photoemission spectroscopy (ARPES) and first-principles calculations. In contrast with previously studied materials, Bi_2Se_3 is shown to have a large bandgap and a single surface Dirac cone associated with the topologically protected state in the material. Concurrent theoretical work using electronic structure calculations, reported by Shou-Cheng Zhang and co-workers² on page 438, shows that Bi_2Se_3 is in fact only one of an emerging class of new large-bandgap topological insulator, providing a simple tight-binding model to capture their physical properties. These results pave the way for the experimental realization and potential application of robust topological phases in a variety of materials.

What makes topological insulators different from ordinary band insulators? In a topological insulator, spin-orbit coupling causes an insulating material to acquire protected edge or surface states that are similar in nature to edge states in the quantum Hall effect. For example, the three-dimensional topological insulator phase^{3,4} recently discovered in BiSb alloys⁵ has surface states that are predicted to remain metallic even under quite strong disorder, as long as no magnetic fields or magnetic impurities break the time-reversal symmetry that protects the phase. Such a surface state in a three-dimensional topological insulator is a higher-dimensional analogue of one-dimensional current-carrying edge states in the quantum spin Hall effect⁶. In its simplest form, it can be viewed as a Dirac fermion metal, similar to that in graphene but without the twofold valley and spin degeneracies (Fig. 1).

Although the phase observed in BiSb alloys is theoretically the same as the one now observed in Bi_2Se_3 , there are three crucial differences that suggest that Bi_2Se_3 may become the reference material for future experiments on this phase. First, access to the topologically protected surface state in BiSb is complicated by the presence of several other surface bands. In contrast, ARPES measurements and theory show that only a single surface state is present in Bi_2Se_3 , and that it has an electronic dispersion almost the same as an idealized Dirac cone (Fig. 1). Second, Bi_2Se_3 is stoichiometric — it is a pure compound rather than an alloy like $\text{Bi}_x\text{Sb}_{1-x}$ — and, hence, can in principle be prepared with higher purity and less disorder. This is important because although the topological insulator phase is predicted to be quite robust to disorder, many

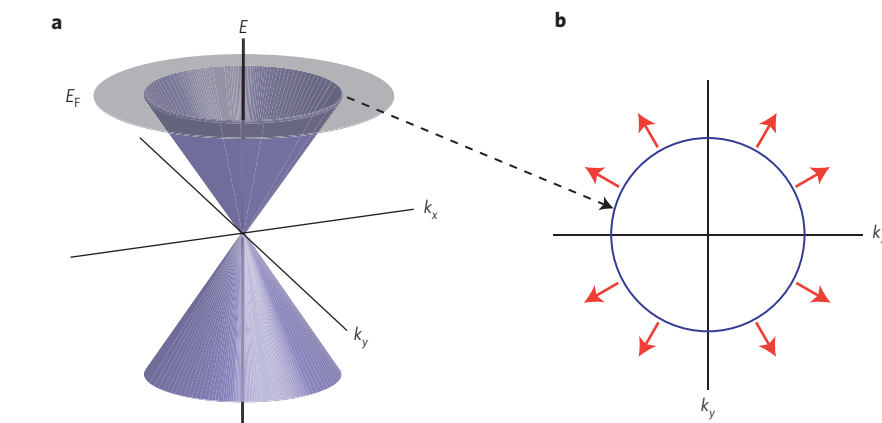


Figure 1 | ‘Light-like’ electrons, protected by time-reversal symmetry, in topological insulators. **a**, A simple model of the surface band structure of a topological insulator with a single Dirac cone. We note that the Fermi level (E_F) does not, in general, pass through the Dirac point. **b**, A distinguishing feature of the topological insulator surface is that there is a single electron spin state at each surface wavevector \mathbf{k} , and that states with opposite wavevector, $-\mathbf{k}$, have opposite spin ‘orientation’.

experimental probes of the phase, including ARPES measurements of the surface band structure, are clearer in high-purity samples. Third, and perhaps most important for applications, Bi_2Se_3 is found to have a large bandgap, of approximately 0.3 eV (equivalent to 3,600 K), which agrees well with theoretical estimates of this quantity.

In combination with the absence of impurity states in the gap, this large bandgap indicates that topological insulator behaviour may be seen at room temperature and greatly increases the potential for applications. To understand the probable impact of these new materials, an analogy can be drawn with the early days of high-temperature superconductivity in the copper oxides: the original cuprate superconductor, lanthanum barium copper oxide, was quickly superseded by second-generation materials such as yttrium barium copper oxide and bismuth strontium copper oxide for most scientific and applications-related purposes. For three-dimensional topological insulators, Bi_2Se_3 is likely to become part of such a second-generation class of material, superseding the first-generation BiSb. Another possible second-generation topological insulator is Bi_2Te_3 . One of the topological insulator materials discussed by Zhang and colleagues², Bi_2Te_3 is already well known to materials scientists working on thermoelectricity — it is a commonly used thermoelectric material in the crucial engineering regime near room temperature.

This second generation of topological insulators is likely to pave the way for a variety of experiments and potential applications. One class of proposed experiment is based on the observation that a weak time-reversal-breaking perturbation

applied to a topological insulator opens a surface bandgap. This results in a quantized magnetoelectric coupling (an applied electrical field induces a magnetic dipole and vice versa, with a proportionality constant of fixed magnitude) resulting from a quantum Hall effect carried by the surfaces⁷. Although some effects resulting from this magnetoelectric coupling — labelled ‘axion electrodynamics’ because of an analogous interaction between the proposed axion particle and electromagnetic fields — were discussed in the 1980s⁸, it was then not clearly understood how such effects may be realized in realistic materials. The improved understanding of magnetoelectric coupling that may result from experiments on topological insulators is also relevant to multiferroic materials, in which axion electrodynamics is a part of the full magnetoelectric response⁹. Consequently, the observation of such a quantized magnetoelectric coupling in the topological insulator is a high priority for experiments, as this emergent property would complement the microscopic band structure observed in photoemission¹.

An even more ambitious experimental direction enabled by these new materials is the study of how correlated-electron physics such as superconductivity is modified in a topological insulator. One particularly intriguing example is the prospect of creating local Majorana fermion excitations, which could be realized through the proximity effect between a topological insulator and an ordinary (fully gapped) superconductor¹⁰. A defining characteristic of a Majorana fermion is that it is its own antiparticle and therefore only has half as many degrees of freedom as a conventional

Dirac fermion such as the electron. So far, Majorana fermions have not been observed clearly in experiment, but it has been predicted that in some situations an electron may split into two Majorana fermions¹⁰, and that several interesting condensed-matter phases are believed to support them as emergent excitations. A direct observation of a Majorana fermion would be a key step on the path to quantum computation using topological phases.

There are important open problems for theory as well. The topological insulator phase can be defined at the single-electron level, manifesting excitations having the quantum numbers (spin and charge) of the

electron, similar to the integer quantum Hall effect. In contrast, the fractional quantum Hall effect is a topological phase displaying excitations with fractional charges and statistics. Our developing understanding of topological insulators may lead us to discover new 'fractionalized' phases of this sort. The two papers in this issue demonstrate that rapid experimental and theoretical progress in the research on topological insulators is both answering and raising fundamental questions pertaining to possible exotic phases of electrons in solids. □

Joel Moore is in the Department of Physics, University of California, and the Materials Sciences

Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

e-mail: jemoore@berkeley.edu

References

1. Xia, Y.-Q. *et al.* *Nature Phys.* **5**, 398–402 (2009).
2. Zhang, H. *et al.* *Nature Phys.* **5**, 438–442 (2009).
3. Fu, L., Kane, C. L. & Mele, E. J. *Phys. Rev. Lett.* **98**, 106803 (2007).
4. Moore, J. E. & Balents, L. *Phys. Rev. B* **75**, 121306 (2007).
5. Hsieh, D. *et al.* *Nature* **452**, 970–974 (2008).
6. König, M. *et al.* *Science* **318**, 766–770 (2007).
7. Qi, X.-L., Hughes, T. L. & Zhang, S.-C. *Phys. Rev. B* **78**, 195424 (2008).
8. Wilczek, F. *Phys. Rev. Lett.* **58**, 1799–1802 (1987).
9. Essin, A. M., Moore, J. E. & Vanderbilt, D. *Phys. Rev. Lett.* **102**, 146805 (2009).
10. Fu, L. & Kane, C. L. *Phys. Rev. Lett.* **100**, 096407 (2008).

NUCLEAR PHYSICS

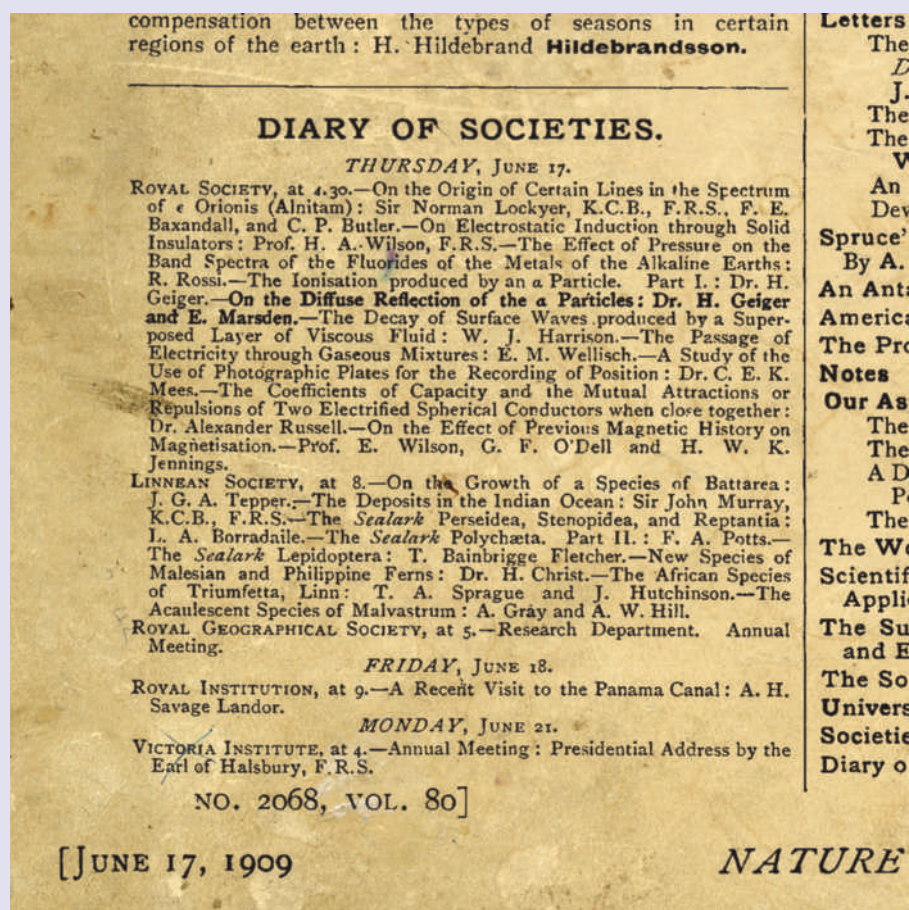
An afternoon's outing

Hidden in the yellowing pages of century-old issues of *Nature* are some scientific gems. They might be fully fledged 'Letters to the Editor', curiosities from 'Notes' or nuggets from 'Our Astronomical Column'. Even the simple listings in 'Diary of Societies', at the end of each issue, can be fascinating — as is this entry (pictured) from the issue of 17 June 1909.

At the behest of their boss, Ernest Rutherford, at the University of Manchester, Hans Geiger and Ernest Marsden had been conducting experiments on the scattering of α particles from a thin gold foil. On that June afternoon — a century ago — they were to present to London's Royal Society their data "On the Diffuse Reflection of the α Particles" (*Proc. R. Soc. A* **82**, 495–500; 1909).

The rest really is history. Geiger and Marsden had observed that, although most α particles passed through the foil pretty much undeflected, very occasionally — and contrary to expectation — an α particle could be scattered right back, through a very large angle. Rutherford had the interpretation: "the atom consists of a central charge supposed concentrated at a point", he wrote later (*Phil. Mag.* **21**, 669–688; 1911); the atom, far from being the 'plum pudding' that had been envisaged, had a nucleus.

Rutherford acknowledged that the essence of his nuclear model had been captured in the 'Saturnian atom' of Japanese physicist Hantaro Nagaoka (*Phil. Mag.* **7**, 445–455; 1904), "which he supposed consisted of a central attracting



mass surrounded by rings of rotating electrons". But it was these data from Geiger and Marsden in 1909, and those that followed, that enabled the detail of the structure of the atom to be drawn more

accurately than ever before. The nucleus was revealed, and a century of nuclear physics began.

ALISON WRIGHT

QUANTUM PHASE TRANSITIONS

Entanglement stirred up

Stirring a two-dimensional quantum fluid at just the right frequency causes the particles to develop strong quantum correlations. This could reveal much about the nature of phase transitions.

Jacob A. Dunningham

Phase transitions have a profound role in physics. Some are familiar, like an ice cube melting in a drink on a warm day. Others are less familiar, but are responsible for fundamental things such as the masses of particles or the clumpy structure of the universe¹. Phase transitions have long been studied in fields like cosmology and condensed-matter physics, but more recently atomic physics experiments with Bose–Einstein condensates (BECs) have started to provide a fresh perspective on the topic². In a theoretical study, on page 431 of this issue, Daniel Dagnino and colleagues³ show that near the critical point of a phase transition in a BEC, strong quantum correlations develop between the particles making up the ultracold atomic vapour. This offers valuable new insights into the physics underlying a class of phenomena known as quantum phase transitions.

Phase transitions can be defined in different ways, but generally involve abrupt changes in the large-scale properties of a system that are often accompanied by symmetry breaking. They fall into two classes: classical transitions, which are driven by thermal noise, and quantum phase transitions, which take place at zero temperature and can only be accessed by varying an external parameter⁴. BECs are an ideal hunting ground for quantum phase transitions because they have extremely low temperatures and their parameters can be changed with relative ease.

Dagnino *et al.*³ consider a BEC in a two-dimensional ‘pancake-shaped’ trap, which is slowly stirred. The geometry of the stirring potential introduces a gap between the energies of the ground and excited states. This is important because it means that the rate of stirring can be varied while always keeping the system in the lowest energy state. Think of a glass of water filled to the brim: if we move it slowly enough, the surface of the water remains flat, but any sudden movements will cause the water to spill.

In contrast to a classical fluid, a BEC cannot undergo rigid-body rotation

when stirred, but can only gain angular momentum by forming quantized vortices beyond some critical velocity. These vortices, which are signatures of superfluidity⁵, are analogous to smoke rings or the behaviour of water when it flows down a sink hole. However, unlike these classical examples, they can only circulate at certain discrete rates. When a superfluid is rotated very rapidly, the number of vortices that appear is comparable to the total number of particles. This case has been studied in some detail because the physics involved is closely related to the quantum Hall effect⁶.

Dagnino *et al.*³ investigate a very different regime, that of slow rotations near the threshold at which the first vortex is formed. Below this threshold, the ground state of the BEC does not rotate, and above it the ground state is a single vortex involving all the particles. Because the system always remains in the ground state, it must undergo a sudden macroscopic symmetry-breaking change as the stirring rate is increased. An interesting question concerns the physics near this critical point. Dagnino *et al.* address that question by considering the relationship between the full quantum ground state and its mean-field approximation.

Mean-field theory is a technique that has been successfully applied in many areas of physics. The idea is that each particle in a many-body system is treated separately and experiences a mean effect due to all the other particles. This simplifies calculations enormously. Dagnino *et al.*³ show that whereas the mean-field approximation describes BECs well far from the critical point, it breaks down near it. This implies that the particles cannot be treated separately and that there are quantum correlations or entanglements between them.

The association between entanglement and phase transitions^{7,8} has been studied in different systems. However, the precise nature of this relationship is not yet fully understood. For example, what role does entanglement play in a

quantum phase transition? And does entanglement provide an efficient way to identify possible phase transitions? Recent work has shown that it is possible to have entanglement without any classical correlations⁹ (between at least three parts of the system). Intriguingly, this means that there could be phase transitions that are revealed only in the quantum correlations but not the classical ones. Studying entanglement could therefore provide a way of detecting phase transitions that is more general than conventional methods.

The scheme of Dagnino *et al.*³ could help to find answers to these questions and an important future development would be to see whether it could be investigated experimentally. One potential obstacle is that the energy gap that protects the ground state from the other levels decreases exponentially with the number of particles. This means that it is likely to be restricted to modest numbers of atoms. Despite that limitation, there should still be scope for uncovering interesting details of the microscopic physics near the critical point. These should lead to general answers to some of the many intriguing questions about phase transitions that are still open. □

Jacob A. Dunningham is in the School of Physics and Astronomy, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, UK.
e-mail: j.a.dunningham@leeds.ac.uk

References

1. Coles, P. & Lucchin, F. *Cosmology: the Origin and Evolution of Cosmic Structure* (Wiley, 2002).
2. Bloch, I., Dalibard, J. & Zwerger, W. *Rev. Mod. Phys.* **80**, 885–964 (2008).
3. Dagnino, D., Barberán, N., Lewenstein, M. & Dalibard, J. *Nature Phys.* **5**, 431–437 (2009).
4. Sachdev, S. *Quantum Phase Transitions* (Cambridge Univ. Press, 2000).
5. Madison, K. W., Chevy, F., Wohlleben, W. & Dalibard, J. *Phys. Rev. Lett.* **84**, 806–809 (2000).
6. Cooper, N. R. *Adv. Phys.* **57**, 539–616 (2008).
7. Wu, L.-A., Sarandy, M. S. & Lidar, D. A. *Phys. Rev. Lett.* **93**, 250404 (2004).
8. Osterloh, A., Amico, L., Falci, G. & Fazio, R. *Nature* **416**, 608–610 (2002).
9. Kaszlikowski, D., Sen(De), A., Sen, U., Vedral, V. & Winter, A. *Phys. Rev. Lett.* **101**, 070502 (2008).

MILLIHERTZ-LINEWIDTH LASERS

A sharper laser

A new approach to lasers that promises optical emission with a spectral linewidth of just 1 mHz could lead to even more accurate and stable atomic clocks.

Uwe Sterr and Christian Lisdat

Optical radiation from trapped atoms can produce laser emission with unprecedented spectral purity and coherence properties — surpassing the linewidth of current lasers, which is limited by the Brownian motion of the optical reference cavity. Writing in *Physical Review Letters*, Dominic Meiser and co-workers¹ show that this idea has the potential to improve the stability of optical lattice clocks by two orders of magnitude.

During recent years, there has been a dramatic improvement in the accuracy of clocks. This progress has mainly been due to atomic-clock technology switching from microwave to optical reference transitions and has led to clocks that tick 10^{15} times per second. Ions or neutral atoms can now also be held undisturbed for periods of several seconds and cooled close to their quantum ground states of motion; therefore, shifts and broadening due to the Doppler effect or from the finite interaction time no longer limit the resolved linewidth.

This progress in the manipulation of absorbers and the control of their motion

makes accessible a host of new optical transitions for use as references in atomic clocks. For example, in neutral alkaline-earth-like atoms, transitions with a natural linewidth in the milli- to microhertz range are available, and in a single Yb^+ ion the upper state of the clock transition has a lifetime of about six years, corresponding to a nanohertz linewidth². To profit fully from these extremely narrow and mostly undisturbed reference transitions, a high-quality laser that serves as a local oscillator must be incorporated into the optical clock, the frequency of which is steered towards the centre of the clock transition by periodical interrogation of the atoms. However, the development of the necessary lasers has not kept pace with the potential of the atomic absorbers. Laser linewidths remain in the neighbourhood of 1 Hz (equivalent to a fractional instability of around 10^{-15}), very similar to the best results observed ten years ago³.

Conventional lasers, such as diode lasers or solid-state lasers, suffer from technical and environmental noise that broadens

their linewidths to tens of kilohertz. To make these suitable for interrogation of narrow atomic transitions, electronic servo loops are used to tame the noisy lasers by forcing their frequencies (or wavelengths) to precisely fit the length of a Fabry–Pérot reference resonator. Thus, the laser wavelength (and frequency) is directly linked to the macroscopic spacing between the mirrors, which is usually around 10 cm. Any disturbance that changes the cavity directly affects the stability of the laser frequency. Therefore, the cavities are isolated from all environmental noise, such as acoustic, seismic and temperature fluctuations. Even with perfect isolation from the environment, however, thermally excited fluctuations of the length of the resonators remain. This Brownian motion of the mirrors, their coatings and the cavity spacer material typically limits the frequency stability of state-of-the-art lasers to a 1 Hz linewidth and to a coherence time of a fraction of a second⁴. There have been several proposals on how to reduce this Brownian noise: apart from the brute-force approach of cooling the cavity to cryogenic temperatures, novel mirrors and other geometries have also been discussed.

Meiser *et al.*¹ present a new solution to the problem that actually reverses the traditional set-up (Fig. 1). Instead of first narrowing the laser radiation with the help of a cavity to interrogate the atoms, in the new approach the atoms themselves emit narrowband optical radiation at the frequency of the optical-clock transition. To avoid any broadening of the emission by atomic motion, about one million atoms are held in an optical lattice nearly at the motional quantum ground state, by a method similar to that nowadays used in optical-lattice clocks.

To obtain a powerful signal, another trick is used. If spontaneously emitted radiation is used, only a vanishingly small amount of power can be extracted because it is spread over the full 4π solid angle. Therefore, Meiser and colleagues proposed to build a high-finesse cavity around the atomic ensemble. Then, as in a laser, induced emission forces the output into a narrow spatial region and ensures very high coherence of the radiation. The authors have

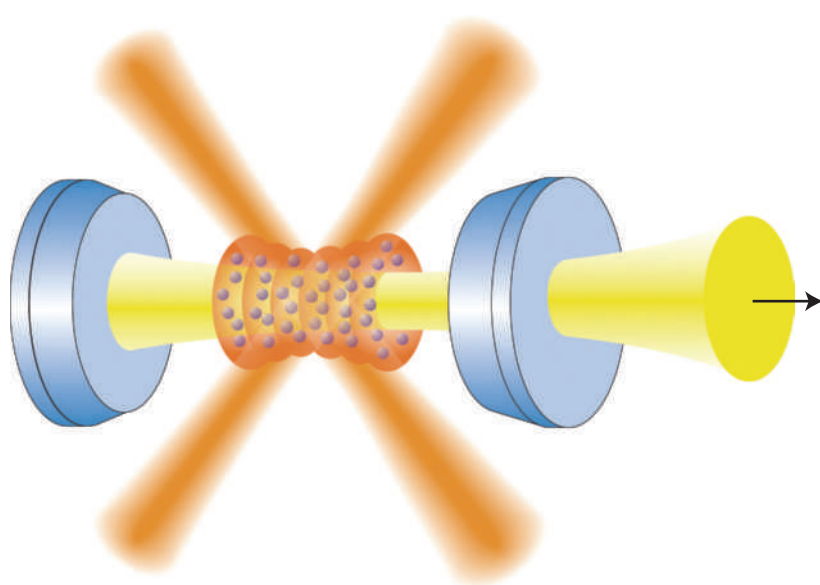


Figure 1 | Using atoms trapped at separate sites of an optical lattice (brown) as an active medium between two highly reflecting mirrors (blue), Meiser *et al.*¹ propose an active laser source that can deliver radiation with a millihertz linewidth.

estimated that up to 10^{-12} W of output power in a linewidth of a few millihertz is emitted from this novel laser, enough to be amplified and used in applications.

The natural linewidth of the atomic transition that is used to produce radiation in this set-up is much less than the linewidth of the cavity. In this sense, rather than to a common laser, the set-up better corresponds to a maser⁵ — the workhorse of frequency metrology where, for example, hydrogen atoms are made to emit narrow-linewidth microwave radiation at 1.4 GHz into a microwave cavity. In the hydrogen maser, a continuous flux of excited atoms replenishes the system with energy lost to emission. In the proposal of Meiser and colleagues¹, this energy is supplied by a rather slow excitation of the atoms by an additional laser field that pumps the trapped atoms back to the excited level of the laser transition. This approach has the benefit that length fluctuations of

the cavity have only a minor effect on the frequency of the emitted radiation.

The proposal is a new approach to the next generation of ultrastable lasers. There are still some open questions: the set-up is experimentally quite demanding, as the atoms have to be held within a small volume defined by the cavity; technical noise might broaden the line; and the back-action of the photon recoil on the cavity and laser fields has so far not been taken into account. Hopefully, an experimental realization will soon show how well these issues can be handled in practice: the race is now on against cryogenic cavities, new materials and other optical designs.

Unquestionably, highly stable lasers will lead to high-stability clocks and enable the further exploitation of atomic reference lines, with a projected stability in lattice clocks of less than 10^{-16} within a one-second averaging time. The ability to compare different clocks leads to improved limits on the possible drift

of fundamental constants in physics, and on tests of local position invariance. The effect of gravity on clocks will then become easily visible (a difference in height of only 1 cm leads to a general-relativistic time dilation of 10^{-18}), making possible fascinating new applications for clocks. □

Uwe Sterr* and Christian Lisdat are at the Physikalisch-Technische Bundesanstalt, Bundesallee 100, 38116 Braunschweig, Germany.

*e-mail: uwe.sterr@ptb.de

References

1. Meiser, D., Ye, J., Carlson, D. R. & Holland, M. J. *Phys. Rev. Lett.* **102**, 163601 (2009).
2. Hosaka, K. *et al. Phys. Rev. A* **79**, 033403 (2009).
3. Young, B. C., Cruz, F. C., Itano, W. M. & Bergquist, J. C. *Phys. Rev. Lett.* **82**, 3799–3802 (1999).
4. Numata, K., Kemery, A. & Camp, J. *Phys. Rev. Lett.* **93**, 250602 (2004).
5. Goldenberg, H. M., Kleppner, D. & Ramsey, N. F. *Phys. Rev. Lett.* **5**, 361–365 (1960).

HISTORY OF QUANTUM THEORY

The short version

"Quantum mechanics is a difficult theory, the history of which is even more difficult." Such is not the conclusion, but the starting point of a study by Olivier Darrigol, in which he sets out to give a simplified account of the complex history of quantum mechanics, and its early history in particular (*Studies in History and Philosophy of Modern Physics* **40**, 151–166; 2009). His "simplified genesis", Darrigol hopes, might serve both physicists and philosophers as a more direct approach to the foundations of quantum theory.

Darrigol considers the period from Max Planck's quantum hypothesis to the first complete mathematical formalism of quantum mechanics, Paul Dirac's transformation theory. During that time, spanning the first quarter of the last century, a number of great minds left their mark, as they investigated a broad spectrum of physical phenomena. Different schools emerged, in terms both of geographical location and of approach, and by the middle of the 1920s, two distinct formulations of quantum mechanics had emerged: matrix mechanics (developed in Germany by Werner Heisenberg, Max Born and Pascual Jordan, and by Dirac in England) and Erwin Schrödinger's wave mechanics.

Darrigol takes these two branches — their formal equivalence was established



© NIELS BOHR ARCHIVE, COPENHAGEN

eventually — as the backbone of his 'simplified history'. He starts the story of matrix mechanics with the failure of classical electrodynamics to describe black-body radiation, leading to the work of Niels Bohr (pictured) on the model of atomic structure and the correspondence principle, and, in what Darrigol says may be regarded as a "necessary consequence" — Heisenberg's quantum mechanics. The developments that eventually led to wave mechanics, on the other hand, he traces back to Einstein's

light-quantum hypothesis and its extension to matter waves, by Louis de Broglie.

Although this overall structure of Darrigol's 'brief history' might be, in itself, not surprising, it is in the selection of key contributions that he chooses to take a new path, so as to construct a coherent sequence of achievements where each step follows as a consequence of previous ones (anything but an easy task in the face of the multilayered history of the field). Also, convoluted derivations are replaced with shorter, more direct reasoning. This approach, Darrigol admits, does leave out important developments, and, in a sense, provides the kind of "linear, great-men accounts" which, in principle, should be avoided in historical writing. But in the light of the already existing large body of work covering the history of quantum mechanics, priority is given to a short and clear account that highlights conceptual connections and key features of quantum theory, in a way that facilitates capturing its foundations, as well as the philosophical stance of some of its fathers. A fuller history, Darrigol argues, would not alter much the basic constructive steps in his simplified genesis, or help to understand why quantum mechanics was born.

ANDREAS TRABESINGER

SUPERCONDUCTOR-METAL HETEROSTRUCTURES

Coherent conductors at a distance

A demonstration that Cooper pairs mediate a non-local coherent coupling between carriers in two normal metal electrodes connected to a superconductor could lead to novel types of superconducting quantum interference devices for studying cross-correlations.

Matthias Eschrig

The flow of current in a superconducting wire is governed by Cooper pairs, which move according to the gradient of the phase of a wavefunction that extends over macroscopic distances, typified by the coherence length that characterizes the separation of the electrons in each pair (on the order of 100 nm for aluminium in its superconducting state). As a consequence, supercurrents exhibit coherent and non-local behaviour. The flow of current in macroscopic, normal metal wires of the sort that we are used to is, in contrast, local — that is, the current density at any given point is determined only by the electric and magnetic fields at that point — and incoherent. So what characterizes the flow of current through a system that consists of both normal and superconducting components? As reported on page 393 of this issue¹, Cadden-Zimansky and colleagues demonstrate that when two normal metal wires are connected to a superconducting wire at a distance within the superconductor's coherence length, the carriers in the

normal wires become coherently and non-locally coupled.

For an electron to pass from a normal metal wire into a superconductor, it must combine with a second electron, of opposite spin and momentum, from the normal metal wire to form a Cooper pair within the superconductor. To conserve the total spin, charge and momentum at the interface, this generates a hole that travels back into the metal in the opposite direction to the incident electron — a process that is known as Andreev reflection². If a second normal wire is connected to the superconductor within a coherence length of the first, electrons from the first (wire A) can combine to form Cooper pairs with electrons from the second (wire B), in a process referred to as crossed (or non-local) Andreev reflection^{3,4} (Fig. 1a).

If wire A is kept at a finite voltage with respect to the superconductor and wire B remains electrically isolated (apart from being connected to the superconductor), the occurrence of crossed Andreev reflection can be measured by the development of a voltage in wire B, caused

by its electron loss to the pairing process. The size of this voltage decays rapidly with the distance between the wires exceeding the coherence length, and is further reduced by the elastic co-tunnelling of electrons from wire A to wire B through intermediate excited states above the gap of the superconductor (Fig. 1b).

Although such effects have been demonstrated before^{5,6}, the nature of the non-local coupling and the degree of coherence between carriers in the two metals have not previously been explored. To do so, Cadden-Zimansky *et al.*¹ use an Andreev interferometer arrangement (Fig. 1c) in which a superconducting wire and the metal wire A are connected in a ring and both ends of the superconducting wire extended beyond the ring to enable contact with wires at a distance. By applying an external magnetic field perpendicular to the ring, the authors maintain a coherent motion of the Cooper pairs around the ring, while measuring variations in the voltages arising from Andreev processes taking place at the contacts.

The magnetic flux through the ring is (nearly) quantized in multiples of the

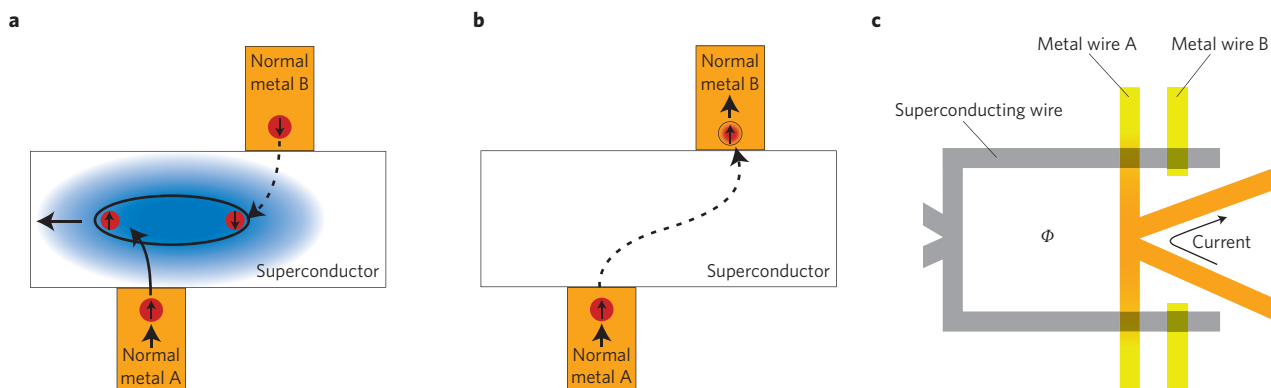


Figure 1 | Competing tunnelling processes involving two normal metal contacts with a superconductor. **a**, Non-local Andreev reflection involves an electron entering the superconductor through the lower normal metal contact and forming a Cooper pair with a second electron from the upper normal metal contact. The process is non-local on the scale of the spatial size of a Cooper pair. **b**, Elastic co-tunnelling, a process in which an electron tunnels directly between metallic contacts through intermediate excited states above the gap in the superconductor, counteracts the non-local Andreev process. **c**, Andreev interferometer set-up used in the experiment. A current is injected through the V-shaped leads on the right. A finite magnetic flux, Φ , penetrates the interferometer loop that is used to modulate a circulating supercurrent. The injected current in combination with the supercurrent leads to voltages at both contacts between the superconductor and the normal metal wires A and B. The non-local voltages are modulated using the flux Φ , showing the coherent nature of the non-local effects involved.

magnetic flux quantum. This causes the magnitude of the supercurrent around the ring to oscillate as a function of the external magnetic field. As the supercurrent itself does not create a voltage at the contact between wire A and the superconductor, Cadden-Zimansky *et al.*¹ use an ingenious method^{7,8,9} that exploits non-locality in a different way: they inject non-equilibrium excitations into wire A that give rise to a current that adds to the supercurrent and leads to the voltage needed at the contact. Owing to current conservation, each current influences the other, and together they allow for tuning of the contact voltage by changing the flux through the ring that controls the supercurrent. This leads to oscillation in the voltage measured at the contact of the superconducting wire with wire A as a function of the flux through the ring. Crucially, Cadden-Zimansky *et al.* show that this in turn induces a non-local voltage in wire B that exhibits flux-dependent oscillations with the same period as those measured in wire A,

thereby demonstrating the coherence of the non-local coupling between them.

Although the experiment does not allow the relative contributions of the crossed Andreev reflection and the elastic co-tunnelling to be determined (in any event, these can be considered independent processes only in the tunnelling limit), the use of Andreev interferometry to investigate such effects does represent a solid starting point. For example, the use of ferromagnetic electrodes to better control the spin polarization of injected carriers, or performing cross-correlation measurements, could provide valuable new information to answer this question.

By demonstrating the ability to tune the coherent interaction of carriers between spatially separated metal contacts, Cadden-Zimansky *et al.*¹ make an important step towards the development of useful devices that exploit crossed Andreev processes, such as control circuits based on quantum interference devices of the kind shown in Fig. 1c, or devices that measure cross-correlations by linking

two or more such devices by means of a common non-local lead. In particular, a further miniaturization of functional electronics will be possible through non-local effects of the kind studied in this work. □

Matthias Eschrig is at the Institut für Theoretische Festkörperphysik, Universität Karlsruhe, D-76128 Karlsruhe, Germany.
e-mail: eschrig@tfp.uni-karlsruhe.de

References

1. Cadden-Zimansky, P., Wei, J. & Chandrasekhar, V. *Nature Phys.* **5**, 393–397 (2009).
2. Andreev, A. F. *Zh. Eksp. Teor. Fiz.* **46**, 1823–1828 (1964); *Sov. Phys. JETP* **19**, 1228–1231 (1965).
3. Byers, J. M. & Flatté, M. E. *Phys. Rev. Lett.* **74**, 306–309 (1995).
4. Deutscher, G. & Feinberg, D. *Appl. Phys. Lett.* **76**, 487–489 (2000).
5. Beckmann, D., Weber, H. B. & von Löhneysen, H. *Phys. Rev. Lett.* **93**, 197003 (2004).
6. Russo, S., Kroug, M., Klapwijk, T. M. & Morpurgo, A. F. *Phys. Rev. Lett.* **95**, 027002 (2005).
7. Volkov, A. F. *Phys. Rev. Lett.* **74**, 4730–4733 (1995).
8. Morpurgo, A. F., Klapwijk, T. M. & van Wees, B. J. *Appl. Phys. Lett.* **72**, 966–968 (1998).
9. Baselmans, J. J., Morpurgo, A. F., van Wees, B. J. & Klapwijk, T. M. *Nature* **397**, 43–45 (1999).

QUANTUM GRAVITY

Progress at a price

The publication of a potentially testable quantum field theory that can accommodate gravity is causing excitement — but it comes at the expense of Lorentz invariance.

Matt Visser

Two papers^{1,2} by Petr Hořava, published in *Physical Review D* and *Physical Review Letters*, seem to have ignited a minor firestorm on the arXiv preprint server, where related submissions have been appearing at the rate of one or two per day (for examples, see refs 3–13). This activity has been prompted by Hořava doing something that was totally unexpected: he has developed a toy model — a proof of principle — that demonstrates that it is, after all, possible to write down a well-behaved (3+1)-dimensional quantum field theory that leads to a reasonable approximation to classical general relativity.

To do this — to build some kind of link, long sought, between quantum physics and general relativity — it is necessary to give something up: the feature Hořava has abandoned is exact Lorentz symmetry^{1–3}. Given the foundational position of Lorentz invariance in the

theory of relativity, this would have been impossible 25 years ago, but so many of the quantum gravity models developed over the past decade have hinted at violations of Lorentz invariance at ultrahigh energies that this is no longer the taboo it once was.

Giving up exact Lorentz symmetry, and implicitly reintroducing a ‘preferred frame’, is certainly a major step³. Although exact Lorentz symmetry is not an *a priori* logical necessity in developing a quantum field theory, it has been one of our major guidelines for the past century. In the longer term, the research community will have to perform some very detailed phenomenological checks to compare Hořava’s model (or its extensions) with observations. Meanwhile, in the short term, there are several key questions that must be asked about the specifics of the model.

Technically, what Hořava has done^{1,2} is, first, to split spacetime into ‘space

plus time’, and to make sure that the Lagrangian of his theory contains at most two time derivatives (this prevents the occurrence of negative-probability ‘ghosts’ and related unitarity violations). Second, he has added terms with up to six space derivatives (this ensures that the graviton propagator falls off sufficiently rapidly at large spatial momenta to make the individual Feynman diagrams of the theory well behaved). By combining these two key features, Hořava’s toy model seems to have done the impossible, namely constitute a renormalizable (and arguably finite) field theory of quantum gravity. That the number of space derivatives in the Lagrangian is not equal to the number of time derivatives leads to the occurrence of ‘anisotropic scaling’, a phrase that in the past has more typically been associated with condensed-matter physics: the model is anisotropic in spacetime, but perfectly isotropic in space.

The questions arise firstly with a symmetry that Hořava introduces, called ‘detailed balance’¹, which restricts the number of terms in the Lagrangian (and so helps keep calculations tractable). Unfortunately, when adding lower-order derivative terms to the most idealized version of his toy model (terms that are needed to obtain an Einstein–Hilbert term in the Lagrangian, and so reproduce classical gravity), Hořava has to adopt a non-zero cosmological constant of the wrong sign to be compatible with observation^{6,9}, and is forced to explicitly break parity invariance in the purely gravitational sector of the model. This naturally leads to the question (perhaps the key question from the quantum field theorist’s point of view) of just how important detailed balance is — is it an essential feature of the model or is it just a simplifying assumption? Are there more general models that allow the gravitational constant and the cosmological constant to be tuned independently? Can explicit parity violation in the pure-gravity sector be eliminated?

Hořava has also introduced an explicit constraint that he calls the ‘projectability condition’¹. From the relativist’s point of view, this is the condition that a certain part of the space-time metric, the ‘lapse function’, can be set globally to unity. Although at first glance this seems a significant constraint, the key solutions of the vacuum and cosmological Einstein equations (the Schwarzschild, Reissner–Nordström, Kerr, Kerr–Newman and Friedmann–

Lemaître–Robertson–Walker metrics) can all be put into this form, at least for the physically interesting parts of those spacetimes⁹. The key question from the point of view of a relativist concerns the importance of this second constraint: is the projectability condition an essential feature of the model or is it just a simplifying assumption?

A further issue is that Hořava’s toy model seems to contain a scalar graviton in addition to the standard spin-2 graviton^{1,2}. Phenomenologically, this is potentially risky, and might, for instance, run into constraints from the gravity-wave-dominated evolution of binary pulsar systems. Should that scalar mode be tuned to zero? Is there any symmetry that would protect this? Moreover, the toy model is purely gravitational, and the question of just how to embed the standard model of particle physics within it¹⁰ will need to be investigated carefully. Because the gravitational field is still completely geometrical, albeit with a preferred frame, it may be that there is no intrinsic difficulty in maintaining a ‘universal’ coupling to the gravitational field. Would there be non-zero signals in Eötvös-type experiments?

Apart from these questions, which I personally view to be fundamental to the whole enterprise, researchers have already begun to investigate the consequences for ultrahigh-energy cosmic rays, the generation of chiral gravitational waves, the effect on cosmological solutions and perturbations^{5,11–13}, modifications of black-hole physics^{4,6,7}, the question of absolute time, emergent gravity⁸ and

much more. Hořava’s work has attracted a lot of attention because it has broken through an informal ‘no-go’ theorem and opened up a vista of ideas that can be tackled with reasonably well known and standard theoretical tools. Furthermore, although in this original incarnation it is only a toy model, it is one of few models of quantum gravity whose variants have any realistic hope of direct comparison with experiment or observation. That certainly makes it well worth a very careful look. □

*Matt Visser is in the School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand.
e-mail: matt.visser@msor.vuw.ac.nz*

References

1. Hořava, P. *Phys. Rev. D* **79**, 084008 (2009).
2. Hořava, P. *Phys. Rev. Lett.* **102**, 161301 (2009).
3. Visser, M. Preprint at <<http://arxiv.org/abs/0902.0590>> (2009).
4. Lu, H., Mei, J. & Pope, C. N. Preprint at <<http://arxiv.org/abs/0904.1595>> (2009).
5. Brandenberger, R. Preprint at <<http://arxiv.org/abs/0904.2835>> (2009).
6. Nastase, H. Preprint at <<http://arxiv.org/abs/0904.3604>> (2009).
7. Cai, R.-G., Cao, L.-M. & Ohta, N. Preprint at <<http://arxiv.org/abs/0904.3670>> (2009).
8. Volovik, G. E. Preprint at <<http://arxiv.org/abs/0904.4113>> (2009).
9. Sotiriou, T., Visser, M. & Weinfurtner, S. Preprint at <<http://arxiv.org/abs/0904.4464>> (2009).
10. Chen, B. & Huang, Q.-G. Preprint at <<http://arxiv.org/abs/0904.4565>> (2009).
11. Mukohyama, S., Nakayama, K., Takahashi, F. & Yokoyama, S. Preprint at <<http://arxiv.org/abs/0905.0055>> (2009).
12. Calcagni, G. Preprint at <<http://arxiv.org/abs/0904.0829>> (2009).
13. Kiritis, E. & Kofinas, G. Preprint at <<http://arxiv.org/abs/0904.1334>> (2009).

IRON ARSENIDE SUPERCONDUCTORS

What is the glue?

Is superconductivity in the iron arsenides conventional? The large isotope effect on both the magnetic and superconducting transitions may indicate that magnetic fluctuations are involved in the superconducting pairing.

D. G. Hinks

Nearly half a century passed between the discovery of superconductivity by H. Kamerlingh Onnes and its theoretical description by Bardeen, Cooper and Schrieffer (BCS theory) in 1954. The isotope effect — the change in the superconducting transition temperature, T_c , caused by a change in the ion mass — was a strong indicator

that phonons were responsible for superconductivity, and the discovery of a large isotope effect in mercury was one of the catalysts leading to the BCS theory. Bardeen, Cooper and Schrieffer showed that the ‘glue’ that binds together electrons to form Cooper pairs in the superconducting state is indeed lattice vibrations. So far, no

mediating agent other than phonons has been definitively shown to occur in any superconductor; however, there is strong evidence that in the copper oxides and other unconventional superconductors magnetism might be a factor. Writing in *Nature*, Rong Hua Liu *et al.*¹ report measurement of the iron isotope effect coefficient (IEC) in the new iron

arsenide superconductors and argue that the superconductivity might be mediated by magnetic fluctuations.

Just over a year ago, a new variety of superconducting material containing iron arsenide layers was discovered². In the FeAs layer, iron atoms sit on a square two-dimensional lattice with arsenic layers above and below arranged such that the iron atoms are all tetrahedrally coordinated. Depending on the nature of the layer separating the FeAs layers, transition temperatures of up to 55 K are possible³. As in the layered copper oxides, the FeAs layer must be 'doped' to become superconducting, usually by ion substitution in the separation layer. In both of these classes of superconductor, the undoped material is magnetic. On doping, the ordered magnetic state is destroyed and superconductivity sets in. Unlike in the copper oxide materials, it is possible in the same iron arsenide compound to dope either electrons or holes into the FeAs layer to induce superconductivity, as both electron and hole Fermi surfaces are present.

A great deal of information has been collected on these materials in a relatively short time, but, as for the copper oxides, there is still no clear-cut agreement over the mediating agent, although there are strong indications that magnetism may have an important role. It has been shown that the order parameter is not a simple *s*-wave state but is unconventional and changes sign between different parts of the Fermi surface.

Liu and co-workers¹ measure the iron IEC for both the magnetic (T_{SDW}) and superconducting transition temperatures in a hole-doped material ($\text{SmO}_{1-x}\text{F}_x\text{FeAs}$, which is of type 1111; Fig. 1) and an electron-doped material ($\text{Ba}_{1-x}\text{K}_x\text{Fe}_2\text{As}_2$, which is of type 122). The IEC for the magnetic transition is measured in the undoped ($x = 0$) compounds and that for the superconducting transition is measured in the doped materials ($x = 0.4$ for potassium and $x = 0.15$ for iron). What they find is somewhat surprising: the IECs (α) for the magnetic and superconducting transitions in both compounds are essentially the same: $0.37(3) = \alpha = -d(\ln T)/d(\ln M)$, where T is either T_{SDW} (spin-density-wave temperature) or T_c and M is the iron mass. Moreover, the authors show that on substitution of a heavier iron isotope, the magnetic transition temperature, T_{SDW} , is suppressed; modifying the lattice phonon modes affects the magnetism. This observation experimentally confirms that the lattice

phonons and magnetism are intimately coupled in the iron arsenides.

On cooling the undoped iron arsenide materials, the nearest-neighbour iron moments order ferromagnetically or antiferromagnetically into a stripe pattern with alternating rows of iron atoms ordered antiparallel. The ferromagnetic ordering along the stripe direction and antiferromagnetic ordering perpendicular to the stripes lead to a frustrated magnetic system with the frustration being ultimately lifted by an orthorhombic distortion of the plane. Unlike the copper oxides, the undoped system is metallic and the magnetic ordering is an itinerant spin-density-wave (SDW) state. Experimentally, the magnetic SDW ordering and phase transition occur nearly simultaneously for the 122 systems and within several kelvin of each other for the 1111 system, indicating a strong coupling between the ordered magnetism and lattice. Calculations⁴ have also shown that a large magnetoelastic coupling effect exists and strongly indicates that the tetragonal–orthorhombic phase transition is most likely driven by the magnetic transition.

As the materials are doped at low temperatures from the magnetic orthorhombic phase, the material becomes non-magnetic, tetragonal and superconducting. The order of these transitions is still not completely clear for all of the various materials, as the synthesis of accurately doped compounds is difficult. Superconductivity is thought to occur only in the non-magnetic tetragonal state. Substitution of ^{18}O for ^{16}O in the separating layer of the type-1111 material has little effect on the magnetic or superconducting transition temperatures, indicating that magnetism and superconductivity occur primarily in the FeAs planes and are relatively independent of the separation layers.

For a simple isotropic BCS superconductor (which the arsenides are not) the maximum α is 0.5 when summed over all atoms in the compound. The large iron isotope effect found by Liu *et al.*¹ for the superconducting transition would, at first glance, seem to indicate that phonons are not only important but are the dominant mediator for superconductivity. However, the superconducting transition temperatures determined using calculated values of the electron–phonon coupling constant and the phonon spectra, are less than 0.1 K. The calculated transition temperatures are so low that it is unlikely that phonons

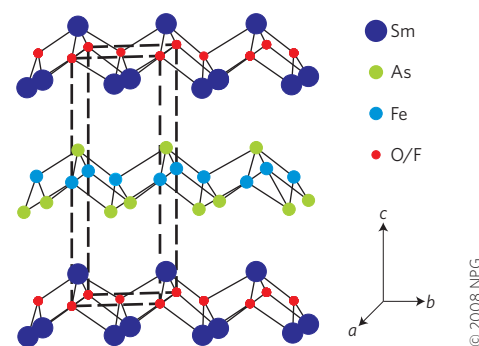


Figure 1 | The structure of $\text{SmFeAsO}_{1-x}\text{F}_x$, with eight atoms per unit cell, as indicated by the dashed lines.

alone mediate the pairing. Even though the long-range magnetic order is removed by doping, short-range spin fluctuations will still be present. The coupling of these spin fluctuations to electrons is thought to be a possible mediator in the copper oxides and now in the iron arsenides as well. That both T_{SDW} and T_c are affected by the iron mass (and with roughly the same sign and magnitude) indicates that they are somehow connected.

The glue for superconductivity in the iron arsenides could be magnetism (that is, the spin fluctuations) or possibly a combination of phonons and spin fluctuations. Although this result is suggestive, the interpretation of the IEC is not straightforward in these complex systems. The copper oxides show almost no oxygen IEC at optimal doping (maximum T_c) but show a value much larger than 0.5 in underdoped, low- T_c material. More work — both experimental, by measuring the IEC as a function of doping, and theoretical, by understanding how magnetism affects the superconducting IEC — will be required to fully understand these data. In the meantime, the results of Liu *et al.*¹ underscore the fact that the excitement generated by the new iron arsenide superconductors is just beginning. □

D. G. Hinks is in the Materials Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439, USA. e-mail: hinks@anl.gov

References

1. Liu, R. H. *et al.* *Nature* **459**, 64–67 (2009).
2. Kamihara, Y. *et al.* *J. Am. Chem. Soc.* **130**, 3296–3297 (2008).
3. Ren, A.-Z. *et al.* *Chin. Phys. Lett.* **25**, 2215–2216 (2008).
4. Boeri, L. *et al.* *Phys. Rev. Lett.* **101**, 026403 (2008).

High-fidelity transmission of entanglement over a high-loss free-space channel

Alessandro Fedrizzi^{1*}, Rupert Ursin¹, Thomas Herbst¹, Matteo Nespoli¹, Robert Prevedel^{1,2}, Thomas Scheidl¹, Felix Tiefenbacher¹, Thomas Jennewein¹ and Anton Zeilinger^{1,2*}

Quantum entanglement enables tasks not possible in classical physics. Many quantum communication protocols¹ require the distribution of entangled states between distant parties. Here, we experimentally demonstrate the successful transmission of an entangled photon pair over a 144 km free-space link. The received entangled states have excellent, noise-limited fidelity, even though they are exposed to extreme attenuation dominated by turbulent atmospheric effects. The total channel loss of 64 dB corresponds to the estimated attenuation regime for a two-photon satellite communication scenario. We confirm that the received two-photon states are still highly entangled by violating the Clauser-Horne-Shimony-Holt inequality by more than five standard deviations. From a fundamental point of view, our results show that the photons are subject to virtually no decoherence during their 0.5-ms-long flight through air, which is encouraging for future worldwide quantum communication scenarios.

Entanglement is at the heart of many peculiarities encountered in quantum mechanics and has enabled many groundbreaking tests on the fundamentals of nature. Entangled photons are ideal tools to investigate the laws of quantum mechanics over long distances and timescales because they are not subject to decoherence. Furthermore, photons can be easily generated, manipulated and transmitted over large distances through optical fibres or free-space links. As the maximal distance for the distribution of quantum entanglement in optical fibres is limited to the order^{2–5} of ~ 100 km with state-of-the-art technology, the most promising option for testing quantum entanglement on a global scale at present is free-space transmission, ultimately using satellites and ground stations⁶.

In recent years, various free-space quantum communication experiments with weak coherent laser pulses^{7–11} and entangled photons^{12–15} have been carried out on ever larger distance scales and with increasing bit rates. So far, the most advanced test bed for free-space distribution of entanglement is a 144 km free-space link between two Canary Islands, where the successful transmission of one photon of an entangled pair was recently achieved¹⁶. In the present experiment, we demonstrate a fundamentally more interesting scenario by sending both photons of an entangled pair over this free-space channel. By violating a Clauser-Horne-Shimony-Holt (CHSH) Bell inequality¹⁷, we find that entanglement is highly stable over long time spans—the photon-pair flight time of ~ 0.5 ms is the longest lifetime of photonic Bell states reported so far, almost twice as long as the previous high^{4,5} of ~ 250 μ s.

The achieved noise-limited fidelity paves the way for free-space implementations of quantum communication protocols that require the transmission of two photons, such as, quantum dense

coding¹⁸, entanglement purification¹⁹, quantum teleportation²⁰ and quantum key distribution without a shared reference frame²¹. From a technological perspective, the overall two-photon loss bridged in our experiment is significantly higher than the current 40 dB limit²² for alternative set-ups relying on weak coherent laser pulses. The attenuation of 64 dB corresponds to the expected attenuation for a satellite scenario with two ground stations⁶, proving the feasibility of quantum communication on a global scale.

The experiment was conducted between La Palma and Tenerife, two Canary Islands situated in the Atlantic Ocean off the West African coast. An overview of the experimental scheme is shown in Fig. 1. At the transmitter station at La Palma, photon pairs at a wavelength of 810 nm and a bandwidth (full-width at half-maximum) of 0.6 nm were generated in a 10-mm-long, periodically poled KTiOPO₄ crystal that was bidirectionally pumped by a grating-stabilized 405 nm diode laser. The photon pairs were coherently combined in a polarization Sagnac interferometer^{23,24} and emitted in the maximally entangled state:

$$|\psi\rangle = 1/\sqrt{2}(|HV\rangle + e^{i\varphi}|VH\rangle) \quad (1)$$

where H/V denote the photons polarization state (horizontal/vertical) and φ is an unknown phase.

At 20 mW of pump power, the source produced $\sim 10^7$ photon pairs per second of which $\sim 3.3 \times 10^6$ single photons per second and $\sim 10^6$ pairs per second were detected locally. These pairs were coupled into single-mode fibres with a length difference of 10 m, which introduced a time delay of $\Delta t = 50$ ns between the two photons. The two photons were then transmitted by two telescopes mounted on a motorized platform and a common receiver telescope—the European Space Agency's Optical Ground Station (OGS) located on Tenerife. The transmitters consisted of single-mode fibre couplers and $f/4$ best form lenses (focal length $f = 280$ mm) that had a lateral separation of 10 cm. To actively compensate the transmitter platform and receiver pointing directions for drifts of the optical path through the time-dependent atmosphere, a bidirectional closed-loop tracking mechanism was used: at the transmitter platform, the virtual position of a 532 nm beacon laser attached to the OGS was monitored in the focus of a third telescope by a CCD (charge-coupled device) camera. Likewise, the OGS monitored the position of a beacon laser mounted at the transmitter (see Fig. 1 and ref. 11 for details).

At the OGS, the incoming photons were collected by a 1 m mirror ($f = 38$ m). To ensure that turbulence-induced beam

¹Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, Boltzmanngasse 3, 1090 Vienna, Austria, ²Quantum Optics, Quantum Nanophysics and Quantum Information, Faculty of Physics, University of Vienna, Boltzmanngasse 5, 1090 Vienna, Austria.

*e-mail: alessandro.fedrizzi@univie.ac.at; zeilinger-office@univie.ac.at.

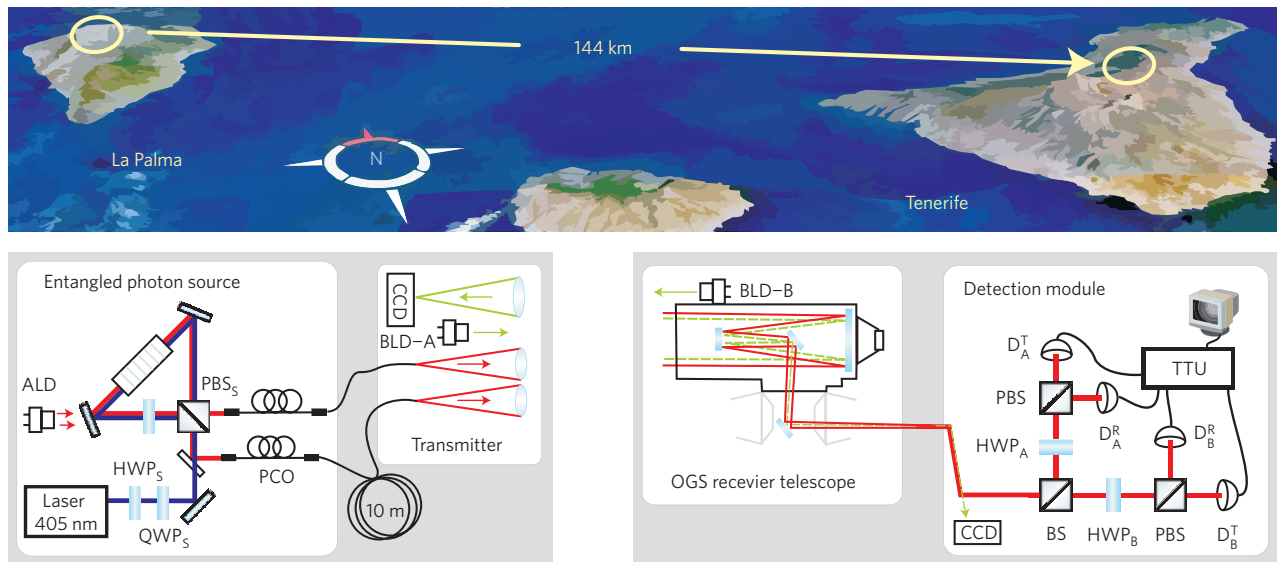


Figure 1 | Satellite image (NASA World Wind) of the Canary Islands of Tenerife and La Palma and overview of the experimental scheme. At La Palma, a Sagnac down-conversion source²⁴ created narrow-band entangled photon pairs. Manual polarization controllers (PCO) and an auxiliary laser diode (ALD) were used for polarization alignment. The photon pairs were transmitted by a pair of telescopes mounted on a rotatable platform to the 144 km distant receiver (OGS) on Tenerife. The transmitter telescope platform was actively locked to a 532 nm beacon laser diode (BLD-B) attached to the OGS. Likewise, the OGS receiver telescope tracked the virtual position of a 532 nm beacon laser attached to the transmitter (BLD-A). At the OGS, the overlapping photon beams were collected and guided to the detection module by a system of mirrors. This module consisted of a 50/50 beamsplitter cube (BS), and two polarization analysers (A, B). Each of these analysers was formed by one half-wave plate (HWP_A , HWP_B), a polarizing beamsplitter cube (PBS) and two single-photon avalanche photodiodes (D_A^T , D_A^R , D_B^T , D_B^R) placed in the transmitted (T) and the reflected (R) output port of the respective PBS.

wander did not divert the beam off the detectors, the photons were collimated ($f = 400$ mm) before they were guided to a polarization analysis module. Here, a symmetric 50/50 beam splitter directed impinging photons randomly to one of two polarization analysers (A, B), each consisting of a half-wave plate (HWP_A , HWP_B), mounted in a motorized rotation stage, and a polarizing beam splitter (PBS). The polarized light was then refocused ($f = 50$ mm) onto four single-photon avalanche diodes (SPADs). A time-stamping unit recorded clicks in the four SPADs and encoded and stored their respective channel number and arrival time relative to a common internal clock with 156 ps resolution. Figure 2a shows the cross-correlations of these time-stamps for two exemplary measurements. One can clearly identify the two coincidence peaks at ± 50 ns around zero delay, which corresponds to the fibre delay Δt introduced at the transmitter. The average width (full-width at half-maximum) of the coincidence peaks was 560 ps, dominated by the timing jitter of the SPADs. To obtain the number of coincident photons, we summed up the number of correlations in a time window of 1.25 ns centred at the coincidence peak positions (Fig. 2b).

As a witness for the presence of entanglement between the received photons, we tested the CHSH Bell inequality¹⁷:

$$S(\alpha, \beta, \alpha', \beta') = |E(\alpha, \beta) - E(\alpha, \beta')| + |E(\alpha', \beta') + E(\alpha', \beta)| < 2 \quad (2)$$

where $E(\theta_A, \theta_B) = (C_{TT}(\theta_A, \theta_B) + C_{RR}(\theta_A, \theta_B) - C_{TR}(\theta_A, \theta_B) - C_{RT}(\theta_A, \theta_B))/N$ is the normalized correlation value of polarization measurement results on photon pairs. $C_{ab}(\theta_A, \theta_B)$ is the number of coincidences measured between detectors at the (T/R) output ports (Fig. 1) of the polarization analysers A and B set to angles (θ_A, θ_B) and N is the sum of these coincidences. Whenever S exceeds the classical bound $S > 2$, the polarization correlations cannot be explained by local hidden variable models¹⁷. For the maximally entangled state $|\psi^-\rangle$,

quantum theory predicts a value of $S_{QM} = 2\sqrt{2}$ for the settings $(\alpha, \beta, \alpha', \beta') = (0, \pi/8, \pi/4, 3\pi/8)$.

The detection module in our experiment enabled us to directly measure the expectation values $E(\theta_A, \theta_B)$ in equation (2) with four different sets of angles of HWP_A and HWP_B (Table 1). We first aligned the system to obtain a $|\psi^-\rangle$ state at the receiver (see details in the Methods section). For each setting (θ_A, θ_B) , we repeatedly accumulated data for typically 900 s, which eventually amounted to a total of 10,800 s acquired in three consecutive nights. Each detector registered an intrinsic dark count rate of ~ 200 s⁻¹, and additionally background light of ~ 200 s⁻¹. In total, we received an average signal of 2,500 single photons per second and 0.071 photon pairs per second. Even though the final single-photon-to-coincidence ratio at the receiver was just 1.7×10^{-5} , the coincidence signal-to-noise ratio was as high as $\sim 15:1$. Compared with the count rates detected at the source, the single-photon attenuation was 34 dB, of which 2 dB was due to the lower efficiency of the detectors used at the receiver ($\sim 25\%$) compared with those at the source ($\sim 40\%$). The measured total photon-pair loss was 71 dB, of which 3 dB was contributed by the beam splitter in the receiver module. The average net attenuation experienced by single photons along the free-space link was therefore 32 dB and the photon-pair attenuation of 64 dB was exactly twice as large, which results from the fact that a pair of photons has double the extinction of a single photon transiting the path. As we have previously ruled out any other adverse effects such as depolarization or timing jitter, which might occur between photons in independent channels¹⁶, we can compare our results to a scenario with two separate free-space links. A detailed analysis²² of the error sources in our system enabled us to estimate the expected background and multiphoton-pair emissions limited quantum visibility to 94.4%. Combined with the source visibility (99.2%) and the polarization contrast of the detection module (99.5%), the upper bound for the overall system visibility was $V_{tot} = 93.2\%$. As the observed CHSH Bell parameter is connected to the set-up visibility through $S_{max} = V_{tot} \times S_{QM}$, this implies a maximum achievable Bell parameter of $S_{max} = 2.636$.

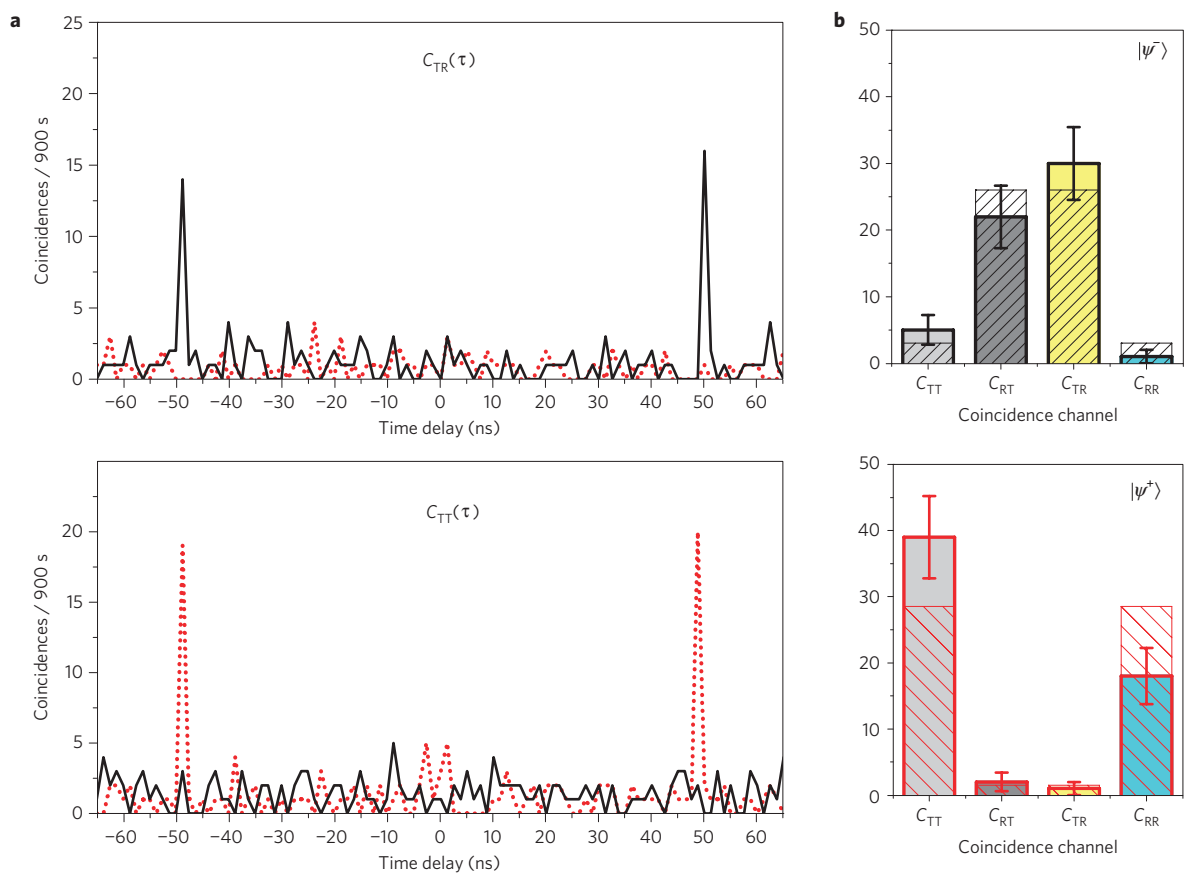


Figure 2 | Coincidence histograms and the respective accumulated coincidence events for measurements on two different Bell states. a, Timing distribution of two out of four coincidence channels. $C_{TR}(\tau)$ and $C_{TT}(\tau)$ between detectors $D_A^I-D_B^R$ (top) and $D_A^I-D_B^I$ (bottom) for a $|\psi^-\rangle$ state (black line) and a $|\psi^+\rangle$ state (dotted red line). The analyser wave plates HWP_A and HWP_B in the detection module were set to $(\pi/8, \pi/8)$. For each detector combination, there are two coincidence peaks at ± 50 ns that can be clearly distinguished from the accidental background. **b,** Total coincidence counts and Poissonian standard deviations for all four relevant coincidence channels integrated over a 1.25 ns time window centred at the peak positions. They show distinct $|\psi^-\rangle$ (top) and $|\psi^+\rangle$ (bottom) signatures. The coloured columns show the joint results of the four corresponding detector combinations necessary to fully characterize the state. The hatched columns show the ideal expectation. The accumulation time for each measurement was 900 s.

Table 1 Experimental polarization correlations $E(\theta_a, \theta_b)$ for the CHSH inequality.		
(θ_a, θ_b)	$E(\theta_a, \theta_b)$	$\Delta E(\theta_a, \theta_b)$
$(0, \pi/8)$	-0.604	0.059
$(\pi/4, \pi/8)$	0.672	0.055
$(0, 3\pi/8)$	0.638	0.056
$(\pi/4, 3\pi/8)$	0.697	0.058

The total integration time was 10,800 s. The standard deviations $\Delta E(\theta_a, \theta_b)$ were calculated assuming Poissonian photon count statistics.

The accumulated coincident events for the different detector pairs yielded the correlation values shown in Table 1. According to equation (2), we measured a CHSH Bell parameter S_{exp} of:

$$S_{\text{exp}} = 2.612 \pm 0.114$$

which is in excellent agreement with our estimate S_{max} . Our result violates the CHSH inequality by 5.4 standard deviations and convincingly proves the successful transmission of entanglement. The fact that S_{exp} is so close to S_{max} shows that the fidelity between the transmitted and received entangled states was essentially noise-limited. Therefore, the entanglement was not

affected by decoherence, even though the photons were subject to extreme attenuation that was dominated by turbulent atmospheric fluctuation¹⁶. Note that our set-up contains some of the basic building blocks of a quantum communication system. However, the set-up could not be used to carry out an actual quantum key distribution experiment²⁵ owing to the lack of a second independent analyser module. In addition, a fully-fledged implementation would have required classical post-processing protocols, that is, error correction and privacy amplification. Nevertheless, from the measured S_{exp} , we can infer a qubit error ratio of $3.85 \pm 2.2\%$. Further note that the photon-pair creation rate of $\sim 10^6$ pairs per second at the source is necessary⁶ for the long-distance distribution of quantum entanglement in the demonstrated high-attenuation regime. The compact entangled photon source, being pumped by a low-power diode laser, can readily be integrated into a satellite-borne photonic terminal, which was previously^{14,16} not the case. We expect that this will enable fundamental tests of the laws of quantum mechanics on a global scale²⁶.

Methods

Before measuring the polarization correlations for equation (2), we had to establish a common polarization reference frame between the individual transmitters and the receiver and to adjust the phase φ of the quantum state in equation (1) such that the detected coincidence signature was consistent with one of the desired Bell states. For the polarization compensation, we used an auxiliary 808 nm laser diode, which was directed at the entangled source such that linearly polarized light was

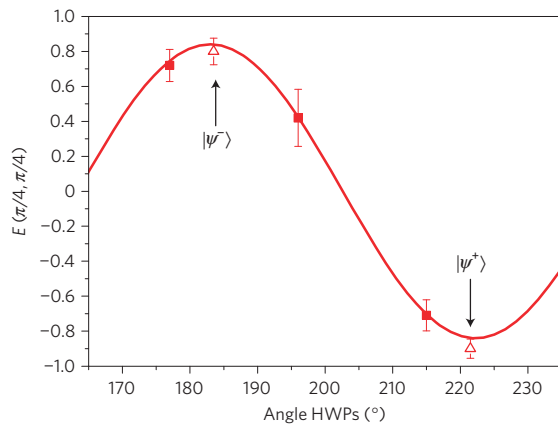


Figure 3 | Scan of the phase ϕ of the entangled two-photon state in one measurement night. We measured the visibility of the entangled states in the $|\pm\rangle$ basis for three settings of the wave plates controlling the pump laser (squares). After fitting a cosine function ($V_0 = 84 \pm 2.4\%$) to these data points, we were able to adjust the source to emit either $|\psi^-\rangle$ or $|\psi^+\rangle$ states. We prepared these states and observed, in this example, a visibility of $V_{\pm} = 80 \pm 7.6\%$ for the $|\psi^-\rangle$ and $V_{\pm} = 90 \pm 5.5\%$ for the $|\psi^+\rangle$ state (triangles). The error bars were determined assuming Poissonian photon count statistics.

coupled into the fibres at a well-defined single-photon level (Fig. 1). We set HWP_A and HWP_B in the detection module to $(0^\circ, 0^\circ)$, measuring in the $|H/V\rangle$ basis, and manually adjusted the fibre polarization controllers at the source to maximize the single-photon polarization visibility $V_{H/V}$ in the remote detectors. The achieved visibility in this basis was typically 95%.

Once the linear polarization was set, the auxiliary laser was switched off and ϕ was tuned using entangled photons. The visibility V_{\pm} in the $|\pm\rangle$ basis depends on ϕ as $V_{\pm} = V_0 \cos(\phi)$. To determine the relation between ϕ and the wave plates controlling the pump laser in the source (HWP_S, QWP_S), we set HWP_A and HWP_B in the detection module to $(\pi/8, \pi/8)$, measured V_{\pm} for three different settings of HWP_S and QWP_S in the equatorial plane of the Poincaré sphere that represents the pump laser polarization (see refs 23, 24) and then numerically fitted a cosine function to the obtained data points. The fitted two-photon fringes in Fig. 3 correspond to a visibility of $V_0 = 84 \pm 2.4\%$. From this fit, we deduced the wave-plate settings to obtain either a $|\psi^-\rangle$ or a $|\psi^+\rangle$ state in the detection module. This procedure was repeated each night at the beginning of a measurement.

We obtained the coincidence measurements in Fig. 2 after we prepared $|\psi^-\rangle$ and $|\psi^+\rangle$ states using this method. The measured respective visibility of $V_{\pm} = 80 \pm 7.6\%$ and $V_{\pm} = 90 \pm 5.5\%$ at the receiver is depicted in Fig. 3. As the linear polarization of the individual photons could be adjusted arbitrarily at the source, we were thus able to prepare, transmit and distinguish any of the four Bell states at the receiver.

Received 7 February 2008; accepted 25 March 2009;
published online 3 May 2009

References

- Gisin, N. & Thew, R. Quantum communication. *Nature Photon.* **1**, 165–171 (2007).
- Takesue, H. *et al.* Quantum key distribution over a 40-dB channel loss using superconducting single-photon detectors. *Nature Photon.* **1**, 343–348 (2007).
- Hübel, H. *et al.* A high-fidelity transmission of polarization encoded qubits from an entangled source over 100 km of fiber. *Opt. Express* **15**, 7853–7862 (2007).
- Honjo, T. *et al.* Long-distance distribution of time-bin entangled photon pairs over 100 km using frequency up-conversion detectors. *Opt. Express* **15**, 13957–13964 (2007).
- Zhang, Q. *et al.* Distribution of time-energy entanglement over 100 km fiber using superconducting single-photon detectors. *Opt. Express* **16**, 5776–5781 (2008).
- Aspelmeyer, M., Jennewein, T., Pfennigbauer, M., Leeb, W. & Zeilinger, A. Long-distance quantum communication with entangled photons using satellites. *IEEE J. Sel. Top. Quant. Electron.* **9**, 1541–1551 (2003).
- Kurtsiefer, C. *et al.* A step towards global quantum key distribution. *Nature* **419**, 450 (2002).
- Buttler, W. T. *et al.* Daylight quantum key distribution over 1.6 km. *Phys. Rev. Lett.* **84**, 5652–5655 (2000).
- Rarity, J. G., Tapster, P. R. & Gorman, P. M. Secure free-space key exchange to 1.9 km and beyond. *J. Mod. Opt.* **48**, 1887–1901 (2001).
- Bienfang, J. *et al.* Quantum key distribution with 1.25 Gbps clock synchronization. *Opt. Express* **12**, 2011–2016 (2004).
- Schmitt-Manderbach, T. *et al.* Experimental demonstration of free-space decoy-state quantum key distribution over 144 km. *Phys. Rev. Lett.* **98**, 10504 (2007).
- Aspelmeyer, M. *et al.* Long-distance free-space distribution of quantum entanglement. *Science* **301**, 621–623 (2003).
- Peng, C. *et al.* Experimental free-space distribution of entangled photon pairs over 13 km: Towards satellite-based global quantum communication. *Phys. Rev. Lett.* **94**, 150501 (2005).
- Resch, K. *et al.* Distributing entanglement and single photons through an intra-city, free-space quantum channel. *Opt. Express* **13**, 202–209 (2005).
- Marcikic, I., Lamas-Linares, A. & Kurtsiefer, C. Free-space quantum key distribution with entangled photons. *Appl. Phys. Lett.* **89**, 101122 (2006).
- Ursin, R. *et al.* Entanglement-based quantum communication over 144 km. *Nature Phys.* **3**, 481–486 (2007).
- Clauser, J., Horne, M., Shimony, A. & Holt, R. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880–884 (1969).
- Mattle, K., Weinfurter, H., Kwiat, P. G. & Zeilinger, A. Dense coding in experimental quantum communication. *Phys. Rev. Lett.* **76**, 4656–4659 (1996).
- Pan, J.-W., Gasparoni, S., Ursin, R., Weihs, G. & Zeilinger, A. Experimental entanglement purification of arbitrary unknown states. *Nature* **423**, 417–422 (2003).
- Bouwmeester, D., Ekert, A. & Zeilinger, A. *The Physics of Quantum Information: Quantum Cryptography, Quantum Teleportation, Quantum Computation* (Springer, 2001).
- Boileau, J., Gottesman, D., Laflamme, R., Poulin, D. & Spekkens, R. Robust polarization-based quantum key distribution over a collective-noise channel. *Phys. Rev. Lett.* **92**, 17901 (2004).
- Ma, X., Fung, C. & Lo, H. Quantum key distribution with entangled photon sources. *Phys. Rev. A* **76**, 12307 (2007).
- Kim, T., Fiorentino, M. & Wong, F. N. C. Phase-stable source of polarization-entangled photons using a polarization Sagnac interferometer. *Phys. Rev. A* **73**, 12316 (2006).
- Fedrizzi, A., Herbst, T., Poppe, A., Jennewein, T. & Zeilinger, A. A wavelength-tunable fiber-coupled source of narrowband entangled photons. *Opt. Express* **15**, 15377–15386 (2007).
- Bennett, C. H., Brassard, G. & Mermin, N. D. Quantum cryptography without Bell's theorem. *Phys. Rev. Lett.* **68**, 557–559 (1992).
- Armengol, J. M. P. *et al.* Quantum communications at ESA: Towards a space experiment on the ISS. *Acta Astronaut.* **63**, 165–178 (2008).

Acknowledgements

We are grateful to H. Weinfurter, J. G. Rarity, T. Schmitt-Manderbach, C. Barbieri, F. Sanchez, A. Alonso, J. Perdigues and Z. Sodnik, T. Augusteijn and the staff of the Nordic Optical Telescope in La Palma for their support at the trial sites. This work was supported by ESA under the General Studies Programme (No. 18805/04/NL/HE), the European Commission through Project QAP (No. 015846), the DTO-Funded US Army Research Office, the Austrian Science Foundation (FWF) under project number SFB1520 and the ASAP-Programme of the Austrian Space Agency (FFG).

Additional information

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>. Correspondence and requests for materials should be addressed to A.F. or A.Z.

Cooper-pair-mediated coherence between two normal metals

P. Cadden-Zimansky, J. Wei and V. Chandrasekhar*

Two electrons bound in a singlet state have long provided a conceptual and pedagogical framework for understanding the non-local nature of entangled quantum objects. As bound singlet electrons separated by a coherence length of up to several hundred nanometres occur naturally in conventional Bardeen-Cooper-Schrieffer superconductors in the form of Cooper pairs, recent theoretical investigations^{1–9} have focused on whether electrons in spatially separated normal-metal probes placed within a coherence length of each other on a superconductor can be quantum mechanically coupled by the singlet pairs. This coupling is predicted to occur through the non-local processes of elastic cotunnelling and crossed Andreev reflection. In crossed Andreev reflection, the constituent electrons of a Cooper pair are sent into different normal probes while retaining their mutual coherence. In elastic cotunnelling, a sub-gap electron approaching the superconductor from one normal probe undergoes coherent, long-range tunnelling to the second probe that is mediated by the Cooper pairs in the condensate. Here, we present experimental evidence for coherent, non-local coupling between electrons in two normal metals linked by a superconductor. The coupling is observed in non-local resistance oscillations that are periodic in an externally applied magnetic flux.

Three key predicted signatures of the elastic cotunnelling and crossed Andreev reflection (CAR) processes, shown schematically in Fig. 1, are (1) the coupling created between the normal probes is non-local—no current need be sent between them, (2) the resultant non-local signals decay rapidly as the probe separation is increased, exponentially over a superconducting coherence length ξ_S or faster^{4,5,7} and (3) the processes create quantum phase coherence between the two probes. Previous experiments looking for evidence of elastic cotunnelling and CAR have used normal–superconductor–normal and ferromagnet–superconductor–ferromagnet devices. A current is sent across one normal–superconductor or ferromagnet–superconductor interface and non-local voltages are measured on the other normal or ferromagnet probe located less than a coherence length from the interface^{10–12}. Although the measured non-local signals exhibit behaviour consistent with predictions (1) and (2), the coherent nature of the non-local signals has not been demonstrated. To establish the presence of non-local coherence, we have attached a hybrid normal–superconducting loop known as an Andreev interferometer^{13–15} to one of the normal probes in a normal–superconductor–normal device. Modifying the phase of the electrons in the normal part of the Andreev interferometer by threading a magnetic flux through the loop leads to oscillations in the resistance of the loop with a period equal to the superconducting flux quantum $\Phi_0 = h/2e$. Periodic oscillations can also be seen in the non-local resistance measured using normal probes placed off the interferometer, but coupled to it by a superconducting

wire of length comparable to ξ_S . These oscillations are the result of CAR/elastic-cotunnelling processes that couple the non-local probes to flux-dependent quasiparticle currents in the normal arm of the Andreev interferometer.

Figure 2a shows a scanning electron micrograph of a hybrid loop device. Although three devices with slightly different geometries were measured (see Supplementary Fig. S1), all show consistent results and here we concentrate on a single device with the geometry shown in Fig. 2a. At the centre of the device is a square loop $\sim 1.7 \mu\text{m}$ on each side formed from lines of 80 nm width. The loop is composed of 80-nm-thick Al lines on three sides and a 50-nm-thick Au wire on one side. Two superconducting leads attached to the loop on the left and two normal leads on the right serve as current and voltage probes for four-terminal measurements of the loop. Extra normal-metal voltage probes are placed on the right side of the loop at the top and bottom corners, and non-locally, just off the loop. The non-local probes are coupled to the normal part of the loop through a superconducting section of length 110 nm, comparable to the superconducting coherence length of Al in the dirty limit¹⁶.

Figure 2b and the inset of Fig. 2a show data for the resistance of the loop using the current and voltage lead configuration marked in Fig. 2a. This is the local measurement configuration. As a function of temperature, the resistance undergoes a sharp drop at 1.2 K as the Al sections become superconducting, and the remaining resistance of the normal side gradually decreases as the loop is cooled further owing to the superconducting proximity effect. At 70 mK, the resistance begins to decline more sharply, although even at 14 mK, the base temperature of our refrigerator, the loop has a residual resistance above 1Ω . The energy scale at which the superconducting proximity effect encompasses the normal section of the loop is given by the Thouless energy $E_c = \hbar D/L^2$, where L is the length of the normal section and D is the electron diffusion coefficient. Using measurements of a simultaneously fabricated Au wire to determine $D = 110 \text{ cm}^2 \text{ s}^{-1}$ gives $E_c = 2.5 \mu\text{eV}$ for the $L = 1.7 \mu\text{m}$ normal side of the loop. This energy corresponds to a temperature of 30 mK, indicating that the low-temperature resistance drop of the loop is due to its transitioning into the correlation regime, where the time it takes electrons to traverse the normal-metal section is less than the decoherence time due to thermal energy. This regime can also be seen in the differential resistance measurements of the loop at 14 mK, where a sharp increase can be seen over a $\pm 1 \mu\text{A}$ range. Robust coherent behaviour in the hybrid loop such as the large $h/2e$ periodic resistance oscillations shown as an inset to Fig. 2a is observed in this correlation regime only below $\pm 1 \mu\text{A}$ d.c. current and 70 mK.

As shown in Fig. 3a, to measure a phase-coherent non-local signal due to elastic cotunnelling and CAR, a phase-dependent quasiparticle current $I_{qp}(\Phi)$ needs to be injected from one normal probe into the superconductor while a non-local voltage is

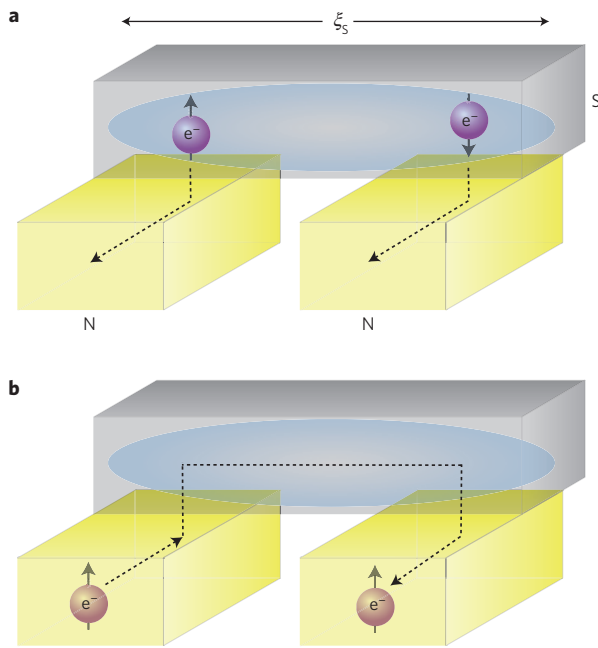


Figure 1 | CAR and elastic cotunnelling. If two normal metals attached to a superconductor are separated by less than the superconducting coherence length ξ_S , the wavefunction of the entangled Cooper pair of two electrons in the superconductor can interact with both normal metals simultaneously. In this geometry, two coherent processes are possible. **a**, CAR, wherein a Cooper pair is broken up into two phase-coherent electrons with one electron entering each of the spatially separated normal metals. **b**, Elastic cotunnelling, wherein an electron in one normal metal undergoes Cooper-pair-mediated tunnelling through the superconductor into the second normal metal. For the same configuration, these processes produce voltages of opposite signs, but both allow phase modulations of electrons in one normal metal to be coherently communicated to the other.

measured on a second probe V_N relative to the superconductor potential V_S . If an observed non-local voltage is due to a phase-coherent process, tuning the phase of $I_{qp}(\Phi)$ will reveal a phase-dependent non-local voltage. $I_{qp}(\Phi)$ is established by embedding the first probe in our interferometer (Fig. 3b) and adding a small d.c. current below $1 \mu\text{A}$ that is modulated by an a.c. measurement current. As discussed below, the small but finite d.c. current creates a non-equilibrium quasiparticle distribution in the normal arm of the loop that is necessary in these devices to establish $I_{qp}(\Phi)$ and the resulting phase-dependent non-local voltages. Similar to the local measurement above, the phase of $I_{qp}(\Phi)$ is tuned by an external flux through the interferometer loop. Although the current path of the local configuration marked in Fig. 2a may be used, any path along or intersecting the normal arm of the interferometer produces a non-equilibrium quasiparticle distribution in the arm, leading to similar qualitative effects. These effects are also independent of any non-equilibrium distribution induced in the non-local voltage probes (see Supplementary Fig. S2). We present results for the current path of Fig. 3b, where the current crosses only the centre of the loop, as it helps to emphasize the non-local nature of the measurement and illuminates the physics of the device.

Using this current configuration, we measure the differential voltage signals (normalized by the a.c. measurement current amplitude) on the non-local probes off the loop (subscript N) as well as the leads on the corner of the loop (subscript C) relative to the superconductor potential V_S . We also compare the signals seen at the top and bottom of the loop (superscripts T and B). Figure 3 exhibits the flux-dependent voltages on these four probes. The signals on the non-local probes, V_N^T and V_N^B , show oscillations

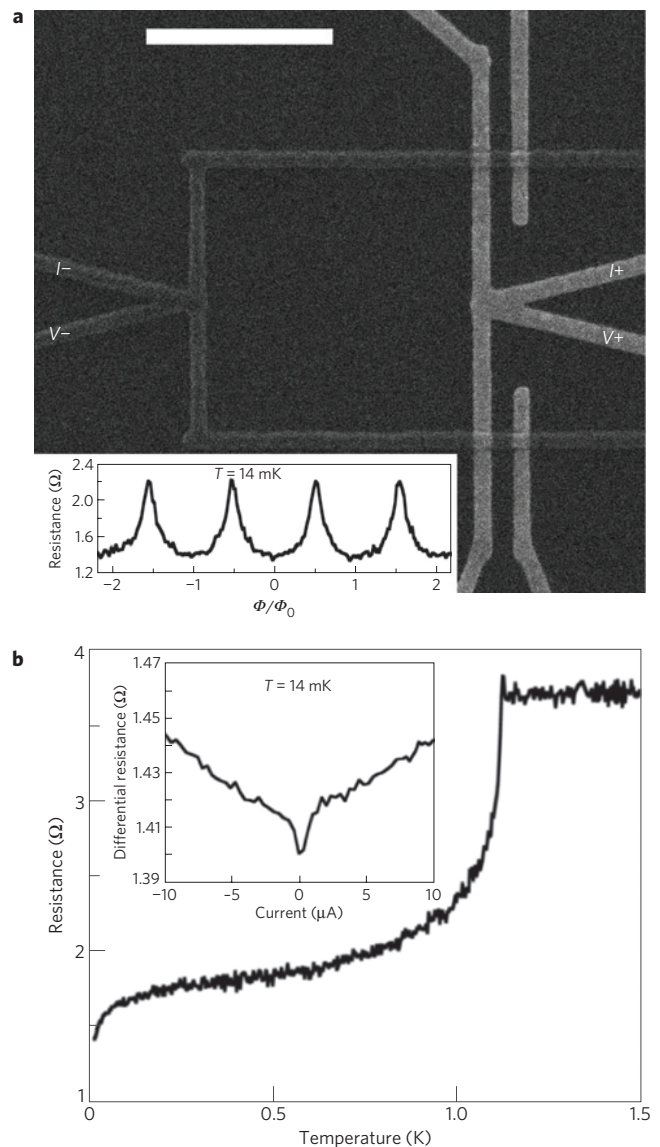


Figure 2 | Local interferometer measurements. **a**, Scanning electron micrograph of a hybrid normal-metal/superconducting loop with non-local normal leads 110 nm to the right of the loop. The scale bar is $1 \mu\text{m}$. The bright wires are normal Au leads and the darker wires superconducting Al. The labelled current/voltage configuration is used to measure the local resistance of the loop. In addition to a small a.c. measurement current, a d.c. current can be applied to measure differential resistance at different d.c. biases. An external field can also be applied to thread a flux through the loop with the convention that a positive flux corresponds to a flux pointing into the page. The inset shows the resistance oscillations of the loop as a function of flux, which have a $\Phi_0 = h/2e$ period. **b**, Local resistance of the hybrid loop as a function of temperature. The Al section of the loop undergoes a superconducting transition at 1.2 K ; the remaining normal-arm resistance gradually decreases as the temperature is lowered owing to an increased superconducting proximity effect. At temperatures below 70 mK , the residual resistance starts to sharply decrease, although the normal section retains a finite resistance at the base temperature of our dilution refrigerator. Inset: The differential resistance of the loop at 14 mK as a function of current. The sharp decrease in the temperature-dependent resistance at low temperature is mirrored by a sharp change in the differential resistance for d.c. currents at $\sim 1 \mu\text{A}$. Robust phase-coherent behaviour, such as shown in the inset of **a**, is observable only in this low-current, low-temperature regime.

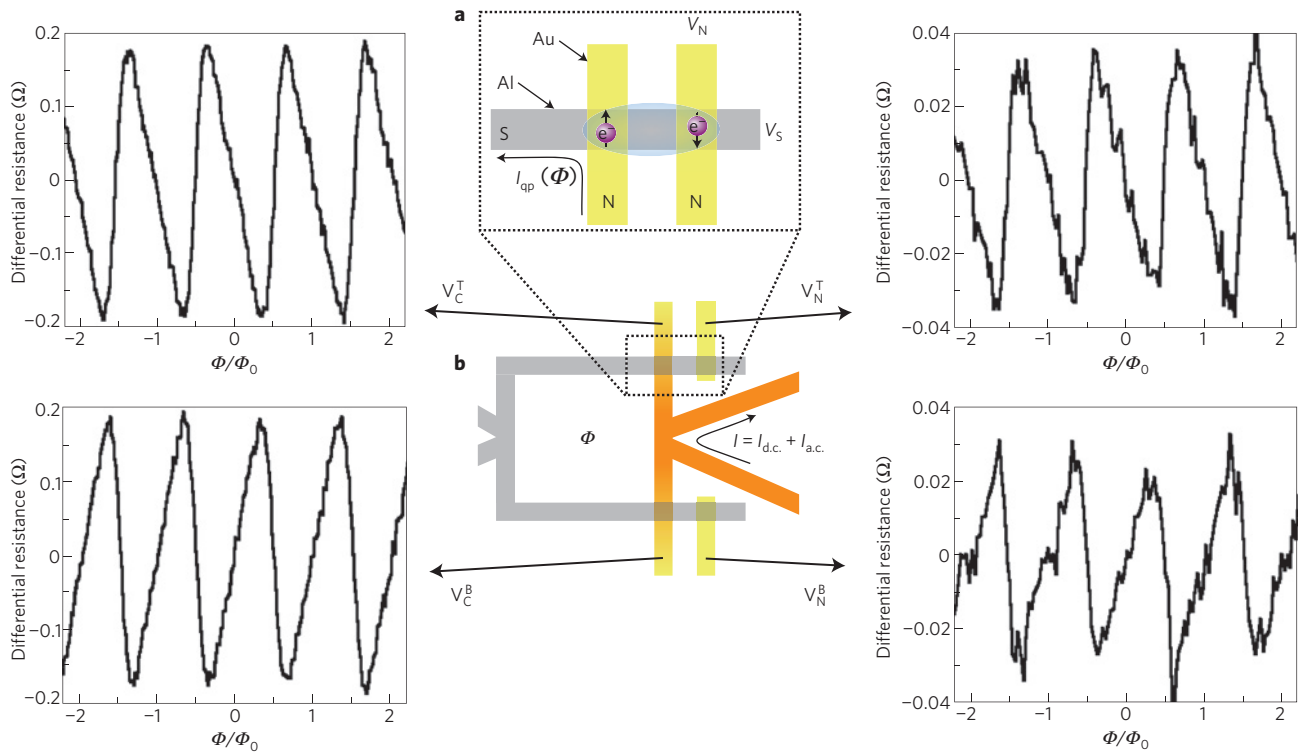


Figure 3 | Coherent non-local signals. **a**, Schemata of the configuration used to measure non-local Cooper-pair-mediated phase coherence. A phase-dependent quasiparticle current $I_{qp}(\Phi)$ is sent along a normal-metal probe into a superconductor. The elastic cotunnelling and CAR processes couple the voltages created by this current into a second normal probe V_N located $\sim \xi_S$ from the first, although no current is sent between the two normal probes. The voltage V_N is measured relative to the superconducting condensate potential V_S as the phase of $I_{qp}(\Phi)$ is varied. **b**, To create $I_{qp}(\Phi)$, the first probe is embedded in an Andreev interferometer through which a 250 nA d.c. current is driven to create a non-equilibrium quasiparticle distribution in the normal section of the loop. The resultant differential voltages, normalized by a low-frequency (<100 Hz) a.c. measurement current added to the d.c. current, are measured on both the corner probes on the loop (subscript C) and non-local probes off the loop (subscript N). These voltages are modulated by an external flux through the loop that can tune $I_{qp}(\Phi)$. Taken at 14 mK, the oscillating signals seen on the probes at the corners of the loop are picked up on the non-local probes through elastic cotunnelling and CAR. The signals are greatly attenuated on the non-local probes, consistent with the expected distance dependence of these coherent processes. For a given flux, the voltages on the leads at the top and bottom of the loop (superscripts T and B) have opposite polarities.

with the Φ_0 flux period dictated by the geometry of the loop. Thus, the quasiparticles in the non-local probes off the loop are being coupled to the phase-coherent processes on the loop, consistent with the predicted elastic cotunnelling and CAR effects. In addition to the coherent nature of these non-local signals, the decay of this signal along the superconductor confirms that they are due to elastic cotunnelling and CAR. The oscillations on the V_C^T and V_C^B corner leads are identical in behaviour to the oscillations on their respective non-local leads, but with an amplitude six times as large. The strong attenuation of the signals along the 110 nm from the corner to non-local leads is consistent with an expected dependence on the superconducting coherence length ξ_S for elastic cotunnelling and CAR processes^{4,5,7}.

As we noted above, the non-local signals shown in Fig. 3 arise from a phase-dependent quasiparticle current $I_{qp}(\Phi)$ in the normal arm of the loop. The origin of $I_{qp}(\Phi)$ can be understood by examining in detail the effect of the non-equilibrium quasiparticle distribution induced by the d.c. current. Within the framework of the quasiclassical theory of superconductivity (see the Methods section), the current between the two normal–superconductor interfaces can be expressed as¹⁷

$$j(R, T) = eN_0 D \int dE (M_{33} \partial_R h_T + Q h_L + M_{03} \partial_R h_L) \quad (1)$$

Here, N_0 is the quasiparticle density of states at the Fermi energy, T is the temperature, M_{33} and M_{03} are normalized diffusion

coefficients that depend on the energy E and spatial coordinates R and Q is the energy-dependent spectral supercurrent. h_T and h_L are space-dependent quasiparticle distribution functions that have the equilibrium values

$$h_{L,T} = \frac{1}{2} \left[\tanh\left(\frac{E + eV}{2k_B T}\right) \pm \tanh\left(\frac{E - eV}{2k_B T}\right) \right] \quad (2)$$

at a reservoir (either superconducting or normal) with an applied voltage V . In particular, at the superconducting reservoirs at the normal–superconductor interface, where we assume V is 0 with no loss of generality, $h_T = 0$ and $h_L = \tanh(E/2k_B T)$.

At low temperatures, where the quasiparticle inelastic scattering length is very long, application of a d.c. current as shown in Fig. 3b will result in non-equilibrium space-dependent distribution functions h_T and h_L in the normal arm of the interferometer. As has already been demonstrated experimentally in superconductor–normal–superconductor junctions, this can affect the total supercurrent through the second term in the brackets in equation (1), even leading to reversal in sign of the critical current, the so-called tunable π -junction^{18–20}. In our sample, the application of the d.c. current results in the value of h_L at the midpoint of the normal arm of the Andreev interferometer being different from its value at either normal–superconductor interface. As the spectral supercurrent Q is conserved, the total supercurrent, given by the

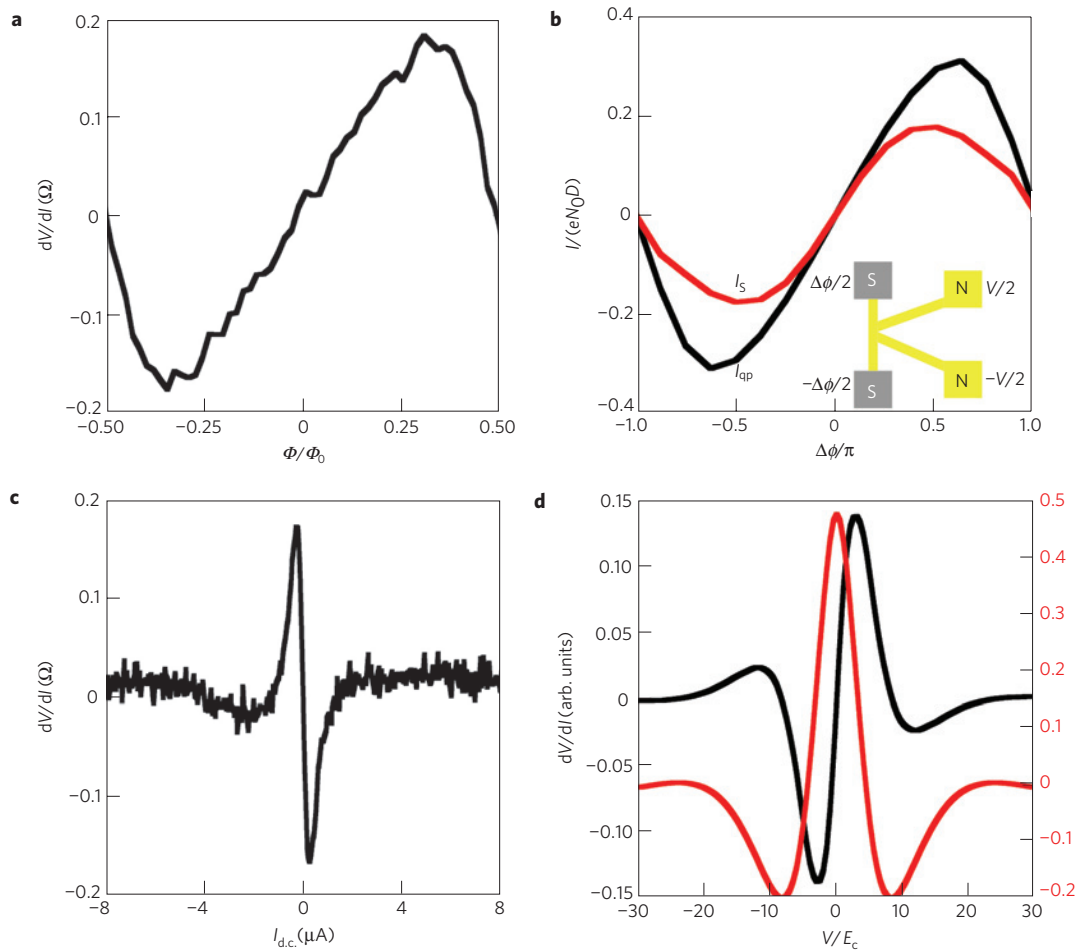


Figure 4 | Comparison with quasiclassical simulations. **a**, Single differential voltage oscillation on the V_C^B probe in Fig. 3 as a function of flux through the interferometer. **b**, Supercurrent and quasiparticle current in the normal arm as a function of the phase difference $\Delta\phi$ between the two superconducting interfaces, calculated by solving the quasiclassical equations of superconductivity for diffusive systems for the model system shown in the inset. The geometry is a normal-metal cross connected to two superconducting reservoirs and two normal reservoirs, with a voltage difference V applied between the normal reservoirs. The calculation in this figure is for $eV = 3E_c$. **c**, Differential resistance measured on the V_C^T probe as a function of the applied d.c. current with a flux of $\Phi_0/4$ threading the Andreev interferometer. **d**, Supercurrent (red curve) calculated at the midpoint of the normal arm of the interferometer, and differential resistance (black curve) as a function of applied voltage, at a phase difference of $\Delta\phi = \pi/2$, corresponding to a flux of $\Phi_0/4$ through the loop.

integral of Qh_L over the energy E also varies as a function of position along the normal arm. The conservation of the total current then implies that there is a compensating quasiparticle current arising from the first and third terms in the brackets in equation (1) (ref. 21).

Simulations based on the quasiclassical theory (see the Methods section) show that h_L varies appreciably only in the vicinity of the normal–superconductor interfaces, so that these quasiparticle currents are also appreciable only in the vicinity of the normal–superconductor interfaces. As the supercurrent is periodic in the applied flux, the induced I_{qp} is also periodic in the applied flux, giving rise to a periodic non-local resistance through CAR/elastic-cotunnelling processes⁷. Figure 4b shows I_{qp} arising from the first term in the brackets in equation (1) near the normal–superconductor interface as a function of the phase difference $\Delta\phi$ between the two normal–superconductor interfaces at a finite voltage $V = 3E_c$ applied between the normal reservoirs. The calculated I_{qp} is periodic and antisymmetric in $\Delta\phi$. Whereas the calculated circulating currents are approximately sinusoidal in $\Delta\phi$, the measured non-local voltages have a slightly more sawtooth dependence on the applied flux; this is due to self-flux effects associated with the finite inductance of the Andreev interferometer^{22–24}.

Furthermore, the fact that the signals measured at the top and bottom corners are of opposite sign is consistent with the fact that I_{qp} goes from the superconductor to the normal arm of the Andreev interferometer at one normal–superconductor interface, whereas it goes from the normal arm into the superconductor at the second normal–superconductor interface.

Figure 4c shows V_C^T as a function of the injected d.c. current with an applied flux of $\Phi_0/4$ through the interferometer, approximately the flux at which the maximum signal is observed. The signal oscillates as a function of the d.c. current, vanishing at zero d.c. current and high d.c. currents. This dependence on the d.c. current is strongly reminiscent of the behaviour of the supercurrent in multi-terminal normal–superconductor devices that has been predicted and observed experimentally^{18–20}. Figure 4d shows the calculated differential resistance as a function of the voltage V (proportional to the applied d.c. current), which reproduces qualitatively the behaviour we see in our experiments. Our experiments show, and the simulations predict, that the quasiparticle current I_{qp} is present only when there is a non-equilibrium distribution in the normal arm of the loop. It is only under non-equilibrium conditions that a conversion of supercurrent to quasiparticle current can occur.

Methods

Experimental. Our devices are fabricated using photo and electron-beam lithography on Si substrates with 300 nm SiO₂ insulating layers. The critical features are patterned in spun methyl methacrylate/polymethyl methacrylate bilayers using a Tescan MIRA electron microscope. The 99.999% pure Au and Al films are deposited in an Edwards thermal evaporator with a base pressure of 3×10^{-7} torr. Before the Au deposition, *in situ* O₂⁺ plasma etching is carried out to clean the substrate, and before the Al deposition *in situ* Ar⁺ plasma etching is carried out to ensure clean interfaces between the two materials. After the Al deposition, the devices are loaded into an Oxford dilution refrigerator and cooled to 77 K within a few hours to prevent degradation of the interfaces. Measurements of our interfaces show a barrier resistivity of $1.9 \mu\text{m}^2 \Omega$, comparable to the Sharvin resistivity.

Measurements were carried out using PAR 124 lock-in amplifiers with modified Adler–Jackson resistance bridges²⁵. The bridges can introduce small, $\lesssim 100$ mΩ, offsets to the data, which are removed by cross-checking the measured resistance using a true four-terminal measurement with a current source and lock-in amplifier. Low-frequency (< 100 Hz) a.c. measurement currents at two different frequencies and amplitudes of 20 and 100 nA were used for local and non-local measurements respectively, with 5 nA currents used to check against the possibility of effects due to heating. To remove extraneous noise signals, first-stage amplification of the measurement signals occurred in a mu-metal-shielded enclosure attached to the cryostat.

Theoretical The simulations were done by solving the Usadel equations of quasiclassical superconductivity in the so-called θ parametrization for the geometry shown in the inset to Fig. 4b. Using the formalism of ref. 17, these equations can be written in the normal wires as

$$\partial_R j_s(E, R) = 0 \quad (3)$$

and

$$D\partial_R^2\theta - \frac{D}{2}\sinh 2\theta(\partial_R\chi)^2 + 2E\sinh\theta = 0 \quad (4)$$

Here, χ is the complex gauge-invariant phase, and

$$j_s(E, R) = \sinh^2\theta(E, R)\partial_R\chi(E, R)$$

where the spectral supercurrent Q is related to j_s by $Q(E, R) = -\Im(j_s(E, R))$. In the normal wire, $\partial_R Q = 0$. In addition, we have the condition that the spectral electronic current $j(E, R) = M_{33}\partial_R h_T + Qh_L + M_{03}\partial_R h_L$ and the spectral thermal current $j_{th}(E, R) = M_{00}\partial_R h_L + Qh_T + M_{30}\partial_R h_T$ are conserved, that is, $\partial_R j(E, R) = 0$ and $\partial_R j_{th}(E, R) = 0$. Here, the normalized diffusion coefficients M are given by $M_{00,33} = [1 + \cosh\theta \cosh\theta^* \pm \sinh\theta \sinh\theta^* \cosh(2\Im(\chi))]/2$, $M_{03} = [\sinh\theta \sinh\theta^* \sinh(2\Im(\chi))]/2$ and $M_{30} = -M_{03}$.

The solutions to these coupled equations are obtained by first solving equations (3) and (4) for the complex θ and χ for the geometry of the inset to Fig. 4b along all four normal wires of the cross, using a numerical relaxation technique and appropriate boundary conditions. The boundary conditions at a normal reservoir are that θ and χ vanish. The phase difference $\Delta\phi$ is dropped symmetrically between the superconducting reservoirs, being set to $\Delta\phi/2$ at one superconducting reservoir, and $-\Delta\phi/2$ at the other, as shown in the inset to Fig. 4b. The boundary condition for θ at a superconducting reservoir is given by $\theta_{s0} = -i(\pi/2) + (1/2)\ln[(|\Delta| + E)/(|\Delta| - E)]$, if $E < |\Delta|$, and $\theta_{s0} = (1/2)\ln[(E + |\Delta|)/(E - |\Delta|)]$, if $E > |\Delta|$, where Δ is the gap in the superconducting reservoir. The boundary conditions at the node where all four wires meet are that θ and χ are continuous, and that the sum of their derivatives is equal to 0.

After θ and χ are obtained, the kinetic equations $\partial_R j(E, R) = 0$ and $\partial_R j_{th}(E, R) = 0$ are then solved using a numerical relaxation technique, subject to the boundary conditions on h_T and h_L at the reservoirs given by equation (2). We assume the voltage applied to the normal reservoirs is dropped symmetrically between them, as shown in the inset to Fig. 4b. The quasiparticle current shown in Fig. 4b is given by the integral of $M_{33}\partial_R h_T$ over the energy E , that is, the first term in the brackets in equation (1); the supercurrent is given by the second term. (The contribution of the third term is much smaller than that of the first term, but has a similar qualitative dependence.) To calculate the differential resistance dV/dI shown in Fig. 4d, we take the derivative of I_{qp} calculated as above with respect to V . As the voltage V is linearly related to the d.c. current by the resistance of the normal wires connected to the normal reservoirs, and the non-local voltage measured should also be linearly related to the quasiparticle current in the non-local normal arm through elastic-cotunnelling/CAR, this quantity is the same as dV/dI to within a numerical factor.

Received 2 September 2008; accepted 20 March 2009;
published online 26 April 2009

References

- Byers, J. M. & Flatté, M. E. Probing spatial correlations with nanoscale two-contact tunneling. *Phys. Rev. Lett.* **74**, 306–309 (1995).
- Deutscher, G. & Feinberg, D. Coupling superconducting-ferromagnetic point contacts by Andreev reflections. *Appl. Phys. Lett.* **76**, 487–489 (2000).
- Falci, G., Feinberg, D. & Hekking, F. W. J. Correlated tunneling into a superconductor in a multiprobe hybrid structure. *Europhys. Lett.* **54**, 255–261 (2001).
- Feinberg, D. Andreev scattering and cotunneling between two superconductor-normal metal interfaces: The dirty limit. *Euro. Phys. J. B* **36**, 419–422 (2003).
- Chitchev, N. M. Superconducting spin filter. *JETP Lett.* **78**, 230–235 (2003).
- Mélin, R. & Feinberg, D. Sign of the crossed conductances at a ferromagnet/superconductor/ferromagnet double interface. *Phys. Rev. B* **70**, 174509 (2004).
- Brinkman, A. & Golubov, A. A. Crossed Andreev reflection in diffusive contacts: Quasiclassical Keldysh–Usadel formalism. *Phys. Rev. B* **74**, 214512 (2006).
- Morten, J. P., Brataas, A. & Belzig, W. Circuit theory for crossed Andreev reflection and nonlocal conductance. *Appl. Phys. A* **89**, 609–612 (2007).
- Levy Yeyati, A., Bergeret, F. S., Martín-Rodero, A. & Klapwijk, T. M. Entangled Andreev pairs and collective excitations in nanoscale superconductors. *Nature Phys.* **3**, 455–459 (2007).
- Beckmann, D., Weber, H. B. & Löhneysen, H. v. Evidence for crossed Andreev reflection in superconductor-ferromagnet hybrid structures. *Phys. Rev. Lett.* **93**, 197003 (2004).
- Russo, S., Kroug, M., Klapwijk, T. M. & Morpurgo, A. F. Experimental observation of bias-dependent nonlocal Andreev reflection. *Phys. Rev. Lett.* **95**, 027002 (2005).
- Cadden-Zimansky, P. & Chandrasekhar, V. Nonlocal correlations in normal-metal superconducting systems. *Phys. Rev. Lett.* **97**, 237003 (2006).
- Nakano, H. & Takayanagi, H. Quasiparticle interferometer controlled by quantum-correlated Andreev reflection. *Phys. Rev. B* **47**, 7986–7994 (1993).
- de Vegvar, P. G. N., Fulton, T. A., Mallison, W. H. & Miller, R. E. Mesoscopic transport in tunable Andreev interferometers. *Phys. Rev. Lett.* **73**, 1416–1419 (1994).
- Pothier, H., Gueron, S., Esteve, D. & Devoret, M. H. Flux-modulated Andreev current caused by electronic interference. *Phys. Rev. Lett.* **73**, 2488–2491 (1994).
- Fujiki, H. *et al.* Nonlinear resistivity in the mixed state of superconducting aluminum films. *Physica C* **297**, 309–316 (1998).
- Chandrasekhar, V. in *Superconductivity: Vol. 1: Conventional and High Temperature Superconductors* (eds Bennemann, K. H. & Ketterson, J. B.) 279–313 (Springer, 2008).
- Baselmans, J. J. A., Morpurgo, A. F., van Wees, B. J. & Klapwijk, T. M. Reversing the direction of the supercurrent in a controllable Josephson junction. *Nature* **397**, 43–45 (1999).
- Shaikhaidarov, R., Volkov, A. F., Takayanagi, H., Petrashov, V. T. & Delsing, P. Josephson effects in a superconductor normal-metal mesoscopic structure with a dangling superconducting arm. *Phys. Rev. B* **62**, R14 649–652 (2000).
- Huang, J., Pierre, F., Heikkilä, T. T., Wilhelm, F. K. & Birge, N. O. Observation of a controllable π junction in a 3-terminal Josephson device. *Phys. Rev. B* **66**, 020507(R) (2002).
- Kogan, V. R., Pavlovskii, V. V. & Volkov, A. F. Electron–hole imbalances in superconductor/normal-metal mesoscopic structures. *Europhys. Lett.* **59**, 875–881 (2002).
- Büttiker, M. & Klapwijk, T. M. Flux sensitivity of a piecewise normal and superconducting metal loop. *Phys. Rev. B* **33**, 5114–5117 (1986).
- Cayssol, J., Kontos, T. & Montambaux, G. Isolated hybrid normal/superconducting ring in a magnetic flux: From persistent current to Josephson current. *Phys. Rev. B* **67**, 184508 (2003).
- Zou, J. *et al.* Influence of supercurrents on low-temperature thermopower in mesoscopic N/S structures. *J. Low Temp. Phys.* **146**, 193–212 (2007).
- Adler, J. G. & Jackson, J. E. System for observing small nonlinearities in tunnel junctions. *Rev. Sci. Instr.* **37**, 1049–1054 (1966).

Acknowledgements

This research was conducted with support from the National Science Foundation under grant No. DMR-0604601. We thank A. A. Golubov, A. D. Zaikin and M. R. Norman for their discussions.

Additional information

Supplementary information accompanies this paper on www.nature.com/naturephysics. Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>. Correspondence and requests for materials should be addressed to V.C.

Observation of a large-gap topological-insulator class with a single Dirac cone on the surface

Y. Xia^{1,2}, D. Qian^{1,3}, D. Hsieh^{1,2}, L. Wray¹, A. Pal¹, H. Lin⁴, A. Bansil⁴, D. Grauer⁵, Y. S. Hor⁵, R. J. Cava⁵ and M. Z. Hasan^{1,2,6*}

Recent experiments and theories have suggested that strong spin-orbit coupling effects in certain band insulators can give rise to a new phase of quantum matter, the so-called topological insulator, which can show macroscopic quantum-entanglement effects^{1–7}. Such systems feature two-dimensional surface states whose electrodynamic properties are described not by the conventional Maxwell equations but rather by an attached axion field, originally proposed to describe interacting quarks^{8–15}. It has been proposed that a topological insulator² with a single Dirac cone interfaced with a superconductor can form the most elementary unit for performing fault-tolerant quantum computation¹⁴. Here we present an angle-resolved photoemission spectroscopy study that reveals the first observation of such a topological state of matter featuring a single surface Dirac cone realized in the naturally occurring Bi_2Se_3 class of materials. Our results, supported by our theoretical calculations, demonstrate that undoped Bi_2Se_3 can serve as the parent matrix compound for the long-sought topological device where in-plane carrier transport would have a purely quantum topological origin. Our study further suggests that the undoped compound reached via n-to-p doping should show topological transport phenomena even at room temperature.

It has been experimentally shown that spin-orbit coupling can lead to new phases of quantum matter with highly non-trivial collective quantum effects^{4–6}. Two such phases are the quantum spin Hall insulator⁴ and the strong topological insulator^{5–7}, both realized in the vicinity of a Dirac point but yet quite distinct from graphene¹⁶. The strong-topological-insulator phase contains surface states (SSs) with novel electromagnetic properties^{7–15}. It is currently believed that the $\text{Bi}_{1-x}\text{Sb}_x$ insulating alloys realize the only known topological-insulator phase in the vicinity of a three-dimensional Dirac point⁵, which can in principle be used to study topological electromagnetic and interface superconducting properties^{8–10,14}. However, a particular challenge for the topological-insulator $\text{Bi}_{1-x}\text{Sb}_x$ system is that the bulk gap is small and the material contains alloying disorder, which makes it difficult to gate for the manipulation and control of charge carriers to realize a device. The topological insulator $\text{Bi}_{1-x}\text{Sb}_x$ features five surface bands, of which only one carries the topological quantum number⁶. Therefore, there is an extensive world-wide search for topological phases in stoichiometric materials with no alloying disorder, with a larger gap and with fewer yet still odd-numbered SSs that may

work as a matrix material to observe a variety of topological quantum phenomena.

The topological-insulator character of BiSb ^{5,6} led us to investigate the alternative Bi-based compounds Bi_2X_3 ($\text{X} = \text{Se}, \text{Te}$). The undoped Bi_2Se_3 is a semiconductor that belongs to the class of thermoelectric materials Bi_2X_3 with a rhombohedral crystal structure (space group $D_{3d}^5(R\bar{3}m)$; refs 17, 18). The unit cell contains five atoms, with quintuple layers ordered in the $\text{Se}(1)\text{--Bi--Se}(2)\text{--Bi--Se}(1)$ sequence. Electrical measurements report that, although the bulk of the material is a moderately large-gap semiconductor, its charge transport properties can vary significantly depending on the sample preparation conditions¹⁹, with a strong tendency to be n-type^{20,21} owing to atomic vacancies or excess selenium. An intrinsic bandgap of approximately 0.35 eV is typically measured in experiments^{22,23}, whereas theoretical calculations estimate the gap to be in the range of 0.24–0.3 eV (refs 20, 24).

It has been shown that spin-orbit coupling can lead to topological effects in materials that determine their spin Hall transport behaviours^{4–7}. Topological quantum properties are directly probed from the nature of the electronic states on the surface by studying the way surface bands connect the material's bulk valence and conduction bands in momentum space^{5–7}. The surface electron behaviour is intimately tied to the number of bulk band inversions that exist in the band structure of a material⁷. The origin of topological \mathbb{Z}_2 order in $\text{Bi}_{1-x}\text{Sb}_x$ is bulk-band inversions at three equivalent L-points^{5,7}, whereas in Bi_2Se_3 only one band is expected to be inverted, making it similar to the case in the two-dimensional quantum spin Hall insulator phase. Therefore, a much simpler surface spectrum is naturally expected in Bi_2Se_3 . All previous experimental studies of Bi_2Se_3 have focused on the material's bulk properties; nothing is known about its SSs. It is this key experimental information that we provide here that, for the first time, enables us to determine its topological quantum class.

The bulk crystal symmetry of Bi_2Se_3 fixes a hexagonal Brillouin zone (BZ) for its (111) surface (Fig. 1d) on which \bar{M} and $\bar{\Gamma}$ are the time-reversal invariant momenta (TRIMs) or the surface Kramers points. We carried out high-momentum-resolution angle-resolved photoemission spectroscopy (ARPES) measurements on the (111) plane of naturally grown Bi_2Se_3 (see the Methods section). The electronic spectral weight distributions observed near the $\bar{\Gamma}$ point are presented in Fig. 1a–c. Within a narrow binding-energy window, a clear V-shaped band pair is observed to approach the Fermi level (E_F). Its dispersion or intensity had no measurable time dependence within the duration of the

¹Joseph Henry Laboratories of Physics, Department of Physics, Princeton University, Princeton, New Jersey 08544, USA, ²Princeton Center for Complex Materials, Princeton University, Princeton, New Jersey 08544, USA, ³Department of Physics, Shanghai Jiao Tong University, Shanghai 200030, China, ⁴Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA, ⁵Department of Chemistry, Princeton University, Princeton, New Jersey 08544, USA, ⁶Princeton Institute for the Science and Technology of Materials, Princeton University, Princeton, New Jersey 08544, USA.

*e-mail: mzh Hasan@Princeton.edu.

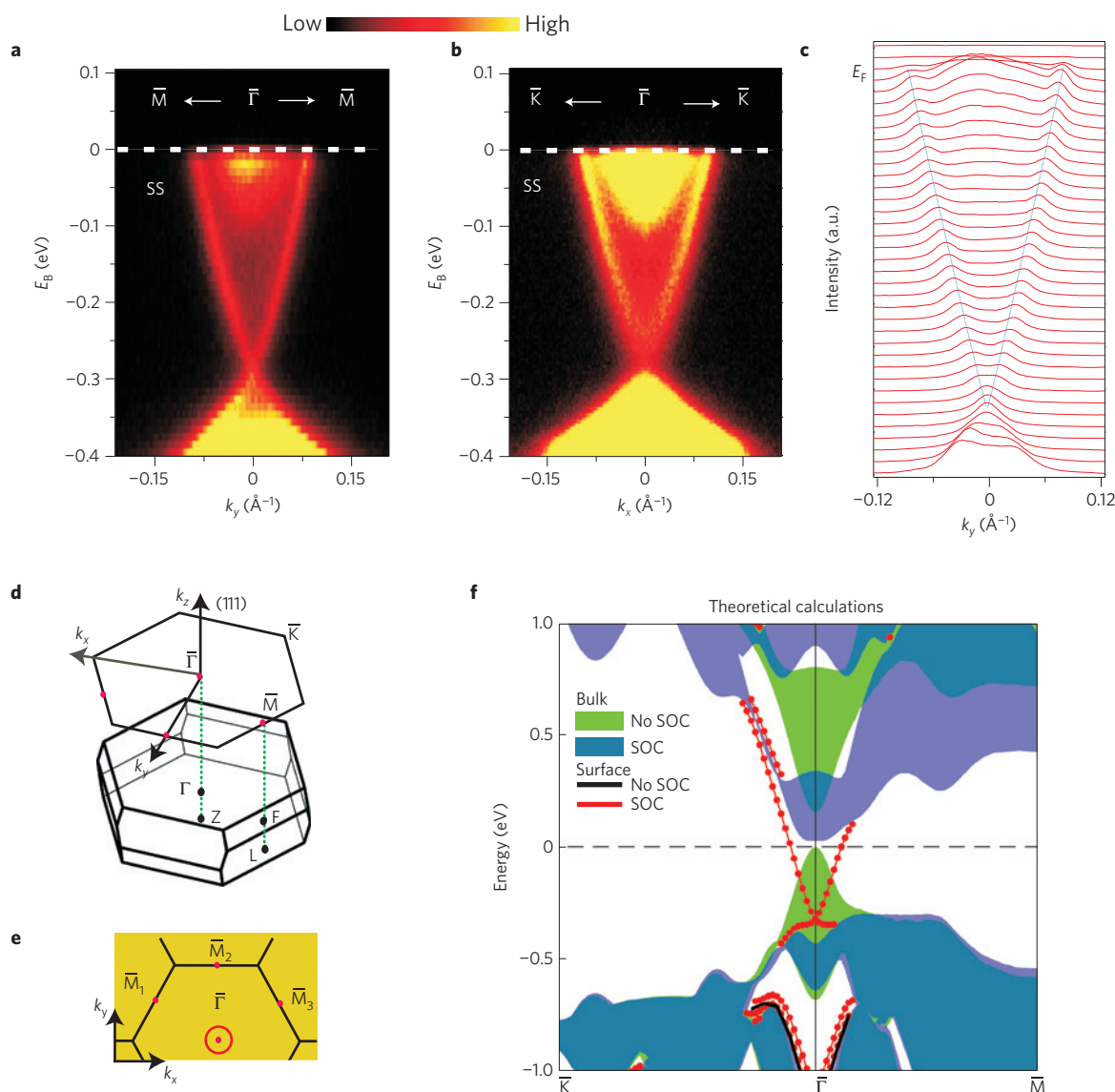


Figure 1 | Strong spin-orbit interaction gives rise to a single SS Dirac cone. Theory (see the Methods section) versus experiments. **a, b**, High-resolution ARPES measurements of surface electronic band dispersion on Bi₂Se₃(111). Electron dispersion data measured with an incident photon energy of 22 eV near the $\bar{\Gamma}$ -point along the $\bar{\Gamma}$ - \bar{M} (**a**) and $\bar{\Gamma}$ - \bar{K} (**b**) momentum-space cuts. **c**, The momentum distribution curves corresponding to **a** suggest that two surface bands converge into a single Dirac point at $\bar{\Gamma}$. The V-shaped pure SS band pair observed in **a-c** is nearly isotropic in the momentum plane, forming a Dirac cone in the energy- k_x - k_y space (where k_x and k_y are in the $\bar{\Gamma}$ - \bar{K} and $\bar{\Gamma}$ - \bar{M} directions, respectively). The U-shaped broad continuum feature inside the V-shaped SS corresponds roughly to the bottom of the conduction band (see the text). **d**, A schematic diagram of the full bulk three-dimensional BZ of Bi₂Se₃ and the two-dimensional BZ of the projected (111) surface. **e**, The surface Fermi surface (FS) of the two-dimensional SSs along the \bar{K} - $\bar{\Gamma}$ - \bar{M} momentum-space cut is a single ring centred at $\bar{\Gamma}$ if the chemical potential is inside the bulk bandgap. The band responsible for this ring is singly degenerate in theory. The TRIMs on the (111) surface BZ are located at $\bar{\Gamma}$ and the three \bar{M} points. The TRIMs are marked by the red dots. In the presence of strong spin-orbit coupling (SOC), the surface band crosses the Fermi level only once between two TRIMs, namely $\bar{\Gamma}$ and \bar{M} ; this ensures the existence of a π Berry phase on the surface. **f**, The corresponding local density approximation (LDA) band structure (see the Methods section). Bulk band projections are represented by the shaded areas. The band-structure topology calculated in the presence of SOC is presented in blue and that without SOC is in green. No pure surface band is observed to lie within the insulating gap in the absence of SOC (black lines) in the theoretical calculation. One pure gapless surface band is observed between $\bar{\Gamma}$ and \bar{M} when SOC is included (red dotted lines).

experiment. The 'V' bands cross E_F at 0.09 \AA^{-1} along $\bar{\Gamma}$ - \bar{M} and at 0.10 \AA^{-1} along $\bar{\Gamma}$ - \bar{K} , and have nearly equal band velocities, approximately $5 \times 10^5 \text{ m s}^{-1}$, along the two directions. A continuum-like manifold of states—a filled U-shaped feature—is observed inside the V-shaped band pair. All of these experimentally observed features can be identified, to first order, by a direct one-to-one comparison with the LDA band calculations. Figure 1f shows the theoretically calculated (see the Methods section) (111)-surface electronic structure of bulk Bi₂Se₃ along the \bar{K} - $\bar{\Gamma}$ - \bar{M} k -space cut.

The calculated band structure with and without SOC are overlaid together for comparison. The bulk band projection continuum on the (111) surface is represented by the shaded areas, blue with SOC and green without SOC. In the bulk, time-reversal symmetry demands $E(\mathbf{k}, \uparrow) = E(-\mathbf{k}, \downarrow)$ whereas space inversion symmetry demands $E(\mathbf{k}, \uparrow) = E(-\mathbf{k}, \uparrow)$. Therefore, all the bulk bands are doubly degenerate. However, because space inversion symmetry is broken at the terminated surface in the experiment, SSs are generally spin-split on the surface by spin-orbit interactions except

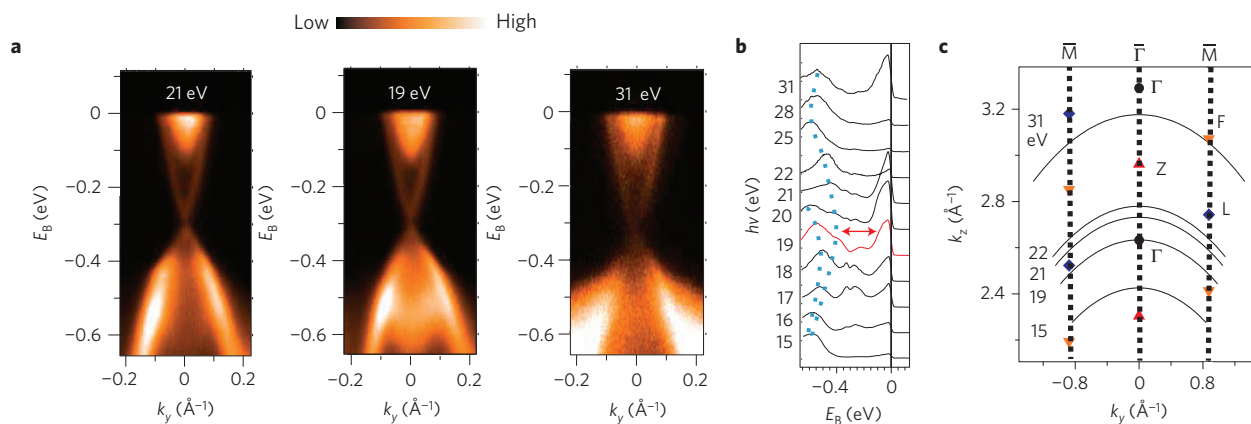


Figure 2 | Transverse-momentum k_z dependence of Dirac bands near $\bar{\Gamma}$. **a**, The energy dispersion data along the $\bar{\Gamma}$ – \bar{M} cut, measured with the photon energy of 21 eV (corresponding to 0.3 k -space length along Γ – $Z \parallel k_z$), 19 eV (Γ) and 31 eV (–0.4 k -space length along Γ – Z of the bulk three-dimensional Γ BZ) are shown. Although the bands below –0.4 eV binding energy show strong k_z dependence, the linearly dispersive Dirac-like bands and the U-shaped broad feature show weaker k_z dispersion. The Dirac point is observed to lie inside the bulk bandgap. A careful look at the individual curves reveals some k_z dependence of the U-shaped continuum (see **b** for details). **b**, The energy distribution curves obtained from the normal-emission spectra measured using 15–31 eV photon energies reveal two dispersive bulk bands below –0.3 eV (blue dotted lines). This is in addition to the two non-dispersive peaks from the Dirac-cone bands inside the gap. The Dirac band intensity is strongly modulated by the photon energy changes due to the matrix-element effects (which is also observed in BiSb; ref. 5). **c**, A k -space map of locations in the bulk three-dimensional BZ scanned by the detector at different photon energies over a theta (θ) range of $\pm 30^\circ$. This map ($k_z, k_y, E_{\text{photon}}$) was used to explore the k_z dependence of the observed bands.

at particular high-symmetry points—the Kramers points on the surface BZ. In our calculations, the SSs (red dotted lines) are doubly degenerate only at $\bar{\Gamma}$ (Fig. 1f). This is generally true for all known spin–orbit-coupled material surfaces such as gold^{25,26} or $\text{Bi}_{1-x}\text{Sb}_x$ (ref. 5). In Bi_2Se_3 , the SSs emerge from the bulk continuum, cross each other at $\bar{\Gamma}$, pass through the Fermi level (E_F) and eventually merge with the bulk conduction-band continuum, ensuring that at least one continuous band-thread traverses the bulk bandgap between a pair of Kramers points. Our calculated result shows that no surface band crosses the Fermi level if SOC is not included in the calculation, and only with the inclusion of the realistic values of SOC (based on atomic Bi) does the calculated spectrum show singly degenerate gapless surface bands that are guaranteed to cross the Fermi level. The calculated band topology with realistic SOC leads to a single ring-like surface FS, which is singly degenerate so long as the chemical potential is inside the bulk bandgap. This topology is consistent with the $Z_2 = -1$ class in the Fu–Kane–Mele classification scheme⁷.

A global agreement between the experimental band structure (Fig. 1a–c) and our theoretical calculation (Fig. 1f) is obtained by considering a rigid shift of the chemical potential by about 200 meV with respect to our calculated band structure (Fig. 1f) of the formula compound Bi_2Se_3 . The experimental sign of this rigid shift (the raised chemical potential) corresponds to an electron doping of the Bi_2Se_3 insulating formula matrix (see Supplementary Information). This is consistent with the fact that naturally grown Bi_2Se_3 semiconductor used in our experiment is n-type, as independently confirmed by our transport measurements. The natural doping of this material, in fact, comes as an advantage in determining the topological class of the corresponding undoped insulator matrix, because we would like to image the SSs not only below the Fermi level but also above it, to examine the way surface bands connect to the bulk conduction band across the gap. A unique determination of the surface band topology of purely insulating $\text{Bi}_{1-x}\text{Sb}_x$ (refs 5, 6) was clarified only on doping with a foreign element, Te. In our experimental data on Bi_2Se_3 , we observe a V-shaped pure SS band to be dispersing towards E_F , which is in good agreement with our calculations. More remarkably, the experimental band velocities are also close to our calculated values. By comparison with calculations combined with a general set of arguments presented above, this

V-shaped band is singly degenerate. Inside this ‘V’ band, an electron-pocket-like U-shaped continuum is observed to be present near the Fermi level. This filled U-shaped broad feature is in close correspondence to the bottom part of the calculated conduction-band continuum (Fig. 1f). Considering the n-type character of the naturally occurring Bi_2Se_3 and by correspondence to our band calculation, we assign the broad feature to correspond roughly to the bottom of the conduction band.

To systematically investigate the nature of all the band features imaged in our data, we have carried out a detailed photon-energy-dependence study, of which selected data sets are presented in Fig. 2a,b. A modulation of incident photon energy enables us to probe the k_z dependence of the bands sampled in an ARPES study (Fig. 2c), allowing for a way to distinguish surface from bulk contributions to a particular photoemission signal⁵. Our photon-energy study did not indicate a strong k_z dispersion of the lowest-lying energy bands on the ‘U’, although the full continuum does have some dispersion (Fig. 2). Some variation of the quasiparticle intensity near E_F is, however, observed owing to the variation of the electron–photon matrix element. In light of the k_z -dependence study (Fig. 2b), if the features above –0.15 eV were purely due to the bulk, we would expect to observe dispersion as k_z moved away from the Γ -point. The lack of strong dispersion yet close one-to-one correspondence to the calculated bulk band structure suggests that the inner electron pocket continuum features are probably a mixture of surface-projected conduction-band states, which also includes some band-bending effects near the surface and the full continuum of bulk conduction-band states sampled from a few layers beneath the surface. Similar behaviour is also observed in the ARPES study of other semiconductors²⁷. In our k_z -dependent study of the bands (Fig. 2b) we also observe two bands dispersing in k_z that have energies below –0.3 eV (blue dotted bands), reflecting the bulk valence bands, in addition to two other non-dispersive features associated with the two sides of the pure SS Dirac bands. The red curve is measured right at the Γ -point, which suggests that the Dirac point lies inside the bulk bandgap. Taking the bottom of the ‘U’ band as the bulk conduction-band minimum, we estimate that a bandgap of about 0.3 eV is realized in the bulk of the undoped material. Our ARPES estimated bandgap is in good agreement with the value deduced from bulk physical measurements²³ and from

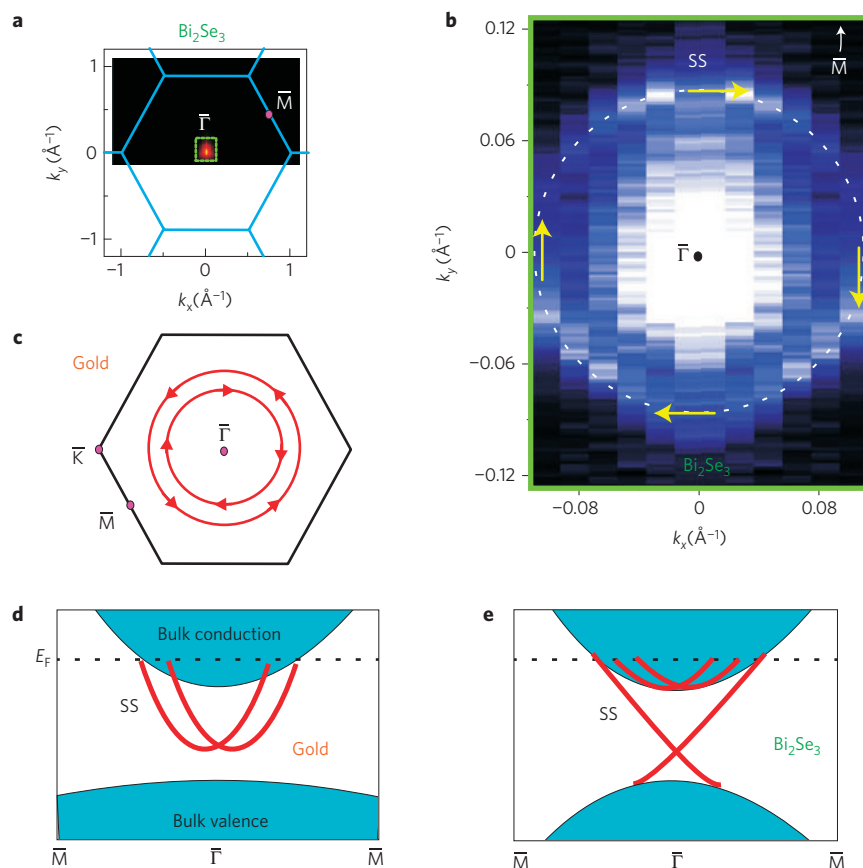


Figure 3 | The topology of the surface Dirac cone FS. **a**, The observed surface FS of Bi_2Se_3 consists of a small electron pocket around the centre of the BZ, $\bar{\Gamma}$. **b**, High-momentum-resolution data around $\bar{\Gamma}$ reveal a single ring formed by the pure SS V-shaped Dirac band. For the naturally occurring Bi_2Se_3 , the spectral intensity in the middle of the ring is due to the presence of the 'U' feature, which roughly images the bottom of the conduction-band continuum (see the text). The observed topology of the pure surface FS of Bi_2Se_3 is different from that of most other spin-orbit materials such as gold ($\text{Au}(111)$). **c**, The $\text{Au}(111)$ surface FS features two rings (each non-degenerate) surrounding the $\bar{\Gamma}$ point. An electron encircling the gold FS carries a Berry phase of zero, characteristic of a trivial band insulator or metal, and can be classified by $Z_2 = +1$ (ref. 7). The single surface FS observed in Bi_2Se_3 is topologically distinct from that of gold. The single non-degenerate surface FS enclosing a Kramers point ($\bar{\Gamma}$) constitutes the key signature of a topological-insulator phase characterized by $Z_2 = -1$. **d**, **e**, Schematic SS topologies in gold and Bi_2Se_3 for direct comparison. In gold, the chemical potential can be continuously tuned to be placed inside the interband gap, making the SSs fully gapped. In Bi_2Se_3 , however, the Dirac structure of the SSs required by the Kramers degeneracy and time-reversal invariance ensures that they remain gapless independent of the location of the chemical potential within the bulk gap. If the chemical potential is placed inside the gap, as it would naturally lie in the purely insulating undoped Bi_2Se_3 , the surface transport would be dominated by the single Dirac fermion, which is of purely topological origin.

other calculations that report the bulk band structure^{20,24}. This suggests that the magnitude of band bending near the surface is not larger than 0.05 eV. We note that in purely insulating Bi_2Se_3 the Fermi level should lie deep inside the bandgap and only pure surface bands will contribute to surface conduction. Therefore, in determining the topological character of the insulating Bi_2Se_3 matrix the 'U' feature is not relevant.

We therefore focus on the pure SS part. The complete surface FS map is presented in Fig. 3. Figure 3a presents electron distribution data over the entire two-dimensional (111) surface BZ. All the observed features are centred around $\bar{\Gamma}$. None of the three TRIMs located at \bar{M} are enclosed by any FS, in contrast to what is observed in $\text{Bi}_{1-x}\text{Sb}_x$ (ref. 5). The detailed spectral behaviour around $\bar{\Gamma}$ is shown in Fig. 3b, which was obtained with high momentum resolution. A ring-like feature formed by the outer 'V' pure SS band (a horizontal cross-section of the upper Dirac cone in Fig. 1) surrounds the conduction-band continuum centred at $\bar{\Gamma}$. This ring is singly degenerate from its one-to-one correspondence to band calculation. An electron encircling the surface FS that encloses a TRIM or a Kramers point obtains a geometrical quantum phase (Berry phase) of $\pi \bmod 2\pi$ in its wavefunction⁷. Therefore, if the

chemical potential (Fermi level) lies inside the bandgap, as it should in purely insulating Bi_2Se_3 , its surface must carry a global $\pi \bmod 2\pi$ Berry phase. In most spin-orbit materials, such as gold ($\text{Au}(111)$), it is known that the surface FS consists of two spin-orbit-split rings generated by two singly degenerate parabolic (not Dirac-like) bands that are shifted in momentum space from each other, with both enclosing the $\bar{\Gamma}$ -point^{25,26}. The resulting FS topology leads to a 2π or 0 Berry phase because the phases from the two rings add or cancel. This makes gold-like SSs topologically trivial despite their spin-orbit origin.

Our theoretical calculation supported by our experimental data suggests that in insulating Bi_2Se_3 there exists a singly degenerate surface FS which encloses only one Kramers point on the surface Brillouin zone. This provides evidence that insulating Bi_2Se_3 belongs to the $Z_2 = -1$ topological class in the Fu-Kane-Mele topological classification scheme for band insulators. On the basis of our ARPES data we suggest that it should be possible to obtain the fully undoped compound by chemically hole-doping the naturally occurring Bi_2Se_3 , thereby shifting the chemical potential to lie inside the bulk bandgap. The surface transport of Bi_2Se_3 , prepared as such would therefore be dominated by topological

effects as it possesses only one Dirac fermion that carries the non-trivial Z_2 index. The existence of a large bulk bandgap (0.3 eV) within which the observed Z_2 Dirac fermion state lies suggests the realistic possibility for the observation of topological effects even at room temperature in this material class. Because of the simplest possible topological surface spectrum realized in Bi_2Se_3 , it can be considered as the ‘hydrogen atom’ of strong topological insulators. Its simplest topological surface spectrum would make it possible to observe and study many exotic quantum phenomena predicted in topological field theories, such as the Majorana fermions¹⁴, magnetic monopole image^{9,10} or topological exciton condensates¹⁵, by transport probes.

Methods

Theoretical calculations. The theoretical band calculations were performed with the LAPW method in slab geometry using the WIEN2K package²⁸. The generalized gradient approximation of Perdew, Burke and Ernzerhof²⁹ was used to describe the exchange–correlation potential. SOC was included as a second variational step using scalar-relativistic eigenfunctions as basis after the initial calculation was converged to self-consistency. The surface was simulated by placing a slab of 12 quintuple layers in vacuum. A grid of $21 \times 21 \times 1$ points was used in the calculations, equivalent to 48 k points in the irreducible BZ and 300 k points in the first BZ. To calculate the k_z of the ARPES measurements ($k_z = (1/\hbar)\sqrt{2m(E_{\text{kin}}\cos^2\theta + V_0)}$), an inner potential V_0 of approximately 11.7 eV was used, given by a fit on the ARPES data at normal emission.

Experimental methods. Single crystals of Bi_2Se_3 were grown by melting stoichiometric mixtures of high-purity elemental Bi and Se in a 4-mm-inner-diameter quartz tube. The sample was cooled over a period of two days, from 850 to 650 °C, and then annealed at this temperature for a week. Single crystals were obtained and could be easily cleaved from the boule. High-resolution ARPES measurements were then performed using 17–45 eV photons on beamline 12.0.1 of the Advanced Light Source at the Lawrence Berkeley National Laboratory and beamline 5–4 at the Stanford Synchrotron Radiation Laboratory. The energy and momentum resolutions were 15 meV and 1.5% of the surface BZ respectively using a Scienta analyser. The samples were cleaved *in situ* between 10 and 55 K under pressures of less than 5×10^{-11} torr, resulting in shiny flat surfaces. The surface band quasiparticle signal is stable throughout the entire measurement duration.

Received 26 December 2008; accepted 2 April 2009;
published online 10 May 2009

References

1. Fu, L., Kane, C. L. & Mele, E. J. Topological insulators in three dimensions. *Phys. Rev. Lett.* **98**, 106803 (2007).
2. Moore, J. E. & Balents, L. Topological invariants of time-reversal-invariant band structures. *Phys. Rev. B* **75**, 121306(R) (2007).
3. Zhang, S.-C. Topological states of quantum matter. *Physics* **1**, 6 (2008).
4. König, M. *et al.* Quantum spin Hall insulator state in HgTe quantum wells. *Science* **318**, 766–770 (2007).
5. Hsieh, D. *et al.* A topological Dirac insulator in a quantum spin Hall phase. *Nature* **452**, 970–974 (2008).
6. Hsieh, D. *et al.* Observation of unconventional quantum spin textures in topological insulators. *Science* **323**, 919–922 (2009).
7. Fu, L. & Kane, C. L. Topological insulators with inversion symmetry. *Phys. Rev. B* **76**, 045302 (2007).
8. Wilczek, F. Two applications of axion electrodynamics. *Phys. Rev. Lett.* **58**, 1799–1802 (1987).
9. Franz, M. High-energy physics in a new guise. *Physics* **1**, 36 (2008).
10. Qi, X.-L., Li, R., Zang, J. & Zhang, S.-C. Inducing a magnetic monopole with topological surface states. *Science* **323**, 1184–1187 (2009).
11. Qi, X.-L. *et al.* Topological field theory of time-reversal invariant insulators. *Phys. Rev. B* **78**, 195424 (2008).
12. Ran, Y., Zhang, Y. & Vishwanath, A. One-dimensional topologically protected modes in topological insulators with lattice dislocations. *Nature Phys.* doi:10.1038/nphys1220 (2009).
13. Moore, J. E., Ran, Y. & Wen, X.-G. Topological surface states in three-dimensional magnetic insulators. *Phys. Rev. Lett.* **101**, 186805 (2008).
14. Fu, L. & Kane, C. L. Superconducting proximity effect and Majorana fermions at the surface of a topological insulator. *Phys. Rev. Lett.* **100**, 096407 (2008).
15. Seradjeh, B., Moore, J. E. & Franz, M. Exciton condensation and charge fractionalization in a topological insulator film. Preprint at <http://arxiv.org/abs/0902.1147v1> (2009).
16. Geim, A. K. & Novoselov, K. S. The rise of graphene. *Nature Mater.* **6**, 183–191 (2007).
17. DiSalvo, F. J. Thermoelectric cooling and power generation. *Science* **285**, 703–706 (1999).
18. Wyckoff, R. W. G. *Crystal Structures* (Krieger, 1986).
19. Hyde, G. R. *et al.* Electronic properties of Bi_2Se_3 crystals. *J. Phys. Chem. Solids* **35**, 1719–1728 (1974).
20. Larson, P. *et al.* Electronic structure of Bi_2X_3 (X = S, Se, Te) compounds: Comparison of theoretical calculations with photoemission studies. *Phys. Rev. B* **65**, 085108 (2002).
21. Greanya, V. A. *et al.* Determination of the valence band dispersions for Bi_2Se_3 using angle resolved photoemission. *J. Appl. Phys.* **92**, 6658–6661 (2002).
22. Mooser, E. & Pearson, W. B. New semiconducting compounds. *Phys. Rev.* **101**, 492–493 (1956).
23. Black, J. *et al.* Electrical and optical properties of some $\text{M}_2^{\text{VI-B}}\text{N}_3^{\text{VI-B}}$ semiconductors. *J. Phys. Chem. Solids* **2**, 240–251 (1957).
24. Mishra, S. K., Satpathy, S. & Jepsen, O. Electronic structure and thermoelectric properties of bismuth telluride and bismuth selenide. *J. Phys. Condens. Matter* **9**, 461–479 (1997).
25. LaShell, S., McDougal, B. A. & Jensen, E. Spin splitting of an Au(111) surface state band observed with angle resolved photoelectron spectroscopy. *Phys. Rev. Lett.* **77**, 3419–3422 (1996).
26. Hoesch, M. *et al.* Spin structure of the Shockley surface state on Au(111). *Phys. Rev. B* **69**, 241401 (2004).
27. Hufner, S. *Photoelectron Spectroscopy* (Springer, 1995).
28. Blaha, P. *et al.* *Computer Code WIEN2K* (Vienna Univ. Technology, 2001).
29. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

Acknowledgements

We thank N. P. Ong, B.A. Bernevig, D. Haldane and D.A. Huse for discussions. The synchrotron X-ray experiments are supported by the DOE-BES (contract DE-FG02-05ER46200) and materials synthesis is supported by the NSF-MRSEC (NSF-DMR-0819860) at Princeton Center for Complex Materials at Princeton University. Theoretical work is supported by the US Department of Energy, Office of Science, Basic Energy Sciences contract DEFG02-07ER46352, and benefited from the allocation of supercomputer time at NERSC and Northeastern University’s Advanced Scientific Computation Center (ASCC). D.Q. was partly supported by the NNSF-China (grant No. 10874116).

Author contributions

Y.X., D.Q. and D.H., carried out the experiment with the assistance of L.W. and A.P. D.G., Y.S.H. and R.J.C. provided the samples. H.L., Y.X. and A.B. carried out the theoretical calculations and the data analysis. M.Z.H. conceived the idea for the Bi_2X_3 topological class before any theoretical proposal and was responsible for overall project direction, planning and management.

Additional information

Supplementary information accompanies this paper on www.nature.com/naturephysics. Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>. Correspondence and requests for materials should be addressed to M.Z.H.

Collective excitations of composite fermions across multiple Λ levels

Dwipesh Majumder¹, Sudhansu S. Mandal¹ and Jainendra K. Jain^{2*}

The fractional quantum Hall state¹ is a quintessential system for the study of collective quantum behaviour. In such a system, the collective behaviour results in the creation of so-called composite fermions, quasi-particles formed by electrons attached to magnetic flux quanta. Recently, a new collective mode was unexpectedly observed in Raman scattering experiments² on such a system as it was found to split off from the familiar ‘fundamental’ long-wavelength mode on increase of the wave vector. Here, we present results from extensive theoretical calculations that make a compelling case that this mode corresponds to an excitation of a composite fermion across two Λ levels—effective kinetic energy levels resembling Landau levels for such particles. In addition to explaining why this excitation merges with the fundamental mode in the long-wavelength limit, our theory also provides a good quantitative account of the amount of splitting, and makes several experimentally verifiable predictions.

Unlike the well-known quantum phenomena of superfluidity and superconductivity, the fractional quantum Hall effect¹ does not entail any Bose–Einstein condensation but occurs as a result of the formation of topological electron–vortex bound states called composite fermions³. Transport⁴, light scattering^{5,6} and phonon scattering^{7,8} have been extensively used during the past quarter of a century to probe its numerous excitations. Of particular significance is the neutral collective mode, which was first studied theoretically at Landau-level filling $\nu = 1/3$ in a single-mode approximation⁹, wherein, following Feynman’s theory of the phonon–roton mode of helium superfluid, the excitation is modelled as a density wave. The neutral collective mode was detected by Raman scattering⁵, with the observations generally consistent with the predictions of the single-mode approximation in the long-wavelength limit. More recently, however, Hirjibehedin *et al.*² have discovered that this mode is not a single mode, as believed earlier, but splits into two as the wave vector is increased. By definition, the single-mode approximation cannot accommodate a doublet. A hydrodynamic approach has been proposed¹⁰ to account for the experimental observation, but does not take into account the microscopic physics of the fractional quantum Hall effect (FQHE), does not naturally explain the merging of the two modes in the long-wavelength limit and also greatly underestimates the splitting.

We show here that this new mode finds a natural explanation within the composite-fermion theory³. Composite fermions are bound states of electrons and an even number ($2p$) of quantized vortices. Because of the Berry phases produced by the bound vortices, composite fermions effectively experience a much reduced magnetic field $B^* = B - 2p\rho\phi_0$ (B is the external magnetic field, ρ is the two-dimensional density and $\phi_0 = hc/e$ is called the flux quantum). Composite fermions form their own Landau-like

kinetic energy levels in this reduced magnetic field, called Λ levels, and their filling factor ν^* is related to the electron filling factor ν through the relation $\nu = \nu^*/(2p\nu^* + 1)$. In particular, at $\nu = n/(2pn + 1)$, the ground state consists of n filled Λ levels. In the composite-fermion theory, the lowest-energy neutral excitation is a particle–hole pair, or an exciton, of composite fermions, wherein a single composite fermion from the topmost occupied Λ level is excited into the lowest unoccupied Λ level (Fig. 1b shows the fundamental composite-fermion exciton at $\nu = 2/5$). The validity of this description has been confirmed for fractions of the form $\nu = n/(2pn + 1)$ by comparison to exact diagonalization results as well as to experiment^{11,12}. This physical explanation for the neutral collective excitations is distinct from the single-mode approximation, and, in particular, suggests the possibility of extra collective modes, in which a composite fermion is excited across two or more Λ levels, as shown schematically in Fig. 1c, in complete analogy to the collective modes of an integral quantum Hall state¹³. However, it is far from obvious that the composite-fermion collective modes across different Λ levels should merge in the long-wavelength limit. In fact, a model that takes composite fermions as non-interacting produces collective modes spaced by the effective cyclotron energy in the long-wavelength limit, as also found for the dispersions obtained in the composite-fermion Chern–Simons approach^{14,15}; if correct, this would make such physics irrelevant to the new collective mode discovered in ref. 2. For a more definitive test, however, the composite-fermion exciton-mode spectrum must be evaluated in a microscopic approach that includes effects of inter-composite-fermion interactions.

Exact diagonalization studies of the FQHE state do not by themselves provide an understanding of the underlying physics, and are not useful in the present context, because, as seen below, systems as large as 200 particles are required for investigating the experimentally relevant wave vectors; the Fock space increases exponentially with the number of particles, and at present exact diagonalization is possible only for 10–12 particles for the filling factors of interest here. Our quantitative investigations below exploit accurate trial wavefunctions for composite fermions³. The standard spherical geometry is used in our calculations, which considers electrons moving on the surface of a sphere, subjected to a radial magnetic field. The magnetic field can be thought to emanate from a ‘magnetic monopole’ of strength Q at the centre, which produces a total magnetic flux of $2Q\phi_0$ through the surface of the sphere. This maps into a system of composite fermions at an effective flux $Q^* = Q - N + 1$, with Q chosen so that the state at Q^* is an integral quantum Hall state at filling $\nu^* = n$. The wavefunction $\Psi^{\text{CF-g}}$ for the FQHE ground state at $\nu = n/(2n + 1)$ is obtained by composite-fermionizing the $\nu^* = n$ integral quantum Hall state Φ^{g} . To model neutral collective excitations, we first construct wavefunctions of the excitons of the

¹Department of Theoretical Physics, Indian Association for the Cultivation of Science, Jadavpur, Kolkata 700 032, India, ²104 Davey Laboratory, Physics Department, Pennsylvania State University, University Park, Pennsylvania 16802, USA. *e-mail: jain@phys.psu.edu.

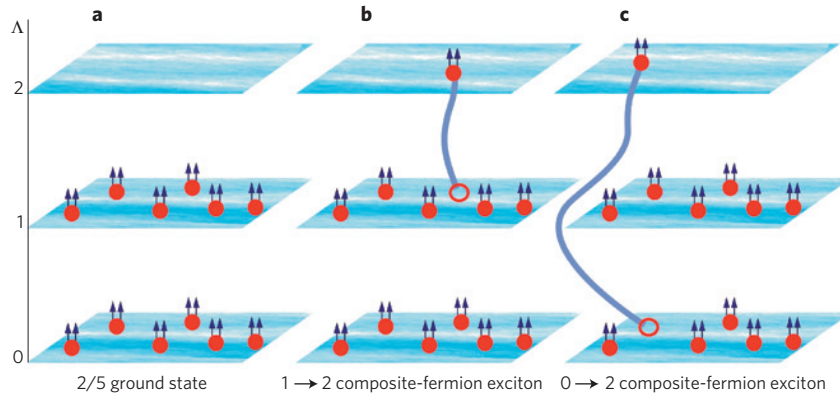


Figure 1 | Schematic diagram of composite-fermion excitons. Each composite fermion is shown as an electron carrying vortices represented by arrows. **a**, Representation of the ground state at $\nu = 2/5$ as two filled Λ levels. **b, c**, $1 \rightarrow 2$ (**b**) and $0 \rightarrow 2$ (**c**) composite-fermion excitons.

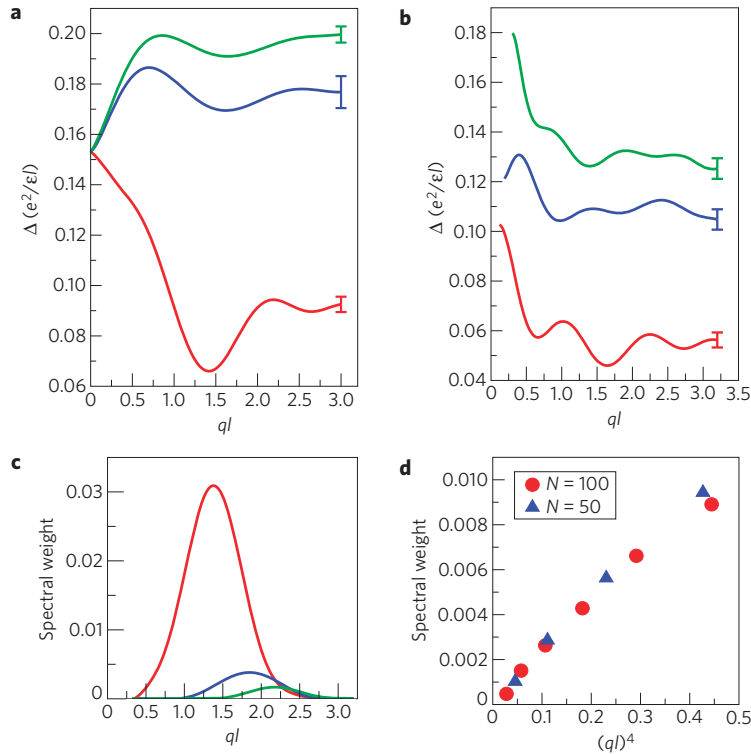


Figure 2 | Dispersions of several composite-fermion excitons and their spectral weights. **a**, The three lowest composite-fermion exciton modes at $\nu = 1/3$, obtained from $0 \rightarrow 1$, $0 \rightarrow 2$ and $0 \rightarrow 3$ excitons. The error bar at the end of each curve represents the typical statistical error in the energy determined by the Monte Carlo method. The energies are quoted in units of $e^2/\epsilon\ell$, where ϵ is the dielectric constant of the background semiconductor and $\ell = \sqrt{\hbar c/eB}$ is the magnetic length. **b**, The three lowest composite-fermion exciton modes at $\nu = 2/5$, obtained from $1 \rightarrow 2$, $0 \rightarrow 2$ and $1 \rightarrow 3$ excitons. **c**, Spectral weights for the three modes at $\nu = 1/3$ for $N = 100$. The curves from top to bottom correspond to the three modes shown in **a** respectively from bottom to top. **d**, The spectral weight for the lowest-energy composite-fermion exciton at $\nu = 1/3$ at small ql for $N = 50$ and $N = 100$. All results in this figure are for the Coulomb eigenstates $\chi_{\lambda,L}^{\text{CF-ex}}$.

integral quantum Hall state, denoted by $\{\Phi_{\lambda,L}^{\text{ex}}\}$, where L is the total orbital angular momentum of the exciton and λ labels different excitons of the type shown in Fig. 1. We composite-fermionize this basis to obtain $\{\Psi_{\lambda,L}^{\text{CF-ex}}\}$, which gives a set of basis functions for composite-fermion excitons. We orthonormalize this basis and diagonalize the Coulomb Hamiltonian to obtain the energies of the physical excitations, and also their wavefunctions $\{\chi_{\lambda,L}^{\text{CF-ex}}\}$. The scalar products of various basis functions and the Hamiltonian matrix elements are evaluated by the Metropolis Monte Carlo method. Blocks of different L are not coupled by the interaction, so can be diagonalized separately. More details can be found in refs 16, 17 and Supplementary Information.

Even though our immediate interest is in understanding the splitting of the collective mode at $\nu = 1/3$, we consider, for completeness, the three lowest collective modes at two filling factors: $0 \rightarrow 1$, $0 \rightarrow 2$ and $0 \rightarrow 3$ modes at $\nu = 1/3$, and $1 \rightarrow 2$, $1 \rightarrow 3$ and $0 \rightarrow 2$ modes at $\nu = 2/5$. We have studied systems with 50, 100 and 200 particles at both $\nu = 1/3$ and $\nu = 2/5$. The results shown in Fig. 2a, b refer to the 200-particle system, which we believe accurately represents the thermodynamic limit. The exciton dispersions are quoted as a function of the wave vector q , defined as $q = L/R$, where $R = \sqrt{Q}$ is the radius of the sphere in the unit of magnetic length $\ell = \sqrt{\hbar c/eB}$. The dispersion curves are obtained by averaging over 1.2×10^7 Monte Carlo iterations.

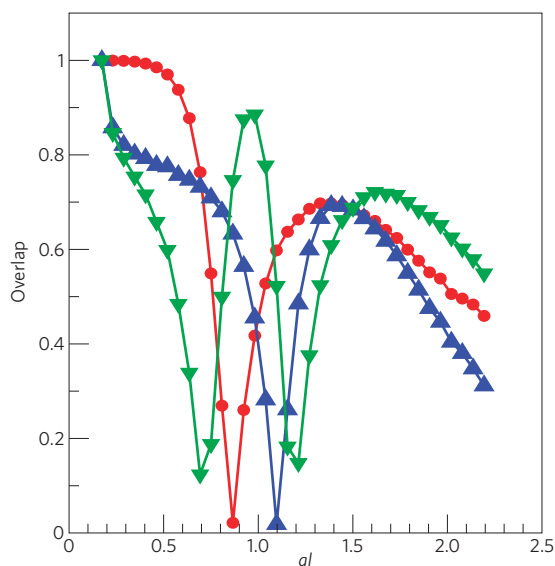


Figure 3 | Overlaps between the ‘unorthogonalized’ exciton basis states.

Overlaps $\langle \Psi_{1,L}^{\text{CF-ex}} | \Psi_{2,L}^{\text{CF-ex}} \rangle$ (red circles), $\langle \Psi_{1,L}^{\text{CF-ex}} | \Psi_{3,L}^{\text{CF-ex}} \rangle$ (blue uptriangles) and $\langle \Psi_{2,L}^{\text{CF-ex}} | \Psi_{3,L}^{\text{CF-ex}} \rangle$ (green downtriangles), where $\Psi_{\lambda,L}^{\text{CF-ex}}$ are composite-fermion exciton wavefunctions at $\nu = 1/3$ before orthogonalization. (The dispersions in Fig. 2a are obtained after orthogonalization of these modes and diagonalizing the Coulomb Hamiltonian in the orthogonal basis.) The overlaps are evaluated for $N = 200$.

The qualitative and quantitative features of experiment at $\nu = 1/3$ are nicely reproduced by our theory. Most strikingly, in Fig. 2a all modes are seen to merge in the long-wavelength limit. We note that this does not require any fine tuning of parameters (the wavefunctions $\Psi^{\text{CF-g}}$ and $\Psi^{\text{CF-ex}}$ do not contain any adjustable parameters) but is a robust effect in the composite-fermion theory. To gain an insight into this merging, we compute the overlap between $\Psi_{1,L}^{\text{CF-ex}}$, $\Psi_{2,L}^{\text{CF-ex}}$ and $\Psi_{3,L}^{\text{CF-ex}}$ at $\nu = 1/3$, shown in Fig. 3 for $N = 200$; surprisingly, the overlaps approach unity at small L , becoming precisely equal to 1 for $L = 2$ and $L = 3$. (A comparison is not possible for $L = 1$ because, in the spherical geometry, the smallest total orbital angular momentum available for the $n \rightarrow n + \lambda$ exciton is λ .) We have tested this result for many system sizes and believe that it holds in the thermodynamic limit. This demonstrates that all three modes become identical at small wave vectors, explaining why they merge into a single mode at small wave vectors; the single mode at $q\ell = 0$ is to be distinguished from two degenerate modes. When the wavefunctions are linearly independent, multiple modes are produced after orthogonalization.

Even though the composite-fermion exciton states obtained from orthogonal collective modes of the integral quantum Hall effect become exactly equal at $L = 2$ and 3 in our numerical study, we have not succeeded in deriving this result analytically. The wavefunctions are very complicated and the lowest-Landau-level projection does not lend itself to simple analytical treatments. We note that a Fock space reduction for excited states in going from the integral quantum Hall effect to the fractional quantum Hall effect has been found previously in another related context: for total orbital angular momentum $L = 1$, the wavefunction for the $0 \rightarrow 1$ composite-fermion excitation is identically annihilated by the lowest-Landau-level projection operator¹¹.

For a quantitative confirmation, we compare the theoretical splitting to the observed one. The light scattering experiments probe only very small wave-vector exchanges, but with our large systems we are able to make meaningful quantitative comparisons. The experimental splitting at $q\ell = 0.15$, which is the largest wave vector

accessible in experiments, is approximately $0.009(3) e^2/(\epsilon\ell)$. With 200 particles, the smallest wave vector directly accessible is $q\ell = 0.17$, but from a smooth extrapolation of the theoretical dispersion (assuming convergence at $q\ell \rightarrow 0$), we obtain at $q\ell = 0.15$ a splitting of $0.013(5) e^2/(\epsilon\ell)$, which is in very good agreement with experiment, especially given the smallness of the energy difference and the neglect, in our calculations, of disorder, Landau-level mixing and finite-width effects, which are all expected to slightly reduce the splitting.

The Raman intensity is proportional to the spectral weight, given by $S_q = (1/N) |\langle \chi_{\lambda,L}^{\text{ex}} | \rho_L | \Psi^{\text{CF-g}} \rangle|^2$, where the density operator at $q = L/R$ is defined as $\rho_L = \sum_i Y_{L0}(\theta_i, \phi_i)$ in terms of spherical harmonics. Figure 2c shows our calculated spectral weights for various orthogonal modes, suggesting that the collective modes involving excitations across several Λ levels are expected to be weaker than the fundamental mode. It is well known⁹ that in the long-wavelength limit the leading term in the spectral weight, proportional to $(q\ell)^2$, is exhausted by the inter-Landau-level Kohn mode at the cyclotron energy. We have explicitly verified that the spectral weight of the lowest mode at $\nu = 1/3$ goes as $(q\ell)^4$ at small wave vectors (Fig. 2d), which serves as an independent test of the accuracy of the method. We believe that the same is true for the other modes at $\nu = 1/3$ as well, but have not confirmed it explicitly because their much smaller spectral weights would require substantially more computational time. In spite of the small spectral weight, the intra-Landau-level collective modes can be observable in resonant Raman scattering owing to a strong resonant enhancement of the matrix elements¹⁸, although a detailed theory is not yet available.

The relation of the results presented above to the so-called two-roton mode is worth discussing. Motivated by a discrepancy between the experimental and theoretical energies in the long-wavelength limit^{5,12}, previous theoretical studies modelled the long-wavelength mode as a two-roton bound state^{9,19,20}, which was shown in explicit calculations^{19,20} to have a slightly lower energy than the single exciton, bringing the theoretical energy closer to the experimental one. The two-roton mode can also be interpreted as the single exciton mode hybridized with excitations consisting of two composite-fermion excitons; this is a natural interpretation given that the energy of the two-roton bound state is close to the single exciton energy (in the long-wavelength limit) and the wavefunctions of the two are not orthogonal. The physics discussed in the present work is distinct, in the sense that whereas the two-roton physics deals with corrections to the energy of the fundamental mode in the long-wavelength limit, the present work is concerned with the splitting of that mode at finite wave vectors. In fact, a mixing with excitations consisting of two composite-fermion excitons will lower the energies of all of the collective modes considered here, but we have not studied that effect because it is unlikely to alter the amount of the splitting significantly.

The results presented above have a number of experimental implications. The collective mode at $\nu = 1/3$ is in fact seen to split into several modes (rather than just two), although the higher modes are very close in energy at small wave vectors and may not be readily resolvable. Multiple modes are predicted also at $\nu = 2/5$. However, in contrast to $\nu = 1/3$, the two lowest modes at $\nu = 2/5$ do not merge in the long-wavelength limit; the composite-fermion theory thus predicts an absence of splitting of the long-wavelength collective mode at $\nu = 2/5$. (This is to be contrasted with the hydrodynamic approach¹⁰, which obtains similar behaviour for all fractions.) Our theory also obtains the full dispersion of the new collective modes, which is outside the range of Raman experiments, but possibly observable in photoluminescence experiments in the presence of a ‘grating’ produced by piezoelectric coupling to certain frozen-in phonons, which picks out certain wave vectors in absorption spectra²¹.

Received 13 November 2008; accepted 14 April 2009;
published online 10 May 2009

References

1. Tsui, D. C., Stormer, H. L. & Gossard, A. C. Two-dimensional magnetotransport in the extreme quantum limit. *Phys. Rev. Lett.* **48**, 1559–1562 (1982).
2. Hirjibehedin, C. F. *et al.* Splitting of long-wavelength modes of the fractional quantum Hall liquid at $\nu = 1/3$. *Phys. Rev. Lett.* **95**, 066803 (2005).
3. Jain, J. K. Composite fermion approach for the fractional quantum Hall effect. *Phys. Rev. Lett.* **63**, 199–202 (1989).
4. Du, R. R., Stormer, H. L., Tsui, D. C., Pfeiffer, L. N. & West, K. W. Experimental evidence for new particles in the fractional quantum Hall effect. *Phys. Rev. Lett.* **70**, 2944–2947 (1993).
5. Pinczuk, A., Dennis, B. S., Pfeiffer, L. N. & West, K. W. Observation of collective excitations in the fractional quantum Hall effect. *Phys. Rev. Lett.* **70**, 3983–3986 (1993).
6. Kang, M., Pinczuk, A., Dennis, B. S., Pfeiffer, L. N. & West, K. W. Observation of multiple magnetorotons in the fractional quantum Hall effect. *Phys. Rev. Lett.* **86**, 2637–2640 (2001).
7. Mellor, C. J. *et al.* Phonon absorption at the magnetoroton minimum in the fractional quantum Hall effect. *Phys. Rev. Lett.* **74**, 2339–2342 (1995).
8. Zeitler, U. *et al.* Ballistic heating of a two-dimensional electron system by phonon excitation of the magnetoroton minimum at $\nu = 1/3$. *Phys. Rev. Lett.* **82**, 5333–5336 (1999).
9. Girvin, S. M., MacDonald, A. H. & Platzman, P. M. Collective-excitation gap in the fractional quantum Hall effect. *Phys. Rev. Lett.* **54**, 581–583 (1985).
10. Tokatly, I. V. & Vignale, G. New collective mode in the fractional quantum Hall liquid. *Phys. Rev. Lett.* **98**, 026805 (2007).
11. Dev, G. & Jain, J. K. Band structure of the fractional quantum Hall effect. *Phys. Rev. Lett.* **69**, 2843–2846 (1992).
12. Scarola, V. W., Park, K. & Jain, J. K. Rotons of composite fermions: Comparison between theory and experiment. *Phys. Rev. B* **61**, 13064–13072 (2000).
13. Kallin, C. & Halperin, B. I. Excitations from a filled Landau-level in the two-dimensional electron-gas. *Phys. Rev. B* **30**, 5655–5668 (1984).
14. Lopez, A. & Fradkin, E. Response functions and spectrum of collective excitations of fractional-quantum-Hall-effect systems. *Phys. Rev. B* **47**, 7080 (1993).
15. Simon, S. H. & Halperin, B. I. Finite-wave-vector electromagnetic response of fractional quantized Hall states. *Phys. Rev. B* **48**, 17368 (1993).
16. Jain, J. K. & Kamilla, R. K. Composite fermions in the Hilbert space of the lowest electronic Landau level. *Int. J. Mod. Phys. B* **11**, 2621–2660 (1997).
17. Mandal, S. S. & Jain, J. K. Theoretical search for the nested quantum Hall effect of composite fermions. *Phys. Rev. B* **66**, 155302 (2002).
18. Pinczuk, A. Resonant inelastic light scattering from quantum Hall systems. in *Perspectives in Quantum Hall Effects* (eds S., Das Sarma & A., Pinczuk) (Wiley-Interscience, 1997).
19. Park, K. & Jain, J. K. Two-roton bound state in the fractional quantum Hall effect. *Phys. Rev. Lett.* **84**, 5576–5579 (2000).
20. Ghosh, T. K. & Baskaran, G. Modeling two-roton bound state formation in the fractional quantum Hall system. *Phys. Rev. Lett.* **87**, 186803 (2001).
21. Kukushkin, I. V., Smet, J. H., Schuh, D., Wegscheider, W. & von Klitzing, K. Dispersion of the composite-fermion cyclotron-resonance mode. *Phys. Rev. Lett.* **98**, 066403 (2007).

Acknowledgements

We acknowledge the Computer Centre of IACS for providing its computing facility. D.M. is supported by CSIR, Government of India.

Additional information

Supplementary information accompanies this paper on www.nature.com/naturephysics. Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>. Correspondence and requests for materials should be addressed to J.K.J.

Breakdown of the nuclear-spin-temperature approach in quantum-dot demagnetization experiments

P. Maletinsky^{*}, M. Kroner and A. Imamoglu^{*}

The physics of interacting nuclear spins arranged on a crystalline lattice is generally described using a thermodynamic framework¹ and the concept of spin temperature. In the past, experimental studies in bulk solid-state systems have proven this concept to be not only correct^{2,3} but also vital for the understanding of experimental observations⁴. Here we show, using demagnetization experiments, that the concept of spin temperature in general fails to describe the mesoscopic nuclear-spin ensemble of a quantum dot. We associate the observed deviations from a thermal spin state with the presence of strong quadrupolar interactions within the quantum dot, which cause significant anharmonicity in the spectrum of the nuclear spins. Strain-induced, inhomogeneous quadrupolar shifts also lead to a complete suppression of angular-momentum exchange between the nuclear-spin ensemble and its environment, resulting in nuclear-spin relaxation times exceeding an hour. Remarkably, the position-dependent axes of the quadrupolar interactions render magnetic-field sweeps inherently non-adiabatic, thereby causing an irreversible loss of nuclear-spin polarization.

The study of nuclear-spin physics by optical orientation experiments in bulk semiconductor materials has been an active field of research over recent decades^{5–7}. These research efforts have shown that, using the electron as a mediator, it is possible to transfer angular momentum from light onto nuclei, thereby establishing a nuclear-spin polarization that is orders of magnitude higher than the equilibrium nuclear polarization at cryogenic temperatures. As a result, the effective nuclear-spin temperature in such an optically pumped system can be pushed to the microkelvin regime. Combining these optical pumping schemes with nuclear adiabatic demagnetization techniques borrowed from bulk nuclear magnetic resonance experiments³ would be a natural extension to these experiments that could lead to a significant further reduction of the nuclear-spin temperature. This approach, previously demonstrated in bulk semiconductors^{5,8}, suffers from the fact that, in most systems where optical orientation of nuclear spins is possible, nuclear-spin relaxation is too fast to allow for a significant reduction of magnetic fields in an adiabatic way. Here, we use the exceedingly long nuclear-spin relaxation time in self-assembled quantum dots (QDs) (ref. 9) to implement an 'adiabatic' demagnetization experiment on the system of $\sim 10^5$ nuclear spins.

The mesoscopic ensemble of nuclear spins in a QD can be conveniently polarized and measured by optical means^{5,9–12}. To this end, we use the photoluminescence of the negatively charged exciton (X^{-1}) under resonant excitation of an excited QD state. It has been shown previously¹³ that, under appropriate excitation conditions, 20–50% of the QD nuclear spins can be efficiently

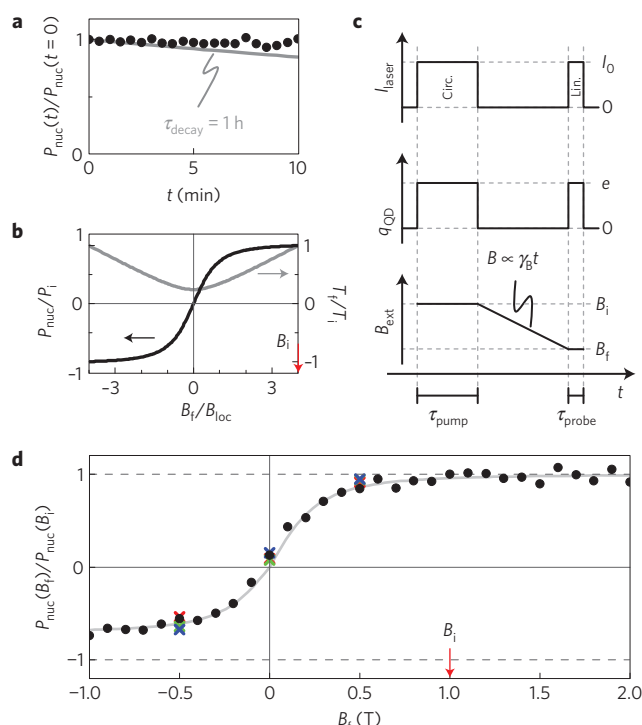


Figure 1 | Demagnetization of QD nuclear spins. **a**, Free decay of P_{nuc} at $B_{\text{ext}} = 2$ T for an uncharged QD after optical pumping of the nuclear spins for $\tau_{\text{pump}} = 600$ ms. The grey curve shows an exponential decay with a time constant of 1 h for comparison. **b**, Theoretical prediction of nuclear-spin temperature and polarization during adiabatic demagnetization from a field B_i (red arrow) to B_f . **c**, Schematic diagram of the experimental procedure for adiabatic demagnetization of QD nuclear spins. The nuclei are optically pumped at $B_{\text{ext}} = B_i$ ($T_{\text{spin},i} \sim \text{mK}$). Directly after the pumping pulse, the electron is ejected from the QD. B_{ext} is then linearly ramped at a rate γ_B to a value B_f , at which we measure P_{nuc} . **d**, Experimental (de)magnetization of QD nuclear spins. $B_i = 1$ T as indicated by the red arrow, $\gamma_B = 10 \text{ mT s}^{-1}$ and $\Delta E_{\text{OS}}(B_i) = 57 \mu\text{eV}$. The grey curve is a fit according to the theoretical predictions shown in **b**; we find $B_{\text{loc}} = 290$ mT. Blue, green and red crosses show a similar experiment, with $\gamma_B = 5, 2.5$ and 0.8 mT s^{-1} , respectively ($B_f = 0.5$ T for these data points).

polarized in a timescale of a few milliseconds. The resulting dynamical nuclear-spin polarization can then be measured through a change in the Zeeman splitting, ΔE_{OS} , of the X^{-1} recombination line¹³; this energy shift due to the spin-polarized nuclei is commonly referred to as the Overhauser shift¹⁴.

A remarkable feature of the QD nuclear-spin system is the excellent isolation from its environment if the QD is uncharged. Figure 1a shows the corresponding free evolution of the nuclear-spin polarization P_{nuc} (proportional to ΔE_{OS}) in a QD subject to an external magnetic field $B_{\text{ext}} = 2$ T. The nuclear-spin relaxation time clearly exceeds one hour and does not vary appreciably over the magnetic-field range relevant to this work⁹. As the bulk material surrounding the QD remains unpolarized during the experiment (see the Methods section), the long nuclear-spin lifetime indicates that nuclear-spin diffusion between the QD and its environment is strongly suppressed. We attribute this quenching of spin diffusion to the structural and chemical mismatch between the InGaAs QD and its GaAs surroundings^{12,15}. The very slow nuclear-spin relaxation leaves room for further manipulation of the QD nuclear-spin system after optical pumping. In particular, we can study how P_{nuc} behaves under slow variations of external parameters and thereby study the validity of spin thermodynamics for the QD nuclear-spin system.

If the QD nuclei were describable using a thermodynamic approach, P_{nuc} would be aligned with B_{ext} and would be described by Curie's law $\gamma P_{\text{nuc}} = B_{\text{ext}} C / T_{\text{spin}}$ (ref. 3) (here, γ is the nuclear gyromagnetic ratio, C the Curie constant and T_{spin} the nuclear-spin temperature). An adiabatic lowering of B_{ext} from an initial value B_i to a final value B_f would conserve P_{nuc} and lead to a reduction of T_{spin} by a factor B_i/B_f . In general, cooling by adiabatic demagnetization is possible for any system where the spin entropy S is conserved and a function of $B_{\text{ext}}/T_{\text{spin}}$ only. The ultimate limit to the achievable cooling is determined by nuclear-spin interactions, which give the dominant contribution to S at low magnetic fields. The strength and nature of these interactions can be phenomenologically described by a random local magnetic field B_{loc} . In most cases, B_{loc} is given by the nuclear dipolar couplings (≈ 0.1 mT). As soon as $B_{\text{ext}} \approx B_{\text{loc}}$, the local fields randomize an established nuclear-spin polarization and thereby limit the efficiency of the adiabatic spin cooling to B_{loc}/B_i . The resulting behaviour of nuclear-spin temperature and polarization as a function of B_f is sketched in Fig. 1b: for $B_{\text{ext}} = 0$, the spin temperature remains finite and the nuclear spins are completely depolarized. Amazingly, this depolarization is a reversible process, provided that S is a conserved quantity at all fields. When the spins are re-magnetized to a magnetic field exceeding B_{loc} , their polarization recovers along the direction of the magnetic field and in particular conserves the sign of its initial spin temperature.

To test the validity of spin thermodynamics for the QD nuclear spins and to study the possibility of adiabatic cooling in this system, we performed demagnetization experiments on a QD, as illustrated in Fig. 1c. A circularly polarized 'pump' pulse of length τ_{pump} is used to polarize the nuclear spins. After ejecting the electron from the QD, we linearly ramp B_{ext} from B_i to B_f with a rate $\gamma_B = 10$ mT s⁻¹. At the final field B_f , the remaining degree of nuclear-spin polarization is measured using a linearly polarized 'probe' pulse of length τ_{probe} (ref. 9). This experiment is repeated at various values of B_f to record the process of 'adiabatic' (de)magnetization.

Figure 1d shows the result of a demagnetization experiment performed on the nuclear-spin system of an individual QD. The nuclei are polarized with a pump pulse $\tau_{\text{pump}} = 300$ ms at $B_i = 1$ T and measured at B_f with a probe pulse $\tau_{\text{probe}} = 5$ ms. At a rough glance, this measurement qualitatively follows the behaviour depicted in Fig. 1b. A closer inspection, however, reveals significant deviations: on ramping the external field to $B_f = -1$ T we recover only 63% of the initial P_{nuc} . In addition, by measuring $P_{\text{nuc}}(B_f)$ we determined the value of the local field to be $B_{\text{loc}} = 290$ mT: this value is about three orders of magnitude larger than typical nuclear dipolar fields. Finally, we observe that even for $B_f = 0$ the QD has a remanent nuclear-spin polarization $P_{\text{nuc}}^{\text{rem}}$. To verify that we do not induce an unwanted increase of spin entropy by sweeping B_{ext}

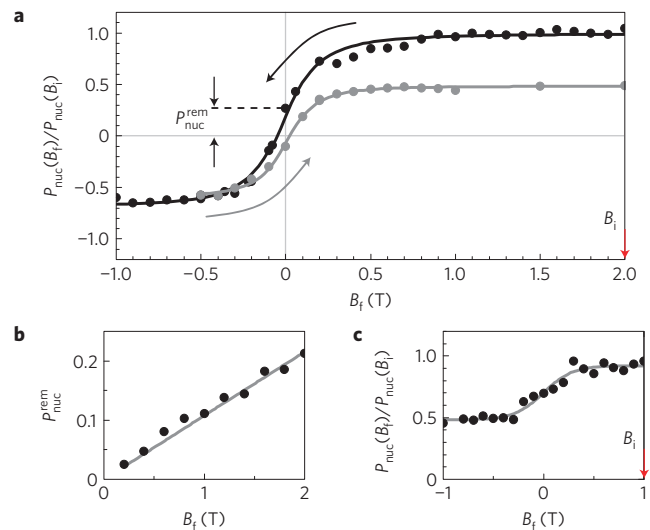


Figure 2 | Irreversibility and hysteresis in the demagnetization experiment.

a, Black circles, the same experiment as in Fig. 1d, with $B_i = 2.0$ T as indicated by the red arrow ($\Delta E_{\text{OS}}(B_i) = 89.5$ μeV). After reaching $B_f = -1$ T, we reverse the magnetic-field sweep direction and bring the nuclei back to the initial field (grey circles). **b**, The remanent nuclear-spin polarization $P_{\text{nuc}}^{\text{rem}}$ (normalized to the value $P_{\text{nuc}}(B_i)$ found in **a**) as a function of B_i . As $P_{\text{nuc}}(B_i) \propto B_i$, the nuclear-spin temperature after optical pumping is roughly constant for all values of B_i in this measurement. **c**, After polarization of nuclear spins at $B_i = 1$ T (red arrow), we sweep B_{ext} to B_f and then back to B_i , where P_{nuc} is measured. The magnetic field sweeps become partly irreversible as soon as $|B_f| \lesssim 0.3$ T $\approx B_{\text{Q}}$. The lines in the figures are guides to the eye.

too fast, we repeated our experiment for values of γ_B of 5, 2.5 and 0.8 mT s⁻¹ (crosses in Fig. 1d). Within the experimentally accessible range, γ_B has no influence on our observations.

The discrepancy between our experimental findings and the predictions from a thermodynamical treatment of nuclear spins becomes even more pronounced if we increase $P_{\text{nuc}}(B_i)$ (which can be achieved by first increasing B_{ext} to 2.2 T (refs 12, 16)). Figure 2a shows an experiment where we then demagnetize the polarized nuclear spins starting from $B_i = 2.0$ T to a final field $B_f = -1$ T (black data points). We then reverse the sweep direction of the magnetic field and ramp B_{ext} back to B_i (grey data points). This experiment shows a considerable hysteresis of the nuclear-spin polarization as a function of B_f . In particular, $P_{\text{nuc}}^{\text{rem}}$ changes sign for the two sweep directions of B_{ext} . Furthermore, the magnitude of $P_{\text{nuc}}^{\text{rem}}$, and respectively the width of the observed hysteresis curve, depends linearly on the initial degree of nuclear-spin polarization and on B_i (Fig. 2b).

To obtain more information about the source of irreversibility of P_{nuc} during magnetic-field sweeps, we performed a further experiment, where we optically orient the nuclear spins at $B_i = 1$ T and ramp the field to a value $B_f < B_i$ and then back to $B_i = B_f$, where we measure the remaining degree of nuclear-spin polarization. The result of this experiment (Fig. 2c) indicates that the magnetic-field sweeps start to induce irreversibilities in P_{nuc} as soon as $|B_{\text{ext}}| \lesssim B_{\text{loc}} \approx 300$ mT.

Finally we note that the experimental observations described here do not depend on the sign of the initial nuclear-spin temperature ($T_{\text{spin},i}$). We have repeated the demagnetization experiments for $T_{\text{spin},i} < 0$ (that is, σ^- laser excitation at $B_i > 0$, not shown here) and observed values of B_{loc} and $P_{\text{nuc}}^{\text{rem}}$ consistent with the measurements presented in Figs 1 and 2. These measurements are complicated by the fact that for $T_{\text{spin}} < 0$ nuclear-spin pumping is rather inefficient¹², leading to a low degree of dynamical

nuclear-spin polarization and therefore a smaller signal-to-noise ratio than for $T_{\text{spin}} > 0$.

The three principal features of our experiments, the existence of $P_{\text{nuc}}^{\text{rem}}$, the hysteretic behaviour of P_{nuc} and the partial irreversibility of our demagnetization experiment, result from a violation of the nuclear (Zeeman) spin temperature approximation^{1,3}. We explain these features by taking into account the strong inhomogeneous quadrupolar interaction (QI) of the nuclear spins in a QD (refs 17–19). The self-assembled growth of InGaAs QDs is driven by a strong lattice mismatch between InGaAs and its surrounding GaAs matrix, which results in a heavily strained QD lattice. As a consequence, QD nuclei experience large electric-field gradients, which couple to the nuclear quadrupolar moment. The resulting quadrupolar Hamiltonian²⁰,

$$\hat{H}_Q = \frac{h\nu_Q}{2} \left(\hat{I}_{z'}^2 - \frac{1}{3}I(I+1) \right)$$

is characterized by a nuclear quadrupolar frequency ν_Q (proportional to the local strain at the nuclear site) and a quadrupolar axis z' (with corresponding unit vector $\mathbf{e}_{z'}$ along the main axis of the local electric-field-gradient tensor). $\hat{\mathbf{I}}$ is the nuclear-spin angular-momentum operator with quantum number I and $\hat{I}_{z'} = \hat{\mathbf{I}} \cdot \mathbf{e}_{z'}$. For typical strain values of 2% (ref. 21), we find $\nu_Q \approx 2.8$ MHz for As and 1.2 MHz for In (ref. 22). For comparison of the interaction strength of \hat{H}_Q with a pure nuclear Zeeman Hamiltonian $\hat{H}_Z = \gamma \hat{\mathbf{I}} \cdot \mathbf{B}_{\text{ext}}$, it is convenient to express the QI strength by an equivalent magnetic field $B_Q = h\nu_Q/\gamma$. For As and In, we find $B_Q = 388$ mT and 125 mT, respectively; the corresponding mean value agrees well with our experimental estimate for B_{oc} .

The spectrum of a nuclear spin with quadrupolar frequency ν_Q depends strongly on the angle θ between $\mathbf{e}_{z'}$ and the external magnetic field (directed along \mathbf{e}_z , Fig. 3a). Figure 3b shows the eigenenergies of a nuclear spin with $I = 3/2$, as a function of B_{ext}/B_Q . At $B_{\text{ext}} = 0$, the spectrum is governed by \hat{H}_Q , which pairs the nuclear-spin states into doublets with angular-momentum projections $\pm m_{z'}$ on $\mathbf{e}_{z'}$. The doublets are split by an energy $|\hbar\omega_{m_{z'}, m_{z'}+1}| = (m_{z'} + 1/2)h\nu_Q$, respectively. Conversely, in a high magnetic field, the spectrum is determined by \hat{H}_Z with nuclear angular momentum being quantized along the axis \mathbf{e}_z . Even at arbitrarily high fields, however, the spectrum is significantly perturbed by \hat{H}_Q and never becomes perfectly harmonic.

We modelled our demagnetization experiment using the steady-state solution of a rate equation for the populations $p_{|m\rangle}$ of spin states $|m\rangle$, which are mutually coupled through dipolar interactions (Fig. 3b and c). The nuclear spins are initialized with a Boltzmann distribution at $B_{\text{ext}} = B_i$ (see the Methods section) and the evolution of the $p_{|m\rangle}$ is calculated as a function of B_{ext} . Owing to the unequal nuclear-spin level spacings, only nuclear-spin flip-flops that preserve $p_{|m\rangle}$ ($\forall |m\rangle$) are energetically allowed in general and therefore the spin populations remain invariant as a function of B_{ext} . Varying B_{ext} will change the relative nuclear-spin level spacings in the nonlinear way depicted in Fig. 3c. As the $p_{|m\rangle}$ remain invariant while B_{ext} is reduced, the nuclear spins are driven into a state that is out of thermal equilibrium (that is, not Boltzmann distributed). At specific values of B_{ext} (red markers in Fig. 3c), transition energies between distinct pairs of nuclear-spin states can coincide—a situation denoted as a ‘crossover’ of nuclear-spin transitions¹. At these fields, the $p_{|m\rangle}$ are no longer constant and the nuclear-spin levels involved in the crossover can relax to a Boltzmann distribution. The irreversibility observed in our magnetic-field sweeps is a consequence of this partial relaxation of nuclear spins to thermal equilibrium. We speculate that the resulting increase of the nuclear-spin entropy is induced by an energy-conserving coupling to the

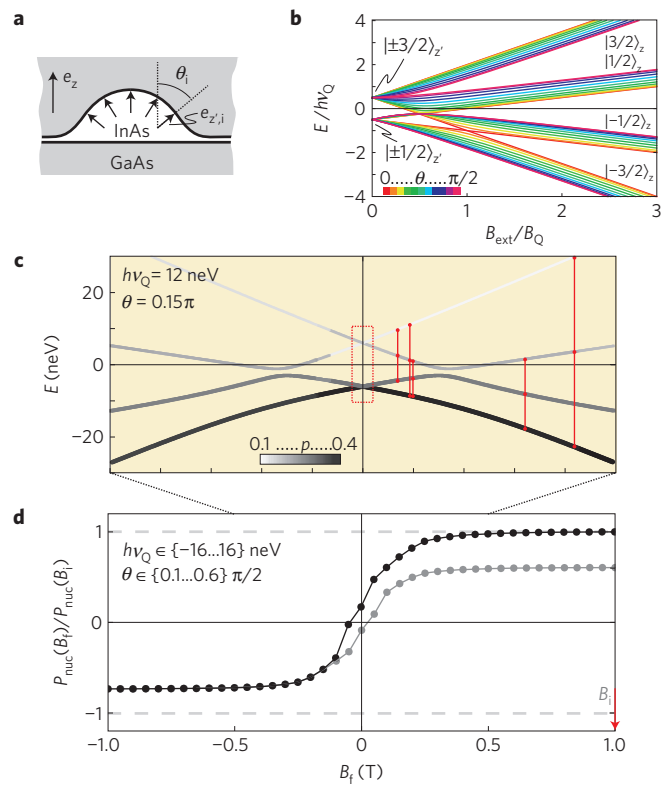


Figure 3 | Modelling of the demagnetization experiment. Local electric-field gradients induced by strain in self-assembled QDs result in strong QI for the nuclear spins. **a**, Model of local strain axis distribution $\mathbf{e}_{z'}$ within a QD. **b**, Spectrum of nuclear spins ($I = 3/2$) under the influence of both \hat{H}_Q and \hat{H}_Z for a variety of angles θ between $\mathbf{e}_{z'}$ and \mathbf{e}_z . **c**, Simulation of QD nuclear-spin demagnetization for a particular setting $\nu_Q = 3$ MHz and $\theta = 0.15\pi$. Nuclear-spin populations p are represented both by line thickness and greyscale of the lines that indicate the energy of the nuclear-spin states. At $B_i = 1$ T, the nuclei are initialized with a Boltzmann distribution over their spin states. The populations remain constant for most values of B_i . Only if a crossover of nuclear-spin transitions occurs (red markers for $B_i > 0$) do the occupations of the involved spin states evolve to a (local) thermal equilibrium distribution (see the text). We simulate this process for a set of configurations $\{\theta, \nu_Q\}$ and calculate the corresponding magnetization $P_{\text{nuc}} \propto \langle I_z \rangle$. **d**, The resulting nuclear-spin polarization as a function of B_i starting at B_i (red arrow), which qualitatively reproduces the experimental findings shown in Fig. 2.

environment of the nuclear spins. If the minimal energy gap of the anticrossing induced by the dipolar coupling between two interacting nuclear spins at their crossover is smaller than the coupling to the environment, pure dephasing of the nuclear-spin transitions will induce irreversible crossover transitions and S will increase.

On sweeping B_{ext} through zero (red box in Fig. 3c), dipolar interactions will couple the states $m_{z'} = \pm 1/2$. The associated passage through the avoided crossing between these single-spin states is adiabatic and preserves the respective populations in the two lowest-lying spin states. In contrast, nuclear dipolar interaction cannot couple any of the states with $|m_{z'}| > 1/2$ owing to conservation of energy and angular momentum. The spin states in the $|m_{z'}| = 3/2$ manifold will therefore cross and in particular preserve their populations $p_{3/2}$ and $p_{-3/2}$. The imbalance between these populations ($p_{3/2} < p_{-3/2}$ in Fig. 3) will result in a remnant polarization $P_{\text{nuc}}^{\text{rem}}$, even if B_{ext} is strictly zero.

We averaged our model over a set of parameters θ and ν_Q to account for the strong inhomogeneity of QI over the QD

(see the Methods section). The result of this full simulation is shown in Fig. 3d. We highlight that the good qualitative agreement with our experimental results (Fig. 2a) is rather insensitive to the set of parameters used in our simulation. In particular, the choice of the distribution for the parameters θ and ν_Q did not affect our results significantly. Furthermore, our simulation treats the QD spin system as a pure spin-3/2 system, whereas for In $I = 9/2$. A numerical treatment of the full InGaAs nuclear-spin system is beyond the scope of this paper and would most probably not alter the qualitative behaviour of our simulations (see the Methods section).

Our results show that the nuclear-spin system of a self-assembled QD provides a rare example for a solid-state nuclear-spin ensemble that cannot be described by a nuclear-spin temperature²³. We note that, if we could assign a spin temperature to the QD nuclear-spin system, optical pumping combined with adiabatic demagnetization of the nuclear spins would be a novel and efficient means of nuclear-spin cooling in QDs without QI: possible systems include nuclear spin-1/2 systems, such as ¹³C-nanotube QDs (ref. 24), where QI is inherently absent, or strain-free semiconductor nanostructures²⁵, such as epitaxially grown droplet QDs (ref. 26). There, adiabatic nuclear-spin cooling would be limited only by nuclear dipolar interactions resulting in $B_{\text{loc}} \approx 0.1$ mT. Achieving nuclear-spin cooling to temperatures of ≈ 100 nK should be feasible in these systems, opening ways to study the remnants of nuclear magnetic phase transitions in the mesoscopic system of QD nuclear spins²⁷.

Methods

Sample and experimental techniques. Individual QDs were studied using the photoluminescence of X^{-1} under resonant excitation of an excited QD state. The QD sample was grown by molecular-beam epitaxy on a (100) semi-insulating GaAs substrate. The approximate composition of the QDs after self-assembled growth and postgrowth annealing was $\text{In}_{0.5}\text{Ga}_{0.5}\text{As}$. For individual optical addressing, the QDs were grown at a low density of $\lesssim 0.1 \mu\text{m}^{-2}$. The QDs were spaced by 25 nm of GaAs from a doped n^{++} -GaAs layer, followed by 30 nm of GaAs and 29 periods of an AlAs/GaAs (2/2 nm) barrier, which was capped by 4 nm of GaAs. A bias voltage applied between the top Schottky and back Ohmic contacts controls the charging state of the QD. Optical pumping of QD nuclear spins was performed at the centre of the X^{-1} stability plateau in gate voltage, where photoluminescence counts as well as the resulting Overhauser shift were maximized¹³.

The QD sample was immersed in a liquid helium bath cryostat equipped with a superconducting magnet and was held at the cryostat base temperature of 1.7 K. The photoluminescence emitted by the QD was analysed in a 750 nm monochromator, allowing for the determination of spectral shifts of the QD emission lines with a precision of $\sim 1 \mu\text{eV}$ (ref. 12). A combination of an optical ‘pump–probe’ technique, together with linear ramps of the applied magnetic field, was used to adiabatically demagnetize the QD nuclear spins (see Fig. 1c); technical details of the pump–probe set-up are given elsewhere⁹. The ‘pump’ pulse consists of a circularly polarized laser pulse of duration τ_{pump} , which is used to optically orient the QD nuclear spins¹². We typically achieve an Overhauser shift of $\Delta E_{\text{OS}} = 60 \mu\text{eV}$ at $B_i = 1$ T, corresponding to nuclear-spin polarization $P_{\text{nuc}} \approx 35\%$ or $T_i \approx 1.5$ mK (for $B_i = 2.2$ T, $\Delta E_{\text{OS}} = 89.5 \mu\text{eV}$ and $P_{\text{nuc}} \approx 50\%$). In the range of B_{ext} relevant to our experiment, $P_{\text{nuc}} \propto B_i$ (ref. 12) such that the initial nuclear-spin temperature T_i is roughly constant and of the order of few millikelvin (ref. 6) for all values of B_i .

Directly after applying the pump pulse to the QD, the gate voltage is switched to a value where the QD is charge neutral. In this regime, nuclear-spin polarization has an exceedingly long relaxation time of the order of hours⁹ (see Fig. 1a). We note that we can exclude any significant nuclear polarization of the bulk material surrounding the QD. The observation of dynamical nuclear-spin polarization in our experiment depends sensitively on the excitation laser energy, which we tune to an intra-dot (p -shell) excitation resonance with a width of $\approx 300 \mu\text{eV}$ and located ≈ 36 meV above the photoluminescence emission energy. The sharpness and energy of this excitation resonance makes any excitation processes that involve the creation of free electrons in the bulk very unlikely²⁸. Furthermore, the pumping time $\tau_{\text{pump}} = 600$ ms used in our experiment is much too short to lead to a significant bulk nuclear-spin polarization, even if some free electrons were created during laser illumination.

Details of the model. The model we developed to explain our experimental findings is based on the steady-state solution of a rate equation for the populations of a nuclear-spin $I = 3/2$ system. The nuclear spins are initialized with a Boltzmann

distribution over the spin states at $B_{\text{ext}} = B_i$. The assumption of a thermal distribution of nuclear-spin levels at $B_{\text{ext}} = B_i$ is justified by the fact that nuclear spins are polarized by hyperfine interaction with the QD electron: optical pumping of the electron leads to a broadening of its spin states by several μeV (ref. 12), allowing for electron–nuclear flip-flops between the electron and any two given nuclear-spin states which are coupled by the hyperfine interaction. It is therefore reasonable to assume that the occupations of nuclear-spin levels at B_i follow a Boltzmann distribution.

We then change the magnetic field by keeping the populations of spin levels fixed. Only at the specific fields where cross-relaxation is permitted (Fig. 3c) do we allow for a local thermal equilibrium to be established between the spin levels involved in the cross-relaxation transitions. All other populations and the total energy of the nuclear-spin system remain constant. On sweeping through $B_{\text{ext}} = 0$, we assume that the levels $m_z = \pm 1/2$ undergo an adiabatic passage through an anticrossing induced by the coupling of these two states by dipolar interactions. Spin states with $m_z = \pm 3/2$ however remain uncoupled and undergo an adiabatic level crossing, which preserves their populations.

The result of our simulations is shown in Fig. 3c,d. We illustrate the evolution of the occupations of the individual nuclear-spin states in Fig. 3c, where we show the spectrum of a nuclear spin for the parameters $\nu_Q = 3$ MHz, $\theta = 0.3\pi/2$ and $\gamma = 10$ MHz T^{−1}. The occupations of the individual levels are encoded by the thickness and grey shade of the corresponding lines. Magnetic fields where cross-relaxation processes take place are indicated by red lines. We repeated this calculation for a set of angles $\theta \in \{0.1, 0.2, \dots, 0.6\} \frac{\pi}{2}$ and quadrupolar frequencies $\nu_Q \in \{-4, -3, -2, 2, 3, 4\}$ MHz, over which we average our results. As the local strain in our QDs can be both tensile and compressive, positive and negative values for ν_Q are possible. By solving the complete Hamiltonian $\hat{H}_Q + \hat{H}_Z$, we can relate the occupancies of the spin levels to our experimentally observed nuclear-spin polarization—the expectation value $\langle \hat{I}_z \rangle$ of the nuclear-spin polarization along the direction of B_{ext} . Figure 3d of the main paper shows the result of our simulation in the form of the calculated evolution of P_{nuc} as a function of B_i .

We note that our model is a great simplification of the actual experimental situation. First, we completely ignore cross-relaxation events between nuclei of different (θ, ν_Q) values. Second, our calculation was performed for a spin-3/2 system for simplicity, whereas the actual QD nuclear-spin system consists of a mixture of spin 3/2 (Ga, As) and spin 9/2 (In), which further complicates the situation. Although a numerical treatment of the full InGaAs nuclear-spin system is beyond the scope of this paper, we argue that such a treatment would not alter the physical picture conveyed by our simulation. Including $I = 9/2$ spins would lead to a nuclear-spin spectrum similar to the one illustrated in Fig. 3b. The number of magnetic-field values where cross-relaxation events would be energetically allowed would increase compared with the case of $I = 3/2$, but these events would still be singular in the sense that for most values of B_{ext} the nuclear spins could not thermalize. The system would thus still be driven out of thermal equilibrium and the relaxation events during cross-relaxation would lead to an increase of nuclear-spin entropy. Including flip-flop events between In and As nuclear spins would have a similar effect: these transitions would be allowed for a subset of close nuclei and would allow for partial thermalization only at specific values of B_{ext} .

Received 16 January 2009; accepted 3 April 2009;
published online 10 May 2009

References

- Goldman, M. *Spin Temperature and Nuclear Magnetic Resonance in Solids* (Oxford Univ. Press, 1970).
- Abragam, A. & Proctor, W. G. Experiments on spin temperature. *Phys. Rev.* **106**, 160–161 (1957).
- Slichter, C. P., Holton, W. C. & Fellow, A. P. S. Adiabatic demagnetization in a rotating reference system. *Phys. Rev.* **122**, 1701–1708 (1961).
- Purcell, E. M. & Pound, R. V. A nuclear spin system at negative temperature. *Phys. Rev.* **81**, 279–280 (1951).
- Meier, F. *Optical Orientation* (North-Holland, 1984).
- Dyakonov, M. I. & Perel, V. I. Optical orientation in a system of electrons and lattice nuclei in semiconductors—theory. *Sov. Phys. JETP* **38**, 177–183 (1974).
- Page, D. Optical-detection of NMR in high-purity GaAs—direct study of the relaxation of nuclei close to shallow donors. *Phys. Rev. B* **25**, 4444–4451 (1982).
- Kalevich, V. K., Kul'kov, V. D. & Fleisher, V. G. Onset of a nuclear polarization front due to optical spin orientation in a semiconductor. *JETP Lett.* **35**, 20–24 (1982).
- Maletinsky, P., Badolato, A. & Imamoglu, A. Dynamics of quantum dot nuclear spin polarization controlled by a single electron. *Phys. Rev. Lett.* **99**, 056804 (2007).
- Gammon, D. *et al.* Nuclear spectroscopy in single quantum dots: Nanoscopic Raman scattering and nuclear magnetic resonance. *Science* **277**, 85–88 (1997).
- Eble, B. *et al.* Dynamic nuclear polarization of a single charge-tunable InAs/GaAs quantum dot. *Phys. Rev. B* **74**, 081306 (2006).
- Maletinsky, P., Lai, C. W., Badolato, A. & Imamoglu, A. Nonlinear dynamics of quantum dot nuclear spins. *Phys. Rev. B* **75**, 035409 (2007).

13. Lai, C. W., Maletinsky, P., Badolato, A. & Imamoglu, A. Knight-field-enabled nuclear spin polarization in single quantum dots. *Phys. Rev. Lett.* **96**, 167403 (2006).
14. Overhauser, A. W. Polarization of nuclei in metals. *Phys. Rev.* **92**, 411–415 (1953).
15. Malinowski, A., Brand, M. A. & Harley, R. T. Nuclear effects in ultrafast quantum-well spin-dynamics. *Physica E* **10**, 13–16 (2001).
16. Braun, P.-F., Urbaszek, B., Amand, T. & Marie, X. Bistability of the nuclear polarization created through optical pumping in $\text{In}_{1-x}\text{Ga}_x\text{As}$ quantum dots. *Phys. Rev. B* **74**, 245306 (2006).
17. Dzhirov, R. I. & Korenev, V. L. Stabilization of the electron–nuclear spin orientation in quantum dots by the nuclear quadrupole interaction. *Phys. Rev. Lett.* **99**, 037401 (2007).
18. Maletinsky, P. *Polarization and Manipulation of a Mesoscopic Nuclear Spin Ensemble Using a Single Confined Electron Spin*. PhD thesis, ETH Zürich (2008).
19. Deng, C. X. & Hu, X. D. Selective dynamic nuclear spin polarization in a spin-blocked double dot. *Phys. Rev. B* **71**, 033307 (2005).
20. Slichter, C. P. *Principles of Magnetic Resonance* (Springer, 1996).
21. Williamson, A. J. & Zunger, A. InAs quantum dots: Predicted electronic structure of free-standing versus GaAs-embedded structures. *Phys. Rev. B* **59**, 15819–15824 (1999).
22. Sundfors, R. K., Tsui, R. K. & Schwab, C. Experimental gradient elastic tensors: Measurement in I–VII semiconductors and the ionic contribution in III–V and I–VII compounds. *Phys. Rev. B* **13**, 4504–4508 (1976).
23. Rhim, W.-K., Pines, A. & Waugh, J. S. Violation of the spin-temperature hypothesis. *Phys. Rev. Lett.* **25**, 218–220 (1970).
24. Churchill, H. O. H. *et al.* Electron–nuclear interaction in ^{13}C nanotube double quantum. *Nature Phys.* **5**, doi:10.1038/NPHYS1247 (2008).
25. Feng, D. H., Akimov, I. A. & Henneberger, F. Nonequilibrium nuclear–electron spin dynamics in semiconductor quantum dots. *Phys. Rev. Lett.* **99**, 036604 (2007).
26. Belhadj, T. *et al.* Optically monitored nuclear spin dynamics in individual GaAs quantum dots grown by droplet epitaxy. *Phys. Rev. B* **78**, 205325 (2008).
27. Simon, P. & Loss, D. Nuclear spin ferromagnetic phase transition in an interacting two dimensional electron gas. *Phys. Rev. Lett.* **98**, 156401 (2007).
28. Vasanelli, A., Ferreira, R. & Bastard, G. Continuous absorption background and decoherence in quantum dots. *Phys. Rev. Lett.* **89**, 216804 (2002).

Acknowledgements

We thank A. Högele, J. Elzerman and S. D. Huber for help with the manuscript, and T. Amand and O. Krebs for discussions. We acknowledge A. Badolato for sample growth. This work is supported by NCCR-Nanoscience and an ERC Advanced Investigator Grant.

Additional information

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>. Correspondence and requests for materials should be addressed to P.M. or A.I.

Atomic wavefunctions probed through strong-field light-matter interaction

D. Shafir¹, Y. Mairesse^{2,3}, D. M. Villeneuve³, P. B. Corkum^{3*} and N. Dudovich^{1,3*}

Strong-field light-matter interactions can encode the spatial properties of the electronic wavefunctions that contribute to the process^{1–4}. In particular, the broadband harmonic spectra, measured for a series of molecular alignments, can be used to create a tomographic reconstruction of molecular orbitals⁵. Here, we present an extension of the tomography approach to systems that cannot be naturally aligned. We demonstrate this ability by probing the two-dimensional properties of atomic wavefunctions. By manipulating an electron-ion recollision process⁶, we are able to resolve the symmetry of the atomic wavefunction with high contrast.

The basic route to spatially probe a molecular orbital involves four main steps⁵. First, the orbital axis is aligned in the laboratory frame⁷, this being accomplished by aligning the molecule. Second, an electron is ionized by a strong laser field through tunnelling ionization, after which the electron oscillates in the laser electric field and may recollide with the parent ion. Third, if the electron recollides, the recollision projects the ground state into the spatial frequencies that compose the free-electron wavefunction. The whole process, induced during less than one optical cycle, leads to the emission of extreme ultraviolet pulses with attosecond duration⁸. The projected ground state is therefore obtained from the broadband spectrum of the emitted pulses. In the final step, the molecule is aligned at different angles to the recolliding electron momentum, which permits tomographic reconstruction of the orbital.

Despite its importance and generality, it has been impossible to extend this approach, known as orbital tomography, beyond simple molecular orbitals such as the N₂ highest occupied molecular orbital. There are two fundamental limitations. The first arises from the coupling between ionization and recollision. When the molecule is rotated, both the tunnelling and recollision probabilities can be very strongly modulated^{9,10}. This couples the angle dependence of tunnelling, recollision and recombination in the harmonic spectrum. These processes must be disentangled before tomography can be extended beyond sigma orbitals (where tunnelling is relatively insensitive to angle). The second limitation arises from the requirement to fix the orbital in the laboratory frame—using molecular alignment. Tomography cannot resolve degenerate orbitals that are not fixed within a molecular structure, or molecules that are difficult to align.

We overcome both limitations, generalizing tomography to atoms and by extension, to degenerate molecular orbitals. In addition, by removing the necessity to rotate the molecule, we remove the deleterious effects of tunnelling, without affecting its benefits. In fact, in the future aligning molecules will provide a new use. Alignment will enable a specific orbital to be selected for study from among a set of ionizing orbitals in complex molecules.

There are two key steps in our approach. Tunnel ionization selects the wavefunction to be probed¹¹. Next, we manipulate the two-dimensional trajectory of free electrons. Such manipulation is achieved by adding a second-harmonic field polarized orthogonally to the fundamental field. By scanning the relative delay between the two colours, we control the angle between ionization and recollision^{12–14}. To measure the angle, we exploit the harmonic spectrum itself. We establish that for a spherically symmetric reference atom, the angle is determined by the relative strength of the even and odd harmonics. Finally, we apply this scheme to probe the spatial properties of the *p* state in neon atoms.

Manipulation of the free electron using a two-colour field is illustrated in Fig. 1a. The fundamental field is polarized along the *x* axis, whereas the second-harmonic field is polarized along the *y* axis. Tunnel ionization occurs along the instantaneous electric field direction. The free electron accelerates in the electric field and is then driven to recollide with the parent ion at an angle θ . The motion of the electron is schematically described by the blue dashed line. If the ground state is spherically symmetric then the pulse emitted during each half cycle of the laser field will be polarized along the recollision direction (purple arrow).

We generate attosecond pulses with a multi-cycle pulse and therefore repeat the process at each half cycle of the laser field¹⁵. From symmetry considerations, attosecond pulses induced by the negative half cycle are polarized along the $\pi - \theta$ direction. The pulses interfere in the high-harmonic spectrum. Owing to the periodicity of the laser field, the interference can be described as:

$$E_{n\omega_0} \propto E_x \hat{x} + E_y \hat{y} - e^{-in\pi} (E_x \hat{x} - E_y \hat{y}) \quad (1)$$

where *n* is the harmonic number and $E_x = E \cos(\theta)$ and $E_y = E \sin(\theta)$ are the extreme-ultraviolet field's projections along the \hat{x} and \hat{y} axes, respectively. Equation (1) shows that odd and even harmonics are orthogonally polarized and depend on the recollision angle according to:

$$E_{\text{odd}} = E_x \hat{x} = E \cos(\theta) \hat{x} \quad E_{\text{even}} = E_y \hat{y} = E \sin(\theta) \hat{y} \quad (2)$$

The second-harmonic field leads to symmetry breaking between adjacent half cycles. Therefore, even harmonics are polarized along the second-harmonic field polarization \hat{y} , whereas odd harmonics are polarized along the fundamental field polarization \hat{x} . Relying on the symmetry properties of the interaction, the polarization of the attosecond pulse can be analysed in a single measurement. Furthermore, as odd and even harmonics take two different

¹Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel, ²CELIA, Université Bordeaux I, UMR 5107 (CNRS, Bordeaux 1, CEA), 351 Cours de la Libération, 33405 Talence Cedex, France, ³National Research Council of Canada, 100 Sussex Drive, Ottawa, Ontario K1A 0R6, Canada. *e-mail: Paul.Corkum@nrc.ca; fnirtd@wisemail.weizmann.ac.il.

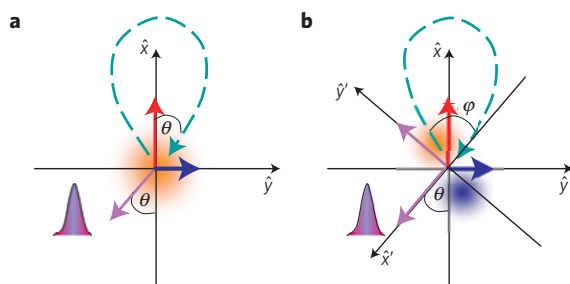


Figure 1 | Schematic diagrams of high-harmonic generation in the two colour fields. **a**, Spherically symmetric ground state. The red and blue arrows correspond to the fundamental and second-harmonic fields in the laboratory frame (\hat{x}, \hat{y}), respectively. Recollision with the ion occurs at an angle θ . Extreme-ultraviolet emission is indicated by the purple arrow and is polarized along the recollision angle. **b**, The recollision is induced by a p state. The angle φ is defined as the angle between ionization and recollision. The probing is read in the free electron's frame (\hat{x}', \hat{y}'). The extreme-ultraviolet polarization is composed of two vectorial components, polarized along \hat{x}' and \hat{y}' , respectively.

projections of the electric field vector, their relative intensities reflect the recollision angle itself.

The free-electron trajectory is determined by the coherent superposition of fundamental and second-harmonic fields. As we change the two fields' delay, we modify significantly both the angle of recollision and the lateral displacement of the electron from its parent ion as it recollides. For some delays, the electron is shifted by the field such that it misses the atom. In this case, the recollision probability and therefore the emitted signal intensity are significantly reduced¹². This mechanism is closely related to the reduction of the harmonic signal induced by an elliptically polarized single colour field¹⁶.

We calibrate the angle of recollision, θ , by measuring high harmonics from the spherically symmetric $1s$ state of helium atoms (see Fig. 2a). The experimental set-up is described in the Methods section. Low-signal areas (dark areas) correspond to a large displacement of the free electron from the parent ion. Owing to the symmetry breaking induced by the second-harmonic field, the spectrum contains both even and odd harmonics. A more careful examination shows that in the low-energy regime, odd harmonics disappear, whereas in the high-frequency regime even harmonics disappear. This spectral response indicates that the polarization of the high harmonics changes markedly across the harmonic spectrum. More importantly, according to equation (2), such modification results from a large variation of the recollision angle.

The range of recollision angles is controlled by the two fields' intensities and relative delay. However, we do not need to know these parameters precisely as the symmetry of the ground state enables us to measure them directly. We assume that the emitted electric field polarization changes slowly as a function of the harmonic number. Applying equation (2), we can directly extract the recollision angle. A detailed analysis is described in the Methods section. Figure 2b presents the measured recollision angle as a function of the harmonic number and the two fields' delay. At the low-frequency regime, the recollision angle exceeds 60° . In this regime, the large values of recollision angle are accompanied by large error bars. The angle changes markedly as a function of the harmonic order and approaches zero in the high-frequency regime. According to the strong-field approximation¹⁷ (SFA), the low-frequency regime corresponds to short electron trajectories. In these trajectories, the electron is born when the fundamental field is rather low; therefore, its dynamics are dominated by the orthogonally polarized second-harmonic field. The high-energy regime, which corresponds to longer electron trajectories,

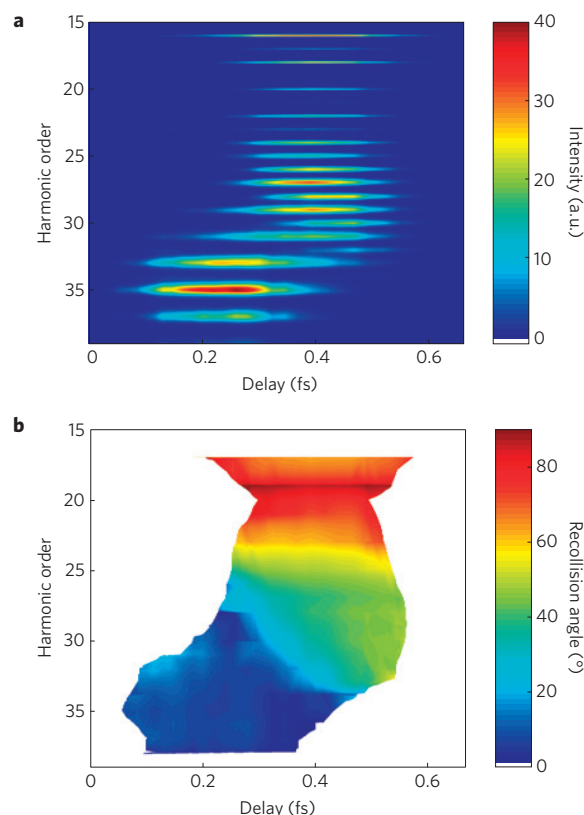


Figure 2 | Calibration of the recollision angle. **a**, High-harmonic spectrum from helium atoms as a function of the delay between the two colours. **b**, Measured electron recollision angle in the laboratory frame.

is dominated by the strong \hat{x} -polarized fundamental field and therefore results in low recollision angles.

At this stage, having measured the angle between the electric field direction of the fundamental beam and the recollision angle, we replace the spherically symmetric ground state by a more complex wavefunction. Although the dynamic range is limited, Fig. 2b shows that we can probe the wavefunction from different angles, obtaining the information needed to identify the orbital^{1–5,13}.

Figure 1b illustrates the process for ground states that have more complex symmetries (for example, the p state). Although the ionization angle did not select a unique orbital in the spherically symmetric case, it has an important role for more complex symmetries. Tunnel ionization occurs along the instantaneous electric field direction. Tunnelling theory¹⁸ suggests that in mixed p states of closed-shell atoms, the state parallel to the electric field is more efficiently ionized^{11,19} creating an effective alignment of the atomic wavefunction. In the figure, we placed the orbital along the polarization of the field at the time of ionization and the angle φ is defined as the angle between ionization and recollision. φ changes with the electron's trajectory length and therefore with the harmonic number. In Fig. 2b, we have determined the recollision angle, θ ; however, the angle φ is an unknown parameter in our experiment. It is uniquely related to θ in the SFA. It serves, in fact, as the effective probing angle. If the quantization axis is not modified during the optical cycle, we probe the atomic wavefunction that was selected by the tunnel ionization process. Therefore, our mechanism enables us to carry out a two-dimensional probing of the pre-selected wavefunction.

A convenient frame to read the pre-selected wavefunction is the probe's frame, or in our case, the free electron's frame. We can easily transfer the measurement from the laboratory frame to the free-electron frame by rotating the original set of axes \hat{x}, \hat{y}

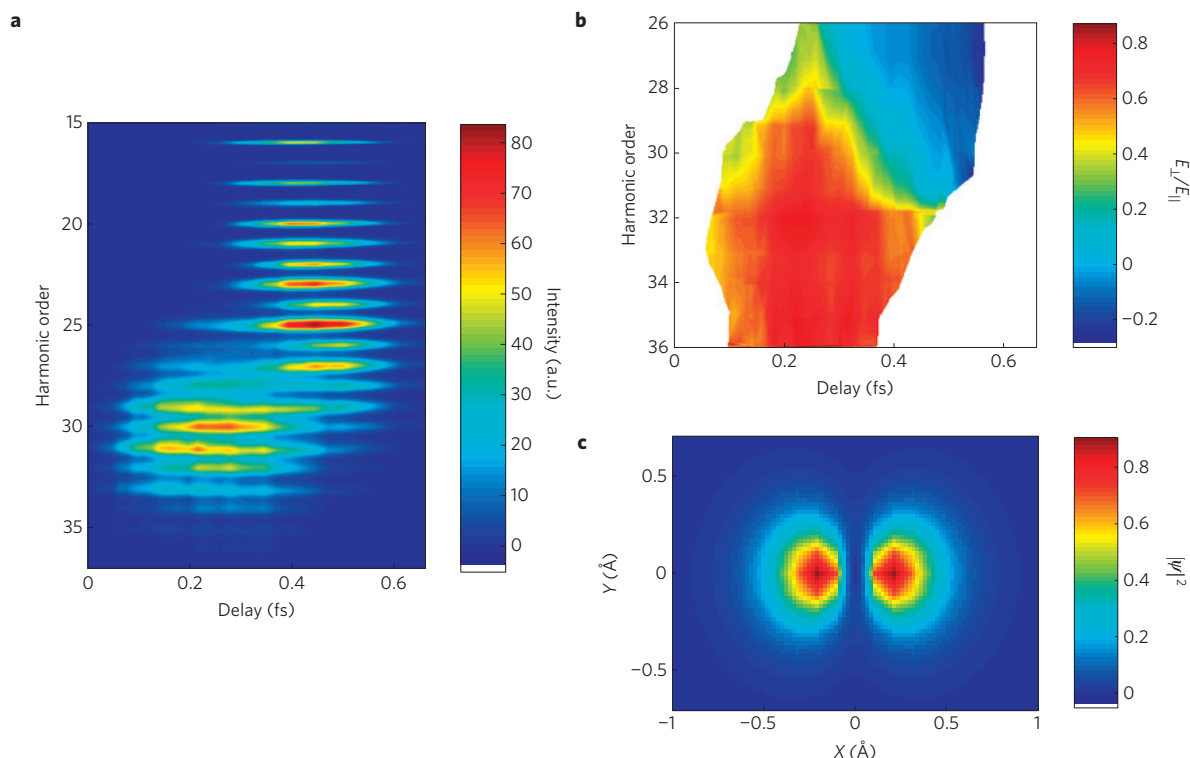


Figure 3 | Probing the atomic wavefunction in neon. **a**, High-harmonic spectrum from neon gas. One major difference from the helium spectrum is that here, even harmonics appear at high energies as well. **b**, The ratio E_{\perp}/E_{\parallel} measured in the electron recollision frame for neon. **c**, Retrieved neon mixed $2p$ orbital ($|\Psi|^2$) with a p_x amplitude of 0.95.

by an angle θ , illustrated in Fig. 2. The new frame is represented in Fig. 1b by the axes \hat{x}' and \hat{y}' , directed along and perpendicular to the re-collision angle, respectively. The polarization of the emitted pulses is described as being composed of two components: E_{\parallel} along \hat{x}' and E_{\perp} along \hat{y}' . The polarization components are dictated by the two matrix elements: $E_{\parallel} \propto \langle \Psi_g(\varphi) | \mathbf{x}' | \exp(ik_{\omega}x') \rangle$, $E_{\perp} \propto \langle \Psi_g(\varphi) | \mathbf{y}' | \exp(ik_{\omega}x) \rangle$, where $\Psi_g(\varphi)$ is the pre-selected ground state, aligned along φ . The vectorial properties of the emitted radiation reflect the wavefunction's symmetry. This idea has been recently demonstrated by measuring the polarization state of the high harmonics with aligned molecules²⁰. In symmetric atomic wavefunctions, $E_{\perp} \equiv 0$. Generalizing equation (2), we describe the spectral components as:

$$\begin{aligned} E_{\text{odd}}(\omega) &= -E_{\parallel} \cos(\theta) + E_{\perp} \sin(\theta) \\ E_{\text{even}}(\omega) &= -E_{\parallel} \sin(\theta) - E_{\perp} \cos(\theta) \end{aligned} \quad (3)$$

where both even and odd harmonics contain a coherent addition of the two polarization components.

A natural candidate that lacks inversion symmetry is neon with its highest occupied orbital being a $2p$ state. Figure 3a describes the high-harmonic spectrum as a function of the two colours' delay. As in the helium experiment (see Fig. 2), we measure both odd and even harmonics. Comparing the two measurements, we observe a striking difference. Whereas even harmonics measured from helium have disappeared in the high-frequency regime, in the neon measurement they are as strong as the odd harmonics.

The neon spectrum was separated into odd- and even-harmonic spectra and interpolated. We now rely on the recollision angle calibration (θ) to transfer the measurement from the laboratory frame (\hat{x}, \hat{y}) to the recolliding electron's frame (\hat{x}', \hat{y}'). The relative signs of E_{odd} and E_{even} were chosen by theoretical considerations. Each point in the neon even/odd spectrum was rotated by the

corresponding angle. The frequency scale of the θ measurement was shifted by 3.03 eV to compensate for differences in ionization potentials. Using equation (3), we extract the ratio E_{\perp}/E_{\parallel} . Figure 3b presents the analysed ratio, as a function of the harmonic number and the two fields' delay.

The extracted ratio reflects the symmetry of the probed wavefunction as seen by the free electron. This is equivalent to the polarization rotation observed with aligned molecules²⁰. There are two conditions required to probe symmetries that are more complex than the spherically symmetric one. The first requires that tunnel ionization selects a preferable quantization axis¹¹. The second is a large angle between the quantization axis and the probing direction. The measured ratio approaches one in the high-frequency range, which indicates that these two conditions are met in this regime. The first is reached by selective ionization, and the second through manipulation of the free-electron trajectory¹⁴. Semiclassical simulation shows that in this spectral regime, large φ values are indeed obtained.

Finally, we analysed the measured ratio E_{\perp}/E_{\parallel} to resolve the selected orbital. Specifically, we compared the measured symmetry with the theoretically calculated one to resolve the degree of selectivity. Using a fitting algorithm, we extracted the relative population of each of the three orthogonal p states. A detailed analysis is described in the Methods section. The wavefunction was found to be almost a pure aligned state, with a relative amplitude weight of 0.95 ± 0.05 . This measurement agrees with tunnelling theory, which predicts a relative amplitude weight of 0.98 (ref. 18). Figure 3c describes the density function of the selected orbital.

We have demonstrated a new concept in orbital tomography—the electronic structure is probed in a correlated manner. In our experiment, the electronic wavefunction is not well defined in the laboratory frame. Its quantization axis rotates within the optical cycle, following the instantaneous direction of the electric field. However, the electronic wavefunction is strongly correlated with the

free electron and therefore well defined in its frame. In other words, we measure a field-selected wavefunction, which provides a direct insight into the outcome of tunnelling ionization—one of the most fundamental processes in strong-field light–matter interactions. We have demonstrated that the probed wavefunction is selected by the ionization step and preserves its orientation during the optical cycle.

In our experiment, the range of probing angles was limited. However, interchanging the role of the fundamental and second-harmonic fields enables all probing angles—for example, a strong 400 nm field combined with a weaker 800 nm field. With sufficient angular information, tomographic reconstruction of atomic orbitals will be achieved.

Although this letter focuses on the measurements of atomic wavefunctions, our approach will have a significant impact on molecular tomography. First, previously inaccessible orbitals, such as degenerate states or orbitals in molecules with small polarizability can now be probed. Second, as it is no longer necessary to rotate the molecules relative to the ionizing field, we decouple tunnelling and recombination. Alignment can now serve a new and even more valuable purpose than before. We can use the sensitivity of tunnelling and recollision to select any orbital that preferentially ionizes at a specific alignment angle. For example, in CO₂ the Σ_g orbital preferentially ionizes at 45° (ref. 9), the Π_u orbital at 90° and the Σ_u orbital at 0°. Once selected, with our two-colour approach, each could be individually imaged. Finally, if ionization preferentially selects an orientation in heteronuclear molecules²¹, then these orbitals can be imaged without orienting the molecule. As orbital tomography is extended to a wide range of systems, the potential for dynamic imaging that is inherent in the technology becomes increasingly significant.

Our approach is not limited to the measurement of static wavefunctions. The high temporal resolution provided by the free electron can be combined with the spatial properties of the measurement. Dynamics that occurs between ionization and recollision will be probed by the free electron with sub-cycle resolution. Such dynamics include for example, spin–orbit coupling²², sub-cycle Stark shifts or multi-electron dynamics in molecules²³.

Methods

Experimental methods. High harmonics are generated with 30 fs, 50 Hz, 2 mJ, 800 nm laser pulses in an atomic gas jet. We estimated the pulse intensity to be 1.8×10^{14} W cm⁻² according to the cutoff harmonic. The second-harmonic field (on a 25% intensity level) is produced using a 100 μ m type-I BaB₂O₄ crystal. The second-harmonic field is orthogonally polarized with respect to the fundamental field. Group-velocity dispersion is compensated using a birefringent crystal (calcite). The phase of the second-harmonic field relative to the fundamental field is controlled with 250 μ m of BK7 glass. High harmonics are generated by focusing the two beams into a pulsed gas jet. The harmonic spectrum is measured by an extreme-ultraviolet spectrometer. The experimental set-up is described in more detail in ref. 24.

Propagation effects can have an important role in this experiment, as in many other experiments that use the free electron as a temporal or spatial probe^{1–5,20}. These effects can be minimized by carefully choosing the focal parameters (see a detailed analysis in ref. 25). We minimize propagation effects by choosing the focal parameters such that the jet length (~ 1 mm) is short compared with the Rayleigh length (~ 2 cm). Furthermore, our approach provides a robust measurement of the free electron's dynamics. The measurement is based on the ratio between even to odd harmonics rather than the absolute signal. By measuring the ratio, we eliminate the sensitivity of the experiment to different parameters such as the detector efficiency or propagation effects.

Extracting the recollision angle. To extract the recollision angle, we assume that in the absence of electronic resonances the dipole moment changes slowly as a function of the harmonic number. On the basis of this assumption, odd and even harmonics were separated and the area between adjacent harmonics was interpolated. We corrected the measured spectra according to the spectrometer efficiencies for the two polarizations. Using equation (2) and relying on the slow variation of the dipole moment, we can directly extract the recollision angle:

$$\tan(\theta)(n, \tau) = \sqrt{\frac{I_{\text{even}}(n, \tau)}{I_{\text{odd}}(n, \tau)}}$$

where I_{even} and I_{odd} are the even- and odd-harmonic intensities respectively, n is the harmonic order and τ is the two colours' delay. This analysis provides an independent measurement of the recollision angle that does not rely on theoretical modelling of the free electron's dynamics.

Resolving the atomic wavefunction. We assume that the atomic ground state in neon is composed of three orthogonal p states

$$\psi_g(x, y, z) = \epsilon_x p_x + \epsilon_y p_y + \epsilon_z p_z$$

where \bar{z} is the propagation axis of the laser beam, \bar{x} and \bar{y} are directed along and perpendicular to the ionization axis respectively, and $\epsilon_{x,y,z}$ are the amplitude coefficients of each of the p states. The free electron recollides with p_z with an angle of 90°. As it is symmetric along the \bar{z} axis, its overlap with the antisymmetric p_z wavefunction cancels out and does not contribute to the harmonics signal. p_y and p_x are quantized orthogonally to the ionization axis and we therefore assume that they carry the same weight. Our reconstruction procedure resolves a single parameter defined as: $s \equiv \epsilon_y/\epsilon_x$. Pure selectivity corresponds to $s = 0$, whereas zero selectivity corresponds to $s = 1$.

In Fig. 3b, we present the measured ratio E_{\perp}/E_{\parallel} . This ratio reflects the symmetry of the probed wavefunction and therefore also of the population ratio s . We resolve s by comparing our experiment with a theoretical calculation. We calculate the ratio E_{\perp}/E_{\parallel} theoretically as a function of s . The atomic wavefunctions were calculated using a standard Hartree–Fock *ab initio* program GAMESS (ref. 26). The probing angles (φ) were calculated according to classical electron's trajectories using SFA. The parameters of the experiment such as the two fields' delay and the relative fields' intensities are calibrated by the helium experiment presented in Fig. 2. The two dipole components are determined according to: $E_{\parallel} \propto \langle \Psi_g(\varphi, s) | \mathbf{x}' | \exp(ik_{\parallel} \mathbf{x}') \rangle$, $E_{\perp} \propto \langle \Psi_g(\varphi, s) | \mathbf{x}' | \exp(ik_{\perp} \mathbf{x}') \rangle$. The ratio $R(s, \tau, n) \equiv E_{\perp}/E_{\parallel}$ is evaluated as a function of the population ratio s , the two fields' delay τ and the harmonic number n . The comparison between theory and experiment is expressed as:

$$\Delta(s) = \sum_{\tau} \sum_n |R(s, \tau, n)_{\text{theory}} - R(\tau, n)_{\text{experiment}}|$$

$\Delta(s)$ represents the integrated difference between a theoretical calculation, based on selectivity s , and the experimental results. Finally, we extract s by minimizing $\Delta(s)$ with respect to s and find that $s = 0.22 \pm 0.15$, that is, $\epsilon_x = 0.95$, $\epsilon_y = \epsilon_z = 0.21$. We conclude that the wavefunction is almost a pure aligned state, which can be expressed as $|\Psi_g(x, y, z)|^2 = |0.95 \times p_x|^2 + |0.21 \times p_y|^2 + |0.21 \times p_z|^2$.

Received 22 January 2008; accepted 18 March 2009;
published online 26 April 2009

References

- Lein, M., Hay, N., Velotta, R., Marangos, J. P. & Knight, P. L. Role of the intramolecular phase in high-harmonic generation. *Phys. Rev. Lett.* **88**, 183903 (2002).
- Vozzi, C. *et al.* Controlling two-center interference in molecular high harmonic generation. *Phys. Rev. Lett.* **95**, 153902 (2005).
- Kanai, T., Minemoto, S. & Sakai, H. Quantum interference during high-order harmonic generation from aligned molecules. *Nature* **435**, 470–474 (2005).
- Torres, R. *et al.* Probing orbital structure of polyatomic molecules by high-order harmonic generation. *Phys. Rev. Lett.* **98**, 203007 (2007).
- Itatani, J. *et al.* Tomographic imaging of molecular orbitals. *Nature* **432**, 867–871 (2004).
- Corkum, P. B. Plasma perspective on strong field multiphoton ionization. *Phys. Rev. Lett.* **71**, 1994–1997 (1993).
- Stapelfeldt, H. & Seideman, T. Aligning molecules with strong laser pulses. *Rev. Mod. Phys.* **75**, 543–557 (2003).
- Corkum, P. B. & Krausz, F. Attosecond science. *Nature Phys.* **3**, 381–387 (2007).
- Pavicic, D. *et al.* Direct measurement of the angular dependence of ionization for N₂, O₂, and CO₂ in intense laser fields. *Phys. Rev. Lett.* **98**, 243001 (2007).
- Meckel, M. *et al.* Laser induced electron tunneling and diffraction. *Science* **320**, 1478–1482 (2008).
- Young, L. *et al.* X-ray microprobe of orbital alignment in strong-field ionized atoms. *Phys. Rev. Lett.* **97**, 083601 (2006).
- Kim, I. J. *et al.* Highly efficient high-harmonic generation in an orthogonally polarized two-color laser field. *Phys. Rev. Lett.* **94**, 243901 (2005).
- Kitzler, M. & Lezius, M. Spatial control of recollision wave packets with attosecond precision. *Phys. Rev. Lett.* **95**, 253001 (2005).
- Kitzler, M., Xie, X., Scrinzi, A. & Baltuska, A. Optical attosecond mapping by polarization selective detection. *Phys. Rev. A* **76**, 011801 (2007).
- Antoine, P., L'Huillier, A. & Lewenstein, M. Attosecond pulse trains using high order harmonics. *Phys. Rev. Lett.* **77**, 1234–1237 (1996).

16. Budil, K. S., Salières, P., Perry, M. D. & L'Huillier, A. Influence of ellipticity on harmonic generation. *Phys. Rev. A* **48**, R3437–R3440 (1993).
17. Lewenstein, M., Balcou, P., Ivanov, M. Y., L'Huillier, A. & Corkum, P. B. Theory of high-harmonic generation by low-frequency laser fields. *Phys. Rev. A* **49**, 2117–2132 (1994).
18. Ammosov, M. V., Delone, N. B. & Krainov, V. P. Tunnel ionization of complex atoms and of atomic ions in an alternating electromagnetic field. *Sov. Phys. JETP* **64**, 1191–1194 (1986).
19. Otobe, T., Yabana, K. & Iwata, J. I. First-principles calculations for the tunnel ionization rate of atoms and molecules. *Phys. Rev. A* **69**, 053404 (2004).
20. Levesque, J. *et al.* Polarization state of high-order harmonic emission from aligned molecules. *Phys. Rev. Lett.* **99**, 243001 (2007).
21. Bandrauk, A. D. & Kamta, G. L. Phase dependence of enhanced ionization in asymmetric molecules. *Phys. Rev. Lett.* **94**, 203003 (2005).
22. Santra, R., Dunford, R. W. & Young, L. Spin–orbit effect on strong-field ionization of krypton. *Phys. Rev. A* **74**, 043403 (2006).
23. Lezius, M. *et al.* Nonadiabatic multielectron dynamics in strong field molecular ionization. *Phys. Rev. Lett.* **86**, 51–54 (2001).
24. Dudovich, N. *et al.* Measuring and controlling the birth of attosecond XUV pulses. *Nature Phys.* **2**, 781–786 (2006).
25. Levesque, J. *et al.* High harmonic generation and the role of atomic orbital wave functions. *Phys. Rev. Lett.* **98**, 183903 (2007).
26. Schmidt, M. W. *et al.* General atomic and molecular electronic structure system. *J. Comput. Chem.* **14**, 1347–1363 (1993).

Acknowledgements

We thank A. Shiner, C. Trallero, N. Kajumba and F. Légaré for carrying out pressure scan experiments to show that phase mismatch effects were negligible.

Additional information

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>. Correspondence and requests for materials should be addressed to N.D. or P.B.C.

Signatures of universal four-body phenomena and their relation to the Efimov effect

J. von Stecher, J. P. D'Incao and Chris H. Greene*

The problem of three interacting quantal bodies, in its various guises, seems deceptively simple, but it has also provided striking surprises, such as the Efimov effect^{1,2}, which was confirmed experimentally³ only more than 35 years after its initial prediction. The importance of understanding the three-body problem was magnified by the explosion of ultracold science following the formation of Bose–Einstein condensates in 1995 (ref. 4). For ultracold gases, three-body recombination (where $B + B + B$ collide to form $B_2 + B$) was quickly recognized as the main loss process and connected^{5–8} with the Efimov effect in the ‘universal’ realm of very large atom–atom scattering lengths a . The problem of four interacting bodies challenges theory far more than the three-body quantal problem. Some key insights have been achieved in recent years^{9–16}. Here, we present a major extension of our understanding of the four-body problem in the universal large- a regime. Our results support a previous conjecture¹⁰ that two resonantly bound four-body states are attached to every universal three-body Efimov resonance and they improve the calculated accuracy of their universal properties. A hitherto unanalysed feature found in ultracold-gas experiments³ supports this universal prediction, and it provides the first evidence of four-body recombination (where $B + B + B + B$ form $B_3 + B$, $B_2 + B + B$ or $B_2 + B_2$).

The experiment³ that observed strong evidence for the long-predicted Efimov effect^{1,2} has spawned a new level of confidence in our theoretical understanding of the three-body problem with short-range forces. However, even though in some respects the three-body problem is beginning to seem ‘almost solved’, the next step in complexity—to four interacting particles—remains at a primitive stage, comparatively speaking. Although a few studies have been pursued^{9–12}, the non-perturbative four-boson problem is still largely uncharted territory, especially for processes that begin or end with three or four free particles. Our present study relates most to the pioneering work of Hammer, Platter and Meißner^{9,10}, and of Yamashita *et al.*¹¹. It concerns a key question in strongly interacting few-body systems: are universal principles of Efimov physics relevant for the four-boson problem? In the early 1970s, the nuclear physicist Vitaly Efimov^{1,2} predicted on very general grounds that three neutral bosons, whose mutual interaction is characterized by a large value for the two-body s -wave scattering length a , can form a large number of weakly bound states whenever $|a| \gg r_0$, where r_0 is the characteristic range of the interaction. Surprisingly, this could happen even when none of the pairs can bind ($a < 0$). Here we provide an analysis that convincingly demonstrates the existence of a class of universal four-boson states that are intimately related to the Efimov effect. Our results connect with and extend previous analyses^{9–11} and provide a more complete landscape of the universal four-boson phenomena. In addition, we demonstrate how four-boson universal states can

be seen (and in retrospect, apparently have already been seen) in ultracold-gas experiments.

Our theoretical model hinges on the tunable interaction strength achievable in ultracold-gas experiments. For alkali atoms, when an external B field is placed near a Fano–Feshbach resonance¹⁷ a small change of B can cause a to vary from $-\infty$ to ∞ , allowing for the exploration of a vast range of interatomic interaction strengths. We mimic such variations in a by explicitly modifying the interatomic interactions⁵. In our framework, the solution of the four-body problem culminates with the solution of the ‘hyper-radial’ Schrödinger equation:

$$-\frac{\hbar^2}{2m} \frac{d^2}{dR^2} F_\nu(R) + \sum_{\nu'} W_{\nu\nu'}(R) F_{\nu'}(R) = E F_\nu(R) \quad (1)$$

where the hyper-radius R describes the overall size of the system. Here, m is the atomic mass, E the total energy and $F_\nu(R)$ the hyper-radial wavefunction, with ν representing the set of quantum numbers needed to label each channel. In the hyperspherical adiabatic representation,^{5,18} most of the complexity of the problem arises in solving the hyperangular equations to determine the effective potentials $W_{\nu\nu}(R)$ and couplings $W_{\nu\nu'}(R)$, in this case using a correlated Gaussian hyperspherical approach¹⁹. The reduction of the problem to the hyper-radial Schrödinger equation (1) then leads to a simple, intuitive picture: the effective potentials $W_{\nu\nu}(R)$ support all bound and quasibound states of the system, and the off-diagonal non-adiabatic couplings $W_{\nu\nu'}(R)$ drive inelastic transitions among different channels.

We explore the universality of the four-boson system and its relation to Efimov physics by solving the Schrödinger equation (1) for different model potentials using two complementary numerical techniques: the adiabatic hyperspherical approximation^{18,20}, and the correlated Gaussian basis set expansion^{14,21}. We base our conclusions on the analysis of energies, potential curves and wavefunctions that describe ground and excited states (see Supplementary Information for details) in the universal regime $|a| \gg r_0$, where r_0 is the characteristic length scale of the two-body interaction, usually associated with the van der Waals length. Figure 1a shows the ‘generalized Efimov plot’¹⁰, with all important features that relate two-, three- and four-body physics. Figure 1a shows our numerical results for the four-boson energies (solid black lines) along with the dimer–dimer and dimer–atom–atom break-up thresholds for $a > 0$ (solid red lines) and the energies of the Efimov states (dashed green lines) representing the atom–trimer break-up threshold. Figure 1c,d conveys the geometrical nature of the four-body states taking into account the extremely ‘floppy’ nature of the Efimov trimer states, in which all possible triangular shapes, and even linear configurations, are comparably probable^{22,23}.

Figure 1a implies that the four-boson spectrum, throughout the range $r_0/|a| \ll 1$, is characterized by precisely two tetramer states

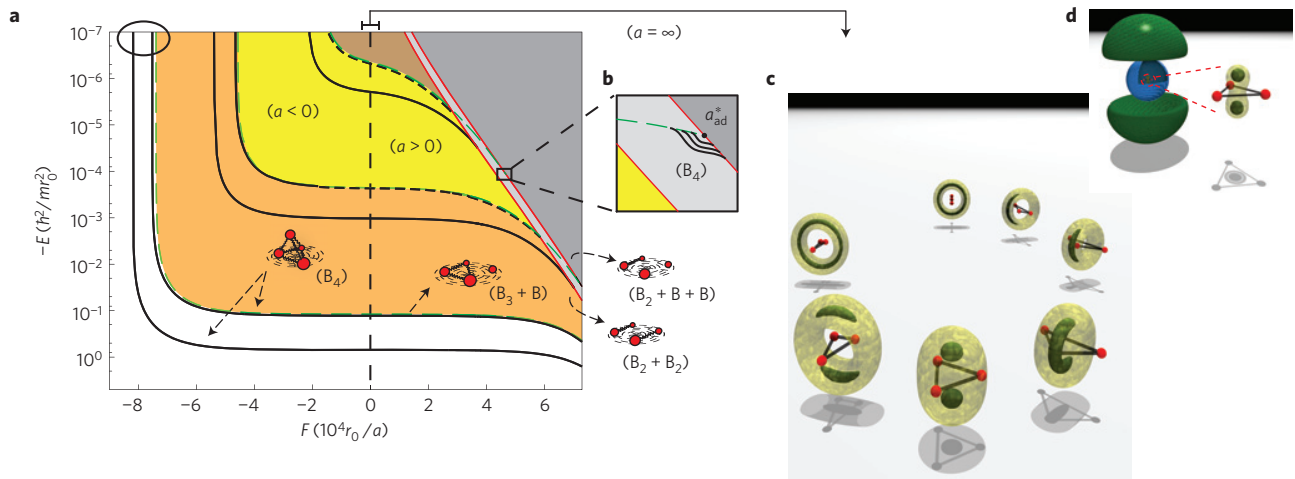


Figure 1 | Spectrum of energies and geometric structure of four-boson states and their connection to the Efimov physics. **a**, The energy spectrum of four bosons as a function of the atom-atom scattering length a using the auxiliary function $F(x) \equiv \text{sgn}(x)\ln(1+|x|)$ designed to aid the visualization of the full energy landscape. Black solid lines correspond to four-body states, dashed green lines represent atom-trimer dissociation thresholds and red lines correspond to dimer-atom-atom (upper) and dimer-dimer (lower) thresholds. **b**, Schematic description of dimer-atom-atom Efimov states. **c, d**, Geometrical nature of ground and excited four-body states, respectively, for different trimer configurations. For the four-boson ground state, the isosurfaces shown are those where the probability of finding the fourth atom is 0.9 and 0.99 of the maximum probability for that particular trimer geometry. The probability isosurface for the first excited four-boson state in the portion of space where the Efimov trimer resides at its most probable equilateral triangular geometry implies that the fourth atom is very weakly bound, and its size exceeds that of the ground state considerably.

that are associated with each three-body Efimov state, confirming the ref. 10 prediction. In fact, our extensive numerical tests show that these four-boson energies obey a universal relationship to the corresponding Efimov state energy, which at unitarity ($|a| = \infty$) can be expressed as

$$E_{4b}^{(n,m)} = c_m E_{3b}^{(n)} \quad (2)$$

where $E_{3b}^{(n)}$ is the energy of the n th Efimov state, $n = 0, 1, 2, \dots$, and $E_{4b}^{(n,m)}$, $m = 1$ and 2 , are the two tetramer energies associated with it (see Supplementary Information for details). Here, we find that the universal relation between three- and four-body energies is characterized by these two universal numbers, equal to: $c_1 \approx 4.58$ and $c_2 \approx 1.01$. For the lowest two four-boson states ref. 10 obtained $c_1 \approx 5$ and $c_2 \approx 1.01$. Similar values, less deep in the universal regime, can be extracted from the small B_2 limit of equations (39) and (41) of ref. 9. Our own calculations differ from the universal values if we consider the lowest four-body states (see Supplementary Information). In fact, we believe that the ability of the current method to calculate many more weakly-bound energy levels than previous techniques has been decisive, and it permits us to verify the universal numbers up to 2% accuracy and resolve a previously-existing disagreement in the literature, between the results of Hammer, Platter and Meißner^{9,10} and those of Yamashita *et al.*¹¹.

The resolution of this controversy is related to the fundamental question of whether or not an additional ‘four-body parameter’, encapsulating non-universal aspects from the details of the interactions, is required to specify the nature of the four-boson spectrum and scattering observables, akin to the usual three-body parameter^{1,2,5–7,24,25}. Specifically, we support the conclusions of refs 9, 10, 12 that no additional four-body parameter enters at leading order in the description of universal four-body spectra and scattering properties. Our result also explains the observations of Yamashita *et al.*¹¹ that the energies of more compact four-boson states can vary depending on the details of the interatomic interactions. Our understanding emerges from Fig. 2, showing the four-body effective potentials calculated at unitarity, $|a| = \infty$. We

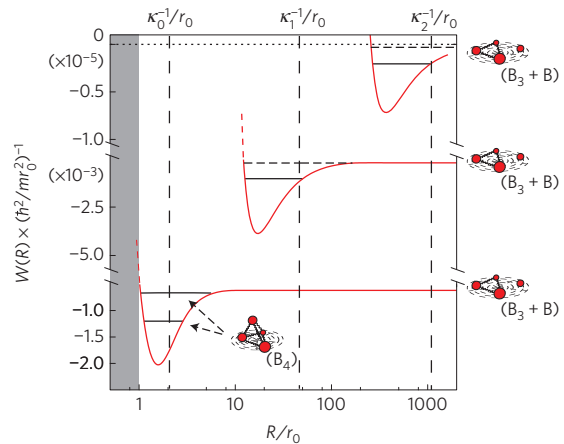


Figure 2 | Effective four-boson potentials for $|a| = \infty$ converging at large R to the atom-trimer thresholds. Solid and dashed black lines represent the four-boson states shown in Fig. 1a. The position in R of the minimum of these potentials scales with the size of the Efimov state, indicated in the figure by κ_i^{-1} (see Supplementary Information for definition). Therefore, when the trimer state is large the four-boson states associated with it will lie at large R , preventing access to the non-universal region $R \lesssim r_0$, whereby the four-boson states are universal.

have verified that the effective-potential curves scale with the size of the trimer state. Therefore, if the lowest Efimov state has a size that exceeds r_0 only marginally, the minimum of the four-body potential is close enough to r_0 and its four-body states are affected by the shape of the two-body interaction, that is, non-universal physics. On the other hand, if the lowest Efimov state is large compared with r_0 , then the minimum of the four-body potential lies at $R \gg r_0$ and the four-boson states probe almost no non-universal effects. Furthermore, the scaling behaviour of the potential curves implies that the three-body Efimov effect controls the four-boson spectrum. As a consequence, the four-boson states follow the same geometric scaling as the three-boson states, with successive energies related by the factor $e^{-2\pi/s_0} \approx 1/515$ and successive radii expanding by the

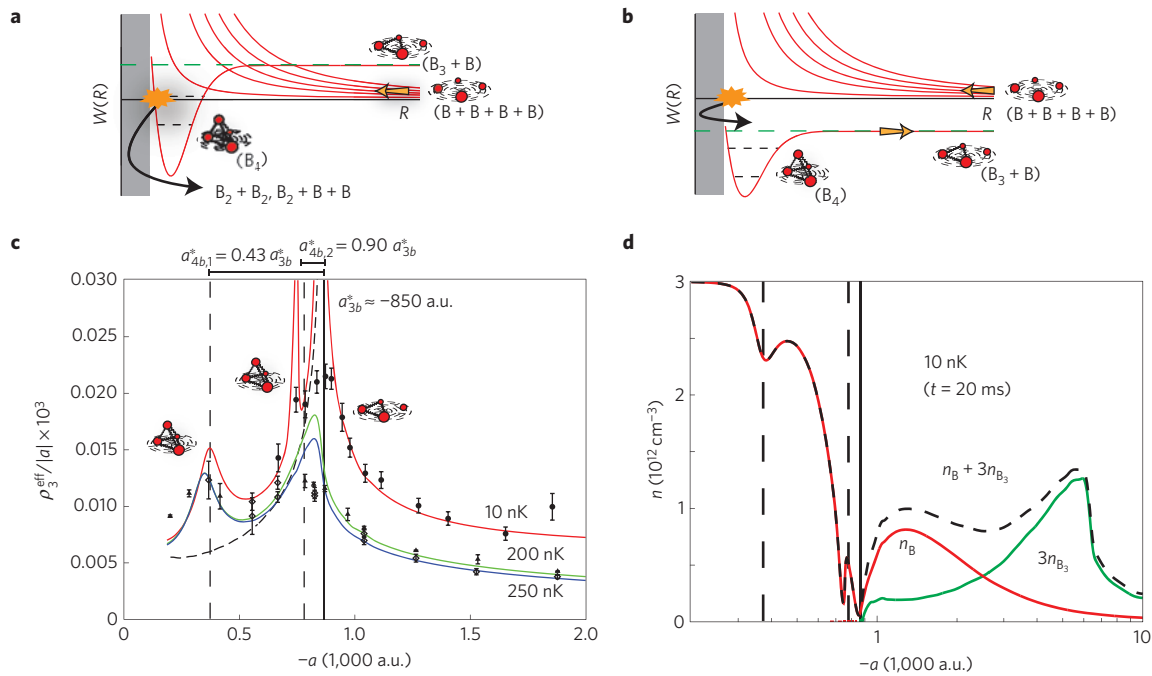


Figure 3 | Evidence of signatures of four-boson states through four-body recombination. **a, b**, Schematic representation of the pathways for four-body recombination when the Efimov state (green dashed line) is above and below the four-body break-up threshold, respectively. In **a** the only possible decay channels are associated with deeply bound (two- and three-body) states, whereas in **b** the main decay channel is to the available Efimov state plus a free atom. When a four-boson state crosses the four-body dissociation threshold, it resonantly affects the four-body recombination rate, enhancing the rate of atom losses in an ultracold gas. **c**, Comparison of ρ_3^{eff} (equation (4)) with the experimental data from ref. 3. The shown error bars refer to statistical uncertainty only (see ref. 3 for details). The vertical lines identify the critical scattering lengths where an Efimov state and its associated four-body states are created, respectively, $a_{3b}^* \approx -850$ a.u., $a_{4b1}^* \approx 0.43a_{3b}^* \approx -370$ a.u. and $a_{4b2}^* \approx 0.90a_{3b}^* \approx -770$ a.u. The dashed curve is the 10 nK contribution from K_3 . **d**, The atom and trimer densities at 10 nK, demonstrating that four-body recombination can be used to form Efimov trimers.

factor $e^{\pi/s_0} \approx 22.7$ (where $s_0 \approx 1.00624$). These four-boson states are not true bound states, of course, as was pointed out in ref. 10, so the preceding discussion relates to the real part of their energies.

Parenthetically, we also confirm the existence of a class of four-body states that represent the Efimov effect for three bodies. As pointed out in ref. 8, these four-atom states occur whenever an Efimov state is created for $a > 0$ (see Fig. 1b). In this case the atom–dimer scattering length, a_{ad} , has a pole and whenever $|a_{\text{ad}}| \gg a$ a series of four-atom, three-body Efimov states, namely dimer–atom–atom states. Numerically, we observe the emergence of an attractive dipole potential ($\propto -1/R^2$) in the dimer–atom–atom channel, for $a \ll R \ll |a_{\text{ad}}|$, confirming the existence of such four-boson states. In fact, the recent observation of an Efimov resonance in ref. 26 in an ultracold atom–molecule mixture could enable the probing of such four-body states experimentally. We also confirm the non-existence of a ‘true Efimov effect’ for four bosons in the spirit of the ref. 27 prediction (see Supplementary Information).

The universal four-body physics discussed above can readily be observed in ultracold quantum gases. In general, weakly bound states deeply affect the collisional properties of ultracold gases, enhancing the atomic and/or molecular losses. The relative importance of four-body processes, however, remains largely unexplored, and we could argue that such processes should be far less likely than two-body or three-body collisions in a typical low-density gas. On the other hand, near the threshold for formation of any four-boson states, the four-body scattering observables should show a resonant enhancement that dramatically affects the collisional behaviour of the gas, even at low densities. The results of the Innsbruck experiment³, realized at atomic densities of $n(0) \approx 3 \times 10^{12} \text{ cm}^{-3}$, were interpreted under the premise that the atom loss stems just from three-body recombination,

$B + B + B \rightarrow B_2 + B + E_{\text{rel}}$, which releases enough kinetic energy to eject the collision products. Sure enough, the experimental data show a resonant peak in the three-body recombination rate K_3 , more specifically at a three-body recombination length of $\rho_3 = [2mK_3/(\sqrt{3}\hbar)]^{1/4}$, at $a = -850$ a.u., in agreement with theoretical expectations for the manifestation of Efimov physics through three-body recombination^{5–7}.

Although the experimental Innsbruck data³ are reasonably well understood^{24,28–30}, distinguishing three- and four-body losses is difficult, and four-body processes could still be embedded in the observed decay rates. Accordingly, we have reanalysed the Innsbruck data³, looking for possible signatures of new four-boson states. The key observation from our results is that for $a < 0$, when an Efimov state is created, say at $a = a_{3b}^*$, it is accompanied by the creation of two four-boson states at slightly less negative values of a , and those states should enhance four-body processes in this experimentally explored region of a . Moreover, our calculations indicate that, once the scattering length a_{3b}^* is known, we know the scattering lengths at which such four-body states appear. This universal relation is determined by the energy spectrum (see Fig. 1a), namely

$$a_{4b1}^* \approx 0.43a_{3b}^*, \quad \text{and} \quad a_{4b2}^* \approx 0.90a_{3b}^* \quad (3)$$

From our numerical calculations we have found that these relations are approximately fulfilled even when a_{3b}^* is not deep in the universal regime, suggesting an insensitivity to non-universal effects.

The main process where such states should appear is four-body recombination, where the four initially free atoms collide to recombine into the dimer–dimer channel, the dimer–atom–atom channel and/or the atom–trimer channel. Figure 3a,b, respectively,

depicts this process through the effective potentials at scattering lengths very close to the threshold a_{3b}^* for formation of an Efimov state (green dashed line) and just past the point ($|a| < |a_{3b}^*|$) of its creation. When a four-boson state resides energetically close to the collision threshold, we expect a resonant enhancement to the four-body recombination rate, K_4 . A straightforward Wigner threshold-law analysis demonstrates that K_4 approaches a constant as the collision energy is tuned to zero^{16,31}, and thus four-body recombination can indeed potentially compete with three-body recombination in causing atomic losses.

To assess the importance of K_4 and quantify our predictions, we have calculated K_4 by numerically solving equation (1) using a formula for K_4 derived elsewhere³¹. The main difficulty in comparing our results with data is that existing experiments are probably unable to distinguish three- and four-body losses. We therefore introduce an effective three-body recombination rate, in which both three- and four-body physics are included:

$$K_3^{\text{eff}}(a, t) = K_3(a) + n(t)K_4(a)/3 \quad (4)$$

where $n(t)$ is the peak atomic density in the trap at time t , calculated by solving the time evolution rate equations. Figure 3c shows our recombination length⁵ $\rho_3^{\text{eff}} = [2mK_3^{\text{eff}}/(\sqrt{3}\hbar)]^{1/4}$ for $t = 20$ ms. For K_3 we use the thermally averaged results of ref. 24 calculated for temperatures of 10, 200 and 250 nK, and adjust it to fit the Efimov resonance at $a = a_{3b}^* = -850$ a.u. and the experimental data for $|a| > |a_{3b}^*|$. Our results show that for this range of a three-body recombination is indeed the dominant loss process. For $|a| < |a_{3b}^*|$, however, we find much better agreement by assuming that four-body recombination is the dominant loss process—the dashed curve in Fig. 3c is the 10 nK contribution from K_3 . For this range of a , as indicated in Fig. 3c by the vertical dashed lines at $a = a_{4b1}^*$ and $a = a_{4b2}^*$, K_4 is resonantly enhanced when the two four-boson states are created (see the circled region in Fig. 1a). This agreement strongly suggests that the 2006 Innsbruck experiment³ also offers the first experimental evidence for the universal four-boson states we discussed here, although the agreement with the second resonance predicted for $a = a_{4b2}^*$ and 10 nK requires some imagination—and for temperatures of 200 and 250 nK this resonance feature is washed out owing to thermal effects. Nevertheless, the verification of the universal constraint between three- and four-body physics (equation (3)) strengthens the conclusion that the main resonant feature³ at -850 a.u. is indeed an Efimov resonance.

Note also that for $|a| > |a_{3b}^*|$ four-body recombination to Efimov states, $B + B + B + B \rightarrow B_3 + B$ (see Fig. 3b), is likely to be the dominant four-body decay pathway. Although three-body recombination tends to dominate the atom loss, the formation of Efimov states through four-body recombination is non-negligible. In Fig. 3d we show the atomic density n_B and the density of trimers n_{B_3} at 10 nK for a up to $-10,000$ a.u. Near the threshold for Efimov state formation little energy is released through four-body recombination (approximately the trimer binding energy) and the Efimov state can remain trapped. In this case, our results indicate that about 10% of the atoms will form trimers. For larger $|a|$, however, the trimer formation is strongly enhanced by a resonance associated with the lowest four-boson state attached to the second Efimov trimer (see Fig. 1a). In this case, we find that about 50% of the atoms will form trimers. Here, however, the energy released through four-body recombination ejects both the atom and the Efimov trimer. Nevertheless, in an experiment where only atoms are visible, the magnetic field could be set to a value such that $|a| > |a_{3b}^*|$, the number of atoms measured and the field ramped back to a value such that $|a| < |a_{3b}^*|$, where no trimers exist. The reappearance of atoms after this ramp would be a convincing signature of the first experimental realization of an ultracold gas of Efimov trimers.

Finally, the Innsbruck group has observed two resonant loss features in four-body recombination³² that satisfy our relations in equation (3), offering stronger experimental evidence of the universal four-body physics demonstrated here.

Methods

Effective three-body recombination. The idea behind the definition of $K_3^{\text{eff}}(a, t)$ is to obtain a quantity that encapsulates both three- and four-body contributions. Accordingly to our findings, the atomic losses can indeed be seen as an effective three-body recombination within a characteristic timescale, as explained below. That can presumably facilitate a comparison with the available experimental data.

The key ingredient in the derivation of $K_3^{\text{eff}}(a, t)$ is the relation between the loss coefficient and the loss rates as given by

$$L_N = N \frac{1}{N!} K_N$$

where N is the number of atoms involved in the recombination event. Note that K_3 (or K_4) represents a fundamental few-body entity, namely, the recombination probability per unit time for a single triad (or tetrad) in a unit volume squared (cubed)¹⁷. The first factor on the right-hand side of this equation represents the number of atoms lost in the N -atom recombination process, whereas the $1/N!$ factor accounts for the indistinguishability of the collision partners. The atoms are assumed not to be in a Bose–Einstein condensate.

The rate equation that governs the time evolution of the atomic density can be written as

$$\frac{d}{dt}n(a, t) = -L_3(a)n(a, t)^3 - L_4(a)n(a, t)^4$$

or alternatively,

$$\begin{aligned} \frac{d}{dt}n(a, t) &= -[L_3(a) + n(a, t)L_4(a)]n(a, t)^3 \\ &= -L_3^{\text{eff}}(a, t)n(a, t)^3 \end{aligned}$$

from which we define the effective $L_3^{\text{eff}}(a, t)$. From this definition, and using the relation between L_N and K_N above, we can easily arrive at our definition of K_3^{eff} following the steps below:

$$L_3^{\text{eff}}(a, t) = L_3(a) + n(a, t)L_4(a)$$

$$3 \frac{L_3^{\text{eff}}(a, t)}{3!} = 3 \frac{K_3(a)}{3!} + n(a, t)4 \frac{K_4(a)}{4!}$$

$$K_3^{\text{eff}}(a, t) = K_3(a) + n(a, t)K_4(a)/3$$

We have in fact verified that for times $t < t_0 = [n(0)^2(K_3 + n(0)K_4/3)]^{-1}$ (≈ 50 ms for our case) the time dependence of $n(t)$ can be described as a result of the effective three-body rate in equation (4), by setting $n(t) = n(0)$. For longer times, $n(t)$ is affected primarily by three- or four-body processes depending on whether or not $K_3 \gg n(0)K_4$.

Received 29 October 2008; accepted 30 March 2009;
published online 26 April 2009

References

1. Efimov, V. Weakly bound states of three resonantly interacting particles. *Yad. Fiz.* **12**, 1080–1091 (1970); *Sov. J. Nucl. Phys.* **12**, 589–595 (1971).
2. Efimov, V. Energy levels of three resonantly interacting particles. *Nucl. Phys. A* **210**, 157–188 (1973).
3. Kraemer, T. *et al.* Evidence for Efimov quantum states in an ultracold gas of caesium atoms. *Nature* **440**, 315–318 (2006).
4. Anderson, M. H., Ensher, J. R., Matthews, M. R., Wieman, C. E. & Cornell, E. A. Observation of Bose–Einstein condensation in a dilute atomic vapor. *Science* **269**, 198–201 (1995).
5. Esry, B. D., Greene, C. H. & Burke, J. P. Jr. Recombination of three atoms in the ultracold limit. *Phys. Rev. Lett.* **83**, 1751–1754 (1999).
6. Nielsen, E. & Macek, J. H. Low-energy recombination of identical bosons by three-body collisions. *Phys. Rev. Lett.* **83**, 1566–1569 (1999).
7. Bedaque, P. F., Braaten, E. & Hammer, H.-W. Three-body recombination in Bose gases with large scattering length. *Phys. Rev. Lett.* **85**, 908–911 (2000).
8. Braaten, E. & Hammer, H. W. Universality in few-body systems with large scattering length. *Phys. Rep.* **428**, 259–390 (2006).

9. Platter, L., Hammer, H. & Meißner, U. Four-boson system with short-range interactions. *Phys. Rev. A* **70**, 52101 (2004).
10. Hammer, H. W. & Platter, L. Universal properties of the four-body system with large scattering length. *Eur. Phys. J. A* **32**, 113–120 (2007).
11. Yamashita, M. T., Tomio, L., Delfino, A. & Frederico, T. Four-boson scale near a Feshbach resonance. *Europhys. Lett.* **75**, 555–561 (2006).
12. Hanna, G. J. & Blume, D. Energetics and structural properties of three-dimensional bosonic clusters near threshold. *Phys. Rev. A* **74**, 063604 (2006).
13. Petrov, D. S., Salomon, C. & Shlyapnikov, G. V. Weakly bound dimers of fermionic atoms. *Phys. Rev. Lett.* **93**, 090404 (2004).
14. von Stecher, J. & Greene, C. H. Spectrum and dynamics of the BCS-BEC crossover from a few-body perspective. *Phys. Rev. Lett.* **99**, 090402 (2007).
15. D’Incao, J. P., Rittenhouse, S. T., Mehta, N. P. & Greene, C. H. Dimer–dimer collisions at finite energies in two-component Fermi gases. *Phys. Rev. A* **79**, 030501 (2009).
16. Wang, Y. & Esry, B. D. Efimov trimer formation via ultracold four-body recombination. *Phys. Rev. Lett.* **102**, 133201 (2009).
17. Köhler, T., Góral, K. & Julienne, P. S. Production of cold molecules via magnetically tunable Feshbach resonances. *Rev. Mod. Phys.* **78**, 1311–1362 (2006).
18. Macek, J. H. Properties of autoionizing states of He. *J. Phys. B* **1**, 831–843 (1968).
19. von Stecher, J. *Trapped Ultracold Atoms With Tunable Interactions*. PhD thesis, Univ. of Colorado, Boulder (2008); <<http://jilawwww.colorado.edu/pubs/thesis/vonstecher/>>.
20. Coelho, H. T. & Hornos, J. E. Proof of basic inequalities in the hyperspherical formalism for the N-body problem. *Phys. Rev. A* **43**, 6379–6381 (1991).
21. Suzuki, Y. & Varga, K. *Stochastic Variational Approach to Quantum-Mechanical Few-Body Problems*. (Springer, 1998).
22. D’Incao, J. P. & Esry, B. D. Manifestations of the Efimov effect for three identical bosons. *Phys. Rev. A* **72**, 032710 (2005).
23. Esry, B. D. & Greene, C. H. Quantum physics: A ménage à trois laid bare. *Nature* **440**, 289–290 (2006).
24. D’Incao, J. P., Greene, C. H. & Esry, B. D. The short-range three-body phase and other issues impacting the observation of Efimov physics in ultracold quantum gases. *J. Phys. B* **42**, 044016 (2009).
25. Danilov, G. S. On the three-body problem in the case of short-range forces. *Zh. Eksp. Teor. Fiz.* **40**, 498–507 (1961); *Sov. Phys. JETP* **13**, 349–355 (1961).
26. Knoop, S. *et al.* Observation of an Efimov-like trimer resonance in ultracold atom–dimer scattering. *Nature Phys.* **5**, 227–230 (2009).
27. Amado, R. D. & Greenwood, F. C. There is no Efimov effect for four or more particles. *Phys. Rev. D* **7**, 2517–2519 (1973).
28. Jonsell, S. Efimov states for systems with negative scattering lengths. *Europhys. Lett.* **76**, 8–14 (2006).
29. Lee, M. D., Koehler, T. & Julienne, P. S. Excited Thomas–Efimov levels in ultracold gases. *Phys. Rev. A* **76**, 012720 (2007).
30. Massignan, P. & Stoof, H. T. C. Efimov states near a Feshbach resonance. *Phys. Rev. A* **78**, 030701 (2008).
31. Mehta, N. P., Rittenhouse, S. T., D’Incao, J. P., von Stecher, J. & Greene, C. H. A general theoretical description of N-body recombination. Preprint at <<http://arxiv.org/abs/0903.4145>> (2009).
32. Ferlaino, F. *et al.* Evidence for universal four-body states tied to an Efimov trimer. *Phys. Rev. Lett.* **102**, 140401 (2009).

Acknowledgements

This work was supported in part by the National Science Foundation. We are indebted to N. Mehta and S. Rittenhouse for extensive discussions and for access to their unpublished derivations before publication. We also thank F. Ferlaino, S. Knoop, H.-C. Nägerl and R. Grimm from the Innsbruck group for discussions about their experimental data.

Author contributions

The authors contributed equally to the manuscript.

Additional information

Supplementary information accompanies this paper on www.nature.com/naturephysics. Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>. Correspondence and requests for materials should be addressed to C.H.G.

A unified explanation of the Kadowaki–Woods ratio in strongly correlated metals

A. C. Jacko¹, J. O. Fjærestad² and B. J. Powell^{1*}

Discoveries of ratios whose values are constant within broad classes of materials have led to many deep physical insights. The Kadowaki–Woods ratio (KWR; refs 1, 2) compares the temperature dependence of a metal's resistivity to that of its heat capacity, thereby probing the relationship between the electron–electron scattering rate and the renormalization of the electron mass. However, the KWR takes very different values in different materials^{3,4}. Here we introduce a ratio, closely related to the KWR, that includes the effects of carrier density and spatial dimensionality and takes the same (predicted) value in organic charge-transfer salts, transition-metal oxides, heavy fermions and transition metals—despite the numerator and denominator varying by ten orders of magnitude. Hence, in these materials, the same emergent physics is responsible for the mass enhancement and the quadratic temperature dependence of the resistivity, and no exotic explanations of their KWRs are required.

In a Fermi liquid the electronic contribution to the heat capacity has a linear temperature dependence, that is, $C_{el}(T) = \gamma T$. Another prediction of Fermi-liquid theory⁵ is that, at low temperatures, the resistivity varies as $\rho(T) = \rho_0 + AT^2$. This is observed experimentally when electron–electron scattering, which gives rise to the quadratic term, dominates over electron–phonon scattering.

In a number of transition metals¹ $A/\gamma^2 \approx a_{TM} = 0.4 \mu\Omega \text{ cm mol}^2 \text{ K}^{-2}$ (Fig. 1), even though γ^2 varies by an order of magnitude across the materials studied. Later, it was found² that in many heavy-fermion compounds $A/\gamma^2 \approx a_{HF} = 10 \mu\Omega \text{ cm mol}^2 \text{ K}^{-2}$ (Fig. 1), despite the large mass renormalization, which causes γ^2 to vary by more than two orders of magnitude in these materials. Because of this remarkable behaviour A/γ^2 has become known as the Kadowaki–Woods ratio. However, it has long been known^{2,6} that the heavy-fermion material UPt_3 has an anomalously large KWR. More recently, studies of other strongly correlated metals, such as the transition-metal oxides^{3,7} and the organic charge-transfer salts^{4,8}, have found surprisingly large KWRs (Fig. 1). It is therefore clear that the KWR is not the same in all metals; in fact, it varies by more than seven orders of magnitude across the materials shown in Fig. 1.

Several important questions need to be answered about the KWR. (1) Why is the ratio approximately constant within the transition metals and within the heavy fermions (even though many-body effects cause large variations in their effective masses)? (2) Why is the KWR larger for the heavy fermions than it is for the transition metals? (3) Why are such large and varied KWRs observed in layered metals such as the organic charge-transfer salts and transition-metal oxides? The main aim of this paper is to resolve question (3). We shall also make some comments on the first two questions, which have been extensively studied previously.

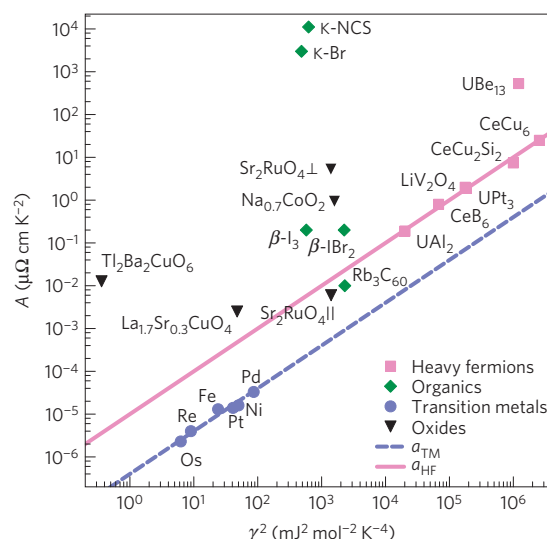


Figure 1 | The standard Kadowaki–Woods plot. It can be seen that the data for the transition metals and heavy fermions (other than UPt_3) fall onto two separate lines. However, a wide range of other strongly correlated metals do not fall on either line or between the two lines. $a_{TM} = 0.4 \mu\Omega \text{ cm mol}^2 \text{ K}^{-2}$ is the value of the KWR observed in the transition metals¹ and $a_{HF} = 10 \mu\Omega \text{ cm mol}^2 \text{ K}^{-2}$ is the value seen in the heavy fermions². In labelling the data points we use the following abbreviations: $\kappa\text{-Br}$ is $\kappa\text{-(BEDT-TTF)}_2\text{Cu}[\text{N}(\text{CN})_2]\text{Br}$; $\kappa\text{-NCS}$ is $\kappa\text{-(BEDT-TTF)}_2\text{Cu}(\text{NCS})_2$; $\beta\text{-I}_3$ is $\beta\text{-(BEDT-TTF)}_2\text{I}_3$; and $\beta\text{-IBr}_2$ is $\beta\text{-(BEDT-TTF)}_2\text{IBr}_2$. For Sr_2RuO_4 we show data for A measured with the current both perpendicular and parallel to the basal plane; these data points are distinguished by the symbols \perp and \parallel respectively. Further details of the data are reported in Supplementary Information.

There have been a number of studies of the KWR based on specific microscopic Hamiltonians (see, for example, refs 9–13). However, if the KWR has something general to tell us about strongly correlated metals, then we would also like to understand which features of the ratio transcend specific microscopic models. Nevertheless, two important points emerge from these microscopic treatments of the KWR: (1) if the momentum dependence of the self-energy can be ignored then the many-body renormalization effects on A and γ^2 cancel, and (2) material-specific parameters are required to reproduce the experimentally observed values of the KWR. Below we investigate the KWR using a phenomenological Fermi-liquid theory; this work builds on previous studies of related models^{3,6}. Indeed, our calculation is closely related to that in ref. 6. The main results reported here are the identification of a ratio (equation (5)) relating A

¹Centre for Organic Photonics and Electronics, School of Physical Sciences, University of Queensland, Brisbane, Queensland 4072, Australia, ²School of Physical Sciences, University of Queensland, Brisbane, Queensland 4072, Australia. *e-mail: bjpowell@gmail.com.

and γ^2 , which we predict takes a single value in a broad class of strongly correlated metals, and the demonstration that this ratio does indeed describe the data for a wide variety of strongly correlated metals (Fig. 2).

It has been argued that the KWR is larger in the heavy fermions than the transition metals because the former are more strongly correlated (in the sense that the self-energy is more strongly frequency dependent) than the latter⁶. Several scenarios have been proposed to account for the large KWRs observed in UBe_{13} , transition-metal oxides and organic charge-transfer salts, including impurity scattering⁶, proximity to a quantum critical point⁷ and the suggestion that electron–phonon scattering in reduced dimensions might give rise to a quadratic temperature dependence of the resistivity¹⁴. It has been previously observed³ that using volumetric (rather than molar) units for γ reduces the variation in the KWRs of the transition-metal oxides. However, even in these units, the organic charge-transfer salts have KWRs orders of magnitude larger than those of other strongly correlated metals. We shall argue that the different KWRs observed across this wide range of materials result from the simple fact that the KWR contains a number of material-specific quantities. As a consequence, when we replace the KWR with a ratio that accounts for these material-specific effects (equation (5)) the data for all of these materials do indeed lie on a single line (Fig. 2).

Many properties of strongly correlated Fermi liquids can be understood in terms of a momentum-independent self-energy^{15,16}. Therefore, following ref. 6, we assume that the imaginary part of the self-energy, $\Sigma''(\omega, T)$, at energy ω , is given by

$$\Sigma''(\omega, T) = -\frac{\hbar}{2\tau_0} - s \frac{\omega^2 + (\pi k_B T)^2}{\omega^{*2}} \quad (1)$$

for $|\omega^2 + (\pi k_B T)^2| < \omega^{*2}$ and

$$\Sigma''(\omega, T) = -[\hbar/2\tau_0 + s]F([\omega^2 + (\pi k_B T)^2]^{1/2}/\omega^*)$$

for $|\omega^2 + (\pi k_B T)^2| > \omega^{*2}$, where $2s/\hbar$ is the scattering rate due to electron–electron scattering in the absence of quantum many-body effects, τ_0^{-1} is the impurity scattering rate, F is a monotonically decreasing function with boundary conditions $F(1) = 1$ and $F(\infty) = 0$ and ω^* is determined by the strength of the many-body correlations. (See the Methods section for further discussion of the self-energy.)

The diagonal part of the conductivity tensor may be written as¹⁷

$$\sigma_{xx}(T) = \hbar e^2 \int \frac{d\mathbf{k}}{(2\pi)^3} v_{0x}^2 \int \frac{d\omega}{2\pi} A_s^2(\mathbf{k}, \omega) \left(\frac{-\partial f(\omega)}{\partial \omega} \right) \quad (2)$$

where $\mathbf{k} = (k_x, k_y, k_z)$ is the momentum, $v_{0x} = \hbar^{-1} \partial \varepsilon_0(\mathbf{k}) / \partial k_x$ is the unrenormalized velocity in the x direction, $f(\omega)$ is the Fermi–Dirac distribution, $A_s(\mathbf{k}, \omega) = -2 \text{Im}\{[\omega - \varepsilon_0(\mathbf{k}) + \mu^* - \Sigma(\omega, T)]^{-1}\}$ is the spectral density, $\varepsilon_0(\mathbf{k})$ is the non-interacting dispersion relation and μ^* is the chemical potential. Note that equation (2) does not contain vertex corrections; the absence of vertex corrections to the conductivity is closely related to the momentum independence of the self-energy¹⁵. Further, the presence of Umklapp processes, which enable electron–electron scattering to contribute to the resistivity in the pure limit¹⁸, is implicit in the above formula.

In a strongly correlated metal, s may be approximated by its value in the unitary scattering limit^{6,16}, $s_u = 2n/3\pi D_0$, where n is the conduction-electron density and D_0 is the bare density of states

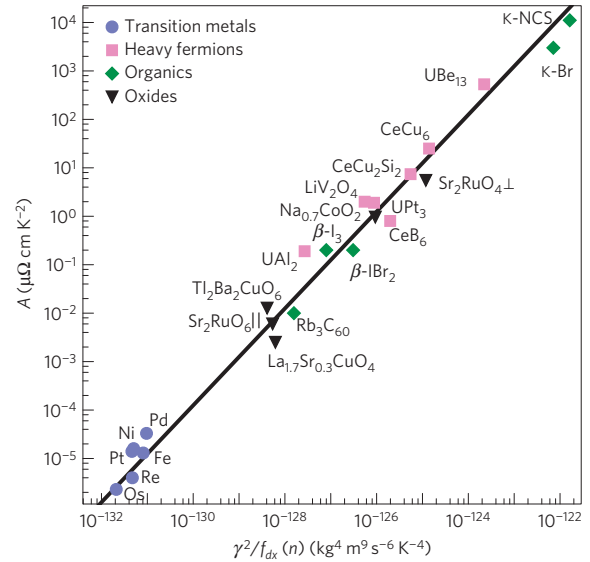


Figure 2 | Comparison of the ratio defined in equation (5) with

experimental data. It can be seen that, in all of the materials studied, the data are in excellent agreement with our prediction (line). The abbreviations in the data-point labels are the same as in Fig. 1. Further details of the data are given in Supplementary Information.

(DOS) at the Fermi energy. In the low-temperature, pure limit we find (see the Methods section) that

$$A = \frac{16nk_B^2}{\pi \hbar e^2 \langle v_{0x}^2 \rangle D_0^2 \omega^{*2}} \quad (3)$$

where $\langle \dots \rangle$ denotes an average over the Fermi surface. Note that neither the DOS nor the Fermi velocity are renormalized in this expression. Indeed, all of the many-body effects are encapsulated by ω^* , which determines the magnitude of the frequency-dependent term in $\Sigma''(\omega, T)$; see equation (1).

The Kramers–Kronig relation for the retarded self-energy^{19,20} can be used to show (see the Methods section) that, in the pure limit,

$$\gamma = \gamma_0 \left(1 - \frac{\partial \Sigma'}{\partial \omega} \right) = \gamma_0 \left(1 + \frac{4s_u \xi}{\pi \omega^*} \right)$$

where $\gamma_0 = \pi^2 k_B^2 D_0 / 3$ is the linear coefficient of the specific heat for a gas of non-interacting fermions, Σ' is the real part of the self-energy and $\xi \approx 1$ is a pure number defined in the Methods section. Thus we see that the renormalization of γ is also controlled by ω^* . For a strongly correlated metal the effective mass, $m^* \gg m_0$, the bare (band) mass of the electron, hence $s_u \gg \omega^*$ and $\gamma \approx (8nk_B^2 \xi) / (9\omega^*)$. The corrections to this approximation are given in the Methods section.

Combining the above results we see that the KWR is

$$\frac{A}{\gamma^2} = \frac{81}{4\pi \hbar k_B^2 e^2} \frac{1}{\xi^2 n D_0^2 \langle v_{0x}^2 \rangle} \quad (4)$$

First, we note that in this ratio the dependence of the individual factors on ω^* has vanished. Hence the KWR is not renormalized. On the other hand, although the first factor contains only fundamental constants, the second factor is clearly material dependent as it depends on the electron density, the DOS and the Fermi velocity of the non-interacting system. An important corollary to this result is that band-structure calculations should give accurate predictions

of the KWR by equation (4), as none of the properties on the right-hand side are renormalized.

The wide range of KWRs found in layered materials suggests that $\xi^2 n D_0^2 \langle v_{0x}^2 \rangle$ varies significantly in these materials. For example, for highly anisotropic materials $\langle v_{0x}^2 \rangle$ may vary by more than an order of magnitude, depending on which direction the resistivity is measured in. This effect needs to be taken into account if we wish to understand what the KWR has to tell us about strongly correlated metals. Further, equation (4) suggests that the reason that the transition metals and the heavy fermions have ‘constant’ KWRs is that $\xi^2 n D_0^2 \langle v_{0x}^2 \rangle$ is roughly constant across each class of materials and that $a_{\text{HF}} \neq a_{\text{TM}}$ because $\xi^2 n D_0^2 \langle v_{0x}^2 \rangle$ is different in the two different classes of materials. Therefore, we propose that a more fundamental ratio is

$$\frac{A f_{dx}(n)}{\gamma^2} = \frac{81}{4\pi\hbar k_B^2 e^2} \quad (5)$$

where $f_{dx}(n) \equiv n D_0^2 \langle v_{0x}^2 \rangle \xi^2$ may be written in terms of the dimensionality, d , of the system, the electron density and, in layered systems, the interlayer spacing or the interlayer hopping integral.

For simplicity we assume that the reasonably isotropic materials (the heavy fermions, the transition metals and Rb_3C_{60}) are isotropic Fermi liquids and the layered materials (that is, the transition-metal oxides and organic charge-transfer salts based on BEDT-TTF) have warped cylindrical Fermi surfaces; $f_{dx}(n)$ is derived for these band structures in the Methods section. This enables us to test explicitly, in Fig. 2, the prediction of equation (5) against previously published experimental data for a variety of strongly correlated metals. It is clear that the new ratio is in good agreement with the data for all of the materials investigated. We therefore see that the observations of constant KWRs for the heavy fermions and for the transition metals are due not only to the profound but also to the prosaic. The renormalization of γ^2 cancels with that of A owing to the Kramers–Kronig relation for the self-energy; but the unrenormalized properties are remarkably consistent within each class of materials. Further, the large KWRs in transition-metal oxides, the organics and UBe_{13} are simply a consequence of the small values of $f_{dx}(n)$ in these materials. Therefore, the absolute value of the KWR does not reveal anything about electronic correlations unless the material-specific effects, described by $f_{dx}(n)$, are first accounted for.

It will be interesting to identify and understand strongly correlated metals that are not described by equation (5). Our calculation already gives some clues as to when this might happen: for example, when the self-energy is strongly momentum dependent or when there are significant vertex corrections to the conductivity. Another outstanding challenge is to understand the KWR in compensated semimetals^{21–23}.

Methods

Resistivity. To calculate A we approximate the spectral density by

$$A_i^2(\mathbf{k}, \omega) \approx \frac{2\pi Z \delta(\omega - Z\xi_0(\mathbf{k}))}{-\Sigma''(\omega, T)} \quad (6)$$

where $\xi_0(\mathbf{k}) = \varepsilon_0(\mathbf{k}) - \mu_0$ with μ_0 the chemical potential of the bare system (that is, in the absence of both electron–electron interactions and impurity scattering) and Z is the quasi-particle weight defined as $Z^{-1} = 1 - (\partial/\partial\omega)\Sigma'(\omega, 0)|_{\omega=0}$. Equation (6) will give the right behaviour in the limit $\Sigma'' \rightarrow 0$ (ref. 17). Inserting equation (6) into equation (2) and noting that at low temperatures $-\partial f(\omega)/\partial\omega$ will be sharply peaked at $\omega = 0$, we replace ω by zero in $\delta(\omega - Z\xi_0(\mathbf{k}))$. The delta function can then be taken outside the ω integration, giving

$$\sigma_{xx}(T) \approx Z \hbar e^2 \langle v_{0x}^2 \rangle \int \frac{d\mathbf{k}}{(2\pi)^3} \delta(Z\xi_0(\mathbf{k})) \int d\omega \frac{1}{-\Sigma''(\omega, T)} \left(\frac{-\partial f(\omega)}{\partial\omega} \right)$$

The \mathbf{k} -space integral here is (half) the renormalized DOS, D^* , at the renormalized Fermi energy. As $D^* = D_0/Z$, where D_0 is the DOS of the bare system at its Fermi energy, we get

$$\sigma_{xx}(T) = \hbar e^2 \langle v_{0x}^2 \rangle D_0 \int d\omega \frac{(-\partial f(\omega)/\partial\omega)}{-2\Sigma''(\omega, T)} \quad (7)$$

Using $-\partial f(\omega)/\partial\omega \rightarrow \delta(\omega)$ as $T \rightarrow 0$ it follows from equation (1) that the zero-temperature resistivity is given by $\rho_0 = (e^2 \tau_0 \langle v_{0x}^2 \rangle D_0)^{-1}$. With the temperature dependence of the resistivity given by $\rho(T) = \rho_0 + AT^2$, equation (7) yields

$$AT^2 = \rho - \rho_0 = \frac{1}{\hbar e^2 \langle v_{0x}^2 \rangle D_0} \left(\left[\int d\omega \frac{(-\partial f(\omega)/\partial\omega)}{-2\Sigma''(\omega, T)} \right]^{-1} - \frac{\hbar}{\tau_0} \right)$$

We now consider the pure limit, $\tau_0 \rightarrow \infty$. At sufficiently low temperatures the contribution to the integral from the region $\omega > \omega^*$ is small, so it is a good approximation to use equation (1) for $\Sigma''(\omega, T)$ for all ω . The resulting integral can then be evaluated analytically,

$$\int_{-\infty}^{\infty} d\omega \frac{(-\partial f(\omega)/\partial\omega)}{-2\Sigma''(\omega, T)} \approx \frac{\omega^{*2}}{2s} \int_{-\infty}^{\infty} d\omega \frac{(-\partial f(\omega)/\partial\omega)}{\omega^2 + (\pi k_B T)^2} = \frac{(\omega^*/k_B T)^2}{24s}$$

Equation (3) follows on taking $s = s_u$.

Real part of the self-energy. To calculate γ we need to know the real part of the self-energy. To evaluate this we apply the Kramers–Kronig relation^{19,20} for the retarded self-energy,

$$\Sigma'(\omega, T) = \Sigma'(\infty, T) + \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{\Sigma''(\omega', T)}{\omega' - \omega} d\omega'$$

where P indicates the principal part of the integral.

For $T = 0$ and taking the pure limit we find that for $|\omega| \ll \omega^*$

$$\begin{aligned} \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{\Sigma''(\omega')}{\omega' - \omega} d\omega' &= -\frac{2s}{\pi} \left[\gamma + \frac{1}{2} \gamma^2 \ln \left| \frac{1 - \gamma}{1 + \gamma} \right| \right. \\ &\quad \left. + \int_1^{\infty} dy' \frac{F(y')}{y'} \sum_{n=0}^{\infty} \left(\frac{\gamma}{y'} \right)^{2n+1} \right] \end{aligned}$$

where $\gamma = \omega/\omega^*$, $y' = \omega'/\omega^*$, and we have used the fact that $F(y)$ is an even function. Therefore, in the limit $|\gamma| \ll 1$

$$\Sigma'(\omega, 0) = \Sigma'(\infty, 0) - \frac{4s_u \xi}{\pi} \frac{\omega}{\omega^*} + O\left(\frac{\omega^3}{\omega^{*3}}\right)$$

where $1 < 2\xi \equiv 1 + \int_1^{\infty} y^{-2} F(y) dy \leq 1 + \int_1^{\infty} y^{-2} dy = 2$ as $F(y) \leq 1$ for $y \geq 1$. Provided that $F(y)$ decreases sufficiently slowly as $y \rightarrow \infty$, we expect $\xi \approx 1$. In general, changes to the exact form of the self-energy will simply lead to small changes in ξ provided that the boundary conditions for Σ remain the same. Note that $\Sigma'(\infty, 0)$ is just the shift in the zero-temperature chemical potential due to many-body interactions.

The form of the self-energy given in equation (1) has a kink at $\omega = \omega^*$. Kinks are found in the self-energies of local Fermi-liquid theories, such as dynamical mean-field theory²⁴. However, the location, and even the existence, of this kink is not important for our results. To calculate A and γ we must integrate over all ω (see above); therefore, any sharp features in $\Sigma''(\omega, T)$ will be washed out.

Band structure. A , γ and n are relatively straightforward to determine experimentally. It is harder to directly measure D_0 and $\langle v_{0x}^2 \rangle$. Therefore, we consider two model band structures. (1) For an isotropic Fermi liquid $\varepsilon_0(\mathbf{k}) = \hbar^2 \mathbf{k}^2 / 2m_0$, $D_0 = m_0 k_F / \hbar^2 \pi^2$ and $\langle v_{0x}^2 \rangle = \hbar^2 k_F^2 / 3m_0^2$, where $k_F = \sqrt{3\pi^2 n}$. Hence, for $\xi = 1$, $f_{3x}(n) = \sqrt{3} \pi^2 / \pi^4 \hbar^6$. (2) To study layered materials we use the simple model dispersion $\varepsilon_0(\mathbf{k}) = \hbar^2 \mathbf{k}_{ab}^2 / 2m_0 - 2t_{\perp 0} \cos c k_{\perp}$, where $\mathbf{k} = (\mathbf{k}_{ab}, k_{\perp})$, \mathbf{k}_{ab} is the in-plane wavevector, k_{\perp} is the wavenumber perpendicular to the plane, c is the interlayer spacing and $t_{\perp 0}$ is the bare interlayer hopping integral. If A is taken from measurements of the resistivity parallel to the plane $\langle v_{0\parallel}^2 \rangle = \hbar^2 k_F^2 / 2m_0^2$. However, for A measured perpendicular to the plane $\langle v_{0\perp}^2 \rangle = 2c^2 t_{\perp 0}^2 / \hbar^2$. In either case, for the warped cylindrical Fermi surface we are now considering $D_0 = m_0 / \pi c \hbar^2$ and $k_F = \sqrt{2\pi c n}$. So, for $\xi = 1$, we find that $f_{\perp}(n) = n^2 / \pi c \hbar^2$ and $f_{\parallel}(n) = 2nm_0^2 t_{\perp 0}^2 / \pi^2 \hbar^6$.

This formalism can straightforwardly be generalized to include other factors known to affect the KWR by extending the definition of $f_{dx}(n)$. For example, it

has been shown¹³ that in the N -fold orbitally degenerate periodic Anderson model $A/\gamma^2 \propto [N(N-1)]^{-1}$, in good agreement with experiments on heavy fermions with orbital degeneracy^{25,26}. This result is specific to this particular model, and the systems so far studied^{25,26} all have rather similar values of $\xi^2 n D_0^2 (v_{0x}^2)$. However, it is clear, from a comparison of the calculation of ref. 13 with ours, that if orbitally degenerate systems with different electron densities or reduced dimensionalities were fabricated $f_{dx}(n)$ would need to be included to understand the relationship between A and γ . It has also been shown³ that the number of sheets of the Fermi surface also affects the KWR. It is straightforward to generalize the above calculations of $f_{dx}(n)$ to Fermi surfaces with any number of sheets. Finally, we note that if we relax the condition $m^* \gg m_0$ then we find that $A f_{dx}(n)/\gamma^2 = (81/4\pi \hbar k_B^2 e^2)(1 - m_0/m^*)^2$, that is, the ratio vanishes as $m^* \rightarrow m_0$. It is therefore particularly interesting that a constant KWR is seen in the transition metals, which do not all have such large effective masses as the other materials discussed above.

Received 30 May 2008; accepted 20 March 2009;
published online 19 April 2009

References

- Rice, M. J. Electron–electron scattering in transition metals. *Phys. Rev. Lett.* **20**, 1439–1441 (1968).
- Kadowaki, K. & Woods, S. B. Universal relationship of the resistivity and specific heat in heavy-fermion compounds. *Solid State Commun.* **58**, 507–509 (1986).
- Hussey, N. E. Non-generality of the Kadowaki–Woods ratio in correlated oxides. *J. Phys. Soc. Japan* **74**, 1107–1110 (2005).
- Dressel, M., Grüner, G., Eldridge, J. E. & Williams, J. M. Optical properties of organic superconductors. *Synth. Met.* **85**, 1503–1508 (1997).
- Nozières, P. & Pines, D. *The Theory of Quantum Liquids* (Perseus Books, 1999).
- Miyake, K., Matsuura, T. & Varma, C. M. Relation between resistivity and effective mass in heavy-fermion and A15 compounds. *Solid State Commun.* **71**, 1149–1153 (1989).
- Li, S. Y. *et al.* Giant electron–electron scattering in the Fermi-liquid state of $\text{Na}_{0.7}\text{CoO}_2$. *Phys. Rev. Lett.* **93**, 056401 (2004).
- Powell, B. J. & McKenzie, R. H. Strong electron correlations in superconducting organic charge transfer salts. *J. Phys. Condens. Matter* **18**, R827–R866 (2006).
- Yamada, K. & Yosida, K. Fermi liquid theory on the basis of the periodic Anderson hamiltonian. *Prog. Theor. Phys.* **76**, 621–638 (1986).
- Auerbach, A. & Levin, K. Universal low-temperature properties of normal heavy-fermion systems. *J. Appl. Phys.* **61**, 3162–3167 (1987).
- Coleman, P. Constrained quasiparticles and conduction in heavy-fermion systems. *Phys. Rev. Lett.* **59**, 1026–1029 (1987).
- Li, T. C. & Rasul, J. W. Gutzwiller dynamic susceptibility: Consequences for the transport properties of transition metals. *Phys. Rev. B* **39**, 4630–4633 (1989).
- Kontani, H. Generalized Kadowaki–Woods relation in heavy fermion systems with orbital degeneracy. *J. Phys. Soc. Japan* **73**, 515–518 (2004).
- Strack, Ch. *et al.* Resistivity studies under hydrostatic pressure on a low-resistance variant of the quasi-two-dimensional organic superconductor κ -(BEDT-TTF)₂Cu[N(CN)₂]Br: Search for intrinsic scattering contributions. *Phys. Rev. B* **72**, 054511 (2005).
- Georges, A., Kotliar, G., Krauth, W. & Rozenberg, M. J. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Rev. Mod. Phys.* **68**, 13–125 (1996).
- Hewson, A. C. *The Kondo Problem To Heavy Fermions* (Cambridge Univ. Press, 1997).
- Mahan, G. D. *Many-Particle Physics* 2nd edn (Plenum Press, 1990).
- Maebashi, H. & Fukuyama, H. Electrical conductivity of interacting fermions II: Effects of normal scattering process in the presence of Umklapp scattering process. *J. Phys. Soc. Japan* **67**, 242–251 (1998).
- Luttinger, J. M. Analytic properties of single-particle propagators for many-fermion systems. *Phys. Rev.* **121**, 942–949 (1961).
- Giuliani, G. & Vignale, G. *Quantum Theory Of The Electron Liquid* (Cambridge Univ. Press, 2005).
- Collan, H. K., Krusius, M. & Pickett, G. R. Specific heat of antimony and bismuth between 0.03 and 0.8 K. *Phys. Rev. B* **1**, 2888–2895 (1969).
- Chopra, V., Ray, R. K. & Bhagat, S. M. Low-temperature resistivity of Bi and its alloys. *Phys. Status Solidi A* **4**, 205–214 (1971).
- Terashima, T. *et al.* Resistivity, Hall effect, and Shubnikov–de Haas oscillations in CeNiSn. *Phys. Rev. B* **66**, 075127 (2002).
- Byczuk, K. *et al.* Kinks in the dispersion of strongly correlated electrons. *Nature Phys.* **3**, 168–171 (2007).
- Tsujii, N., Kontani, H. & Yoshimura, K. Universality in heavy fermion systems with general degeneracy. *Phys. Rev. Lett.* **94**, 057201 (2005).
- Torikachvili, M. S. *et al.* Six closely related YbT₂Zn₂₀ (T = Fe, Co, Ru, Rh, Os, Ir) heavy fermion compounds with large local moment degeneracy. *Proc. Natl Acad. Sci.* **104**, 9960–9963 (2007).

Acknowledgements

It is a pleasure to thank J. Castro, N. Hussey, M. Kennett, R. McKenzie, J. Merino, G. Notley, M. Smith, T. Stace, C. Varma, A. White and J. Wosnitzer for their helpful comments. This research was supported under the Australian Research Council's (ARC's) Discovery Projects funding scheme (project DP0878523). B.J.P. is the recipient of an ARC Queen Elizabeth II Fellowship (DP0878523).

Author contributions

This project was planned and led by B.J.P. All authors contributed equally to the derivation. The analysis of the previously published experimental data was carried out by A.C.J. The paper was written by B.J.P. with significant input from A.C.J. and J.O.F.

Additional information

Supplementary information accompanies this paper on www.nature.com/naturephysics. Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>. Correspondence and requests for materials should be addressed to B.J.P.

Physical forces during collective cell migration

Xavier Trepats^{1,2*}, Michael R. Wasserman¹, Thomas E. Angelini³, Emil Millet¹, David A. Weitz³, James P. Butler^{1,4} and Jeffrey J. Fredberg^{1*}

Fundamental biological processes including morphogenesis, tissue repair and tumour metastasis require collective cell motions^{1–3}, and to drive these motions cells exert traction forces on their surroundings⁴. Current understanding emphasizes that these traction forces arise mainly in ‘leader cells’ at the front edge of the advancing cell sheet^{5–9}. Our data are contrary to that assumption and show for the first time by direct measurement that traction forces driving collective cell migration arise predominately many cell rows behind the leading front edge and extend across enormous distances. Traction fluctuations are anomalous, moreover, exhibiting broad non-Gaussian distributions characterized by exponential tails^{10–12}. Taken together, these unexpected findings demonstrate that although the leader cell may have a pivotal role in local cell guidance, physical forces that it generates are but a small part of a global tug-of-war involving cells well back from the leading edge.

The single adherent cell moves by the action of two synchronized cycles, one involving extension and contraction of its cytoskeleton and the other involving formation and detachment of its adhesions^{13,14}. Although this complex process remains a matter of intense research^{14–16}, it is now well established that a fundamental aspect of the motility mechanism is the transmission of contractile forces to the surrounding matrix at the cell’s leading and trailing edges^{17,18}. In contrast with the case of migration of the single cell studied in isolation^{14–16}, the case of collective migration of cells within a contiguous cell sheet has more physiological relevance but is substantially less well understood¹⁹. Within an advancing epithelial cell sheet, for example, each individual cell is physically constrained by its neighbours, and cell–cell signalling through biochemical and biophysical pathways may influence the collective motion of the group^{20,21}. Do leader cells at the advancing front edge of the sheet exert physical forces locally that are transmitted rearward, from cell-to-cell, and thus act to pull along those cells in the ranks behind^{5,6,8,9}? Or instead is each individual cell in the sheet mechanically self-propelled²¹? Or does cell proliferation expand the cell colony and thereby push the advancing front forward? Or is the correct answer none of the above? For more than a century these fundamental questions have been debated intensively^{5,22} and, using a variety of methods *in vivo*²³, *in vitro*^{4,9,24} and *in silico*²¹, much conflicting evidence has accumulated. This conflicting evidence has been in most cases indirect or inferential, however, because within the cell sheet the physical forces themselves have remained largely inaccessible to direct experimental observation.

Here, we report by direct measurement the first explicit maps of those physical forces and their distribution. To do this within an advancing cell sheet, we used Fourier-transform traction microscopy together with a balance of forces that is

demanding by straightforward application of Newton’s laws of motion (see Supplementary Information S2 and S3). To address the case of an advancing cell sheet, however, traction microscopy as described originally^{18,25} or as modified subsequently^{26–28} is inadequate and therefore required fundamental reformulation (see Supplementary Information S2 and S3). We seeded a small number of Madin–Darby canine kidney epithelial cells (~5,000) at the centre of a soft collagen-coated polyacrylamide gel (Young’s modulus of 1.3 kPa). The cells adhered readily to this substrate and within 24 h formed a confluent colony. With time, the colony expanded radially outward, thus providing a simple model of collective migration without need of damaging the monolayer as in classical scratch-wounding experiments (see Supplementary Information S1). Growth of the colony was largely insensitive to the stiffness of the underlying substrate (see Supplementary Information S4). After allowing the colony to expand for at least 72 h, we mapped the traction forces that marginal and submarginal cells exerted on their underlying matrix.

We first assessed the locus of traction forces in the proximity of the leading edge (Fig. 1, Supplementary Movie S1). Maps of tractions normal (T_{\perp}) and parallel (T_{\parallel}) to the leading edge show that tractions are not restricted to cells at the leading edge or even to cells located 2–3 rows behind it, as is commonly emphasized^{5,6,29}. Instead, large tractions are applied by cells many cell rows behind the edge. Independent of the distance from the edge, both T_{\perp} and T_{\parallel} exhibited broad non-Gaussian distributions characterized by exponential tails (Fig. 2a, b). The distribution of T_{\perp} was skewed towards positive tractions at the leading edge, whereas the distribution of T_{\parallel} was symmetric. Both traction distributions narrowed as the distance from the leading edge increased. Taken together, these data are inconsistent with the existence of two populations of cells, each with a distinct mechanical phenotype, one corresponding to active mechanical leaders at the leading edge and the other to passive mechanical followers. Instead, our data show a single distribution, the tails of which were roughly exponential rather than Gaussian, revealing probabilities of high tractions much larger than would be predicted according to the central limit theorem for independent and identically distributed random variables. Exponential distributions have previously been reported at the level of the single focal adhesion^{30–32}, thereby indicating that this particular kind of distribution might underlie tissue behaviour over multiple length scales.

Submarginal cells have previously been shown to extend cryptic lamellipodia beneath cells in front of them²⁴. Regardless of the extent to which cryptic lamellipodia are mechanically active and represent a locus of force generation, traction forces generated by these submarginal cells are seen to be comparable to those at the leading edge. A more important question, however, is whether these

¹Program in Molecular and Integrative Physiological Sciences, School of Public Health, Harvard University, Boston, Massachusetts 02115, USA, ²Unitat de Biofísica i Bioenginyeria, Universitat de Barcelona, Institute for Bioengineering of Catalonia, and Ciber Enfermedades Respiratorias, 08036 Barcelona, Spain, ³School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA, ⁴Dept. Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. *e-mail: xtrepats@ub.edu; jeffrey_fredberg@hsph.harvard.edu.

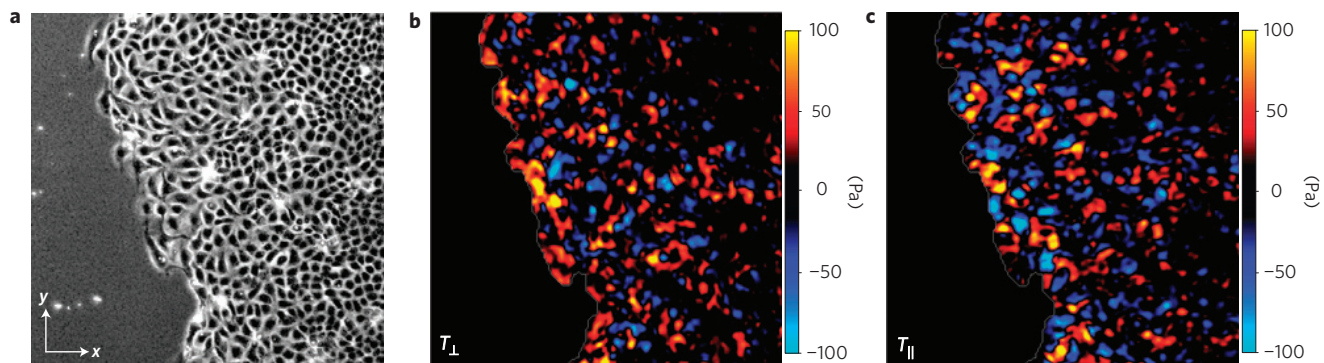


Figure 1 | Traction forces generated by a collectively migrating cell sheet. **a**, Phase contrast image. **b**, Tractions normal to the edge. **c**, Tractions parallel to the edge. The field of view is $750\ \mu\text{m} \times 750\ \mu\text{m}$. T_{\parallel} and T_{\perp} were calculated from T_x and T_y and from the local normal vector to the cell edge (see Supplementary Methods).

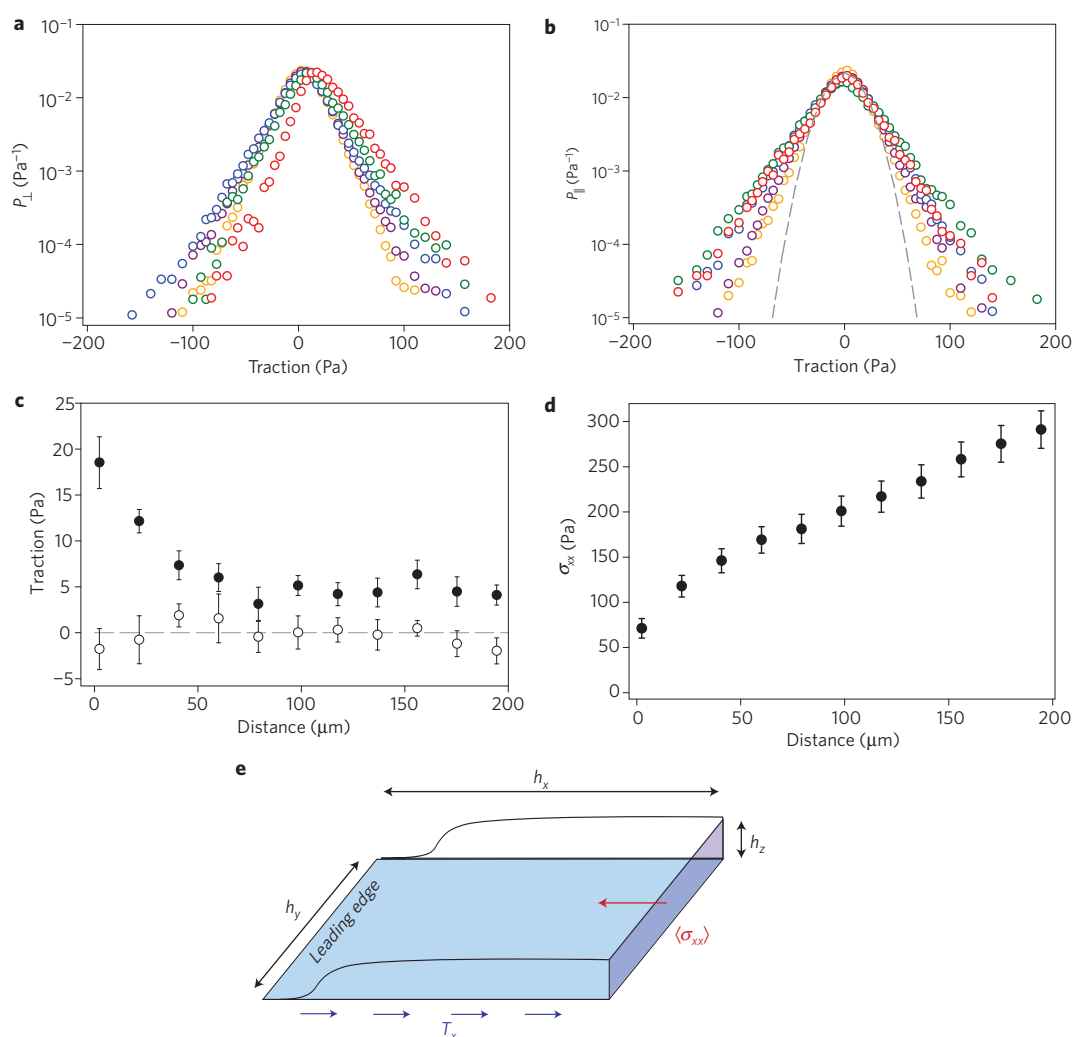


Figure 2 | Traction force distributions at different distances from the leading edge. **a**, Normal tractions. **b**, Parallel tractions. Red circles: first cell row from the leading edge; green circles: row 2; blue circles: rows 3–5; purple circles: rows 6–11; orange circles: rows 12–17. For computational simplicity, each row was assumed to be $19.2\ \mu\text{m}$ in radial dimension. Data were pooled from $n = 4$ different cell sheets at four different time points for each well. The tails of each distribution appear straight in a semilog plot, showing the exponential nature of the distributions. A Gaussian fit to the distribution of parallel tractions for rows 12–17 is plotted as a reference (dashed grey line). **c**, The average normal traction decayed slowly with distance from the edge (filled symbols), whereas the average parallel traction was negligible and independent of the distance from the edge (open symbols). Error bars indicate standard errors. **d**, Stress within the cell sheet increased as a function of the distance from the leading edge. Error bars indicate standard errors. **e**, Schematic diagram illustrating the computation of stress within the cell sheet. The average stress $\langle \sigma_{xx} \rangle$ normal to a plane perpendicular to the substrate and parallel to the leading edge can be calculated by integration of tractions T_x between the edge and the plane.

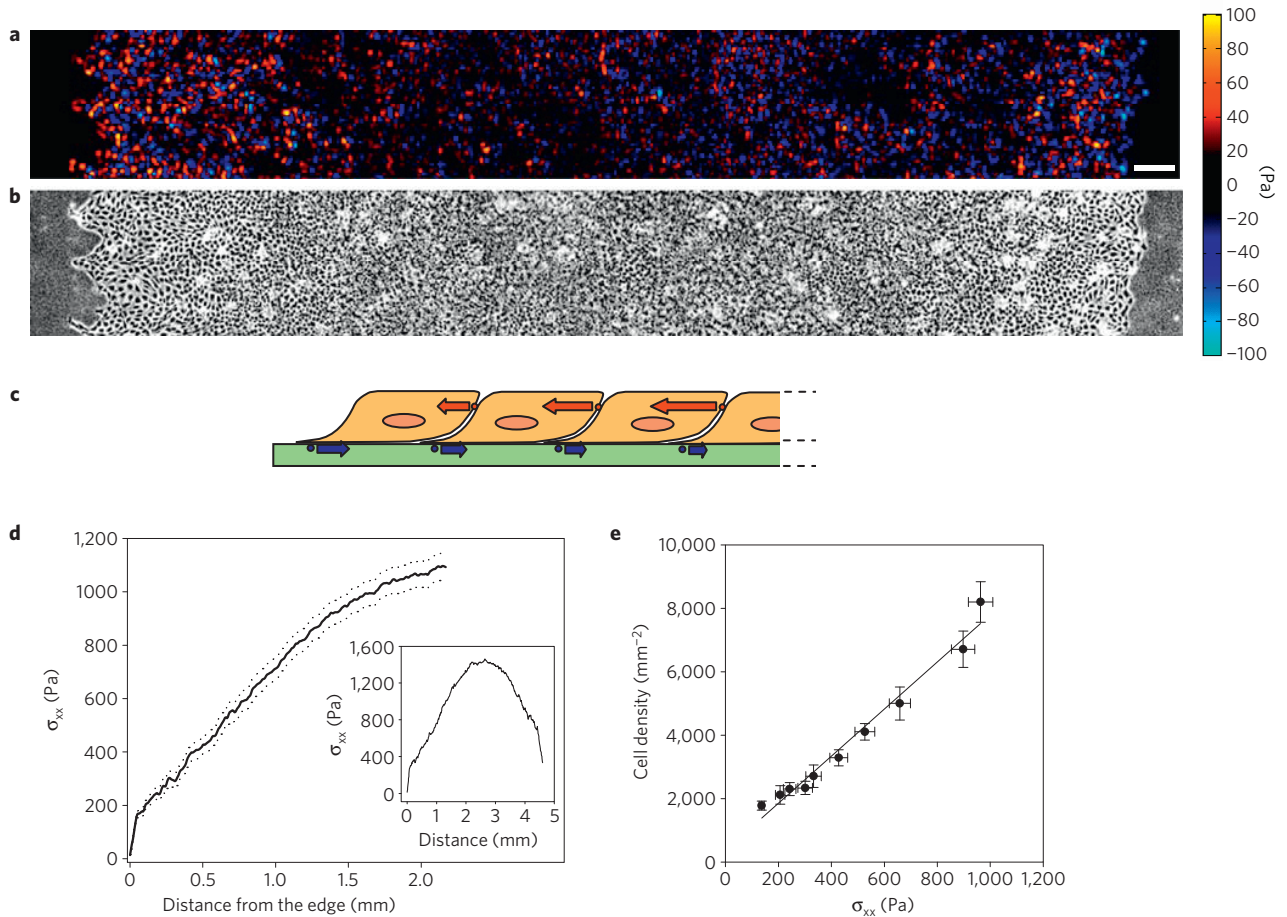


Figure 3 | The state of stress of the whole expanding colony is set by a global tug-of-war. a, Radial component of tractions along a diameter of the expanding cell colony (bar = 200 μm). **b**, Phase contrast image. Note the presence of multicellular protrusions on the left margin of the colony. This particular shape is reminiscent of that caused by fingering instabilities in fluids⁹. **c**, Tug-of-war model. A small portion of the traction that each cell generates is transmitted to the cell behind. As such, tension in the cytoskeleton and cell-cell junctions increases towards the centre of the cell colony (red arrows). **d**, Once integrated over many cell rows, this small portion becomes dominant over traction fluctuations (see Fig. 2). Dotted lines indicate mean \pm standard error. Inset: Representative measurement of σ_{xx} over the whole diameter of a colony. σ_{xx} reaches a maximum roughly at the centre of the colony. The fact that σ_{xx} does not decrease exactly to zero at the end of the monolayer points to the existence of weak shear stresses that also contribute to the force balance. **e**, The ratio between radial stress and cell density across the monolayer was roughly constant. Error bars indicate standard errors.

forces are balanced locally, as if each cell were self-propelled, or whether they are transmitted from cell to cell in a cooperative manner. We note that no amount of kinematic data or structural data, no matter how detailed, and no molecular manipulation, no matter how sophisticated, can ever suffice to resolve this question. Instead, we provide here a conclusive answer based solely on the application of Newton's laws. We begin by computing the spatial averages (denoted by $\langle \rangle$) of T_{\perp} and T_{\parallel} (Fig. 2c). As expected by symmetry, $\langle T_{\parallel} \rangle$ was near zero. $\langle T_{\perp} \rangle$ was maximum at the leading edge and progressively decayed with distance from the edge before changing sign in the opposite half of the sheet. In contrast to $\langle T_{\parallel} \rangle$, however, the spatial average $\langle T_{\perp} \rangle$ far away from the edge remained weakly but systematically greater than zero, with typical tugging tractions of the order of 5 Pa. In the context of cell mechanics, a regional stress of magnitude in this range is often regarded as being small, and is certainly small compared with the fluctuations that we observed. One might therefore conclude that its effects are essentially negligible; we come to a quite different conclusion, however.

Within the field of measurement—here spanning less than half the sheet diameter—the average stress at the interface between the cell base and the cell substrate, $\langle T_{\perp} \rangle$, was regionally positive everywhere. This stress acts systematically in a direction that pulls

the sheet towards the leading edge. The question then arises, how are these stresses balanced, as is required by Newton's laws? The simple and inescapable answer is this: at any arbitrary given distance L remote from the leading edge, the sum of the traction stresses perpendicular to the edge from $x = 0$ up to $x = L$ must be balanced by forces carried within the cell sheet at position L (Fig. 2e). At every position within the sheet, therefore, the accumulated traction must be balanced by local cell-borne stresses that are transmitted along the cell sheet by the cytoskeleton within cells and by cell-cell junctions between cells. Using σ_{xx} to denote stresses within the cell sheet on a plane perpendicular to the substrate and parallel to the edge, as distinct from cell tractions T at the cell-substrate interface, force balance demands that these be related by

$$\langle \sigma_{xx}(x) \rangle = \frac{1}{h_z h_y} \int_0^x \int_0^{h_y} T_x(x', y') dx' dy'$$

where we take cell height, h_z , to be roughly 5 μm (ref. 33), and h_y is the length of the field of view.

We note, first, that if each cell were entirely self-propelled, then $\langle \sigma_{xx} \rangle$ would be identically zero everywhere. This possibility can now be ruled out (Fig. 2d). Instead, $\langle \sigma_{xx} \rangle$ increased steadily with

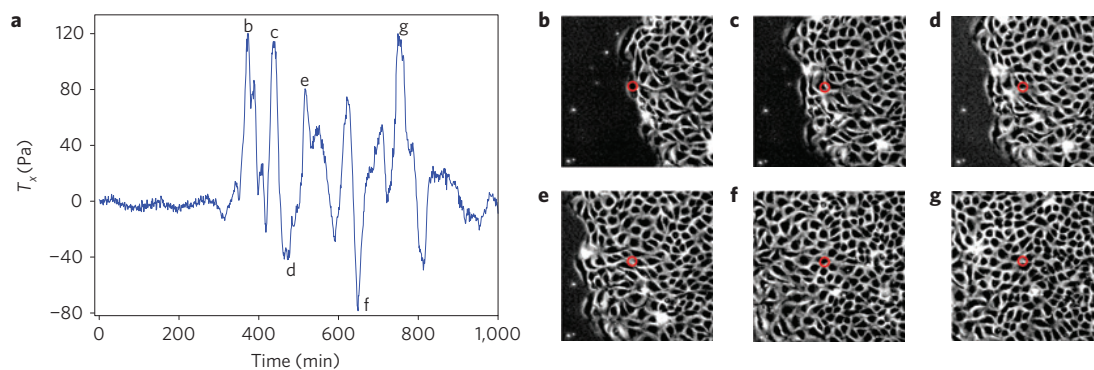


Figure 4 | Time fluctuations of tractions show cell-cell force transmission. **a**, Traction T_x at a fixed point of the gel (red circle in **b–g**) as cells crawl over that point. **b–g**, The right-to-left advance of the cell sheet at the time points labelled in **a**. Note that T_x remains mostly positive throughout the observation period, which implies stress transmission. (See Supplementary Movie S2.)

x , reaching a stress $\langle\sigma_{xx}\rangle \sim 300$ Pa within the first 10 cell rows (Fig. 2d). Second, although they contribute little to the overall balance of forces, the pull that leading cell rows exert may be sufficient to direct the ranks immediately behind; in that sense they may still be ‘leaders’. To study further the spatial extent of this tugging mechanism of force transmission, we obtained traction maps spanning the whole diameter of the cell colony (Fig. 3). Approaching the centre of the cell colony, $\langle\sigma_{xx}\rangle$ increased steadily and reached values that dominated traction fluctuations. Therefore, as the cell sheet grows, it exists in a global state of tensile stress. Such tensile stress establishes definitively that a build-up of pressure caused by proliferation is not the driving mechanism that underlies expansion of the cell sheet, for if this were the case, stresses within the monolayer would be compressive and tractions would point outwards. In both regards, we found systematic evidence to the contrary. Furthermore, our data show that cell density increased proportionally to the tensile stress σ_{xx} ; for reasons that remain unclear, proliferation during tissue growth seems to be regulated in such a way that the ratio between density and tensile stress is invariant (Fig. 3e). Shraiman³⁴ proposed that lateral stresses may act as a local feedback signal to regulate the rate of tissue growth, although his model deals with monolayer compression and buckling as opposed to the tensile stresses applicable here.

To assess further the nature of force transmission, we considered the time evolution of tractions at specific points of the traction map. An illustrative example of such time evolution is shown in Fig. 4 (see also Supplementary Movie S2). As the first leader cell of the sheet advanced over the traction sensing point (that is, a particular pixel of the traction map), a strong pull was first observed followed by a sharp decay. This decay, however, did not drop significantly below zero as would be the case for a self-propelled cell in isolation. Instead, a second sharp pulse was observed with a maximum corresponding to the boundary between the third and fourth cells. In this particular case, the first significantly negative force was observed after the fourth row, implying that the force generated by the first three rows was almost entirely transmitted to the following rows. In every case, traction fluctuations occurred at timescales longer than the time for a whole cell to move one cell length, which further demonstrates the existence of mechanical cooperativity and force transmission at length scales larger than the single cell.

In this connection, an exponential distribution of forces as reported here has been considered to be the signature of the force distributions that arise in jammed granular materials such as a pile of sand, grain in a silo or coffee beans stuck in a chute^{10–12}, although the precise asymptotic form remains an open question³⁵. The physics of jammed materials remains poorly understood, but the exponential nature of the force distribution is thought to arise from the combination of three generic features: close

packing, structural disorder and long-range force transmission³⁶. Both close packing and structural disorder are obvious properties of the cell sheet (Supplementary Movies S1 and S2), and here we provide evidence that transmission of force is long-ranged with exponentially distributed tails. Mechanics of jammed granular matter is governed by compressive stresses, however, whereas that of the cell sheet is governed by tensile stresses as demonstrated here. Although the connection between them remains a matter of speculation, the analogy between jammed inert materials and collective migration of cells is striking and suggests that these systems may share common mechanisms of long-range force transmission that are yet to be fully understood.

In summary, we present here definitive evidence establishing that collective motion in an advancing epithelial cell sheet results neither from leader cells dragging those behind, nor from cells that are individually self-propelled. Instead, each individual cell, both at the leading edge and well inside the sheet, engages in a global tug-of-war that integrates local force generation into a global state of tensile stress. Such a mechanism is innately integrative and would be inapparent in any cell studied in isolation. Whether this integrative mechanism is specific to certain tissues or instead is applicable generally during development, tissue healing and disease remains an open question, but one that is now accessible to direct experimental attack.

Methods

Cell culture. Madin–Darby canine kidney cells (strain II) were cultured on plastic flasks in MEM with Earle’s salts supplemented with 5% FBS, 2 mM L-glutamine, 100 U ml^{−1} penicillin and 100 µg ml^{−1} streptomycin. To seed a cell colony for experiments, a 5 µl drop of supplemented media containing 5,000 cells was added to the centre of the gels. Surface tension limited the drop to approximately 1 mm² of the central region of the gel. The cells were allowed to adhere to the gel for 30 min at 37 °C and 5% CO₂ before 2 ml of supplemented media was added to cover the whole surface of the dish.

Experiments. All experiments were conducted 3–4 days after seeding the cells. Three different images were collected every 60 s, one imaging cells in phase contrast, one imaging the layer of beads located immediately below the cells and one imaging the diffraction rings of the layer of beads attached to the glass underlying the gel (Supplementary Information S2). A typical experiment lasted 6–24 h. At the end of the time course experiment, cells were trypsinized using isotonic 10× trypsin for 1 h and a stack of 30 reference images of both layers of beads was acquired. To obtain traction maps of the whole cell diameter, we merged a series of overlapping images using a correlation-based algorithm. In this case, registration was achieved by equalling displacement fields over the overlapping regions and imposing zero traction outside the colony. All experiments were conducted in the presence of serum, at 37 °C and 5% CO₂.

Preparation of polyacrylamide gel substrates. Polyacrylamide substrate preparation was adapted from previously published protocols^{18,37} to enable image registration and improve resolution of the displacement fields (see Supplementary Information S2). The concentrations of crosslinker and polymer were adjusted for a Young’s modulus of 1.3 kPa (ref. 38). The details of the protocol for polyacrylamide

gel preparation including all modifications from previous publications are provided below. Step 1: A few drops of 0.1 M NaOH were added to the centre of each 35 mm dish (glass bottom, uncoated, no.0; MatTek). The dishes were air-dried overnight. Step 2: 2–3 drops of 97% 3-aminopropyltrimethoxysilane were added over the NaOH-stained circular regions from the previous step. The dishes were then washed and the glass surface was scrubbed with a foam swab to remove debris. The dishes were washed again and 400 μ l of a solution containing 0.0001% yellow fluorescent carboxylate-modified beads (2 μ m diameter, Fluospheres, Invitrogen) was added to each well. These beads were used for image registration. After the dishes dried, 0.5% glutaraldehyde in PBS was added to the central region in each dish for 30 min. The dishes were subsequently washed and air-dried overnight. Step 3: 20 μ l of an acrylamide / bis-acrylamide mixture dissolved in ultrapure water containing 0.01% of 0.5- μ m-diameter red fluorescent carboxylate-modified beads (Fluospheres, Invitrogen), 0.5% of ammonium persulphate and 0.05% TEMED (Bio-Rad) was added to the centre of each dish. This gel mixture was covered with glass cover slips (18 mm diameter; VWR). To ensure that all red beads laid in the top plane of the gel, the dishes were centrifuged at 500 r.p.m. for 15 min during gelation. Once polymerization was completed, the cover slips were removed. The surface was activated by adding 225 μ l of a solution containing 4 μ M sulphosuccinimidyl-6-(4-azido-2-nitrophenylamino) hexanoate (Sulfo-SANPAH; Pierce) dissolved in 0.1 M HEPES buffer. The dishes were then exposed to ultraviolet light for 10 min, washed twice with 0.1 M HEPES solution, washed once with PBS, coated with 1 ml of type-I collagen solution (0.1 mg ml⁻¹; Inamed Biomaterials) and stored at 4 °C. On the day before seeding the cells, the gels were washed, and incubated overnight with 3 ml of MEM with Earle's salts supplemented with 5% FBS. Then the gel surface was allowed to dry at room temperature for 2 h immediately before seeding the cells.

Fourier-transform traction microscopy. A new algorithm of traction microscopy was developed to account for finite substrate thickness and force imbalance within the microscope field of view (see Supplementary Information S3). Gel displacements were computed using correlation-based particle image velocimetry. To reduce systematic biases in subpixel resolution and peak-locking effects, we implemented an iterative process ($n = 4$ iterations) based on a continuous window shift technique³⁹. The interrogation windows were 25.6 μ m on a side and window overlap ranged from 3/4 to 7/8.

Received 18 December 2008; accepted 31 March 2009;
published online 3 May 2009

References

- Lecaudey, V. & Gilmour, D. Organizing moving groups during morphogenesis. *Curr. Opin. Cell Biol.* **18**, 102–107 (2006).
- Martin, P. & Parkhurst, S. M. Parallels between tissue repair and embryo morphogenesis. *Development* **131**, 3021–3034 (2004).
- Friedl, P. & Wolf, K. Tumour-cell invasion and migration: Diversity and escape mechanisms. *Nature Rev. Cancer* **3**, 362–374 (2003).
- du Roure, O. *et al.* Force mapping in epithelial cell migration. *Proc. Natl Acad. Sci. USA* **102**, 2390–2395 (2005).
- Vaughan, R. B. & Trinkaus, J. P. Movements of epithelial cell sheets in vitro. *J. Cell Sci.* **1**, 407–413 (1966).
- Omelchenko, T. *et al.* Rho-dependent formation of epithelial 'leader' cells during wound healing. *Proc. Natl Acad. Sci. USA* **100**, 10788–10793 (2003).
- Friedl, P., Hegerfeldt, Y. & Tusch, M. Collective cell migration in morphogenesis and cancer. *Int. J. Dev. Biol.* **48**, 441–449 (2004).
- Gov, N. S. Collective cell migration patterns: Follow the leader. *Proc. Natl Acad. Sci. USA* **104**, 15970–15971 (2007).
- Poujade, M. *et al.* Collective migration of an epithelial monolayer in response to a model wound. *Proc. Natl Acad. Sci. USA* **104**, 15988–15993 (2007).
- Liu, C. H. *et al.* Force fluctuations in bead packs. *Science* **269**, 513–515 (1995).
- O'Hern, C. S., Langer, S. A., Liu, A. J. & Nagel, S. R. Force distributions near jamming and glass transitions. *Phys. Rev. Lett.* **86**, 111–114 (2001).
- Ostojic, S., Somfai, E. & Nienhuis, B. Scale invariance and universality of force networks in static granular matter. *Nature* **439**, 828–830 (2006).
- Lauffenburger, D. A. & Horwitz, A. F. Cell migration: A physically integrated molecular process. *Cell* **84**, 359–369 (1996).
- Keren, K. *et al.* Mechanism of shape determination in motile cells. *Nature* **453**, 475–480 (2008).
- Hu, K. *et al.* Differential transmission of actin motion within focal adhesions. *Science* **315**, 111–115 (2007).
- Giannone, G. *et al.* Lamellipodial actin mechanically links myosin activity with adhesion-site formation. *Cell* **128**, 561–575 (2007).
- Beningo, K. A. *et al.* Nascent focal adhesions are responsible for the generation of strong propulsive forces in migrating fibroblasts. *J. Cell Biol.* **153**, 881–888 (2001).
- Dembo, M. & Wang, Y. L. Stresses at the cell-to-substrate interface during locomotion of fibroblasts. *Biophys. J.* **76**, 2307–2316 (1999).
- Montell, D. J. Morphogenetic cell movements: Diversity from modular mechanical properties. *Science* **322**, 1502–1505 (2008).
- Matsubayashi, Y., Ebisuya, M., Honjoh, S. & Nishida, E. ERK activation propagates in epithelial cell sheets and regulates their migration during wound healing. *Curr. Biol.* **14**, 731–735 (2004).
- Bindschadler, M. & McGrath, J. L. Sheet migration by wounded monolayers as an emergent property of single-cell dynamics. *J. Cell Sci.* **120**, 876–884 (2007).
- Holmes, S. J. The behaviour of the epidermis of amphibians when cultivated outside the body. *J. Exp. Zool.* **17**, 281–295 (1914).
- Hutson, M. S. *et al.* Forces for morphogenesis investigated with laser microsurgery and quantitative modeling. *Science* **300**, 145–149 (2003).
- Farooqui, R. & Fenteany, G. Multiple rows of cells behind an epithelial wound edge extend cryptic lamellipodia to collectively drive cell-sheet movement. *J. Cell Sci.* **118**, 51–63 (2005).
- Butler, J. P., Tolic-Norrelykke, I. M., Fabry, B. & Fredberg, J. J. Traction fields, moments, and strain energy that cells exert on their surroundings. *Am. J. Physiol. Cell Physiol.* **282**, C595–C605 (2002).
- Sabass, B., Gardel, M. L., Waterman, C. M. & Schwarz, U. S. High resolution traction force microscopy based on experimental and computational advances. *Biophys. J.* **94**, 207–220 (2008).
- Del Alamo, J. C. *et al.* Spatio-temporal analysis of eukaryotic cell motility by improved force cytometry. *Proc. Natl Acad. Sci. USA* **104**, 13343–13348 (2007).
- Merkel, R., Kirchgessner, N., Cesa, C. M. & Hoffmann, B. Cell force microscopy on elastic layers of finite thickness. *Biophys. J.* **93**, 3314–3323 (2007).
- Fenteany, G., Janmey, P. A. & Stossel, T. P. Signaling pathways and cell mechanics involved in wound closure by epithelial cell sheets. *Curr. Biol.* **10**, 831–838 (2000).
- Goffin, J. M. *et al.* Focal adhesion size controls tension-dependent recruitment of alpha-smooth muscle actin to stress fibers. *J. Cell Biol.* **172**, 259–268 (2006).
- Saez, A., Buguin, A., Silberzan, P. & Ladoux, B. Is the mechanical activity of epithelial cells controlled by deformations or forces? *Biophys. J.* **89**, L52–L54 (2005).
- Gov, N. S. Modeling the size distribution of focal adhesions. *Biophys. J.* **91**, 2844–2847 (2006).
- Zegers, M. M. *et al.* Pak1 and PIX regulate contact inhibition during epithelial wound healing. *EMBO J.* **22**, 4155–4165 (2003).
- Shraiman, B. I. Mechanical feedback as a possible regulator of tissue growth. *Proc. Natl Acad. Sci. USA* **102**, 3318–3323 (2005).
- Van Hecke, M. Granular matter: A tale of tails. *Nature* **435**, 1041–1042 (2005).
- Coppersmith, S. N. *et al.* Model for force fluctuations in bead packs. *Phys. Rev. E* **53**, 4673–4685 (1996).
- Wang, N. *et al.* Cell prestress. I. Stiffness and prestress are closely associated in adherent contractile cells. *Am. J. Physiol. Cell Physiol.* **282**, C606–C616 (2002).
- Yeung, T. *et al.* Effects of substrate stiffness on cell morphology, cytoskeletal structure, and adhesion. *Cell. Motil. Cytoskeleton* **60**, 24–34 (2005).
- Gui, L. & Wereley, S. T. A correlation-based continuous window-shift technique to reduce the peak-locking effect in digital PIV image evaluation. *Exp. Fluids* **32**, 506–517 (2002).

Acknowledgements

We thank N. Gavara, R. Sunyer and C. Y. Park for experimental support and D. Tschumperlin and members of the Fredberg laboratory for insightful discussions.

Author contributions

X.T., J.P.B. and J.J.F. designed research. X.T. and M.R.W. carried out experiments. J.P.B. and X.T. conducted theoretical analysis. T.E.A., D.A.W. and E.M. contributed to design protocols and data interpretation. X.T., J.P.B. and J.J.F. wrote the manuscript. J.P.B. and J.J.F. oversaw the project.

Additional information

Supplementary information accompanies this paper on www.nature.com/naturephysics. Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>. Correspondence and requests for materials should be addressed to X.T. or J.J.F.

Vortex nucleation as a case study of symmetry breaking in quantum systems

D. Dagnino¹, N. Barberán^{1*}, M. Lewenstein^{2,3} and J. Dalibard⁴

Mean-field methods are a very powerful tool for investigating weakly interacting many-body systems in many branches of physics. In particular, they describe with excellent accuracy trapped Bose–Einstein condensates. A generic, but difficult question concerns the relation between the symmetry properties of the true many-body state and its mean-field approximation. Here, we address this question by considering, theoretically, vortex nucleation in a rotating Bose–Einstein condensate. A slow sweep of the rotation frequency changes the state of the system from being at rest to the one containing one vortex. Within the mean-field framework, the jump in symmetry occurs through a turbulent phase around a certain critical frequency. The exact many-body ground state at the critical frequency exhibits strong correlations and entanglement. We believe that this constitutes a paradigm example of symmetry breaking in—or change of the order parameter of—quantum many-body systems in the course of adiabatic evolution.

In classical physics, examples of the usefulness of mean-field theory go back to the ‘molecular field theory’ of magnetism¹. In the classical world, symmetry changes (or breaking) are driven by thermal fluctuations, and in the standard Landau–Ginsburg scenario are associated with an increase of classical correlations. In quantum physics, the paradigm example of applicability of the mean field concerns a weakly interacting quantum Bose gas and Bose–Einstein condensation². The mean-field description of the gas assumes that its ground state Ψ is approximated by a product state $\Psi(\mathbf{r}_1, \dots, \mathbf{r}_N) = \psi(\mathbf{r}_1) \dots \psi(\mathbf{r}_N)$, of essentially uncorrelated particles forming a superfluid Bose–Einstein condensate with order parameter ψ .

Of particular interest for quantum gases are quantum phase transitions and symmetry changes/breaking driven by quantum fluctuations. A celebrated example is the superfluid to Mott-insulator transition of bosons in an optical lattice³. Another example yet to be explored experimentally is the case of a fast rotating gas, when the number of vortices is similar to the number of particles, or equivalently angular momentum $L \sim N^2$ (ref. 4). The ground state of the system is then a strongly correlated quantum liquid such as the Laughlin state, analogous to those emerging in quantum Hall physics⁵. Here, we consider another situation, dealing with the case of a relatively slowly rotating gas at the threshold of the nucleation of the first vortex. We show that owing to the symmetries of the system, the many-body state at nucleation is strongly correlated and characterizes its properties.

The symmetry change/breaking that results from vortex nucleation has drawn a lot of attention since the discovery of superfluids⁶. For quantum gases, atoms are usually confined in an isotropic harmonic trap and experience an extra quadratic potential rotating at angular frequency Ω (for a review, see ref. 7). From a theoretical point of view, the vortex nucleation can be tackled by several techniques, ranging from a mean-field approach based on the Gross–Pitaevskii equation^{8–10} to the investigation of the many-body energy eigenstates^{11–17}. Within the mean-field framework, standard textbooks² associate vortex nucleation with

thermodynamic instability. Above a critical rotation frequency Ω_c , the odd solution ψ of the Gross–Pitaevskii equation with a single vortex^{18,19} has a lower energy than the even solution corresponding to the Bose–Einstein condensate at rest²⁰. Here, we go beyond the mean-field approach and study the exact quantum dynamics of a mesoscopic sample of atoms, in the presence of the stirring potential. Our main result is that for a rotation frequency close to Ω_c , the mean-field description is invalid. The system enters a strongly correlated and entangled state, well described by an effective two-mode model. We compare our results with those obtained from a mean-field description and show that the latter exhibits dynamical instability and hysteresis. As we explicitly include here an anisotropic stirring potential, the present mechanism concerns a discrete parity symmetry breaking. Therefore, it differs from the case of the vortex nucleation in axially symmetric traps: in the latter case, breaking of the continuous rotational symmetry involves a gapless Nambu–Goldstone mode²¹, whereas here we deal with a gapped system.

Model

We consider a mesoscopic sample of N bosonic atoms of mass M placed in an axially symmetric harmonic potential V_0 , with frequency ω_\perp in the xy plane and ω_z along the z axis. Here, $\hbar\omega_z$ is large compared with the interaction energy so that the dynamics along z is frozen and the gas is effectively two-dimensional (2D) at sufficiently low temperature. The gas is set in rotation using an anisotropic quadratic potential V in the xy plane, rotating at angular frequency Ω around the z axis. In the rotating frame, this stirring potential reads $V(x, y) = 2AM\omega_\perp^2(x^2 - y^2)$, where the coefficient A ($\ll 1$) measures the strength of the anisotropy.

For $A \ll 1$ and $\Omega \sim \omega_\perp$, the single-particle energy levels in the rotating frame are grouped in Landau levels, separated by $\hbar(\omega_\perp + \Omega)$ (refs 7, 22). We assume that $\hbar(\omega_\perp + \Omega)$ is large compared with the interaction energy, so that the atomic dynamics is restricted to the lowest Landau level (LLL). For $A = 0$, a basis of the LLL

¹Dept. ECM, Facultat de Física, Universitat de Barcelona, E-08028 Barcelona, Spain, ²ICFO - Institut de Ciències Fotòniques, Parc Mediterrani de la Tecnologia, 08860 Barcelona, Spain, ³ICREA-Institució Catalana de Recerca i Estudis Avançats, E-08010 Barcelona, Spain, ⁴Laboratoire Kastler Brossel, CNRS, UPMC, Ecole Normale Supérieure, 24 rue Lhomond, 75005 Paris, France. *e-mail: nuria.barberan@gmail.com.

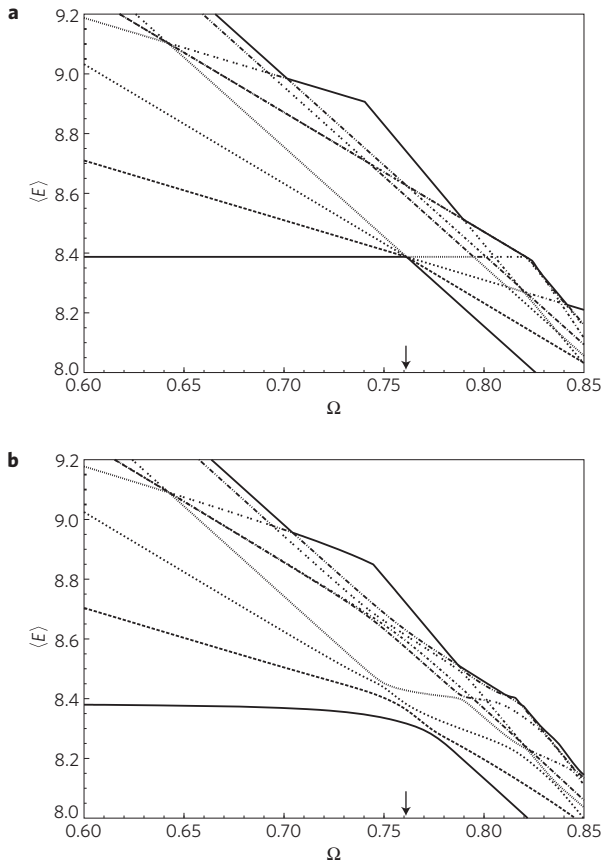


Figure 1 | Energy spectrum as a function of Ω . **a**, Anisotropy parameter $A = 0$. **b**, $A = 0.03$. In both cases, $N = 6$ and $g = 1$. For $A = 0$, the ground state is multiply degenerate at the rotation frequency $\Omega_1 = 1 - gN/(8\pi)$, which corresponds to the nucleation frequency of the first vortex. A non-zero anisotropy parameter lifts the degeneracy of the ground state. Here, we plot only the first nine energy eigenvalues from the subspace formed with even values of the total angular momentum, which are the only relevant ones for the problem addressed in this article. The arrows mark the value of Ω_1 .

single-particle states is the set $\varphi_m(x, y) \propto (x + iy)^m e^{-(x^2 + y^2)/2\lambda_\perp^2}$, where $m \geq 0$ is an integer and $\lambda_\perp = \sqrt{\hbar/M\omega_\perp}$. Each φ_m is an eigenstate of the z -component of the single-particle angular momentum (eigenvalue $m\hbar$) and of the single-particle Hamiltonian without anisotropy (eigenvalue $\hbar(\omega_\perp + m(\omega_\perp - \Omega))$). Within the LLL, we model the atomic interactions by a 2D contact potential $U(\mathbf{r}) = (\hbar^2 g/M) \delta(\mathbf{r})$, where $g = \sqrt{8\pi} a/\lambda_z$ is dimensionless, a is the 3D scattering length and $\lambda_z = \sqrt{\hbar/M\omega_z}$. We choose λ_\perp , $\hbar\omega_\perp$ and ω_\perp as units of length, energy and frequency.

Energy spectrum

We first recall some important properties of the N -particle system in the absence of anisotropy ($A = 0$). In this case, the total angular momentum operator \hat{L} commutes with the Hamiltonian \hat{H} so that one can look for the eigenstates of \hat{H} within subspaces \mathcal{E}_L of fixed L . The lowest-energy state in each \mathcal{E}_L for $2 \leq L \leq N$ is^{12–14}:

$$\Phi_L(\mathbf{r}_1, \dots, \mathbf{r}_N) \propto \sum_{1 \leq i_1, \dots, i_L \leq N} (u_{i_1} - u_c) \dots (u_{i_L} - u_c) \Phi_0$$

where $u_j = x_j + iy_j$, $u_c = \sum_j u_j/N$ and

$$\Phi_0(\mathbf{r}_1, \dots, \mathbf{r}_N) \propto e^{-\sum_j r_j^2/2}$$

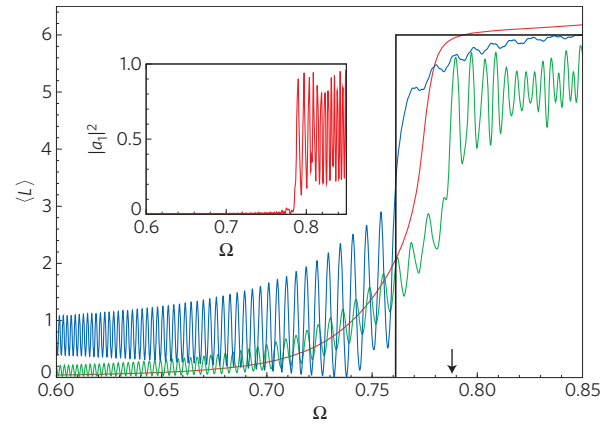


Figure 2 | Variation of the angular momentum with rotation frequency Ω .

The black and red lines show, for an anisotropy $A = 0$ and $A = 0.03$ respectively, the angular momentum of the ground state for a system of $N = 6$ particles and an interaction strength $g = 1$. The green line is the average angular momentum predicted by the mean-field treatment when Ω is linearly ramped from 0 to 0.85 with a slope $\dot{\Omega} = 10^{-4}$. The initial state at $\Omega = 0$ is given by a slight perturbation of the coefficients $a_0 = 1$, $a_1 = a_2 = 0$. It presents a dynamical instability of the zero-vortex mean-field solution for $\Omega = 0.788$ (marked by an arrow). Inset: The evolution of $|a_1|^2$, which explicitly shows the instability. The blue curve of the main figure is the backward evolution corresponding to an initial state at $\Omega = 0.85$ close to the stationary mean-field solution $a_0 = a_2 = 0$ and $a_1 = 1$. This solution ceases to exist for $\Omega < 0.764$, causing the large oscillations in the evolution of the angular momentum.

The energy of the state Φ_L is $N + (1 - \Omega)L + gN(2N - L - 2)/(8\pi)$. At $\Omega_1 = 1 - gN/(8\pi)$, all Φ_L states for $L = 0$ and $2 \leq L \leq N$ are degenerate. The angular momentum of the ground state $L_{GS}(\Omega)$ shows sharp steps at critical values Ω_i , $i = 1, 2, \dots$ (ref. 23). Below Ω_1 , the ground state is the zero angular momentum state Φ_0 . At Ω_1 , L_{GS} jumps from 0 to N . Above Ω_1 , the ground-state angular momentum has a plateau $L = N$ up to Ω_2 , where a second jump takes place. From this value, a sequence of jumps and plateaux emerges up to the last possible L value, $L = N(N - 1)$, corresponding to the Laughlin state. In the following, we focus on the vicinity of the first jump $\Omega \sim \Omega_1$, where the first vortex is nucleated.

We now turn to the case where the rotating anisotropy is present. The many-body energy spectrum is calculated numerically by diagonalization of the Hamiltonian (see the Methods section). We show it in Fig. 1 for both zero anisotropy and for $A = 0.03$, using $N = 6$ for illustration. The interaction coupling $g = 1$ so that $\Omega_1 = 0.761$. For $A \neq 0$, the ground state does not show any degeneracy around Ω_1 , contrary to the case $A = 0$. In Fig. 2, we compare $L_{GS}(\Omega)$ for $A = 0$ and $A = 0.03$. For $A \neq 0$, L_{GS} evolves smoothly from 0 to N around Ω_1 .

Failure of the mean-field approach for $\Omega \sim \Omega_1$

We now explain why a mean-field description must fail at $\Omega \simeq \Omega_1$. We notice that the total Hamiltonian is parity invariant. Consequently, one can look for an eigenbasis of the N -body Hilbert space composed of either even or odd states. From the ground state of the Hamiltonian, we can extract the single-particle density matrix (SPDM) $n^{(1)}(\mathbf{r}, \mathbf{r}')$ (see the Methods section), which is also parity invariant. Hence, the single-particle orbitals ψ_k , which are eigenstates of $n^{(1)}$ with eigenvalues n_k ($\sum_k n_k = N$), can also be chosen with even or odd parity. Suppose that we vary Ω from an initial value Ω_i ($\Omega_i < \Omega_1$) to a final value Ω_f ($\Omega_1 < \Omega_f < \Omega_2$), choosing $\Omega_{i,f}$ in a region where the mean-field description is valid, that is, when the largest eigenvalue n_1 is close to N . For $\Omega_i < \Omega_1$, the most (second most) populated state ψ_1 (ψ_2) has no (has a)

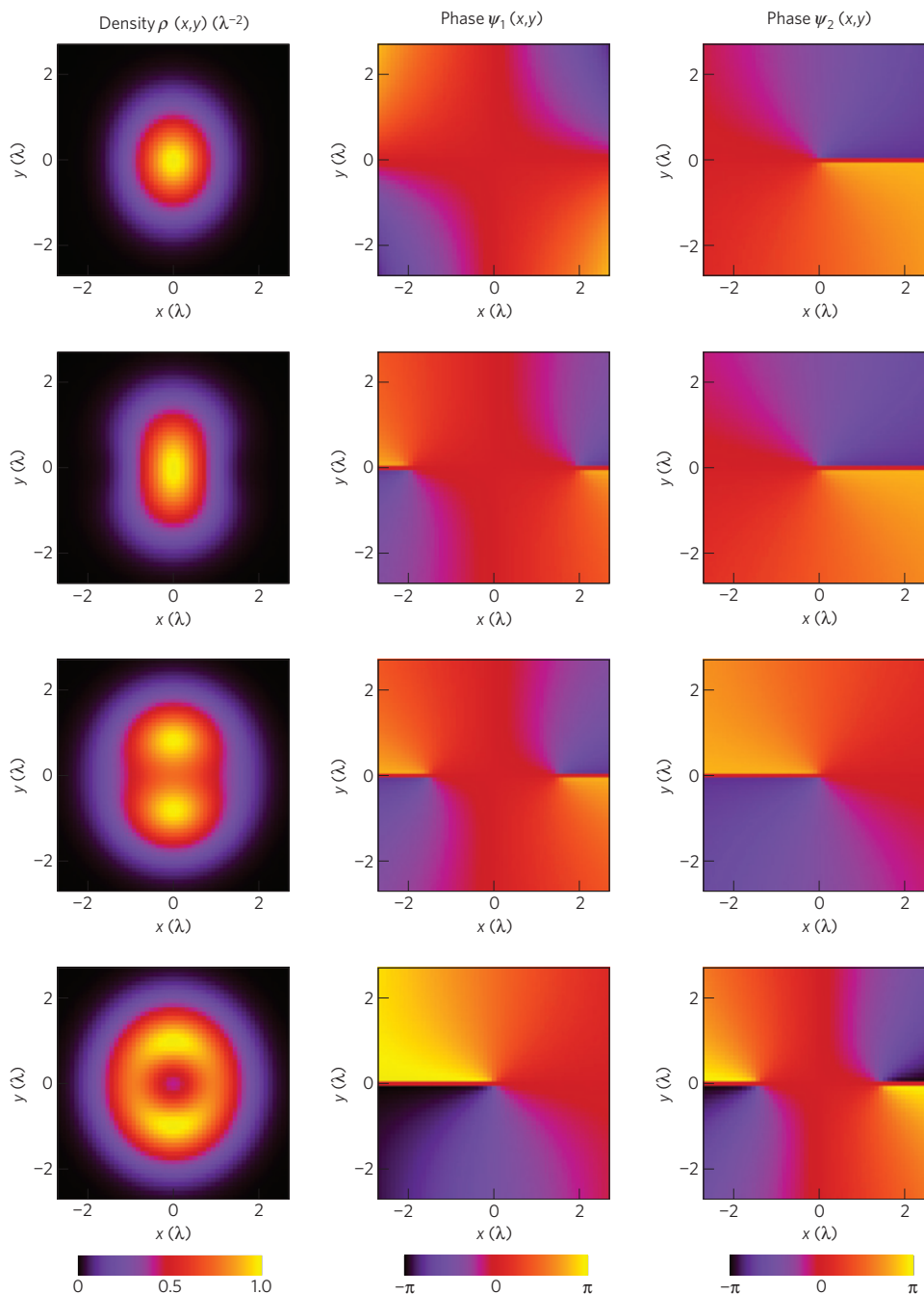


Figure 3 | Density of the ground state and phase maps of ψ_1 and ψ_2 . Four different values of Ω for $N = 6$, $A = 0.03$ and $g = 1$ are considered. First row: $\Omega = 0.7$, $n_1 = 5.85$, $n_2 = 0.12$. Second row: $\Omega = 0.760$, $n_1 = 5.01$, $n_2 = 0.60$. Third row: $\Omega = \Omega_c = 0.776$, $n_1 = n_2 = 2.88$. Fourth row: $\Omega = 0.8$ with $n_1 = 4.24$, $n_2 = 1.07$. The first column is the contour plot of the total density, and the second and third columns show the local phase maps of ψ_1 and ψ_2 , respectively. Vortices are localized at the singularities of the phase maps, surrounded by diffuse change of the phase. This figure shows that the nucleation of the first centred vortex in a rotating condensate by a slow frequency sweep does not occur through a smooth entrance of the vortex. The system passes through a correlated, non-mean-field state where two single-particle states have equal weight. At this point, ψ_1 changes from being a coherent superposition of φ_0 and φ_2 (two off-centred vortices) to the single φ_1 state, which corresponds to a well-centred single vortex. Simultaneously, ψ_2 experiences the inverse change.

vortex in its central region and is even (odd). Choosing $\Omega_i = 0.7$, we plot the phase profiles of $\psi_{1,2}$ in the first row of Fig. 3 for $N = 6$ atoms, $g = 1$ and $A = 0.03$. On the other hand, at Ω_f the ground state has a single well-centred vortex and ψ_1 and ψ_2 are odd and even, respectively (see the last row of Fig. 3 for $\Omega = 0.8$). Hence, the parity of ψ_1 must change at some intermediate Ω_c , which is close (for small A) to the vortex nucleation frequency Ω_1 in the absence of anisotropy. By continuity, the two most populated eigenstates ψ_1

and ψ_2 of $n^{(1)}$ must have equal populations, heralding a failure of the mean field at Ω_c .

We show in Fig. 4 the variation of n_1/N and n_2/N as a function of Ω , for $N = 12$, $g = 0.5$ and $A = 0.03$. These two populations are equal for $\Omega_c = 0.775$. We see that $n_1 + n_2 \simeq N$ over the whole range of frequencies of this figure, indicating that most of the population of the SPDM is concentrated in the first two modes ψ_1 and ψ_2 . We checked up to $N = 20$

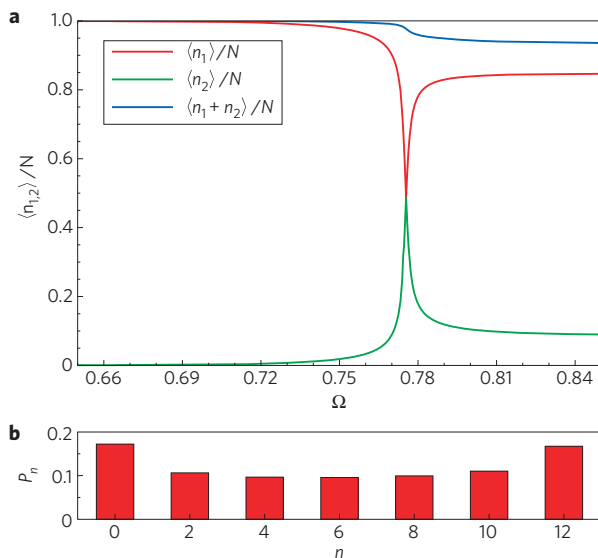


Figure 4 | Structure of the ground state. **a**, Variation of the relative populations n_1/N and n_2/N of the two most occupied states ψ_1 and ψ_2 of the SPDM. When Ω is sufficiently different from Ω_c , $n_1 \simeq N$, the system is well described by a single mode and the mean-field description is valid. Conversely, for $\Omega \simeq \Omega_c$, the two populations are comparable, corresponding to the case where a two-mode approximation is valid even in the entangled region. **b**, Analysis of the state of the system at the critical point where $n_1 = n_2$, in terms of the square of the scalar products $P_n = |\langle n | \psi_1; N-n | \psi_2 \rangle \Psi_0|^2$. We obtain $|\langle E | \Psi_0 \rangle| = 0.92$ (see equation (1)). Both panels are plotted for $N=12$, $g=0.5$ and $A=0.03$.

that this concentration increases with N . Another relevant fact is that only the first three LLL single-particle states ($m=0, 1, 2$) have a significant weight in the expansion of ψ_1 and ψ_2 . More specifically, below Ω_c , ψ_1 is approximately a coherent superposition of φ_0 and φ_2 , corresponding to two off-centred vortices (even parity), whereas ψ_2 is very close to a well-centred single-vortex state φ_1 (odd parity). Above Ω_c , ψ_1 and ψ_2 abruptly exchange their form (see Fig. 3).

The failure of the mean-field description around Ω_c may occur in two ways. A first possibility is that for $\Omega = \Omega_c$, the many-body ground level itself has a two-fold degeneracy with two eigenstates of opposite parity. This scenario corresponds to a first-order transition. It occurs when N is odd, because the ground state evolves from $\sim \psi_1^{\otimes N}$ with ψ_1 even to $\sim \psi_1^{\otimes N}$ with ψ_1 odd. The second possibility is that the many-body ground state $|\Psi_0\rangle$ remains non-degenerate, as is the case in Fig. 1b. In this case, $|\Psi_0\rangle$ is even over the whole range $[\Omega_i, \Omega_f]$. This occurs for even N and will be of interest for the rest of the article.

Quantum correlations for $\Omega \sim \Omega_c$

We have carried out a detailed study of the ground state $|\Psi_0\rangle$ around the critical frequency Ω_c , where the two largest eigenvalues of the SPDM are equal ($n_1 = n_2$). At criticality, the system is very well described by a two-mode approximation implied by Fig. 4a. The two largest eigenvalues of the SPDM are much larger than all of the others, so that $n_1 = n_2 \simeq N/2$. For example, for $N=12$, $g=0.5$ and $A=0.03$, we obtain $n_1 = n_2 = 0.49N$ at $\Omega_c = 0.776$. The ground state is strongly correlated and is well described (for even N) by

$$|E\rangle = [|N, 0\rangle + |N-2, 2\rangle + \dots + |0, N\rangle] / \sqrt{N/2 + 1} \quad (1)$$

where $|n, m\rangle$ is the state with n (respectively m) atoms in ψ_1 (respectively ψ_2). Amazingly, the form of the ground state at

Ω_c is practically independent of A , as long as $A \ll 1$. For a quantitative comparison of the exact ground state with the state (1), we show in Fig. 4b the squared scalar products $\langle n, N-n | \Psi_0 \rangle$ in the case $N=12$. They are all zero for odd values of n (as expected from the parity of $|\Psi_0\rangle$) and approximately constant for even values of n . We also compared our ground state at Ω_c with other celebrated correlated states, such as ‘Schrödinger cat’ states $(|N, 0\rangle + |0, N\rangle)/\sqrt{2}$ or ‘twin’ states $|N/2, N/2\rangle$, and found much smaller overlaps. Although there are various ways of defining entanglement for identical particles (for a review, see ref. 24), according to Zanardi’s concept of mode entanglement²⁵, the state (1) is maximally entangled. This is clearly seen by tracing the state (1) over one of the two modes and observing that the von Neumann entropy of the reduced density matrix reaches the maximal value $S \sim \log(N)$.

At this point we mention related work on rotating ring lattices and Josephson junctions²⁶. There, strongly correlated states are predicted at critical rotation, but the mechanism of their generation, as well as their nature is fundamentally different. The starting situation of these discretized models is that there are two degenerated single-particle states. Interactions lift the degeneracy in the many-body system and favour the ‘cat’ states. In our case, the ground state for $A=0$ is macroscopically degenerated in the presence of interactions. The degeneracy is lifted here by the anisotropy, leading to another kind of strongly correlated ground state.

Vortex nucleation with adiabatic passage

We now study the real-time dynamics of the system using the time-dependent Schrödinger equation. A quasi-adiabatic evolution that brings the system from the zero-vortex to the one-vortex state, is realized by sweeping Ω : $\Omega(t) = \Omega_i + \gamma t$ from the initial frequency Ω_i chosen well below Ω_c (typically $\Omega_i = 0.65$) to the final frequency Ω_f , well above Ω_c (typically $\Omega_f = 0.85$). This evolution produces as an intermediate step the strongly correlated state (1). The key parameter for the success of this quasi-adiabatic evolution is the energy gap Δ between the ground state and the first excited state of the system.

We have carried out a study of this gap for various N , keeping the product Ng constant so that Ω_i also remains constant. We found that for small A values (below 0.1), the gap is roughly constant over the range $10 \leq N \leq 20$, and equal to $\simeq 0.5A$. Knowing the gap, we estimate the largest possible γ compatible with adiabatic evolution following ref. 27 and find $\gamma_{\max} = \xi \Delta^2/N$, where $\xi \ll 1$ (see the Methods section). This criterion agrees well with our results. Defining as successful an adiabatic evolution that leads to an overlap larger than 0.98 between the final state and the ground state at Ω_f , we find $\xi \simeq 0.1$ for $10 \leq N \leq 20$. Such a quasi-adiabatic evolution enables us to attain the correlated state (1) with comparable overlap. For practical implementations, the atoms can be confined in a relatively tight trap at the nodes of an optical lattice with $\omega_{\perp}/2\pi$ in the 10 kHz range. For an anisotropy $A=0.1$ and $N=10$ atoms, the sweep time has to be of the order of one second to ensure adiabaticity.

A natural question is the generalization of the present scheme to large N . Assuming that the gap protecting the ground state remains constant, the mechanism will in principle survive. However, we have neglected here any parity-breaking perturbation in the Hamiltonian. Such a term would couple the subspaces corresponding to even and odd L values. As shown in ref. 17, the lowest energies of these two subspaces are exponentially close when N increases, which affects the robustness of the ground state. This coupling thus constitutes an important decoherence mechanism for large N , whereas our scheme remains valid for N not exceeding a few tens.

Mean-field approach

As our results point out that strongly correlated states may be reached in the course of the time evolution, it is interesting to see what the predictions of usual mean-field theory are. To this aim, we expand the condensate wavefunction $f(\mathbf{r}, t)$ into the relevant single-particle LLL orbitals $\varphi_m(\mathbf{r})$ with angular momentum $m = 0, 1, 2$, $f(\mathbf{r}, t) = \sum_{m=0}^2 a_m(t) \varphi_m(\mathbf{r})$. Using the dynamical variational principle²⁸, we derive Lagrange equations for the complex amplitudes $a_m(t)$ (see the Methods section), and look for the stationary solutions of the form $a_m(t) = \exp(-i\mu t) a_m(0)$ and their stability. Finally, we evolve the mean-field equations and compare the results with the full quantum treatment.

We choose $gN = 6$ and $A = 0.03$. Among the several possible stationary solutions, two of them are relevant. The first one f_a corresponds to the ‘no-vortex’ situation with a small admixture of ‘two-vortex’ orbital, that is, $|a_0| \simeq 1$, $|a_2| \ll 1$ and $a_1 = 0$. This solution is the ground state for $\Omega < \tilde{\Omega} = 0.773$. The second relevant solution f_b contains a non-zero contribution of the one-vortex state ($a_1 \neq 0$) and it is the ground state for $\Omega > \tilde{\Omega}$. Thus, $\tilde{\Omega}$ marks the critical value of Ω for the thermodynamical stability of a centred vortex, and the ‘first-order transition’ within the mean-field approach.

In the frequency range $0.764 < \Omega < 0.788$, both solutions exist and are stable, leading to a bistable and hysteresis behaviour. For $\Omega > 0.788$, f_a becomes dynamically unstable ($|a_1|$ grows exponentially in time, starting from noise, see inset in Fig. 2). For $\Omega < 0.764$, f_b does not exist. The numerical study confirms this hysteresis behaviour, as shown in Fig. 2. The green line shows the angular momentum when Ω is ramped linearly in time from $\Omega_i = 0$ to $\Omega_f = 0.85$, with the rate $\dot{\Omega} = 10^{-4}$. A turbulent behaviour occurs once $\Omega(t)$ reaches the edge of the stability domain of f_a . The blue line shows the reverse evolution in which Ω varies from Ω_f to Ω_i at the same rate. Evidently, the adiabatic character of the dynamics cannot be maintained, in contrast to the result of the exact many-body treatment.

Summary

We conjecture that the scenario presented above is generic for the following situations: (1) it concerns quantum mechanical systems in which the ground state undergoes symmetry change/breaking as some parameter of the system λ crosses a critical value λ_c ; (2) far from λ_c , the systems are well described by the mean-field theory with order parameters reflecting the change of symmetry; (3) in the dynamical mean-field description, the system exhibits dynamical instability and breakdown of adiabaticity.

In such situations we expect the appearance of strongly correlated states. The SPDM shows typically a few relevant single-particle modes that are involved in the symmetry change. They can be guessed by analysing the results of the dynamical mean-field approach. For instance, if this approach exhibits standard signatures of bistability, we can expect two relevant modes as in the case study presented here. Similar insight can be gained from analysis of small Gaussian fluctuations around the mean-field solutions, that is, Bogoliubov–de Gennes equations²⁹. Reduction of the full theory to the quantum modes provides then a very good approximation. Alternatively, it can be viewed as re-quantization of the mean-field theory reduced to the relevant single-particle orbitals¹⁷. The strongly correlated states appearing in such a situation exhibit strong entanglement and this property can be detected in experiments with moderate N .

Methods

Diagonalization of the Hamiltonian. In the frame rotating at angular frequency Ω , the Hamiltonian of the system is $H = H_0 + U$, where H_0 is the sum of one-body Hamiltonians $H_0 = \sum_{j=1}^N H_{0,j}$ and U is the two-body interaction potential,

characterized by the 3D scattering length a . Each one-body Hamiltonian is the sum of kinetic, potential and rotation energy:

$$H_{0,j} = \frac{p_j^2 + p_{z,j}^2}{2M} + \frac{M}{2} (\omega_\perp^2 r_j^2 + \omega_\parallel^2 z_j^2) - \Omega L_{z,j} + V_j$$

where $V_j = 2AM\omega_\perp^2(x_j^2 - y_j^2)$ is the anisotropic potential that sets the gas in rotation. We assume that the interaction energy is much smaller than $\hbar\omega_z$ so that the z motion is frozen and the atoms occupy only the ground state $\exp(-z^2/(2\lambda_z^2))$ of this degree of freedom. The gas is supposed to be rotating sufficiently fast to have $\omega_\perp - \Omega \ll \omega_\perp + \Omega$, which guarantees that the various Landau levels are well separated from each other. The interaction energy is also assumed to be small compared with $\hbar(\omega_\perp + \Omega)$ so that the low-temperature dynamics is restricted to the LLL.

In the absence of anisotropic potential $A = 0$, the eigenstates of the one-body Hamiltonian in the LLL are the functions $\varphi_m(x, y) \propto (x + iy)^m e^{-(x^2 + y^2)/(2\lambda_\perp^2)}$, $m = 0, 1, 2, \dots$. We introduce the creation a_m^\dagger and annihilation a_m operators of an atom in state φ_m , and we write H in the second quantization

$$\hat{H} = \hbar\omega_\perp \hat{N} + \hbar(\omega_\perp - \Omega) \hat{L} + \hat{V} + \hat{U}$$

where $\hat{N} = \sum a_m^\dagger a_m$ and $\hat{L} = \sum m a_m^\dagger a_m$ are the particle number operator and the total z -component angular momentum operator, respectively. The expression of the rotating potential in the second quantization is

$$\hat{V} = \frac{A}{2} \lambda_\perp^2 \sum_m \left(\sqrt{m(m-1)} a_m^\dagger a_{m-2} + \sqrt{(m+1)(m+2)} a_m^\dagger a_{m+2} \right)$$

Finally the contact interaction potential reads

$$\hat{U} = \frac{1}{2} \sum_{m_1 m_2 m_3 m_4} U_{1234} a_{m_1}^\dagger a_{m_2}^\dagger a_{m_3} a_{m_4}$$

where the matrix elements are given by

$$U_{1234} = \langle m_1 m_2 | U | m_3 m_4 \rangle = \frac{g}{\lambda_\perp^2 \pi} \frac{\delta_{m_1+m_2, m_3+m_4}}{\sqrt{m_1! m_2! m_3! m_4!}} \frac{(m_1 + m_2)!}{2^{m_1+m_2+1}}$$

In the absence of anisotropy ($A = 0$), \hat{H} and \hat{L} commute and share a common basis. The first step in the diagonalization of the Hamiltonian is to determine a basis $|\Lambda_p\rangle$ ($p = 1, \dots, n_L$) for each subspace of given total angular momentum L . The dimension n_L of each subspace corresponds to all of the possible configurations of N particles with angular momentum m_j that fulfil the condition $L = \sum_{j=1}^N m_j$. The matrix of the Hamiltonian in the LLL basis then consists of blocks of size $n_L \times n_L$, which we diagonalize using standard codes.

When $A \neq 0$, the anisotropic potential connects the various subspaces of given L . We then choose a maximum angular momentum L_{\max} and write the matrix giving the restriction of the Hamiltonian to the subspace of states with $L \leq L_{\max}$. This $Q \times Q$ matrix, with $Q = \sum_{L=0}^{L_{\max}} n_L$, is again diagonalized using standard codes. In practice the value of L_{\max} is chosen to ensure a good convergence for the energies and the eigenstates of the Hamiltonian. The results given here have been obtained with $L_{\max} = N + 2$.

Note that the anisotropic rotating contribution V can in principle be included within the framework of the LLL approximation in two ways. The first approach has just been described above and consists of keeping the same Landau levels as for $A = 0$ and then diagonalizing \hat{H} within the LLL. The second approach consists of calculating exactly the single-particle eigenstates in the presence of the anisotropy V , and defining a new LLL accordingly³⁰. The Hamiltonian is then diagonalized within this ‘anisotropic’ LLL. We have checked that both methods lead to very similar results for $\Omega \sim \Omega_1$. The results presented here have been obtained with the first approach.

Single-particle density matrix. The SPDM can be regarded as an integral operator with the kernel:

$$n^{(1)}(\mathbf{r}, \mathbf{r}') = \langle \Psi_0 | \hat{\Psi}^\dagger(\mathbf{r}) \hat{\Psi}(\mathbf{r}') | \Psi_0 \rangle$$

with $\hat{\Psi}(\mathbf{r})$ and $\hat{\Psi}^\dagger$ being the annihilation and creation field operators of an atom in \mathbf{r} . The single-particle orbitals are the eigenstates of the SPDM:

$$\int d\mathbf{r}' n^{(1)}(\mathbf{r}, \mathbf{r}') \psi_k^*(\mathbf{r}') = n_k \psi_k(\mathbf{r})$$

If there exists a single relevant eigenvalue such that $n_1 \gg \sum_{k \geq 2} n_k$, then $\sqrt{n_1} \psi_1(\mathbf{r})$ has the role of the order parameter of the system. In particular, the map of the local phase of this function gives precise information on the location of vortices¹⁵.

Adiabatic approximation. The diagonalization of the many-body Hamiltonian provides the eigenstates $|\Psi_j(\Omega)\rangle$ and the eigenenergies $E_j(\Omega)$. In particular, the ground state $|\Psi_0(\Omega)\rangle$ is separated from the first excited state $|\Psi_1(\Omega)\rangle$ by an energy gap $\hbar\omega_{10}(\Omega)$, which is minimal at the avoided crossing close to Ω_1 . We consider here a process where Ω is scanned linearly from $\Omega_i < \Omega_1$ to $\Omega_f > \Omega_1$ and we want to find a criterion on Ω ensuring that the system follows adiabatically the ground state, with a negligible transition rate to the other states.

The probability for a non-adiabatic transition $\Psi_0 \rightarrow \Psi_j$ is given by²⁷:

$$p_{0 \rightarrow j} \leq \max \left(\frac{\alpha_{j0}}{\omega_{j0}} \right)^2$$

where $\alpha_{j0} = \langle \Psi_j | (d|\Psi_0\rangle/dt) \rangle$. We have

$$\frac{d|\Psi_0\rangle}{dt} = \dot{\Omega} \frac{d|\Psi_0\rangle}{d\Omega}$$

From the eigenvalue equation $H|\Psi_0\rangle = E_0|\Psi_0\rangle$, we obtain after a derivative with respect to Ω :

$$-L_z |\Psi_0(\Omega)\rangle + H(\Omega) \frac{d|\Psi_0\rangle}{d\Omega} = \frac{dE_0}{d\Omega} |\Psi_0(\Omega)\rangle + E_0 \frac{d|\Psi_0\rangle}{d\Omega}$$

We now take the scalar product with $\langle \Psi_j |$ ($j \neq 0$) and we get:

$$\langle \Psi_j | L_z | \Psi_0 \rangle = (E_j - E_0) \langle \Psi_j | \frac{d|\Psi_0\rangle}{d\Omega} \rangle$$

We choose $|\Psi_j\rangle$ equal to the first excited state of the system $|\Psi_1\rangle$. The matrix element $\langle \Psi_1 | L_z | \Psi_0 \rangle$ is at most of order $N\hbar$ in the vicinity of the avoided crossing. Therefore,

$$\alpha_{10} = \langle \Psi_1 | \frac{d|\Psi_0\rangle}{dt} \rangle \leq \dot{\Omega} \frac{N\hbar}{\hbar\omega_{10}}$$

hence the condition for $p_{0 \rightarrow 1} \ll 1$:

$$\dot{\Omega} \frac{N}{\omega_{10}^2} \ll 1$$

Mean-field approach. The mean-field approach consists of assuming that all atoms are in the same state $f(\mathbf{r}, t) = \sum_{m=0}^2 a_m(t) \varphi_m(\mathbf{r})$ with $\sum |a_m|^2 = 1$. The average angular momentum per particle is $L = |a_1|^2 + 2|a_2|^2$ and the average energy per particle $E(\psi) = \frac{1}{N} \langle f^{\otimes N} | H | f^{\otimes N} \rangle$ reads (up to an additive constant):

$$\begin{aligned} E(\psi) = & (1 - \Omega)(|a_1|^2 + 2|a_2|^2) + \sqrt{2}A(a_0a_2^* + a_0^*a_2) \\ & + \frac{Ng}{4\pi} \left[|a_0|^4 + \frac{1}{2}|a_1|^4 + \frac{3}{8}|a_2|^4 \right. \\ & + 2|a_0|^2|a_1|^2 + |a_0|^2|a_2|^2 + \frac{3}{2}|a_1|^2|a_2|^2 \\ & \left. + \frac{1}{\sqrt{2}}(a_0a_2(a_1^*)^2 + a_0^*a_2^*a_1^2) \right] \end{aligned}$$

The Lagrange equations associated with this energy are $i\dot{a}_j = \partial E / \partial a_j^*$ (ref. 28), which gives for example:

$$i\dot{a}_0 = \sqrt{2}Aa_2 + \frac{Ng}{2\pi} \left[a_0 \left(|a_0|^2 + |a_1|^2 + \frac{1}{2}|a_2|^2 \right) + \frac{1}{2\sqrt{2}}a_1^2a_2^* \right]$$

and two similar equations for \dot{a}_1 and \dot{a}_2 . Note that in this mean-field approach, N and g have a role only through the product Ng . In particular, the fact that N is even or odd is of no relevance here.

The stationary solutions are obtained by inserting $a_m(t) = a_m(0)e^{-i\mu t}$ in the three Lagrange equations. A detailed analysis of the resulting 3×3 nonlinear system shows that two classes of solution exist. The first class (f_0) corresponds to $a_1 = 0$. Depending on the value of the parameters Ng , A and Ω , there may exist two, three or four solutions of this kind. After some tedious but straightforward calculation,

one can obtain for this first class of solution an analytical relation between Ω and the angular momentum per particle $L = 2|a_2|^2$:

$$\Omega = 1 - \frac{Ng}{8\pi} \left(1 - \frac{3}{8}L \right) \pm \sqrt{2}A \frac{1-L}{\sqrt{L(2-L)}}$$

The second class of solution corresponds to a non-zero value for a_1 and we have not been able to provide an exact analytical expression for the solution in this case. Using a numerical analysis, we have determined the local minima of the energy and we found that one solution of this kind exists if and only if $\Omega > 0.766$. We have compared the energy of this solution with the lowest energy of the solutions in the first class: for $\Omega < \bar{\Omega} = 0.773$ (respectively $\Omega > \bar{\Omega}$), the ground state is obtained with a solution belonging to the first (respectively second) class.

The stability of the solutions of the first class ($a_1 = 0$) can be studied analytically by looking at the equation of evolution of $b_1 = a_1 e^{i\mu t}$. This equation can be linearized around $b_1 = 0$ and written in the form $i\dot{b}_1 = Ab_1 + Bb_1^*$, where the constants A and B are real numbers that can be calculated explicitly in terms of the parameters Ω , A and Ng . The stationary solution corresponds to $b_1 = 0$, and it is stable if $b_1(t)$ stays around 0 when starting from a small non-zero initial value. This happens when $|A| > |B|$, whereas b_1 undergoes an exponential divergence from any initial noise if $|A| < |B|$, signalling a dynamical instability of the solution.

Received 29 September 2008; accepted 16 April 2009;
published online 24 May 2009

References

- Weiss, P. L'hypothèse du champ moléculaire et la propriété ferromagnétique. *J. Phys. Théor. et Appliq.* **6**, 661–690 (1907).
- Pitaevskii, L. & Stringari, S. *Bose–Einstein Condensation* (Oxford Univ. Press, 2003).
- Jaksch, D., Bruder, C., Cirac, J. I., Gardiner, C. W. & Zoller, P. Cold bosonic atoms in optical lattices. *Phys. Rev. Lett.* **81**, 3108–3111 (1998).
- Cooper, N. R. Rapidly rotating atomic gases. *Adv. Phys.* **57**, 539–616 (2008).
- Yoshioka, D. *The Quantum Hall Effect* (Springer, 2002).
- Griffin, A. *Excitations in a Bose–Condensed Liquid* (Cambridge Univ. Press, 1993).
- Fetter, A. L. Rotating trapped Bose–Einstein condensates. *Laser. Phys.* **18**, 1–11 (2008).
- Feder, D. L., Clark, C. W. & Schneider, B. I. Nucleation of vortex arrays in rotating anisotropic Bose–Einstein condensates. *Phys. Rev. A* **61**, 011601 (2000).
- Sinha, S. & Castin, Y. Dynamic instability of a rotating Bose–Einstein condensate. *Phys. Rev. Lett.* **87**, 190402 (2001).
- Kasamatsu, K., Tsubota, M. & Ueda, M. Nonlinear dynamics of vortex lattice formation in a rotating Bose–Einstein condensate. *Phys. Rev. A* **67**, 033610 (2003).
- Butts, D. A. & Rokhsar, D. S. Predicted signatures of rotating Bose–Einstein condensates. *Nature* **397**, 327–329 (1999).
- Bertsch, G. F. & Papenbrock, T. Yrast line for weakly interacting trapped bosons. *Phys. Rev. Lett.* **83**, 5412–5414 (1999).
- Smith, R. A. & Wilkin, N. K. Exact eigenstates for repulsive bosons in two dimensions. *Phys. Rev. A* **62**, 061602 (2000).
- Jackson, A. D. & Kavoulakis, G. M. Analytical results for the interaction energy of a trapped, weakly interacting Bose–Einstein condensate. *Phys. Rev. Lett.* **85**, 2854–2856 (2000).
- Dagnino, D., Barberán, N., Osterloh, K., Riera, A. & Lewenstein, M. Symmetry breaking in small rotating clouds of trapped ultracold Bose atoms. *Phys. Rev. A* **76**, 013625 (2007).
- Romanovsky, I., Yannouleas, C. & Landman, U. Symmetry-conserving vortex clusters in small rotating clouds of ultracold bosons. *Phys. Rev. A* **78**, 011606(R) (2008).
- Parke, M. I., Wilkin, N. K., Gunn, J. M. F. & Bourne, A. Exact vortex nucleation and cooperative tunneling in dilute BECs. *Phys. Rev. Lett.* **101**, 110401 (2008).
- Pitaevskii, L. P. Vortex lines in an imperfect Bose gas. *Sov. Phys. JETP* **13**, 451–454 (1961).
- Gross, E. P. Structure of a quantized vortex in boson systems. *Nuovo Cimento* **20**, 454–477 (1961).
- Stringari, S. Phase diagram of quantized vortices in a trapped Bose–Einstein condensed gas. *Phys. Rev. Lett.* **82**, 4371–4375 (1999).
- Ueda, M. & Nakalima, T. Nambu–Goldstone mode in a rotating Bose–Einstein condensate. *Phys. Rev. A* **73**, 043603 (2006).
- Morris, A. G. & Feder, D. L. Validity of the lowest-Landau-level approximation for rotating Bose gases. *Phys. Rev. A* **60**, 033605 (2006).
- Wilkin, N. K. & Gunn, J. M. Condensation of composite bosons in a rotating BEC. *Phys. Rev. Lett.* **84**, 6–9 (2000).
- Eckert, K., Schliemann, J., Bruß, D. & Lewenstein, M. Quantum correlations in systems of indistinguishable particles. *Ann. Phys. (NY)* **299**, 88–127 (2002).

25. Zanardi, P. Quantum entanglement in fermionic lattices. *Phys. Rev. A* **65**, 042101 (2001).
26. Nunnenkamp, A., Rey, A. M. & Burnett, K. Cat state production with ultracold bosons in rotating ring superlattices. *Phys. Rev. A* **84**, 023622 (2008).
27. Messiah, A. *Quantum Mechanics* Ch. XVII (Courier Dover Publications, 1999).
28. Perez-García, V. M., Michinel, H., Cirac, J. I., Lewenstein, M. & Zoller, P. Low energy excitations of a Bose–Einstein condensate: A time-dependent variational analysis. *Phys. Rev. Lett.* **77**, 5320–5323 (1996).
29. Garay, L. J., Anglin, J. R., Cirac, J. I. & Zoller, P. Sonic analog of gravitational black holes in Bose–Einstein condensates. *Phys. Rev. Lett.* **85**, 4643–4647 (2000).
30. Fetter, A. L. Lowest-Landau-level description of a Bose–Einstein condensate in a rapidly rotating anisotropic trap. *Phys. Rev. A* **75**, 013620 (2007).

Acknowledgements

We acknowledge discussions with I. Cirac and the support of the EU SCALA and ESF Fermix Programs, Spanish MEC grants (FIS 2005-03169/04627, QOIT) and the French programs ANR and IFRAF.

Author contributions

All authors have contributed equally to this work.

Additional information

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>. Correspondence and requests for materials should be addressed to N.B.

Topological insulators in Bi_2Se_3 , Bi_2Te_3 and Sb_2Te_3 with a single Dirac cone on the surface

Haijun Zhang¹, Chao-Xing Liu², Xiao-Liang Qi³, Xi Dai¹, Zhong Fang¹ and Shou-Cheng Zhang^{3*}

Topological insulators are new states of quantum matter in which surface states residing in the bulk insulating gap of such systems are protected by time-reversal symmetry. The study of such states was originally inspired by the robustness to scattering of conducting edge states in quantum Hall systems. Recently, such analogies have resulted in the discovery of topologically protected states in two-dimensional and three-dimensional band insulators with large spin-orbit coupling. So far, the only known three-dimensional topological insulator is $\text{Bi}_x\text{Sb}_{1-x}$, which is an alloy with complex surface states. Here, we present the results of first-principles electronic structure calculations of the layered, stoichiometric crystals Sb_2Te_3 , Sb_2Se_3 , Bi_2Te_3 and Bi_2Se_3 . Our calculations predict that Sb_2Te_3 , Bi_2Te_3 and Bi_2Se_3 are topological insulators, whereas Sb_2Se_3 is not. These topological insulators have robust and simple surface states consisting of a single Dirac cone at the Γ point. In addition, we predict that Bi_2Se_3 has a topologically non-trivial energy gap of 0.3 eV, which is larger than the energy scale of room temperature. We further present a simple and unified continuum model that captures the salient topological features of this class of materials.

Recently, the subject of time-reversal-invariant topological insulators has attracted great attention in condensed-matter physics^{1–12}. Topological insulators in two or three dimensions have insulating energy gaps in the bulk, and gapless edge or surface states on the sample boundary that are protected by time-reversal symmetry. The surface states of a three-dimensional (3D) topological insulator consist of an odd number of massless Dirac cones, with a single Dirac cone being the simplest case. The existence of an odd number of massless Dirac cones on the surface is ensured by the Z_2 topological invariant^{7–9} of the bulk. Furthermore, owing to the Kramers theorem, no time-reversal-invariant perturbation can open up an insulating gap at the Dirac point on the surface. However, a topological insulator can become fully insulating both in the bulk and on the surface if a time-reversal-breaking perturbation is introduced on the surface. In this case, the electromagnetic response of three-dimensional (3D) topological insulators is described by the topological θ term of the form $S_\theta = (\theta/2\pi)(\alpha/2\pi) \int d^3x dt \mathbf{E} \cdot \mathbf{B}$, where \mathbf{E} and \mathbf{B} are the conventional electromagnetic fields and α is the fine-structure constant¹⁰. $\theta = 0$ describes a conventional insulator, whereas $\theta = \pi$ describes topological insulators. Such a physically measurable and topologically non-trivial response originates from the odd number of Dirac fermions on the surface of a topological insulator.

Soon after the theoretical prediction⁵, the 2D topological insulator exhibiting the quantum spin Hall effect was experimentally observed in HgTe quantum wells⁶. The electronic states of the 2D HgTe quantum wells are well described by a 2 + 1-dimensional Dirac equation where the mass term is continuously tunable by the thickness of the quantum well. Beyond a critical thickness, the Dirac mass term of the 2D quantum well changes sign from being positive to negative, and a pair of gapless helical edge states appears inside the bulk energy gap. This microscopic mechanism for obtaining topological insulators by inverting the bulk Dirac gap spectrum can also be generalized to other 2D and 3D systems. The guiding principle is to search for insulators where the

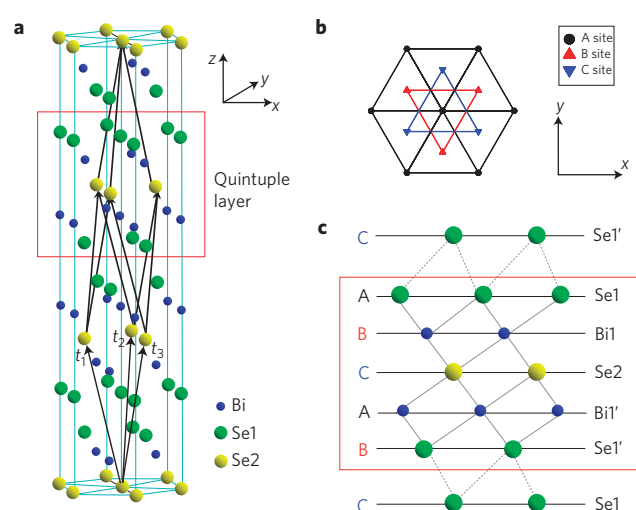


Figure 1 | Crystal structure. **a**, Crystal structure of Bi_2Se_3 with three primitive lattice vectors denoted as $\mathbf{t}_{1,2,3}$. A quintuple layer with $\text{Se1-Bi1-Se2-Bi1'-Se1'}$ is indicated by the red square. **b**, Top view along the z -direction. The triangle lattice in one quintuple layer has three different positions, denoted as A, B and C. **c**, Side view of the quintuple layer structure. Along the z -direction, the stacking order of Se and Bi atomic layers is $\dots\text{-C(Se1')-A(Se1)-B(Bi1)-C(Se2)-A(Bi1')-B(Se1')-C(Se1)-}\dots$. The Se1 (Bi1) layer can be related to the Se1' (Bi1') layer by an inversion operation in which the Se2 atoms have the role of inversion centres.

conduction and the valence bands have the opposite parity, and a 'band inversion' occurs when the strength of some parameter, say the spin-orbit coupling (SOC), is tuned. For systems with inversion symmetry, a method based on the parity eigenvalues of band states at time-reversal-invariant points can be applied¹³. On the basis of this analysis, the $\text{Bi}_x\text{Sb}_{1-x}$ alloy has been predicted

¹Beijing National Laboratory for Condensed Matter Physics, and Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China, ²Center for Advanced Study, Tsinghua University, Beijing 100084, China, ³Department of Physics, McCullough Building, Stanford University, Stanford, California 94305-4045, USA. *e-mail: sczhang@stanford.edu.

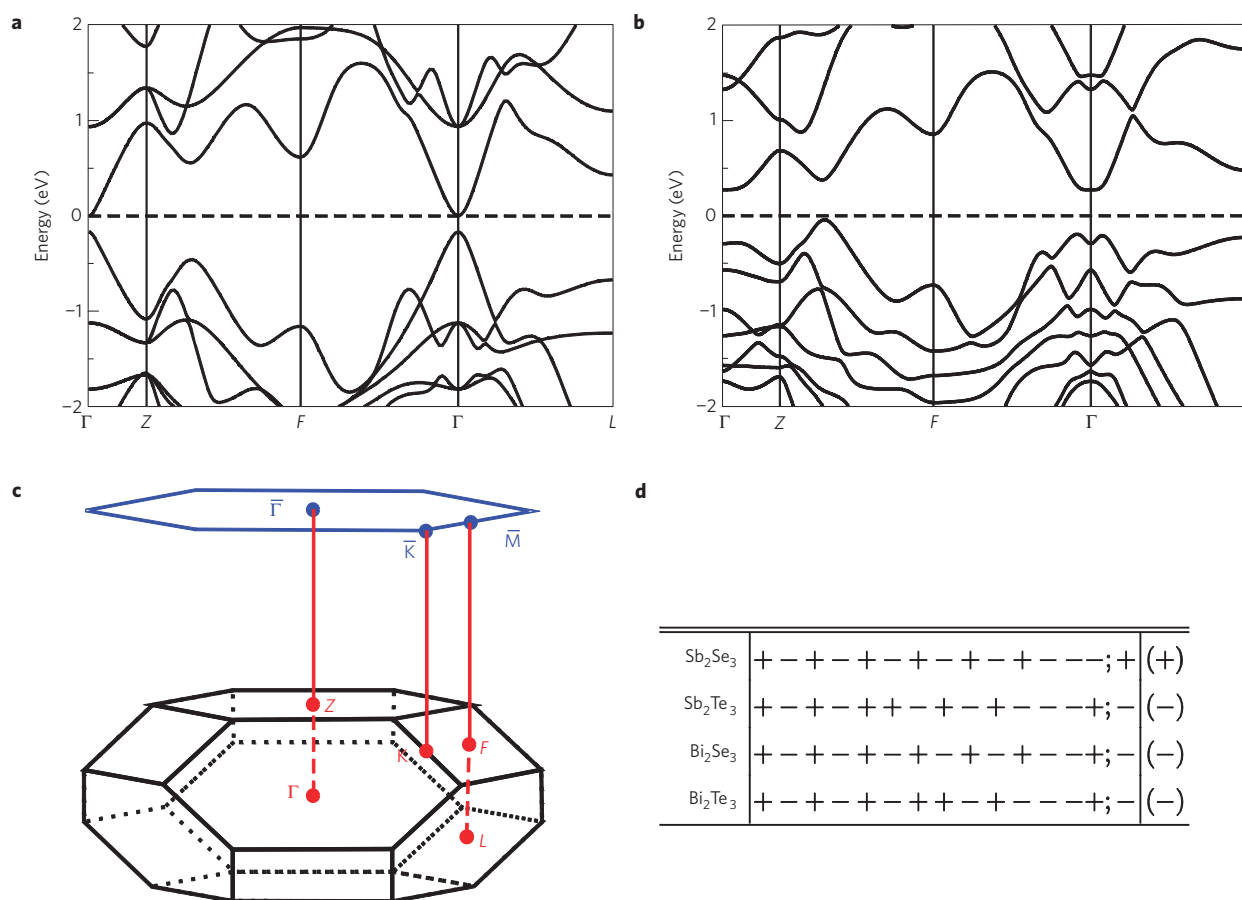


Figure 2 | Band structure, Brillouin zone and parity eigenvalues. **a, b**, Band structure for Bi₂Se₃ without **(a)** and with **(b)** SOC. The dashed line indicates the Fermi level. **c**, Brillouin zone for Bi₂Se₃ with space group $R\bar{3}m$. The four inequivalent time-reversal-invariant points are $\bar{\Gamma}(0,0,0)$, $\bar{L}(\pi,0,0)$, $\bar{F}(\pi,\pi,0)$ and $\bar{Z}(\pi,\pi,\pi)$. The blue hexagon shows the 2D Brillouin zone of the projected $(1,1,1)$ surface, in which the high-symmetry \mathbf{k} points $\bar{\Gamma}$, \bar{K} and \bar{M} are labelled. **d**, The parity of the band at the Γ point for the four materials Sb₂Te₃, Sb₂Se₃, Bi₂Se₃ and Bi₂Te₃. Here, we show the parities of fourteen occupied bands, including five *s* bands and nine *p* bands, and the lowest unoccupied band. The product of the parities for the fourteen occupied bands is given in brackets on the right of each row.

to be a topological insulator for a small range of x , and recently, surface states with an odd number of crossings at the Fermi energy have been observed in angle-resolved photoemission spectroscopy (ARPES) experiments¹².

As Bi_{*x*}Sb_{1-*x*} is an alloy with random substitutional disorder, its electronic structures and dispersion relations are only defined within the mean field, or the coherent potential approximation. Its surface states are also extremely complex, with as many as five or possibly more dispersion branches, which are not easily describable by simple theoretical models. Alloys also tend to have impurity bands inside the nominal bulk energy gap, which could overlap with the surface states. Given the importance of topological insulators as new states of quantum matter, it is important to search for material systems that are stoichiometric crystals with well-defined electronic structures, preferably with simple surface states, and describable by simple theoretical models. Here, we focus on layered, stoichiometric crystals Sb₂Te₃, Sb₂Se₃, Bi₂Te₃ and Bi₂Se₃. Our theoretical calculations predict that Sb₂Te₃, Bi₂Te₃ and Bi₂Se₃ are topological insulators, whereas Sb₂Se₃ is not. Most importantly, our theory predicts that Bi₂Se₃ has a topologically non-trivial energy gap of 0.3 eV, larger than the energy scale of room temperature. The topological surface states for these crystals are extremely simple, described by a single gapless Dirac cone at the $\mathbf{k}=0$ Γ point in the surface Brillouin zone. We also propose a simple and unified continuum model that captures the salient topological features of this class of materials. In this

precise sense, this class of 3D topological insulators shares the great simplicity of the 2D topological insulators realized in the HgTe quantum wells.

Band structure and parity analysis

Bi₂Se₃, Bi₂Te₃, Sb₂Te₃ and Sb₂Se₃ share the same rhombohedral crystal structure with the space group D_{3d}^5 ($R\bar{3}m$) with five atoms in one unit cell. We take Bi₂Se₃ as an example and show its crystal structure in Fig. 1a, which has layered structures with a triangle lattice within one layer. It has a trigonal axis (three-fold rotation symmetry), defined as the *z* axis, a binary axis (two-fold rotation symmetry), defined as the *x* axis, and a bisectrix axis (in the reflection plane), defined as the *y* axis. The material consists of five-atom layers arranged along the *z*-direction, known as quintuple layers. Each quintuple layer consists of five atoms with two equivalent Se atoms (denoted as Se1 and Se1' in Fig. 1c), two equivalent Bi atoms (denoted as Bi1 and Bi1' in Fig. 1c) and a third Se atom (denoted as Se2 in Fig. 1c). The coupling is strong between two atomic layers within one quintuple layer but much weaker, predominantly of the van der Waals type, between two quintuple layers. The primitive lattice vectors $\mathbf{t}_{1,2,3}$ and rhombohedral unit cells are shown in Fig. 1a. The Se2 site has the role of an inversion centre and under an inversion operation, Bi1 is changed to Bi1' and Se1 is changed to Se1'. The existence of inversion symmetry enables us to construct eigenstates with definite parity for this system.

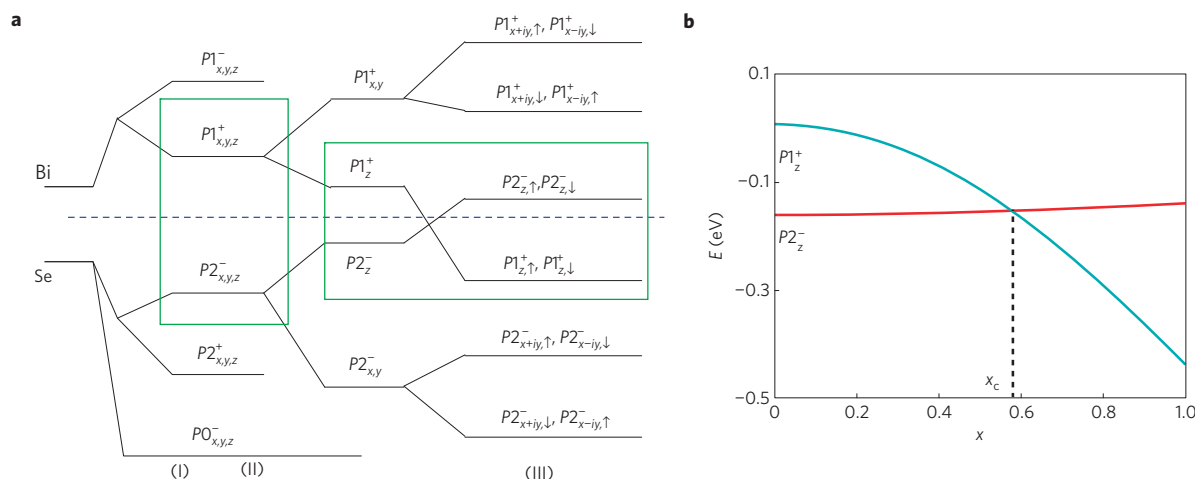


Figure 3 | Band sequence. **a**, Schematic diagram of the evolution from the atomic $p_{x,y,z}$ orbitals of Bi and Se into the conduction and valence bands of Bi_2Se_3 at the Γ point. The three different stages (I), (II) and (III) represent the effect of turning on chemical bonding, crystal-field splitting and SOC, respectively (see text). The blue dashed line represents the Fermi energy. **b**, The energy levels $|P1^+ \rangle$ and $|P2^- \rangle$ of Bi_2Se_3 at the Γ point versus an artificially rescaled atomic SOC $\lambda(\text{Bi}) = x\lambda_0(\text{Bi}) = 1.25x \text{ eV}$, $\lambda(\text{Se}) = x\lambda_0(\text{Se}) = 0.22x \text{ eV}$ (see text). A level crossing occurs between these two states at $x = x_c \approx 0.6$.

Ab initio calculations for Sb_2Te_3 , Sb_2Se_3 , Bi_2Te_3 and Bi_2Se_3 are carried out in the framework of the Perdew–Burke–Ernzerhof-type¹⁴ generalized gradient approximation of the density functional theory. The BSTATE package¹⁵ with the plane-wave pseudopotential method is used with a \mathbf{k} -point grid taken as $10 \times 10 \times 10$ and the kinetic energy cutoff fixed to 340 eV. For Sb_2Te_3 , Bi_2Te_3 and Bi_2Se_3 , the lattice constants are chosen from experiments, whereas for Sb_2Se_3 , the lattice parameters are optimized in the self-consistent calculation for rhombohedral crystal structure ($a = 4.076 \text{ \AA}$, $c = 29.830 \text{ \AA}$), owing to the lack of experimental data.

Our results are consistent with the previous calculations^{16,17}. In particular, we note that Bi_2Se_3 has an energy gap of about 0.3 eV, which agrees well with the experimental data (about 0.2–0.3 eV; refs 18, 19). In the following, we take the band structure of Bi_2Se_3 as an example. Figure 2a and b show the band structure of Bi_2Se_3 without and with SOC, respectively. By comparing the two figure parts, one can see clearly that the only qualitative change induced by turning on SOC is an anti-crossing feature around the Γ point, which thus indicates an inversion between the conduction band and valence band due to SOC effects, suggesting that Bi_2Se_3 is a topological insulator. To firmly establish the topological nature of this material, we follow the method proposed by Fu and Kane¹³. Thus, we calculate the product of the parities of the Bloch wavefunction for the occupied bands at all time-reversal-invariant momenta Γ, F, L, Z in the Brillouin zone. As expected, we find that at the Γ point, the parity of one occupied band is changed on turning on SOC, whereas the parity remains unchanged for all occupied bands at the other momenta F, L, Z . As the system without SOC is guaranteed to be a trivial insulator, we conclude that Bi_2Se_3 is a strong topological insulator. The same calculation is carried out for the other three materials, from which we find that Sb_2Te_3 and Bi_2Te_3 are also strong topological insulators, and Sb_2Se_3 is a trivial insulator. The parity eigenvalues of the highest 14 bands below the Fermi level and the first conduction band at the Γ point are listed in Fig. 2d. From this table we can see that the product of parities of occupied bands at the Γ point changes from the trivial material Sb_2Se_3 to the three non-trivial materials, owing to an exchange of the highest occupied state and the lowest unoccupied state. This agrees with our earlier analysis that an inversion between the conduction band and valence band occurs at the Γ point.

To get a better understanding of the inversion and the parity exchange, we start from the atomic energy levels and consider the effect of crystal-field splitting and SOC on the energy eigenvalues

at the Γ point. This is summarized schematically in three stages (I), (II) and (III) in Fig. 3a. As the states near the Fermi surface are mainly coming from p orbitals, we will neglect the effect of s orbitals and start from the atomic p orbitals of Bi ($6s^26p^3$) and Se ($4s^24p^4$). In stage (I), we consider the chemical bonding between Bi and Se atoms within a quintuple layer, which is the largest energy scale in the current problem. First we can recombine the orbitals in a single unit cell according to their parity, which results in three states (two odd, one even) from each Se p orbital and two states (one odd, one even) from each Bi p orbital. The formation of chemical bonding hybridizes the states on Bi and Se atoms, thus pushing down all of the Se states and lifting up all of the Bi states. In Fig. 3a, these five hybridized states are labelled as $|P1^{\pm}_{x,y,z} \rangle$, $|P2^{\pm}_{x,y,z} \rangle$ and $|P0^{-}_{x,y,z} \rangle$, where the superscripts $+$, $-$ stand for the parity of the corresponding states. In stage (II), we consider the effect of the crystal-field splitting between different p orbitals. According to the point-group symmetry, the p_z orbital is split from the p_x and p_y orbitals whereas the last two remain degenerate. After this splitting, the energy levels closest to the Fermi energy turn out to be the p_z levels $|P1^+_{z} \rangle$ and $|P2^-_{z} \rangle$. In the last stage (III), we take into account the effect of SOC. The atomic SOC Hamiltonian is given by $H_{\text{so}} = \lambda \mathbf{l} \cdot \mathbf{S}$, with \mathbf{l}, \mathbf{S} being the orbital and spin angular momentum, and λ is the SOC parameter. The SOC Hamiltonian mixes spin and orbital angular momenta while preserving the total angular momentum, which thus leads to a level repulsion between $|P1^+_{z}, \uparrow \rangle$ and $|P1^+_{x+iy}, \downarrow \rangle$, and similar combinations. Consequently, the $|P1^+_{z}, \uparrow \rangle$ (\downarrow) state is pushed down by the SOC effect and the $|P2^-_{z}, \uparrow \rangle$ (\downarrow) state is pushed up. If the SOC is large enough ($\lambda > \lambda_c$), the order of these two levels is reversed. To see this inversion process explicitly, we also calculate the energy levels $|P1^+_{z} \rangle$ and $|P2^-_{z} \rangle$ for a model Hamiltonian of Bi_2Se_3 with artificially rescaled atomic SOC parameters $\lambda(\text{Bi}) = x\lambda_0(\text{Bi})$, $\lambda(\text{Se}) = x\lambda_0(\text{Se})$, as shown in Fig. 3b. Here, $\lambda_0(\text{Bi}) = 1.25 \text{ eV}$ and $\lambda_0(\text{Se}) = 0.22 \text{ eV}$ are the realistic values of Bi and Se atomic SOC parameters, respectively²⁰. From Fig. 3b, one can see clearly that a level crossing occurs between $|P1^+_{z} \rangle$ and $|P2^-_{z} \rangle$ when the SOC is about 60% of the realistic value. As these two levels have opposite parity, the inversion between them drives the system into a topological insulator phase. Therefore, the mechanism for the 3D topological insulator in this system is exactly analogous to the mechanism in the 2D topological insulator HgTe. In summary, through the analysis above we find that Bi_2Se_3 is topologically non-trivial due to the inversion between two p_z orbitals with opposite parity at the Γ point. Similar analyses can

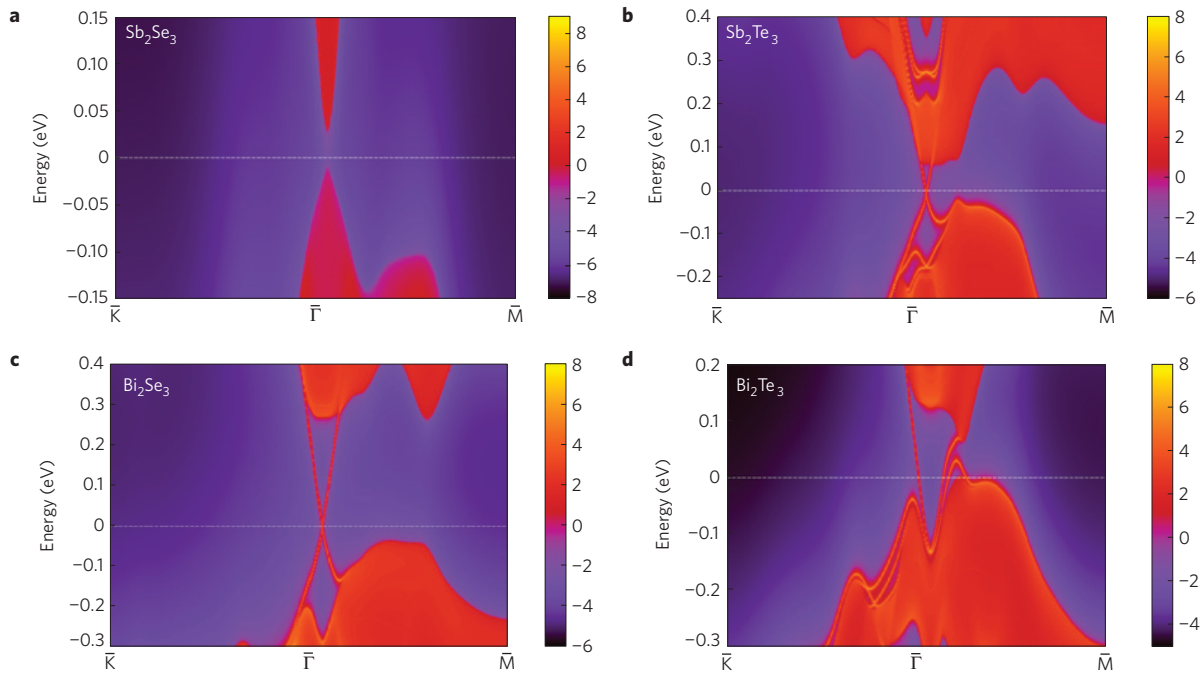


Figure 4 | Surface states. **a–d**, Energy and momentum dependence of the LDOS for Sb_2Se_3 (**a**), Sb_2Te_3 (**b**), Bi_2Se_3 (**c**) and Bi_2Te_3 (**d**) on the [111] surface. Here, the warmer colours represent higher LDOS. The red regions indicate bulk energy bands and the blue regions indicate bulk energy gaps. The surface states can be clearly seen around the Γ point as red lines dispersing in the bulk gap for Sb_2Te_3 , Bi_2Se_3 and Bi_2Te_3 . No surface state exists for Sb_2Se_3 .

be carried out on the other three materials, from which we see that Sb_2Te_3 and Bi_2Te_3 are qualitatively the same as Bi_2Se_3 , whereas the SOC of Sb_2Te_3 is not strong enough to induce such an inversion.

Topological surface states

The existence of topological surface states is one of the most important properties of the topological insulators. To see the topological features of the four systems explicitly, we calculate the surface states of these four systems on the basis of an *ab initio* calculation. First we construct the maximally localized Wannier function (MLWF) from the *ab initio* calculation using the method developed by Marzari and co-workers^{21,22}. We divide the semi-infinite system into a surface slab with finite thickness and the remaining part as the bulk. The MLWF hopping parameters for the bulk part can be constructed from the bulk *ab initio* calculation, and the ones for the surface slab can be constructed from the *ab initio* calculation of the slab, in which the surface correction to the lattice constants and band structure have been considered self-consistently and the chemical potential is determined by the charge neutrality condition. With these bulk and surface MLWF hopping parameters, we use an iterative method^{23,24} to obtain the surface Green's function of the semi-infinite system. The imaginary part of the surface Green's function is the local density of states (LDOS), from which we can obtain the dispersion of the surface states. The surface LDOS on the [111] surface for all four systems is shown in Fig. 4. For Sb_2Te_3 , Bi_2Se_3 and Bi_2Te_3 , one can clearly see the topological surface states that form a single Dirac cone at the Γ point. In comparison, Sb_2Se_3 has no surface state and is a topologically trivial insulator. Thus, the surface-state calculation agrees well with the bulk parity analysis, and confirms conclusively the topologically non-trivial nature of the three materials. For Bi_2Se_3 , the Fermi velocity of the topological surface states is $v_F \simeq 5.0 \times 10^5 \text{ m s}^{-1}$, which is similar to that of the other two materials.

Low-energy effective model

As the topological nature is determined by the physics near the Γ point, it is possible to write down a simple effective Hamiltonian

to characterize the low-energy long-wavelength properties of the system. Starting from the four low-lying states $|P1_z^+, \uparrow (\downarrow)\rangle$ and $|P2_z^-, \uparrow (\downarrow)\rangle$ at the Γ point, such a Hamiltonian can be constructed by the theory of invariants²⁵ for the finite wave vector \mathbf{k} . On the basis of the symmetries of the system, the generic form of the 4×4 effective Hamiltonian can be written down up to the order of $O(\mathbf{k}^2)$, and the tunable parameters in the Hamiltonian can be obtained by fitting the band structure of our *ab initio* calculation. The important symmetries of the system are time-reversal symmetry T , inversion symmetry I and three-fold rotation symmetry C_3 along the z axis. In the basis of $(|P1_z^+, \uparrow\rangle, |P2_z^-, \uparrow\rangle, |P1_z^+, \downarrow\rangle, |P2_z^-, \downarrow\rangle)$, the representation of the symmetry operations is given by $T = \mathcal{K} \cdot i\sigma^y \otimes I_{2 \times 2}$, $I = I_{2 \times 2} \otimes \tau_3$ and $C_3 = \exp(i\pi/3)\sigma^z \otimes I_{2 \times 2}$, where \mathcal{K} is the complex conjugation operator, $\sigma^{x,y,z}$ and $\tau^{x,y,z}$ denote the Pauli matrices in the spin and orbital space, respectively. By requiring these three symmetries and keeping only the terms up to quadratic order in \mathbf{k} , we obtain the following generic form of the effective Hamiltonian:

$$H(\mathbf{k}) = \epsilon_0(\mathbf{k})I_{4 \times 4} + \begin{pmatrix} \mathcal{M}(\mathbf{k}) & A_1 k_z & 0 & A_2 k_- \\ A_1 k_z & -\mathcal{M}(\mathbf{k}) & A_2 k_- & 0 \\ 0 & A_2 k_+ & \mathcal{M}(\mathbf{k}) & -A_1 k_z \\ A_2 k_+ & 0 & -A_1 k_z & -\mathcal{M}(\mathbf{k}) \end{pmatrix} + o(\mathbf{k}^2) \quad (1)$$

with $k_{\pm} = k_x \pm ik_y$, $\epsilon_0(\mathbf{k}) = C + D_1 k_z^2 + D_2 k_{\perp}^2$ and $\mathcal{M}(\mathbf{k}) = M - B_1 k_z^2 - B_2 k_{\perp}^2$. By fitting the energy spectrum of the effective Hamiltonian with that of the *ab initio* calculation, the parameters in the effective model can be determined. For Bi_2Se_3 , our fitting leads to $M = 0.28 \text{ eV}$, $A_1 = 2.2 \text{ eV \AA}$, $A_2 = 4.1 \text{ eV \AA}$, $B_1 = 10 \text{ eV \AA}^2$, $B_2 = 56.6 \text{ eV \AA}^2$, $C = -0.0068 \text{ eV}$, $D_1 = 1.3 \text{ eV \AA}^2$, $D_2 = 19.6 \text{ eV \AA}^2$. Except for the identity term $\epsilon_0(\mathbf{k})$, the Hamiltonian (1) is nothing but the 3D Dirac model with uniaxial anisotropy along the z -direction and \mathbf{k} -dependent mass terms. From the fact

$M, B_1, B_2 > 0$, we can see that the order of the bands $|T_1^+, \uparrow(\downarrow)\rangle$ and $|T_2^-, \uparrow(\downarrow)\rangle$ is inverted around $\mathbf{k} = 0$ compared with large \mathbf{k} , which correctly characterizes the topologically non-trivial nature of the system. Such an effective Dirac model can be used for further theoretical study of the Bi_2Se_3 system, as long as the low-energy properties are considered. For example, as one of the most important low-energy properties of the topological insulators, the topological surface states can be obtained from diagonalizing the effective Hamiltonian equation (1) with an open boundary condition, with the same method used in the study of the 2D quantum spin Hall insulator²⁶. For a surface perpendicular to the z -direction (that is, the $[111]$ direction), k_x, k_y are still good quantum numbers but k_z is not. By substituting $-\partial_z$ for k_z in equation (1), one can write down the 1D Schrödinger equations for the wavefunctions $\psi_{k_x, k_y}(z)$. For $k_x = k_y = 0$, there are two renormalizable surface-state solutions on the half infinite space $z > 0$, denoted by $|\psi_{0\uparrow}\rangle, |\psi_{0\downarrow}\rangle$. By projecting the bulk Hamiltonian (1) onto the subspace of these two surface states, to the leading order of k_x, k_y we obtain the following surface Hamiltonian

$$H_{\text{surf}}(k_x, k_y) = \begin{pmatrix} 0 & A_2 k_- \\ A_2 k_+ & 0 \end{pmatrix} \quad (2)$$

in the basis of $|\psi_{0\uparrow}\rangle, |\psi_{0\downarrow}\rangle$. Here, the surface-state wavefunction $|\psi_{0\uparrow(\downarrow)}\rangle$ is a superposition of the $|P_1^+, \uparrow(\downarrow)\rangle$ and $|P_2^+, \uparrow(\downarrow)\rangle$ states, respectively. For $A_2 = 4.1 \text{ eV \AA}$ obtained from the fitting, the Fermi velocity of the surface states is given by $v_F = A_2/\hbar \simeq 6.2 \times 10^5 \text{ m s}^{-1}$, which agrees reasonably well with the *ab initio* results shown in Fig. 4c. In summary, the effective model of the surface states equation (2) characterizes the key features of the topological surface states, and can be used in the future to study the surface-state properties of the Bi_2Se_3 family of topological insulators.

The topological surface states can be directly verified by various experimental techniques, such as ARPES and scanning tunnelling microscopy. In recent years, evidence of surface states has been observed for Bi_2Se_3 and Bi_2Te_3 in ARPES (ref. 27) and scanning tunnelling microscopy²⁸ experiments. In particular, the surface states of Bi_2Te_3 observed in ref. 27 had a similar dispersion to what we obtained in Fig. 4d, which were also shown to be quite stable and robust, regardless of photon exposure and temperature. Near the completion of this work, we became aware of the ARPES experiment²⁹ on Bi_2Se_3 , which measures a Dirac cone near the Γ point of the surface Brillouin zone. These experimental results support the main conclusion of our theoretical work. Moreover, the 3D topological insulators are predicted to exhibit the universal topological magneto-electric effect¹⁰ when the surface is coated with a thin magnetic film. Compared with the $\text{Bi}_{1-x}\text{Sb}_x$ alloy, the surface states of the Bi_2Se_3 family of topological insulators contain only a single Fermi pocket, making it easier to open up a gap on the surface by magnetization and to observe the topological Faraday/Kerr rotation¹⁰ and image magnetic monopole effect³⁰. If observed, such effects would be an unambiguous experimental signature of the non-trivial topology of the electronic properties.

Received 8 December 2008; accepted 1 April 2009;
published online 10 May 2009

References

- Kane, C. L. & Mele, E. J. Quantum spin hall effect in graphene. *Phys. Rev. Lett.* **95**, 226801 (2005).
- Bernevig, B. A. & Zhang, S. C. Quantum spin hall effect. *Phys. Rev. Lett.* **96**, 106802 (2006).
- Kane, C. L. & Mele, E. J. Z_2 topological order and the quantum spin hall effect. *Phys. Rev. Lett.* **95**, 146802 (2005).
- Murakami, S. Quantum spin hall effect and enhanced magnetic response by spin-orbit coupling. *Phys. Rev. Lett.* **97**, 236805 (2006).
- Bernevig, B. A., Hughes, T. L. & Zhang, S. C. Quantum spin hall effect and topological phase transition in HgTe quantum wells. *Science* **314**, 1757–1761 (2006).
- König, M. *et al.* Quantum spin hall insulator state in HgTe quantum wells. *Science* **318**, 766–770 (2007).
- Fu, L., Kane, C. L. & Mele, E. J. Topological insulators in three dimensions. *Phys. Rev. Lett.* **98**, 106803 (2007).
- Moore, J. E. & Balents, L. Topological invariants of time-reversal-invariant band structures. *Phys. Rev. B* **75**, 121306 (2007).
- Roy, R. On the Z_2 classification of quantum spin hall models. Preprint at <<http://arxiv.org/abs/cond-mat/0604211>> (2006).
- Qi, X.-L., Hughes, T. L. & Zhang, S.-C. Topological field theory of time-reversal invariant insulators. *Phys. Rev. B* **78**, 195424 (2008).
- Dai, X., Hughes, T. L., Qi, X.-L., Fang, Z. & Zhang, S.-C. Helical edge and surface states in HgTe quantum wells and bulk insulators. *Phys. Rev. B* **77**, 125319 (2008).
- Hsieh, D. *et al.* A topological dirac insulator in a quantum spin hall phase. *Nature* **452**, 970–974 (2008).
- Fu, L. & Kane, C. L. Topological insulators with inversion symmetry. *Phys. Rev. B* **76**, 045302 (2007).
- Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
- Fang, Z. & Terakura, K. Structural distortion and magnetism in transition metal oxides: Crucial roles of orbital degrees of freedom. *J. Phys. Condens. Matter* **14**, 3001–3014 (2002).
- Mishra, S. K., Satpathy, S. & Jepsen, O. Electronic structure and thermoelectric properties of bismuth telluride and bismuth selenide. *J. Phys. Condens. Matter* **9**, 461–470 (1997).
- Larson, P. Effects of uniaxial and hydrostatic pressure on the valence band maximum in Sb_2Te_3 : An electronic structure study. *Phys. Rev. B* **74**, 205113 (2006).
- Black, J., Conwell, E. M., Seigle, L. & Spencer, C. W. Electrical and optical properties of some M2-N3-semiconductors. *J. Phys. Chem. Solids* **2**, 240–251 (1957).
- Mooser, E. & Pearson, W. B. New semiconducting compounds. *Phys. Rev.* **101**, 492–493 (1956).
- Wittel, K. & Manne, R. Atomic spin-orbit interaction parameters from spectral data for 19 elements. *Theor. Chim. Acta* **33**, 347–349 (1974).
- Marzari, N. & Vanderbilt, D. Maximally localized generalized wannier functions for composite energy bands. *Phys. Rev. B* **56**, 12847–12865 (1997).
- Souza, I., Marzari, N. & Vanderbilt, D. Maximally localized wannier functions for entangled energy bands. *Phys. Rev. B* **65**, 035109 (2001).
- Sancho, M. P. L., Sancho, J. M. L. & Rubio, J. Quick iterative scheme for the calculation of transfer matrices: Application to $\text{Mo}(100)$. *J. Phys. F* **14**, 1205–1215 (1984).
- Sancho, M. P. L., Sancho, J. M. L., Sancho, J. M. L. & Rubio, J. Highly convergent schemes for the calculation of bulk and surface green functions. *J. Phys. F* **15**, 851–858 (1985).
- Winkler, R. *Spin-Orbit Coupling Effects in Two-Dimensional Electron and Hole Systems* (Springer Tracts in Modern Physics, Vol. 191, Springer, 2003).
- Koenig, M. *et al.* The quantum spin hall effect: Theory and experiment. *J. Phys. Soc. Japan* **77**, 031007 (2008).
- Noh, H.-J. *et al.* Spin-orbit interaction effect in the electronic structure of Bi_2Te_3 observed by angle-resolved photoemission spectroscopy. *Europhys. Lett.* **81**, 57006 (2008).
- Urazhdin, S. *et al.* Surface effects in layered semiconductors Bi_2Se_3 and Bi_2Te_3 . *Phys. Rev. B* **69**, 085313 (2004).
- Xia, Y. *et al.* Electrons on the surface of Bi_2Se_3 form a topologically-ordered two dimensional gas with a non-trivial berry's phase. Preprint at <<http://arxiv.org/abs/0812.2078>> (2008).
- Qi, X.-L., Li, R.-D., Zang, J. & Zhang, S.-C. Inducing a magnetic monopole with topological surface states. *Science* **323**, 1184–1187 (2009).

Acknowledgements

We would like to thank B. F. Zhu for the helpful discussion. This work is supported by the NSF of China, the National Basic Research Program of China (No. 2007CB925000), the International Science and Technology Cooperation Program of China (No. 2008DFB00170) and by the US Department of Energy, Office of Basic Energy Sciences under contract DE-AC02-76SF00515.

Additional information

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>. Correspondence and requests for materials should be addressed to S.-C.Z.

The distribution of spatially averaged critical properties

Steven T. Bramwell*

Critical properties throughout science are popularly associated with heavy-tailed distributions, but experimental evidence indicates several alternative, and very different, functional forms. Until now there has been no clear understanding of why this is, nor any general criterion as to which form to expect in a given practical situation. Here, a general scaling argument is presented, specific to spatially averaged properties, that indicates the following simple rule: if the mean value increases rapidly with system size then a power-law distribution is appropriate; if it changes slowly then a 'generalized Gumbel distribution' is likely, and if it decreases rapidly then an exponentially truncated power-law distribution is appropriate. The three scenarios are connected with the well-established classification of a scaling variable as either irrelevant, marginal or relevant. This result is supported by the current data set and finally renders comprehensible the fact that real critical properties exhibit diverse and apparently unrelated distributions, instead of ubiquitous heavy tails.

Many phenomena in natural and social science may be loosely described as 'highly correlated' or 'critical'. Distributions with 'heavy' or power-law tails are often observed for non-physical properties such as wealth or word frequency¹ and occasionally for physical ones, such as earthquake magnitude². Nevertheless, among physical properties that are spatially averaged, another type of distribution is much more commonly observed, which is unimodal, with exponential tails. There is much empirical and analytical evidence to suggest that this distribution should be treated as the generalized Gumbel distribution of extreme value statistics³, despite the fact that a spatially averaged quantity is not necessarily an extreme value. A recent experimental example of great interest is the distribution of critical order parameter fluctuations at the Fréedericksz transition in liquid crystals⁴. Older experimental examples of Gumbel-like distributions include the power consumption of a turbulent flow^{5–7}, plasma density fluctuations in a tokamak⁸ and the 'roughness' of $1/f$ noise in a resistor⁹. The generalized Gumbel form has also been invoked in discussions of river levels^{10,11}, resistance fluctuations near to electrical breakdown¹², fluctuations in glassy systems¹³, critical fluctuations in granular gases¹⁴, power fluctuations in electroconvection^{15,16}, self-organized critical systems^{17,18}, geophysical phenomena¹⁹ and imbibition in porous media²⁰.

Although there has been significant progress in understanding the occurrence of Gumbel or Gumbel-like distributions from a microscopic point of view^{9,21–26}, some of the broader questions remain completely open. In particular there is no phenomenological argument to indicate under what conditions one should expect a particular distribution—power law, generalized Gumbel or other^{27–29}. This is a problem, as it means that the experimental application of any particular distribution needs to be based more on qualitative arguments than on a definite, refutable hypothesis. Nor is there any clear understanding of why driven systems should show the same distributions as equilibrium systems and what this means. Here, we shed light on these questions by deriving possible forms of the probability density function for a physical quantity that is subject only to the condition that the quantity is spatially averaged and obeys finite-size scaling.

The aim of the current work was to construct a relationship between the functional form of a critical distribution and the size dependence of its moments, the idea being that this relationship might be of practical value in cases when either quantity is difficult to calculate, or to measure. The specific approach was motivated by the observation (discussed in the 'A useful analogy' section) that the probability distribution of a spatially averaged quantity is mathematically analogous to a finite-size order parameter versus temperature curve. In that case, general forms of the curve follow from the characteristic scale invariance of a critical system.

We therefore consider a dimensionless, spatially averaged, scalar random variable X_N (taking values x) that is 'intensively defined' by summing some dimensionless quantity over the system of interest and dividing by a dimensionless system size N (for example, the system might be formally divided into N unit cells). The expression $P_{X_N}(x)$ is used to denote, depending on context, either the cumulative distribution function (CDF) or the complementary CDF (CCDF) of the variable. For simplicity, we consider unimodal distributions. It is proposed that critical behaviour corresponds to distributions with the following scale invariance:

$$P_{X_N}(x) = P_{X_{N/\lambda}}(\mathcal{R}_\lambda(x)) \quad (1)$$

where $\lambda > 1$ and \mathcal{R}_λ is generally a nonlinear transformation. This proposition is discussed in detail in the 'Solution of the scaling equation' section. A general solution is stated (equation (4)), which is then analysed by a renormalization group argument. The results, listed in Table 1, depend on an unspecified function F . Specific examples, treated in the 'Example applications' section, are used to confirm the general validity of the scheme and to identify its boundaries. As discussed in the 'Generalization and conclusions' section, F may generally be approximated by a normal or exponential function, which, when combined with the results of Table 1, leads to the main conclusion quoted in the abstract. The final paragraphs of the 'Generalization and conclusions' section summarize the logic behind the current approach, and identify points that should be subject to more rigorous scrutiny in the future.

A useful analogy

The scaling equation (1) that leads to the main results of this article may be unfamiliar in the present context, so it is useful to first illustrate how an analogous equation arises in a more familiar problem of critical behaviour.

Consider a spin system at equilibrium with coupling K : in simple cases, $K = J/T$, where J is the exchange constant, T is the temperature and Boltzmann's constant is set to unity. If the mean scalar order parameter of a spin system (\bar{m}) is expressed as a function of the coupling then $\bar{m}(K)$ is, in purely mathematical terms, a distribution for K as it increases continuously from zero at $K = 0$ to unity as $K \rightarrow \infty$. Consider, for example, the regime with $T \geq T_c$, where T_c is the critical temperature. In this regime, the renormalization group may be summarized in the equation

$$\bar{m}_L(K) = \bar{m}_{L/\lambda}(K') \quad (2)$$

which is analogous to equation (1) (here L is the system size and K' is a renormalized coupling). Equation (2), as written, is a finite-size scaling equation, as it considers a renormalized system of $1/\lambda$ the size of the original (other possible formulations give equivalent results).

We now consider models that order only at $K = \infty$, and without loss of any generality, the specific case of one-dimensional models. In this case, a plausible ansatz for the finite-size order parameter is to equate it to the square root of the correlation function at a distance L :

$$\bar{m}_L(K) = e^{-L\kappa(K)/2} \quad (3)$$

where κ is the inverse correlation length. As $K \rightarrow \infty$, the correlation length diverges and κ approaches zero. The renormalization group transformation may be linearized near to the fixed point at $\kappa = 0$, which can be shown to lead to either an algebraic or essential singularity in the correlation length as a function of $1/K$ (the mathematical method is equivalent to that described in the next section).

This result is quite general: all simple spin models exhibit one or the other type of behaviour, the precise form depending on the critical exponent that determines the divergence of the correlation length. The finite-size order parameter turns out to be quite accurately determined by equation (3), so if the correlation length exponent is known, then $\bar{m}_L(K)$ may be determined without detailed calculation. In the present context, this is equivalent to the size dependence of the principal moments determining the functional form of the probability distribution.

With a single variable K , equation (2) may be taken to be a definition of critical behaviour (at least if the fixed point is non-trivial: see ref. 30 for a full discussion). Thus, our equation (1) seems a natural definition of a critical distribution. It should, however, be emphasized that this analogy with an equilibrium critical system does not imply that the current argument is restricted to equilibrium systems. Indeed, equation (1) can apply in principle to any system for which P_{X_N} is time independent, which includes both equilibrium systems and driven systems at steady state. The solution of equation (1) is the subject of the next section.

Solution of the scaling equation

Equation (1) expresses the 'scale invariance' that is the defining characteristic of critical behaviour. Thus, through the scale transformation \mathcal{R} , distributions measured at one system size may be superimposed on those measured at another, possibly very different system size. The key feature of \mathcal{R} is that it depends only on the relative size of systems, λ , and not on their absolute size N .

Equation (1) may be solved by the following ansatz:

$$P_{X_N}(x) = F(Ng(x)) \quad (4)$$

where F and g are unspecified functions. We offer no formal proof that this exhausts the analytical solutions of equation (1), but simply note that it is difficult to imagine closed-form solutions that can be expressed otherwise. Equation (4) is equivalent to equation (3) discussed in the last section and may be subjected to a similar renormalization group analysis. The dimensionless function g is assumed to be a monotonically decreasing or increasing function of x , depending on the nature of F and whether P represents a CDF or a CCDF.

We first consider the case where $g \rightarrow 0$ corresponds to $x \rightarrow \infty$. Consider a sequence of renormalization group transformations $N \rightarrow N/\lambda$. The transformation may be linearized near to the fixed point at $g = 0$ such that $(1/x)' \approx (1/x)\lambda^\phi$ or $x' = x\lambda^{-\phi}$ where ϕ is a critical exponent (here, the prime denotes a renormalized variable). After n transformations, it follows from equations (1) and (4) that $g(x) = \lambda^{-n}g(x\lambda^{-n\phi})$, which may be solved by setting $\lambda^{-n\phi} = b/x$, where $b \gg 1$ is a constant³⁰. The result is $g(x) = g_0x^{-1/\phi}$, where $g_0 = g(b)b^{1/\phi}$ is of order unity. For g to be a decreasing function of x , the exponent ϕ must be positive, showing that x is an irrelevant variable with respect to the $x = \infty$ fixed point.

In the case that $\phi = 0$, x is a marginal variable and logarithmic corrections to this algebraic scaling must be considered. The renormalization transformation is written $(1/x)' \approx (1/x) + x_0(1/x)^2 \ln \lambda$, where x_0 is a positive constant of order unity. This corresponds to $x' = x - x_0 \ln \lambda$, which identifies x as a marginally irrelevant variable. By a similar argument to that outlined above, we find $g(x) = e^{-x/x_0}$.

The results of the analysis are given in Table 1. The two solutions already described (irrelevant and marginally irrelevant) are labelled 3 and 2b, respectively. Two further solutions (marginally relevant and relevant) are generated by allowing $g(x)$ to be monotonically increasing rather than monotonically decreasing: these are respectively labelled 2a and 1.

Forms 1 and 3 exhibit 'hyperscaling': the mean $\mu_{X_N} \propto N^{\pm|\phi|}$ is proportional to the scale σ_{X_N} , so size-independent forms of the distribution may be derived by defining a rescaled variable $Z = X/\mu_{X_N}$ that takes values z . Forms 2a and 2b have $\mu_{X_N} \sim c \ln N$ and $\sigma_{X_N} \sim x_0$, so do not exhibit hyperscaling. Nevertheless, size-independent forms result from defining $Z = (X - \mu_{X_N})/\sigma_{X_N}$. Examples are given in Table 1 for the particular case where F is an exponential function, where constants resulting from the rescaling of the variable are set to unity for clarity. Thus, if F is known, then the size dependence of the leading moments fully determines the form of the size-independent distribution (see Table 1) and hence its density function.

Example applications

As already mentioned, the results listed in Table 1 might equally apply to driven systems at steady state as to equilibrium critical systems. However, the latter class of system, being better comprehended than the former, provides the more meaningful test of the theory. Hence in this section, we consider in detail the order parameter distributions of two canonical models of equilibrium critical behaviour: the two-dimensional (2D) Ising and XY models. The latter is equivalent to the 2D Edwards–Wilkinson model of interface growth (a steady-state system that maps onto an equilibrium one), so as a third example we consider the analogous one-dimensional Edwards–Wilkinson model. These examples cover all of the types of distribution (1, 2a, 2b, 3) listed in Table 1.

The key in these applications is to know the function F and the size dependence of the mean or standard deviation. In the spirit of the aims expressed in the introduction, the size dependence may

Table 1 | Summary of the main results.

Type	Distribution*	Scale [†]	Mean [†]	Example [‡]	Variable
1	$F(Ng_0x^{1/ \phi })$	$\sim N^{- \phi }$	$\sim N^{- \phi }$	$1 - e^{-z^{1/ \phi }}$	Relevant
2a	$F(Ne^{x/x_0})$	~ 1	$\sim -\ln N$	$1 - e^{-e^z}$	Marginally relevant
2b	$F(Ne^{-x/x_0})$	~ 1	$\sim \ln N$	$e^{-e^{-z}}$	Marginally irrelevant
3	$F(Ng_0x^{-1/ \phi })$	$\sim N^{ \phi }$	$\sim N^{ \phi }$	$e^{-z^{-1/ \phi }}$	Irrelevant

*The 'distribution' can represent either a CDF or a CCDF.

[†] These columns give the size dependence, not precise values.

[‡] As an example, the CDF is given for the case where F is an exponential function: here z is a rescaled variable (for its precise definition, see the text).

be taken as an 'experimental input', but to know the function F is generally problematic. Nevertheless, simple assumptions for F give meaningful results as we now show.

At the critical point of the Ising model, the scalar order parameter varies as $N^{-1/(\delta+1)}$, where δ is the equation-of-state critical exponent. This indicates form 1 with $\phi = 1/(\delta + 1)$. To find F , we argue that in the high-temperature, or uncorrelated limit, the central limit theorem applies and the distribution is given accurately by equation (4) with $F(Ng) = \mathcal{N} \circ \sqrt{[Ng]}$, where \mathcal{N} is the normal distribution function, and $g(x) = x^2$. Renormalization group arguments suggest a continuous connection between the distributions at different temperatures or couplings, so it is tempting to retain this form for F , allowing $g(x)$ to alter its functional form as the interactions are turned on. This may be further justified by observing that at all temperatures the distribution must at least contain a binomial part, well approximated by the normal, to describe the density of states. Thus, it would not be surprising if the latter were to dominate the true distribution for large deviations from $x = 0$. This assumption leads to the following prediction for the density of the normalized order parameter $Z = X/\mu_{X_N}$,

$$p_Z(z) \sim z^{(\delta-1)/2} e^{-c_\delta z^{(\delta+1)}} \quad (5)$$

which is exactly the large- z asymptotic distribution for Ising models, previously established by elaborate statistical mechanical arguments (c_δ is a constant)²⁷. As anticipated, equation (5) describes the actual distribution well at large z , but poorly as $z \rightarrow 0$ (ref. 31; the same is true in three dimensions³²). The reason for the failure at small z may thus be attributed to imprecise knowledge of the function F . The Ising model is known to have a critical distribution of the form $\sim P(x/\mu_{X_N})$, which is equivalent to equation (4), confirming that the failure lies in the form adopted for F rather than the assumption that F exists.

Turning now to the XY model, equation (5) would be expected to apply on its critical line, where δ depends on temperature^{33–36}, suggesting that the assumption of the normal form for the density of states is also meaningful in this case. However, in reality there is a competition between this form and a form dominant at low temperature that resembles the Gumbel distribution^{33,36}. This behaviour, which has hitherto remained mysterious, is clearly predicted by the current analysis, which shows that the competition of form 5 with a Gumbel form should be completely general (an important conclusion, mentioned in the final paragraph of the 'Generalization and conclusions' section). Thus, for the XY model in the low temperature limit, the exponent $\phi = 1/(\delta + 1)$ decreases towards zero (the anomalous dimension exponent η also approaches zero in this limit). The analysis of the 'Solution of the scaling equation' section then suggests a crossover to a distribution of form 2a with logarithmic dependence of the mean on system size. This is indeed what is observed.

To analyse the form of the distribution in this limit, we could adopt the normal assumption for F , or perhaps more realistically an exponential form: as the scalar order parameter m approaches unity and the system explores the wings of the density of states function, the variable $X = 1 - m$ should have CDF approximated by $e^{-Ng(x)}$, where g is a decreasing function of x . The mean of the new variable X increases as $\ln(N)$, indicating form 2b, and the predicted density is then $e^{-z-e^{-z}}$, the Gumbel distribution. However, the true distribution is not exactly the Gumbel distribution³³ and it is interesting to establish how accurate the Gumbel prediction is. In the Methods section, it is shown that the exact distribution is very nearly a convolution of two Gumbel functions, approaching this form exactly in the wings. Thus, the Gumbel prediction is seen to be meaningful for the XY model, although not exact. It seems that a closed-form solution to the density function does not exist.

The above results apply equally to the roughness distribution of the 2D Edwards–Wilkinson model. The corresponding 1D Edwards–Wilkinson model is not strictly a critical system, but does exhibit a non-Gaussian roughness distribution³⁷. Here, the mean scales as N^1 , indicating form 3, and by analogy with the 2D problem, it seems reasonable to assume that F is exponential. The predicted density function is then $p_Z \sim z^{-2} e^{-c/z}$, which compares quite well with the exact small- z asymptote, $p_Z \sim z^{-5/2} e^{-(3/2)/z}$, that describes the density almost perfectly to well beyond its mode³⁷. However, the prediction fails at large z where the true density has an exponential tail that arises from a single dominant length scale³³, thus marking a departure from critical behaviour.

Generalization and conclusions

The examples given in the previous section suggest the importance of functions F (in equation (4)) based either on the normal or exponential distributions. The resulting predicted distributions are similar in both cases and the corresponding densities may be represented thus:

$$p_{Z_1}(z) \sim z^{a/|\phi|-1} e^{-acz^{1/|\phi|}}$$

$$p_{Z_2}(z) \sim (e^{-z-e^{-z}})^a$$

$$p_{Z_3}(z) \sim z^{-a/|\phi|-1} e^{-acz^{-1/|\phi|}}$$

where a is a parameter of order unity that depends on the system. This result seems highly consistent with the current data set. Thus, densities of type 1 are typical for thermodynamic critical systems exhibiting power-law scaling. However, when the exponent of the power law is small, or the scaling is logarithmic, those of type 2 may be more appropriate. These are the generalized Gumbel densities, commonly observed in experiment (2a and 2b differ in the sense of the skew). The exponent a is further predicted to depend on correlation length^{25,38}. Although less commonly observed for spatially averaged quantities, the densities of type 3 are typical

heavy-tailed functions, which for $z \gg 1$ approach $p \sim z^{-\alpha}$, with $\alpha \approx 1/\phi$. The power-law distribution is therefore suggested to be a possibility for spatially averaged critical properties, but its rarity can be understood as a consequence of the fact that most such properties do not increase as a power law with system size.

The experimental data set thus supports the idea that the function F takes the proposed narrow range of functional forms: this is not proven mathematically, but it is clearly justified *a posteriori*. One possible approach to the mathematical proof of the proposal is large deviations theory, which typically predicts exponential forms reminiscent of equation (4), and, when applied to thermodynamic systems, emphasizes the role of the density of states, as in the heuristic arguments of the 'Example applications' section (see, for example, refs 39–42).

When based on an exponential form for F , the three densities have $a = 1$ and are the three distributions of extreme value theory⁴³, confirming the proposed link^{9,17,25}. Indeed, the Fisher–Tippett formulation of extreme value theory⁴³ has an obvious similarity with the renormalization argument of the 'A useful analogy' section (although is less general). In a recent paper, Györgyi and co-workers present a detailed renormalization group analysis of corrections to extreme value theory⁴⁴ that might be adapted to refine the current approach. More generally, the current method of treating a single variable at a particular fixed point could in principle be adapted to treating more than one variable at more than one fixed point. In this context, it would be interesting to relate the current approach to the renormalization group method described in ref. 45 as well as to the formalism for effective extreme value statistics described by Bertin and Clusel^{25,26}.

The above results illustrate an algorithm by which critical distributions can be predicted, given the size dependence of the mean, mode or scale. This algorithm is remarkably successful, but like all phenomenologies, it cannot predict its own failures. It is therefore no substitute for precise microscopic calculations, but might prove useful in cases where such calculations are unfeasible. The success of the algorithm is clearly related to the special properties of probability distributions, where functional forms are limited by the combined constraints of boundary conditions, scale invariance and few parameters. As outlined in the 'Example applications' section, breakdowns of the algorithm can generally be attributed to a non-universal F , absence of a closed-form expression for the distribution or departures from criticality.

We finally summarize the logical structure of the current argument and highlight areas for future investigation. The argument relies on three assumptions. First, the assumption that the scaling equation (1) represents critical behaviour: this seems highly plausible and should probably remain as a postulate. Second, the proposition that equation (4) is the most general solution to equation (1): this still needs to be justified rigorously. Third, the assumption of exponential or normal forms for the function F in equation (4). This is the main area for future work, as it would be useful to identify, possibly using large deviations theory, conditions under which this assumption is valid: at present, we can say only that it is generally a meaningful, if inexact, approximation. It might also be possible to identify an improved approximation for the function F for particular classes of system.

In summary, in addition to its predictive power, the main success of the method described here is the insight that it gives to the problem posed in the introduction: how to relate the various probability distributions that are commonly exhibited by spatially averaged critical properties. The answer is offered in the form of the simple rule stated in the abstract of this paper: that one should expect one of three universal density functions that are distinguished by their differing dependence of typical values (mode or mean) on system size. This is seen to stem from the well-established classification of a single scaling variable as

relevant, marginal or irrelevant, with respect to the fixed point of a renormalization group transformation. A slow power variation will generally result in a competition between forms 1 or 3 and form 2, so the appearance of Gumbel-like forms in driven systems such as the turbulence experiment⁶ is readily understood. In general, these results enable one to understand the relationship of the generalized Gumbel form to the more established power-law distributions and establish a straightforward and testable criterion for the application of a particular density function to a given set of experimental or numerical data.

Methods

Exact result for 2D models. We consider (1) the Gumbel distribution and (2) the order parameter distribution of the XY model in its low-temperature limit (the 'BHP' distribution), which is equivalent to the roughness distribution of the Edwards–Wilkinson model. In the XY representation, the variable of interest is $X = 1 - M/N$, where M is the scalar magnetic moment. The respective cumulants κ_r can be represented in terms of lattice sums that run over positive and negative integers¹⁰:

$$\kappa_r^{\text{Gumbel}} = \frac{1}{2} \Gamma(r) \sum_{n \neq 0} \frac{1}{|n|^r}$$

$$\kappa_r^{\text{BHP}} = \frac{1}{2} \Gamma(r) \sum_{n_x, n_y \neq 0,0} \frac{1}{(n_x^2 + n_y^2)^r}$$

To relate the two sets of cumulants, we make use of a remarkable exact relation, discussed in refs 46, 47 and attributed there to G. H. Hardy (1919) and L. Lorenz (1871):

$$\sum_{n_x, n_y \neq 0,0} \frac{1}{(n_x^2 + n_y^2)^r} = 4\zeta(r)\beta(r)$$

where $\beta(r)$ is the Dirichlet beta function and $\zeta(r)$ is the Reimann zeta function:

$$\zeta(r) = \sum_{n=1}^{\infty} \frac{1}{n^r}$$

$$\beta(r) = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{(2n-1)^r}$$

We thus arrive at the following exact relations:

$$\kappa_r^{\text{Gumbel}} = \Gamma(r)\zeta(r)$$

$$\kappa_r^{\text{BHP}} = 2\Gamma(r)\zeta(r)\beta(r) = 2\kappa_r^{\text{Gumbel}} \times \beta(r)$$

In this representation, the mean of the Gumbel distribution is shifted by a factor of $\ln N$: $\zeta(1) = \ln N + \gamma_e \rightarrow \infty$, where γ_e is the Euler–Mascheroni constant, 0.57721566... The series for $\zeta(r)$ with $r > 1$ converge to exactly known sums only for even r : the first few values are $\zeta(2) = \pi^2/6$, $\zeta(3) = 1.20256903...$, $\zeta(4) = \pi^4/90$. The series for $\beta(r)$ converge on exactly known sums only for odd r : $\beta(1) = \pi/4$, $\beta(2) = 0.91596559... = K$, where K is Catalan's constant, $\beta(3) = \pi^3/32$, $\beta(4) = 0.988944552...$. There is no combination of $\zeta(r)\beta(r)$ for which both components are exactly known. The β and ζ sums both approach unity for large r (E. W. Weisstein, <http://mathworld.wolfram.com>).

The BHP density function is therefore approximated by the convolution of two Gumbel functions. This approximation is very accurate at all x , and asymptotically exact in the wings, which has been confirmed by comparing it with numerical data (Bramwell, unpublished).

Received 21 July 2008; accepted 30 March 2009;
published online 3 May 2009

References

- Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46**, 323–351 (2005).
- Christensen, K., Danon, L., Scanlon, T. & Bak, P. Universal scaling law for earthquakes. *Proc. Natl Acad. Sci.* **99**, 2509–2513 (2002).
- Gumbel, E. J. *Statistics of Extremes* (Columbia Univ. Press, 1958).
- Joubaud, S., Petrosyan, A., Ciliberto, S. & Garnier, N. B. Experimental evidence of non-Gaussian fluctuations near a critical point. *Phys. Rev. Lett.* **100**, 180601 (2008).

5. Bramwell, S. T., Holdsworth, P. C. W. & Pinton, J.-F. Universality of rare fluctuations in turbulence and critical phenomena. *Nature* **396**, 552–554 (1998).
6. Pinton, J.-F., Holdsworth, P. C. W. & Labbé, R. Power fluctuations in a closed turbulent shear flow. *Phys. Rev. E* **60**, R2452–R2455 (1999).
7. Portelli, B., Holdsworth, P. C. W. & Pinton, J.-F. Intermittency and non-Gaussian fluctuations of the global energy transfer in fully developed turbulence. *Phys. Rev. Lett.* **90**, 104501 (2003).
8. Van Milligen, B. P., Sanchez, R., Carreras, B. A., Lynch, V. E. & LaBombard, B. Additional evidence for the universality of the probability distribution of turbulent fluctuations and fluxes in the scrape-off layer region of fusion plasmas. *Phys. Plasmas* **12**, 052507 (2005).
9. Antal, T., Droz, M., Györgyi, G. & Rácz, Z. 1/f noise and extreme value statistics. *Phys. Rev. Lett.* **87**, 240601 (2001).
10. Bramwell, S. T., Fennell, T., Holdsworth, P. C. W. & Portelli, B. Universal fluctuations of the Danube water level: A link with turbulence, criticality and company growth. *Europhys. Lett.* **57**, 310–314 (2002).
11. Dahlstedt, K. & Jensen, H. J. Fluctuation spectrum and size scaling of river flow and level. *Phys. A* **348**, 596–610 (2005).
12. Pennetta, C., Alfinito, E., Reggiani, L. & Ruffo, S. Non-Gaussianity of resistance fluctuations near electrical breakdown. *Semicond. Sci. Technol.* **19**, S164–S166 (2004).
13. Chamon, C. & Cugliandolo, L. F. Fluctuations in glassy systems. *J. Stat. Mech.* P07022 (2007).
14. Brey, J. J., García de Soria, M. I., Maynar, P. & Ruiz-Montero, M. J. Mesoscopic theory of critical fluctuations in isolated granular gases. *Phys. Rev. Lett.* **94**, 098001 (2005).
15. Goldburg, W. I., Goldschmidt, Y. Y. & Kellay, H. Fluctuation and dissipation in liquid-crystal electroconvection. *Phys. Rev. Lett.* **87**, 245502 (2001).
16. Toth-Katona, T. & Gleeson, J. T. Distribution of injected power fluctuations in electroconvection. *Phys. Rev. Lett.* **91**, 264501 (2003).
17. Bramwell, S. T. *et al.* Universal fluctuations in correlated systems. *Phys. Rev. Lett.* **84**, 3744–3747 (2000).
18. Rypdal, K. *et al.* Scale-free vortex cascade emerging from random forcing in a strongly coupled system. *New J. Phys.* **10**, 093018 (2008).
19. Chapman, S. C., Rowlands, G. & Watkins, N. W. Extreme statistics: A framework for data analysis. *Nonl. Proc. Geophys.* **9**, 409–418 (2002).
20. Planet, R., Santucci, S. & Ortín, J. Avalanches and non-Gaussian fluctuations of the global velocity of imbibition fronts. *Phys. Rev. Lett.* **102**, 094502 (2009).
21. Zheng, B. Generic features of fluctuations in critical systems. *Phys. Rev. E* **67**, 026114 (2003).
22. Bertin, E. Global fluctuations and Gumbel statistics. *Phys. Rev. Lett.* **95**, 170601 (2005).
23. Clusel, M., Fortin, J.-Y. & Holdsworth, P. C. W. Criterion for universality-class-independent critical fluctuations: Example of the two-dimensional Ising model. *Phys. Rev. E* **70**, 046112 (2004).
24. van Wijland, F. Phonon displacement distribution at $T = 0$. *Physica A* **332**, 360–366 (2004).
25. Bertin, E. & Clusel, M. Generalised extreme value statistics and sum of correlated variables. *J. Phys. A* **39**, 7607–7619 (2006).
26. Bertin, E. & Clusel, M. Global fluctuations in physical systems: A subtle interplay between sum and extreme value statistics. *Int. J. Mod. Phys B* **22**, 3311–3368 (2008).
27. Bruce, A. D. Critical finite-size scaling of the free energy. *J. Phys. A* **28**, 3345–3349 (1995).
28. Labit, B. *et al.* Universal statistical properties of drift-interchange turbulence in TORPEX plasmas. *Phys. Rev. Lett.* **98**, 255002 (2007).
29. Farago, J. Injected power fluctuations in Langevin equation. *J. Stat. Phys.* **107**, 781–803 (2002).
30. Goldenfeld, N. *Lectures on Phase Transitions and the Renormalization Group* (Addison–Wesley, 1992).
31. Malakis, A. & Fytas, N. G. Universal features and tail analysis of the order-parameter distribution of the two-dimensional Ising model: An entropic sampling Monte Carlo study. *Phys. Rev. E* **73**, 056114 (2006).
32. Tsypin, M. M. & Blöte, H. W. J. Probability distribution of the order parameter for the three-dimensional Ising-model universality class: A high-precision Monte Carlo study. *Phys. Rev. E* **62**, 73–76 (2000).
33. Bramwell, S. T. *et al.* Magnetic fluctuations in the classical XY model: The origin of an exponential tail in a complex system. *Phys. Rev. E* **63**, 041106 (2001).
34. Berezinskii, V. L. Destruction of long range order on one-dimensional and two-dimensional systems having a continuous symmetry, I—classical systems. *J. Exp. Theor. Phys.* **32**, 493–500 (1971).
35. Banks, S. T. & Bramwell, S. T. Temperature-dependent fluctuations in the two-dimensional XY model. *J. Phys. A* **38**, 5603–5615 (2005).
36. Ricardo Paredes, V. & Botet, R. Scanning the critical fluctuations: Application to the phenomenology of the two-dimensional XY model. *Phys. Rev. E* **74**, 060102 (2006).
37. Foltin, G., Oerding, K., Rácz, Z., Workman, R. L. & Zia, R. K. P. Width distribution for random-walk interfaces. *Phys. Rev. E* **50**, R639–R642 (1994).
38. Portelli, B., Holdsworth, P. C. W., Sellito, M. & Bramwell, S. T. Universal magnetic fluctuations with a field-induced length scale. *Phys. Rev. E* **64**, 036111 (2001).
39. Oono, Y. Large deviation and statistical physics. *Prog. Theor. Phys. Suppl.* **99**, 165–205 (1989).
40. Boucher, C., Ellis, R. S. & Turkington, B. Spatializing random measures: Doubly indexed processes and the large deviation principle. *Ann. Probab.* **27**, 297–324 (1999).
41. Salazar, R., Toral, R. & Plastinoc, A. R. Numerical determination of the distribution of energies for the XY-model. *Physica A* **305**, 144–147 (2002).
42. Barré, J., Bouchet, F., Dauxois, T. & Ruffo, S. Large deviation techniques applied to systems with long-range interactions. *J. Stat. Phys.* **119**, 677–713 (2005).
43. Fisher, R. A. & Tippett, L. H. C. Limiting forms of the frequency distribution of the largest or smallest member of sample. *Proc. Camb. Phil. Soc.* **24**, 180–190 (1928).
44. Györgyi, G., Moloney, N. R., Ozogány, K. & Rácz, Z. Finite-size scaling in extreme statistics. *Phys. Rev. Lett.* **100**, 210601 (2008).
45. Cassandro, M. & Jona-Lasinio, G. Critical point behaviour and probability theory. *Adv. Phys.* **27**, 913–941 (1978).
46. Zucker, I. J. & Robertson, M. M. Exact values of some two-dimensional lattice sums. *J. Phys. A* **8**, 874–881 (1975).
47. McPhedran, R. C., Botten, L. C., Nicorovici, N. A. & Zucker, I. J. Systematic investigation of two-dimensional static array sums. *J. Math. Phys.* **48**, 033501 (2007).

Acknowledgements

It is a pleasure to thank Maxime Clusel and Peter Holdsworth for very valuable comments and criticisms.

Additional information

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>.