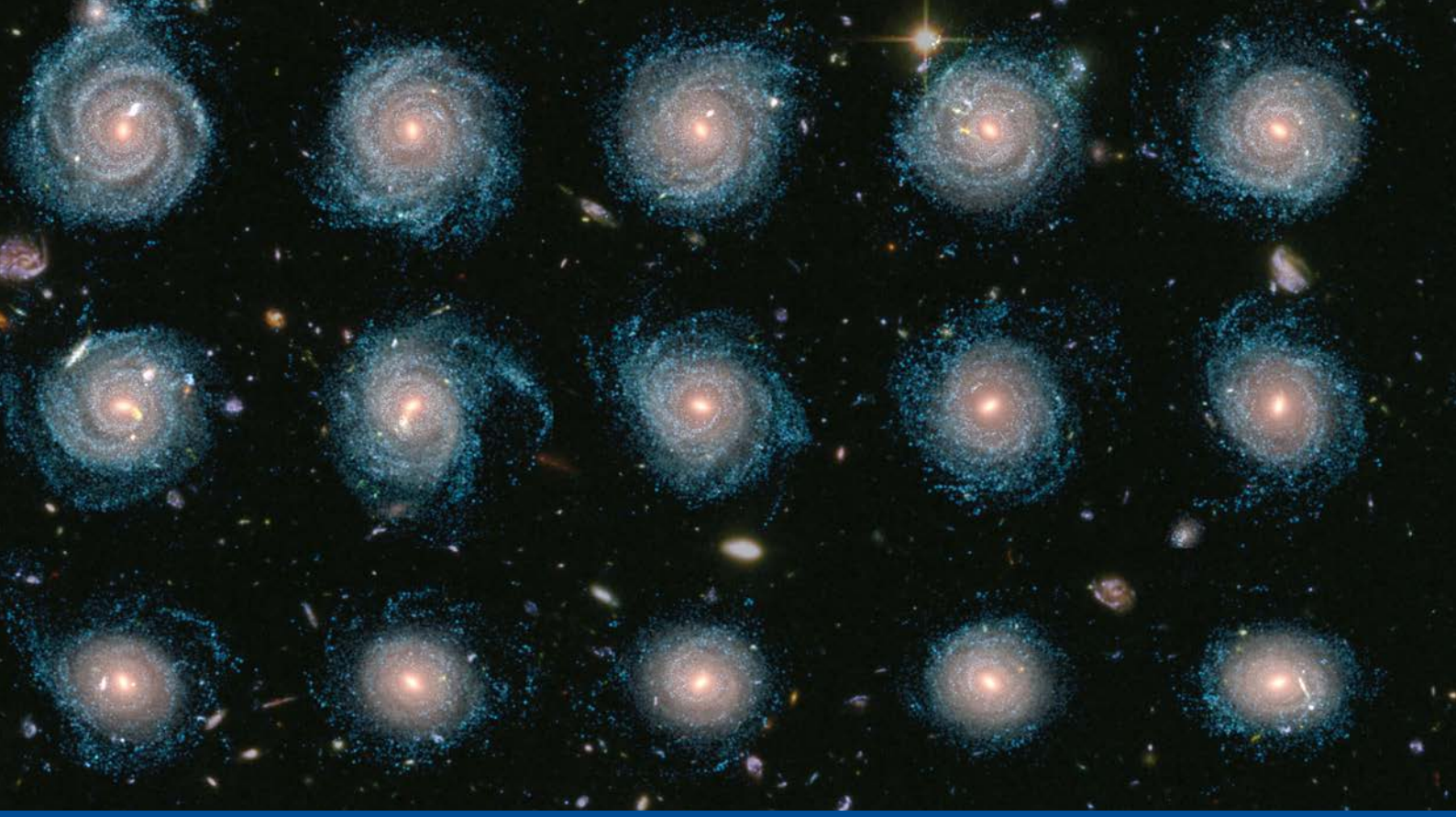


HITS 2013

Annual Report
Jahresbericht

Heidelberg Institute for
Theoretical Studies

HITS

THINK BEYOND THE LIMITS!

Inhalt | Table of Contents

1 Think Beyond the Limits!

Im Jahr 2013 ist der Aufbau des Heidelberger Instituts für Theoretische Studien erfreulich rasch vorangekommen. In der Sprache der Kosmologie könnte man sagen: Das HITS hat eine inflationäre Phase durchlaufen und wird von nun an mit deutlich geringerer Geschwindigkeit weiter expandieren. Weiter wollen wir diese Metapher freilich nicht strapazieren; sonst müssten wir noch diskutieren, was die für diese Expansion verantwortliche dunkle Energie in unserem Fall ist.

Zu den Fakten: Am Anfang des Jahres hat die Nachwuchsgruppe „Computational Biology“ (CBI) unter der Leitung von Dr. Siegfried Schloissnig ihre Arbeit aufgenommen. Sie beschäftigt sich insbesondere mit geeigneten Assemblierungsverfahren für die neueste Generation von Sequenzierern, die durch lange „Reads“ einerseits und relativ hohe Fehlerraten andererseits gekennzeichnet sind. Die Gruppe arbeitet eng zusammen mit Gene Myers vom Max-Planck-Institut für Molekulare Zellbiologie in Dresden. Myers, Autor des weltweit eingesetzten BLAST-Codes, begleitet die Arbeit von CBI als wissenschaftlicher Mentor.

Im Mai wurde die assoziierte Gruppe „Data Mining and Uncertainty Quantification“ (DMQ) eingerichtet. Sie wird geleitet von Prof. Vincent Heuveline, der hauptamtlich an der Universität Heidelberg tätig ist und dort u.a. die wissenschaftliche Leitung des Rechenzentrums übernommen hat. Die Gruppe befasst sich mit der Quantifizierung des Effektes von Ungenauigkeiten in großen Datenmengen bzw. mathematischen Modellen auf die Zuverlässigkeit von Ergebnissen der Analyse sehr großer Datenmengen. Darüber hinaus erwarten wir aus der Kooperation mit Prof. Heuveline wichtige Anstöße für die Weiterentwicklung von Methoden der „eScience“ zusammen mit der Universität Heidelberg.

Im Herbst nahm eine weitere Nachwuchsgruppe ihre Arbeit auf, die den Namen „Astroinformatik“ (AIN) trägt. Ihr Leiter, Dr. Kai Polsterer, ist durch seinen fachlichen Hintergrund, ein Diplom in Informatik und ein Doktorgrad in Physik, sowie durch seine langjährige Kooperation mit beobachtenden Astronomen prädestiniert für die Entwicklung von Methoden des Data Mining für astronomische Daten.

Im November schließlich startete die Gruppe „Computational Statistics“ (CST) unter der Leitung von Prof. Tilmann Gneiting vom Karlsruher Institut für Technologie (KIT). Ihr Forschungsschwerpunkt liegt auf den methodischen Grundlagen von Prognoseverfahren und der räumlich-/zeitlichen Statistik.

Für den Betrieb und die Weiterentwicklung der anspruchsvollen IT-Technik des HITS wurde eine eigene Gruppe mit dem Namen „IT-Services“ (ITS) eingerichtet. Ihr Leiter ist Dr. Bogdan Costescu, der zuvor wissenschaftlicher Mitarbeiter in der MBM-Gruppe war.

Leider gab es in diesem Jahr neben all diesen positiven Entwicklungen ein sehr schmerzliches Ereignis: Im Juli verstarb die langjährige Leiterin der SDBV-Gruppe und erste Trägerin der Auszeichnung „HITS-Fellow“, Dr. Isabel Rojas, nach langer Krankheit. Wir werden ihrem Enthusiasmus für alles, was mit Wissenschaft zu tun hat, und ihren wertvollen Beiträgen zum Aufbau des Institutes stets ein ehrendes Andenken bewahren.

Dr. h.c. Dr. E.h. Klaus Tschira
Prof. Dr.-Ing. Dr. h.c. Andreas Reuter

It is gratifying to see that the development of the Heidelberg Institute for Theoretical Studies has made quick progress in 2013. Speaking in the language of cosmology, one could say that after an inflationary epoch, HITS will from now on continue to expand at a significantly lower rate. We should not, however, overwork this metaphor, for otherwise we might end up having to explain what in our case corresponds to the dark energy responsible for expansion.

So let's take a look at the facts. At the beginning of the year, the junior research group "Computational Biology" (CBI) began its work under the direction of Dr. Siegfried Schloissnig. One of its main focuses is on assembly processes suitable for the latest generation of sequencers characterized both by long "reads" and relatively high error rates. The group works closely with Gene Myers at the Max Planck Institute of Molecular Cell Biology in Dresden, Germany. Myers, the author of the BLAST codes used world-wide, is CBI's scientific mentor.

In May, the associated group "Data Mining and Uncertainty Quantification" (DMQ) was established at HITS. Prof. Vincent Heuveline is the group leader. He is a full-time member of Heidelberg University, one of his tasks being the directorship of the University's computer center. The group investigates the quantification of the impact of uncertainties in large data sets or mathematical models on the reliability of the results obtained from analysis of those data sets. Additionally, we expect cooperation with Prof. Heuveline and the Heidelberg University to bring added impetus to the development of new "eScience" methods.

In the fall, another junior research group started its work. It is called "Astroninformatics" (AIN). Group leader Dr. Kai Polsterer is ideal for the task of developing data-mining methods for astronomical data due to his professional background, his degree in computer science, a PhD in physics, and his long-standing cooperation with observational astronomers.

Finally, the group "Computational Statistics" (CST) under the direction of Prof. Tilmann Gneiting of the Karlsruhe Institute of Technology (KIT) was established in November. Its research focus is on the methodological basis of forecasting methods and spatial statistics.

The group "IT Services" (ITS) was established to take care of the operation and development of HITS' sophisticated IT technology. Its leader is Dr. Bogdan Costescu, who previously worked as a research associate in the MBM group.

Positive as all these developments are, they were overshadowed this year by a very distressing event. Dr. Isabel Rojas, long-standing SDBV group leader and first winner of the "HITS Fellow" award, died in July after a long illness. We will always honor the memory of her enthusiasm for everything science-related and her valuable contributions to the development of our institute.

Dr. h.c. Dr. E.h. Klaus Tschira
Prof. Dr.-Ing. Dr. h.c. Andreas Reuter

2 Research

2.1 Astroinformatics

In 2013, the Astroinformatics group was established at HITS. Our goal is to develop new methods and tools to deal with the exponentially increasing amount of data in astronomy.

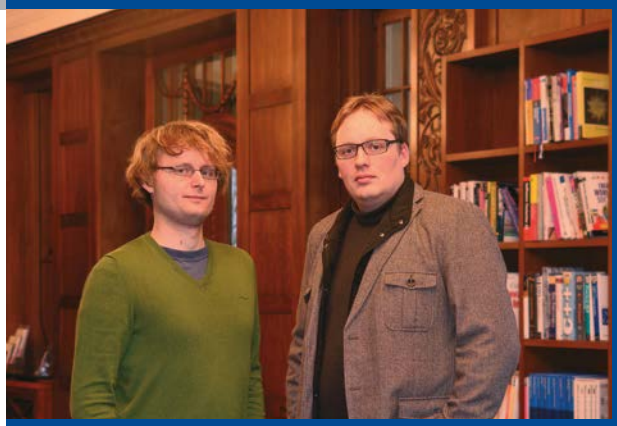
In the last two decades, computers have revolutionized astronomy. Advances in technology have led to the advent of new detectors, complex instruments, and innovative telescope designs. These advances enable today's astronomers to observe objects with unprecedented accuracy and in high spatial/spectral/temporal resolution. In addition, there are new untapped wavelength regimes waiting to be explored. Dedicated survey telescopes map the sky and constantly collect data that are then made available to the community free of charge. Our aim is to help scientists analyze this increasing amount of information.

The group is interested in the development of improved photometric redshift regression models. This is a key tool in the analysis of the huge amounts of data provided by upcoming major survey projects like the Square Kilometer Array (SKA), Gaia, or Euclid. Also of crucial scientific interest are methods and tools for extracting and filtering rare objects for detailed follow-up analysis with 8-m class telescopes. Estimated occurrences of only a few objects per million make manual inspection of the existing catalogs impossible. Morphological classification of galaxies based on imaging data as well as high-dimensional similarity measures are the other research interests of the Astroinformatics group. These are initial stages in providing more exploratory access to the data archives for astronomers.

2013 wurde die Astroinformatik-Gruppe am HITS gegründet. Unser Ziel ist es, neue Methoden und Werkzeuge zu entwickeln, um eine Analyse der exponentiell wachsenden Anzahl an Daten im Bereich der Astronomie zu ermöglichen.

In den letzten zwanzig Jahren hat der Einsatz von Computern die Astronomie stark beeinflusst. Durch technologische Fortschritte wurde es möglich neue Detektoren sowie innovative Instrumente und Teleskopdesigns zu realisieren. Dadurch können Astronomen nun Objekte mit bisher unerreichtem Detailreichtum und in neuen Wellenlängenbereichen beobachten. Mit speziell dafür vorgesehenen Teleskopen wird der Himmel jede Nacht beobachtet und die so gewonnen Daten werden frei zur Verfügung gestellt. Wir möchten es Wissenschaftlern durch unsere Forschung ermöglichen, diese riesigen Datenmengen durch neue Analysemethoden effizienter zu nutzen.

Unsere Gruppe beschäftigt sich mit der Entwicklung photometrischer Rotverschiebungsmodelle. Diese werden für die neuen Generationen von Himmelsdurchmusterungen benötigt. Des Weiteren beschäftigen wir uns mit der Suche nach astronomischen Objekten, die mit einer Häufigkeit von ein paar wenigen pro Million vorkommen. Um solch seltene Objekte für detaillierte Untersuchungen zu finden, scheidet die manuelle Selektion aus. Die morphologische Klassifikation von Galaxien sowie hochdimensionale Ähnlichkeitsmaße sind weitere Forschungsbereiche. Beide Bereiche werden benötigt, um einen explorativeren Datenzugang für die Astronomen zu schaffen.



AIN

Group Leader
Dr. Kai Polsterer

Scholarship Holder
Sven Dennis Kügler (HITS Scholarship, from Oct. 2013)

PHOTOMETRIC REDSHIFT REGRESSION MODELS

Photometric redshift estimation models are an important tool in astronomy. Large-scale all-sky surveys are typically based on broadband imaging. For this reason, only a rough analysis of specific object parameters is possible for most of the objects detected. To verify the true nature of a particular object, spectroscopic follow-up observations (with higher resolution) are usually required. Since obtaining such spectra is much more time-consuming than broadband imaging, detailed information is only available for a relatively small subset of detected objects. Accordingly, regression tasks are common in astronomy, for instance in estimating the redshift or the metallicity of a galaxy. From a data-mining perspective, photometric objects observed spectroscopically form the basis for the generation of new models, which in turn can then be applied to all remaining objects in the catalog. Typical examples of such models are classification or regression models based on photometric features (Bolzonella et al. 2000, Laurino et al. 2011). Alongside the lack of labeled data for specific learning tasks, the heterogenity of the available catalogs often involves the problem of missing data. Detailed information or other wavelength ranges may only be available for some objects in a survey. In addition, the features available may differ in their explanatory power. This immediately poses the question of selection, i.e. how to obtain good features constituting a representative subset of all the features involved. The features extracted for each band extend from plain magnitudes to more complex composite features.

Usually, the colors derived from adjacent filter bands are fed to appropriate regression techniques to deal with the redshift estimation task (O'Mill et al. 2011, Wu et al. 2010). Together with Fabian Gieseke we have been considering an alternative approach. Instead of resorting to these standard features, we consider a large set of combined features, ignoring any particular knowledge about the physical properties of the input parameters. Instead of trying to improve the regression technique itself, we concentrate on selecting the best-performing subset of features. Selecting a subset of features is, however, a combinatorial matter, and it quickly defies feasibility even for moderate-sized subsets. To accelerate the indispensable search, we therefore draw upon the massive computational resources provided by modern-day graphics processing units (GPUs). As an underlying regression model we have recourse to simple nearest-neighbor techniques. These have proved to be an excellent choice for such tasks due to the large number of patterns in a low-dimensional feature space (Stensbo-Smidt et al. 2013). In computing nearest neighbors one can choose any similarity measure more or less at random. We have decided to use the popular Euclidean distance. A parallel implementation can easily be achieved via, say, matrix multiplication (Garcia et al. 2010).

We have used data from the Sloan Digital Sky Survey (SDSS) (York et al. 2000; Ahn et al. 2013) to evaluate our redshift estimation models. Our evaluation criteria are the root mean square (RMS) error and the median absolute deviation (MAD). We first retrieved the photometric features of all quasars from the SDSS with spectroscopically determined redshifts. Then we selected the point spread function (PSF) magnitudes, the model magnitudes, and the petrosian magnitudes, including their measurement errors in all five filter bands (u, g, r, i, z). In addition, the foreground extinction caused by our galaxy was used to deredden the measurements. Typically, color features are built by using adjacent filter bands rather than raw band information. This is designed to reduce object-intrinsic luminosity properties not related to the object's redshift. For this reason, we did not only consider the raw features and the differences between adjacent filter bands, but also all other possible feature differences. All composite features were normalized to the same value range. For the feature selection scheme, a randomly selected subset of 5,000 photometric patterns was considered, while the final performance was calculated on all available quasars, applying cross-fold validation. With a greedy forward selection approach the best-performing features were subsequently added to the selected subset.

By using a massive parallel feature selection approach on a GPU, we were able to determine a set of features. With the nearest-neighbor regression model used for this task, the features selected by our framework led to a significantly better prediction performance for photometric redshifts in comparison with standard features (see Figure 1).

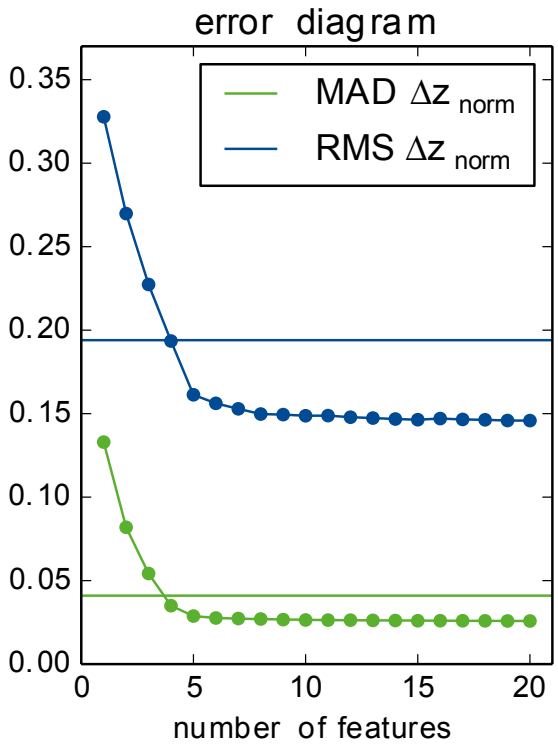


Fig. 1: Comparison of the performance of our photometric redshift model with the best-performing reference model available so far. Depending on the number of selected features, RMSE and MAD fall below the reference values (solid horizontal lines). With only four features, our model equals the performance of the reference model, although this model makes use of eight features.

The method thus derived indicates that optimizing well-known models by systematically testing possible feature combinations can lead to significantly improved results. A well-accepted photometric regression model (Laurion et al. 2011) was used as reference model. With the same reference dataset, the performance with the selected features is approximately 25% better in RMS and 30% in MAD, respectively (see Figure 2).

SEARCH FOR BINARY BLACK HOLES

Observations and numerical simulations suggest that galaxies are growing by merging (Toomre & Toomre 1972). Interacting galaxies such as Whirpool (M51) are nearby examples that confirm this theory. The super massive black holes (SMBHs) found at the center of nearly every large galaxy (Richstone et al. 1998) are also thought to grow over time via hierarchical merging processes.

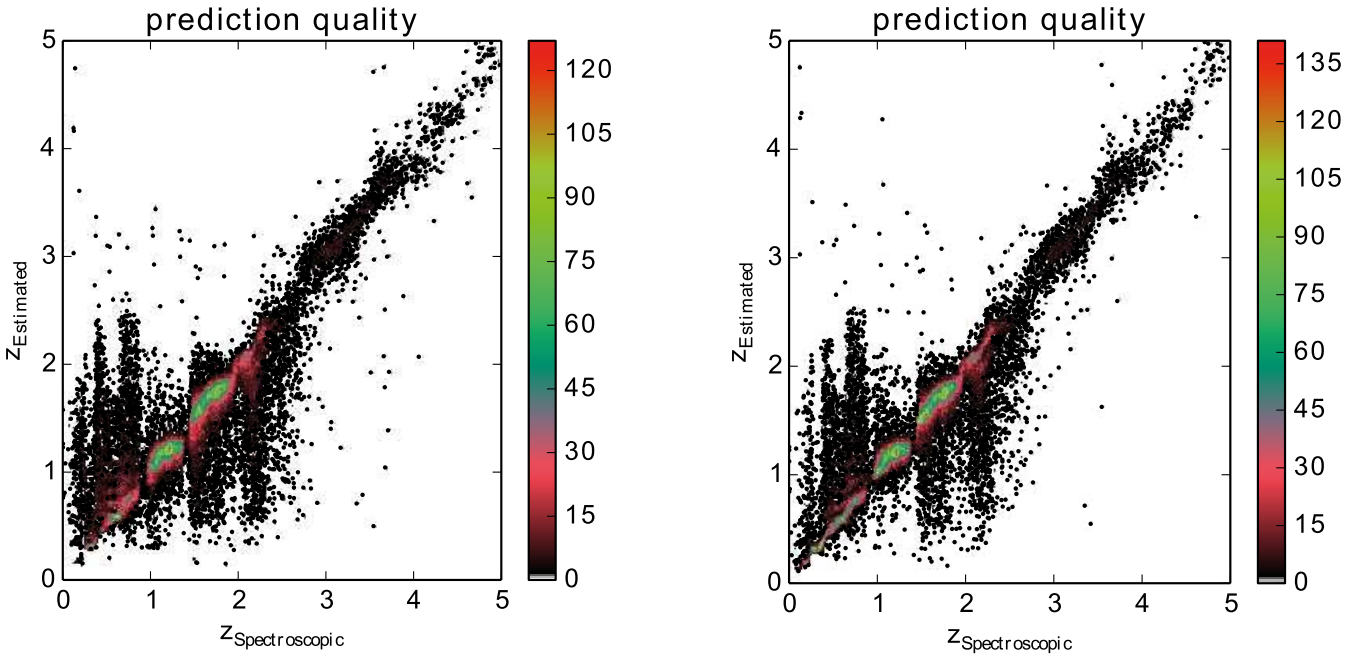


Fig. 2: Left: Redshift estimation performance with four standard features. Note that the high density along the bisecting line is color-encoded. Right: Redshift estimation performance with 20 selected features. In comparison to the performance with four features only, improvement is readily observable, notably the low redshift regime and the density of elements close to the bisecting line

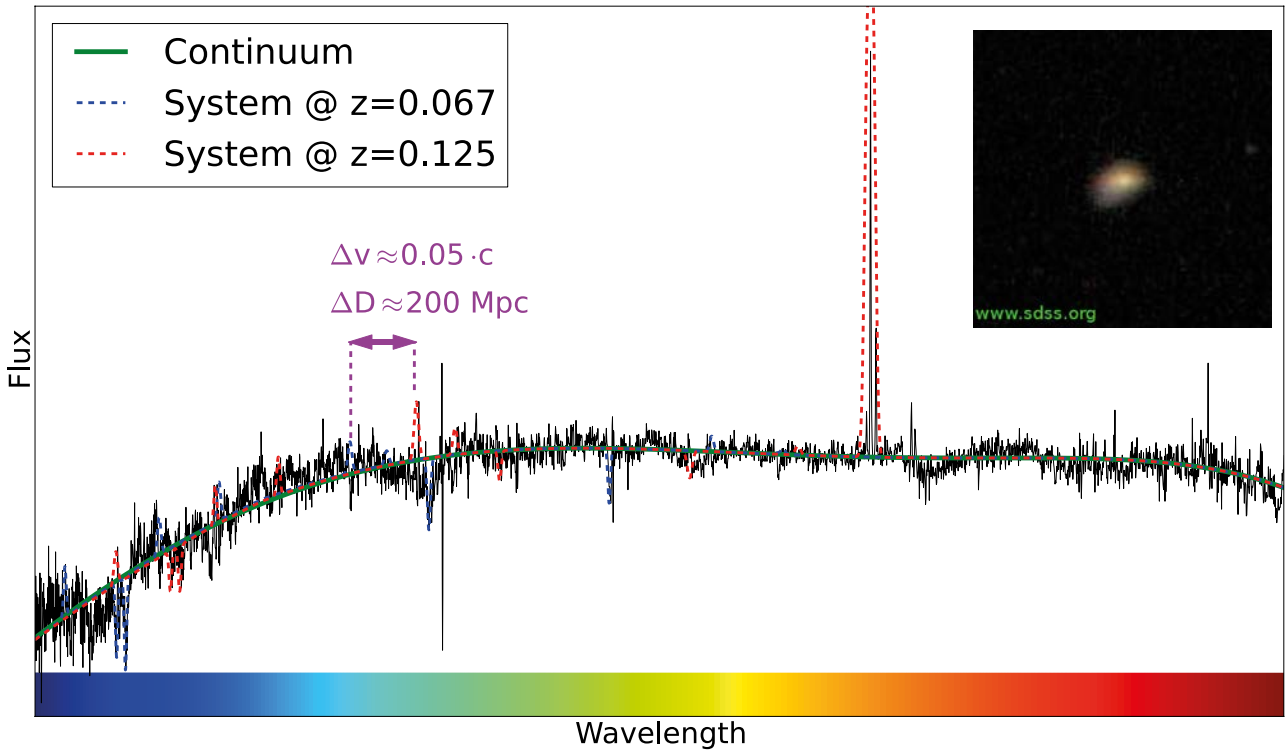
Assuming that nearly all massive galaxies host an SMBH and that those galaxies are merging with each other, it is inevitable that the SMBHs will also merge. The conservation of angular momentum prevents direct infall, so the merging process is preceded by a binary stage. Accordingly, this merging process should be a rather slow process. At this stage, the two SMBHs orbit a common center of mass, where the orbit shrinks due to friction and 3-body interactions. Up to now, we have knowledge of only a mere handful of good candidates for SMBH binaries (NGC 6240 by Komossa et al. 2003, C0402+379 by Rodriguez et al 2006). This flies in the face of what we would expect from the two arguments set out above. In addition, those binaries are still widely separated (>100 light years), which means that they are still in an early merging state. Numerical simulations of these merging phases predict that most of the SMBH binaries achieve their „longest-lived“ state at very tiny separations (a few light years, see Milosavljevic & Merritt 2003). Binaries at this point in the merging process cannot be resolved by imaging, as the distance even to the closest galaxy is extremely high in comparison with their separation. A very naive but straightforward approach is to assume that both of the SMBHs are active. This means that gas and dust in the direct vicinity is accreted by the black hole on a disk and thereby parts of the potential and kinetic energy are converted into radiation. Only about 10% of the galaxy cores are active and emit some very distinct lines that can be classified as broad or narrow. Where two active SMBHs orbit their common center of mass, a relative shift may be observed between their emission lines with respect to the orientation of the system as a whole, a phenomenon caused by the Doppler effect. This approach was followed up by Smith et al. (2010) and Tsalmantza et al. (2011), but it transpired that in most of the selected objects the emission shift was caused by kinematic effects (Fu et al. 2012).

In our work we attempt to improve and extend this approach with tools and methods that are new to astronomy: those of machine learning. We draw upon the physical assumption that about 90% of the cores are actually inactive (do not emit light), so binary black holes with just one active core should occur more often. By analyzing spectra and simply comparing the redshift of the absorption lines of a galaxy with the redshifts of emission lines caused by an embedded active SMBH, we should be able to detect any shift between the host system and the active core in a binary black hole system. Assuming that the likelihood of an inactive-active binary pair is the same as for non-binary galaxies, we would expect to find about a factor of 10 times more candidates than we would expect for active-active pairs. To determine the relative shift between the absorption and emission lines, we cannot rely on standard tools such as model fitting, as they assume a common redshift for the entire spectrum. Just adding two of these models would not provide enough information about a relative shift. The model fit would prefer to match the behavior of the continuum instead of matching real features. Instead of assuming that there must be fitting spectral behavior or a special line shape, we use a statistical approach that compares all spectra with each other on a similarity basis. This enables us to rank the spectra by feature shift and thus to obtain information about the redshifts of distinct emission and absorption frames. This approach can only be applied when a statistically representative training sample exists and when most of the systemic redshifts retrieved from the spectral database are correct. While the former assumption can be easily met (SDSS contains more than $2*10^6$ training objects with approximately 4,000 values each), the latter assumption is harder to comply with. Though some reliability measures are given in the database, the question remains how stable and reproducible the estimates of the applied methods actually are. For this reason, we only select objects

for the training sample that have no detected problems in redshift estimation. Regions with potential emission/absorption features are individually inspected to determine redshift by using a k-nearest-neighbors regression code that takes the selected training sample as a reference. High deviations between the redshift estimate for emission and absorption indicate possible binary black hole candidates.

An initial test run on the training data already revealed a very interesting object displaying signs of a shift between emission and absorption (see Figure 3). Though further investigation revealed that the system is in fact a superposition of two galaxies, it demonstrates that in principle the selection mechanism actually works and is capable of detecting such objects. Given the extremely low likelihood of finding such a superposition in the entire data, only a few objects will figure in our final candidate sample. Next year we will be fine-tuning the preprocessing steps and the selection process and applying our selection mechanism to the entire data set.

Fig. 3: SDSS spectrum of two galaxies in superposition. While the image of the object found with our approach shows just one slightly distorted galaxy, the spectrum reveals two objects that are exactly aligned along the line of sight.



LUCI

The Astroinformatics group is also involved in the LUCI project. LUCI is a pair of near-infrared imagers and spectrographs at the Large Binocular Telescope (LBT), the world's largest optical telescope. The innovative approach embodied by the LBT is to combine two 8.4-m telescopes on a single mount, thus allowing for the interferometric combination of both mirrors. We are responsible both for the management of control software development and for the development of the observation preparation tool. In November 2013, we successfully commissioned LUCI2 and succeeded in running initial binocular observations. Next year will bring two big milestones for the LUCI project. At the end of 2014, we will be implementing new optics to make full use of the telescope's adaptive optics system. Toward mid 2015, the laser guide star system ARGOS (see Figure 4) will be fully functional. This will enable the LUCI instruments to make observations in diffraction-limited resolution.

Fig. 4: First activation of the ARGOS laser guide star system at the Large Binocular Telescope in November 2013. (Photo by Polsterer)

The Computational Biology Junior Group (CBI) started its work at HITS in 2013 and grew over the course of the year to its current size of four. Philipp Kämpfer and Philipp Bongartz, two PhD scholarship holders, joined in June and August respectively, and Martin Pippel joined in July as a postdoc. Furthermore, the group receives mentorship from Gene Myers, one of the pioneers in the field of genome assembly. Gene is a director at the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden and holds the Klaus Tschira Chair of Systems Biology.

The CBI group works at the interface(s) between computer science, mathematics, and the biological sciences. Our research focuses on the computational and algorithmic foundations of genome biology. Of the multitude of issues encountered in that field, we are especially interested in whole-genome assembly, the reconstruction of a genome's sequence from the data produced by a DNA sequencer. The basic principle applied for assembly is to randomly (over-)sample overlapping fragments from the genome, sequence them, and computationally reconstruct the full sequence from those fragments.

The complexity of this task is largely dependent on two characteristics of the fragments, average length and accuracy. The current generation of sequencers produces very long fragments, but with high error rates, so new approaches to the problem of efficient assembly under such conditions are needed. The development of such algorithms and their efficient implementation and application in genome sequencing projects are the main goals of the group.

Die Computational Biology Junior Group (CBI) begann ihre Arbeit am HITS Anfang 2013 und wuchs im Laufe des Jahres zu der aktuellen Größe von vier Mitgliedern. Zwei Promotionsstipendiaten, Philipp Kämpfer und Philipp Bongartz, starten Juni bzw. August und Martin Pippel nahm seine Tätigkeit als PostDoc im Juli auf. Des weiteren hält Gene Myers, einer der Pioniere im Bereich der Genomassemblierung und Direktor am Max-Planck-Institut für Zellbiologie und Genetik in Dresden, die Rolle des Mentors der Gruppe inne.

Die CBI Gruppe arbeitet an der Schnittstelle von Information, Mathematik und Biologie, mit Fokus auf die informatischen und algorithmischen Grundlagen der Genombiologie. Von der Vielzahl an Problemen in diesem Feld, sind wir besonders an der Assemblierung von Genomsequenzen interessiert. Darunter ist die Rekonstruktion der Sequenz (Folge der Nukleotide) eines Genoms, basierend auf Daten die durch einen DNA-Sequenzierer produziert wurden, zu verstehen.

Das Prinzip hinter Assemblierung ist, aus dem Genom zufällig (überlappende) Fragmente auszulesen, diese zu sequenzieren und anschließend aus der Sequenz dieser Fragmente die komplette Genomsequenz mit computergestützten Verfahren zu rekonstruieren.

Die Komplexität dieses Ansatzes wird primär von der Länge der Fragmente und der Fehlerrate des DNA-Sequenzierers bestimmt. Die aktuelle Generation an Sequenzierern, welche sehr lange Fragment aber mit einer hohen Fehlerrate produzieren, erfordert neue algorithmische Ansätze, um Genome effizient unter solchen Bedingungen rekonstruieren zu können. Die Entwicklung solcher Verfahren und deren Anwendung in Genomsequenzierungsprojekten stellen die Hauptaufgaben der Gruppe dar.

A DE NOVO WHOLE-GENOME SHOTGUN ASSEMBLER FOR NOISY LONG-READ DATA

On the face of it, 10Kbp reads from single molecule sequencers are impressive, but an error rate of 15% often makes them difficult to handle. However, truly random error positioning and near-Poisson single-molecule sampling imply that reference-quality reconstructions of gigabase genomes are in fact possible with as little as 30X coverage. Such a capability would resurrect the production of true reference genomes and enhance comparative genomics, diversity studies, and our understanding of structural variations within a population.

We have built a prototype assembler we call the Dazzler that can assemble 1-10Gb genomes directly from a shotgun, long-read data set currently producible only with the PacBio RS II sequencer. It is based on the string graph paradigm, its two most important attributes being:

1) It scales reasonably to gigabase genomes, being roughly 30 times faster than current assemblers for this kind of data.
2) A "scrubbing phase" detects and corrects read artifacts including untrimmed adapter, polymerase strand jumps, and ligation chimers that are the primary impediments to long contiguous assemblies.

Developments in this work have been rapid. The Dazzler on the PacBio E. Coli data set produces a perfect result in 10 minutes on a laptop. We have sequenced C. briggsae and Species 9 to 30X each and with these are currently perfecting assembly at the 100Mbp scale. Two 30-40X data sets for two gigabase genomes were completed in December, and we are currently working on two draft assemblies of those genomes.

A HYBRID GRAPH APPROACH FOR SHORT-READ ASSEMBLY

Though many consider the assembly issue a solved problem, unordered and fragmented genome assemblies with false joins are widespread, significantly hampering any downstream analysis in which they are involved. NGS sequencers produce gigabases very quickly and cheaply but read-lengths are mostly very short. Short read-lengths and short paired-read insert lengths are the primary reasons why despite ~100X sequencing coverage most assembly tools have difficulty producing accurate assemblies with long-range contiguity. Another issue is that most assembly projects choose insert lengths and mixes that are not optimized for the target genome.

In this sector we are currently developing approaches for systematically optimizing paired-end libraries to the repeat structure and composition of the target genome before sequencing. This entails the creation of heuristics that gauge the coverage and paired-end library insert sizes needed before any sequencing actually takes place.

Given close to optimal insert sizes and coverage, the next challenge lies in the reconstruction of the genomic sequence based on the short-read data, a problem for which currently two distinct graph-based approaches are employed. The string graph concept is superior to the deBruijn graph in that the unit of assembly is a read as opposed to a small k-mer, so that the graph and its path structure are simpler.

However, most NGS assemblers rely on the deBruijn graph approach simply because it is more time- and space-efficient. In this project, we are taking advantage of the best of both approaches. We quickly build a deBruijn graph, noting where each read starts and ends within the graph, and employ novel graph algorithms to efficiently find the transitively invariant read overlaps. These

are the edges of the string graph, with traversals to the left and right along paths from each read location. We are currently working toward prototype implementation and refining the algorithmic aspects in order to compute the string graph in linear expected time without computing all the pairwise overlaps, the Achilles heel of the approach in computation terms.

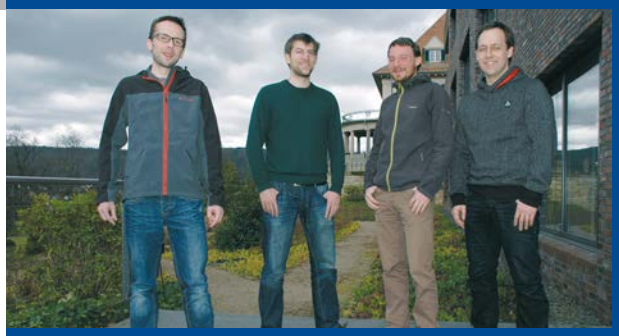
GENOME PROJECTS

PLANARIANS

The ability to regenerate lost body parts is widespread in the animal kingdom. Humans, by contrast, are unable to regenerate even minor extremities. If the “survival of the fittest” principle really holds, regeneration should be the exception rather than the rule and remains a fascinating conundrum in biology. Even amongst planarian flatworms, celebrated for their ability to regenerate complete specimens from random tissue fragments, species exist that have completely lost the ancestral ability to regenerate.

Owing to the lack of physical maps, high AT-content, and high repeat density, not one single high-quality planarian genome is currently available. We are working on a draft assembly of Schmidtea mediterranea, which has defied previous assembly attempts for many years now.

This work is performed in close collaboration with the Max Planck Institute for Molecular Cell Biology and Genetics and the Systems Biology Center in Dresden. These two institutions provide species samples and will subsequently conduct further experiments on the basis of the said genome sequence in an attempt to understand regeneration at a systems level.



CBI

The CBI group in 2013 (f.l.t.r.):
Siegfried Schloissnig, Philipp Bongartz, Philipp Kämpfer, Martin Pippel

Group Leader
Dr. Siegfried Schloissnig

Staff Members
Martin Pippel (from July 2013)

Scholarship Holders
Philipp Bongartz (HITS Scholarship, from Aug.2013)
Philipp Kämpfer (HITS Scholarship, from June 2013)

TIGER FLATWORM (MARITIGRELLA CROZIERI)

Maritigrella crozieri is a member of Policladida, a highly diverse clade within the phylum Platyhelminthes. These marine turbellarian flatworms can be found on the eastern coasts of North America and the Caribbean Sea. Their main advantages are ease of collection, spiral cleavage, biphasic life cycle, and large size (up to 55mm) with many eggs, which can be obtained and raised without eggshells. Accordingly, they represent an interesting system for evolutionary and developmental studies within their phylum.

The diploid genome of M. crozieri is estimated to be highly repetitive and about 2.5 gigabases in length, distributed across three chromosomes. The aim of our work is to generate an initial draft genome of this species using the de novo genome assembly strategies described above.

The assembly is mostly based on multiple short-read, paired-end, and mate-pair libraries originating from the Illumina sequencing platform. First, quality assessment and filtering of the reads was performed. Then a set of assemblers based on both the deBruijn and the string-graph concept was used to produce longer contiguous sequences called contigs. To partly overcome the aforementioned repetitiveness of the genome, additional long-read sequencing will be performed on the PacBio RS II sequencer.

This work is being done in conjunction with Prof. Max Telford of University College London and the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden.

The Computational Statistics group at HITS was established in November 2013, when Tilmann Gneiting was appointed group leader and Professor of Computational Statistics at the Karlsruhe Institute of Technology (KIT). It conducts research in two main areas: (1) the theory and practice of forecasting and (2) spatial and spatio-temporal statistics.

The group's current focus is on probabilistic forecasting. As the future is uncertain, forecasts should be probabilistic in nature, i.e. take the form of probability distributions over future quantities or events. Accordingly, we are witnessing a transdisciplinary change of paradigms from deterministic or point forecasts to probabilistic forecasts. The CST group seeks to provide guidance and leadership in this transition by developing both the theoretical foundations of the science of forecasting and cutting-edge statistical methodology, notably in connection with applications. Weather forecasting is a key example. In this context, the group maintains research contacts and collaborative relations with national and international hydrologic and meteorological organizations like the German Weather Service, the German Federal Institute of Hydrology, and the European Centre for Medium-Range Weather Forecasts.

In 2014, the CST group expects to grow substantially, with new staff members, visiting scientists, and students joining group leader Tilmann Gneiting.

Die Computational Statistics-Grupe am HITS besteht seit November 2013, als Tilmann Gneiting seine Tätigkeit als Gruppenleiter sowie Professor für Computational Statistics am Karlsruher Institut für Technologie (KIT) aufnahm. Sie beschäftigt sich mit zwei wesentlichen Arbeitsgebieten, der Theorie und Praxis der Vorhersage sowie der räumlichen und Raum-Zeit-Statistik.

Der Forschungsschwerpunkt der Gruppe liegt derzeit im Gebiet der probabilistischen Vorhersage. Im Angesicht unvermeidbarer Unsicherheiten sollten Vorhersagen probabilistisch sein, d.h. Prognosen sollten die Form von Wahrscheinlichkeitsverteilungen über zukünftige Ereignisse und Größen annehmen. Dementsprechend erleben wir einen transdisziplinären Paradigmenwechsel von deterministischen oder Punktvorherzusagen hin zu probabilistischen Vorhersagen. Der CST-Gruppe ist es wichtig, diese Entwicklungen nachhaltig zu unterstützen, indem sie theoretische Grundlagen für wissenschaftlich fundierte Vorhersagen begründet, eine Vorreiterrolle in der Entwicklung entsprechender statistischer Methoden einnimmt und diese in wichtigen Anwendungsproblemen, wie etwa in der Wettervorhersage, zum Einsatz bringt. In diesem Zusammenhang bestehen Kooperationen mit nationalen und internationalen hydrologischen und meteorologischen Organisationen, wie etwa dem Deutschen Wetterdienst, der Bundesanstalt für Gewässerkunde und dem Europäischen Zentrum für mittelfristige Wetterprognosen.
Im neuen Jahr 2014 werden PreDocs, PostDocs, Forschungsstudenten und internationale Gäste die Arbeitsgruppe von Gruppenleiter Tilmann Gneiting verstärken.

UNCERTAINTY QUANTIFICATION IN COMPLEX SIMULATION MODELS USING ENSEMBLE COPULA COUPLING

Critical decisions may depend on output from complex computer simulation models. Weather and climate predictions are key examples of this. There is greatly increasing recognition of the need for uncertainty quantification in such settings. To offset the inherent uncertainty in weather forecasts, national and international meteorological centers have developed ensemble systems. The resulting ensemble forecasts are collections of numerical weather prediction (NWP) model runs, where the member runs differ from each other in terms of the two major sources of uncertainty: (a) the initial condition of the atmosphere and (b) the mathematical representation of the respective physical and chemical processes (see Figure 5).

Despite their undisputed successes, NWP ensemble systems are subject to systematic deficiencies such as biases and dispersion errors. It is therefore common practice to statistically postprocess the output of NWP ensemble forecasts with state-of-the-art techniques, like Bayesian model averaging and ensemble model output statistics.

However, these techniques are tailored to any individual weather variable - temperature, precipitation, wind speed, etc. - at single sites and individual look-ahead times. When applied to the full output of an NWP ensemble system comprising multiple weather variables on regional or even global scales and for multiple look-ahead times, these techniques fail to honor multivariate dependence structure and may result in a loss of physical coherence.

To address this challenge, which is ubiquitous in the postprocessing of output from complex simulation models, we have developed and investigated a general procedure called ensemble copula coupling (ECC) proceeding as follows [Schefzik et al. 2013]:

1. Generate a raw ensemble forecast consisting of multiple runs of the computer simulation model in which inputs or model specifications differ in suitable ways.

CST

Prof. Dr. Tilmann Gneiting

Group Leader
Prof. Dr. Tilmann Gneiting (from Nov. 2013)

Student
Patrick Schmidt (from Nov. 2013)

2. Apply statistical postprocessing techniques like Bayesian model averaging or ensemble model output statistics to correct for systematic errors in the ensemble forecast and to obtain a calibrated and well-focused predictive probability distribution for each univariate output variable individually.
3. Draw a sample from each postprocessed predictive distribution.
4. Rearrange the values sampled in the rank-order structure of the raw ensemble to obtain the ECC postprocessed ensemble.

Technically, the critical fourth and final step can be interpreted within the framework of copulas, which have significant roles to play in the description of multivariate statistical dependence structures. The ECC approach is broadly applicable and can be used whenever an ensemble of simulation runs is available, the ensemble is capable of realistically representing multivariate dependence structures, and training data are at hand. In the general setting of uncertainty quantification, ECC can serve to gage incomplete knowledge of current, past, or future

quantities of interest by means of joint probability distributions. These need to be as sharp or focused as possible, subject to them being calibrated, in the broad sense of reality being statistically ompatible with the predictive probability distributions.

USING DIVERGENCE FUNCTIONS TO EVALUATE CLIMATE MODELS

Arguably, environmental change on our planet is the most challenging problem faced by humankind. Mitigation strategies depend on our ability to project future climate on the basis of climate-model output. Accordingly it is crucially important to assess climate models. In this context it has been argued persuasively that the probability distribution (over time and/or space) of climate-model output needs to be compared to the corresponding empirical distributions of observed data (over time and/or space). Distance measures between probability distributions, also called divergence functions, can be used for this purpose. However, there are vast numbers of divergence functions, so which of them should be used? In a recent paper [Thorarinsdottir et al. 2013] we argue that divergence functions should be proper, in the technical sense that acting on a climate modelers' true beliefs is a performance-optimizing strategy. Among the divergences commonly used, the integrated quadratic distance and the Kullback-Leibler divergence turn out to be proper, and we recommend their use for practical climate-model evaluation.

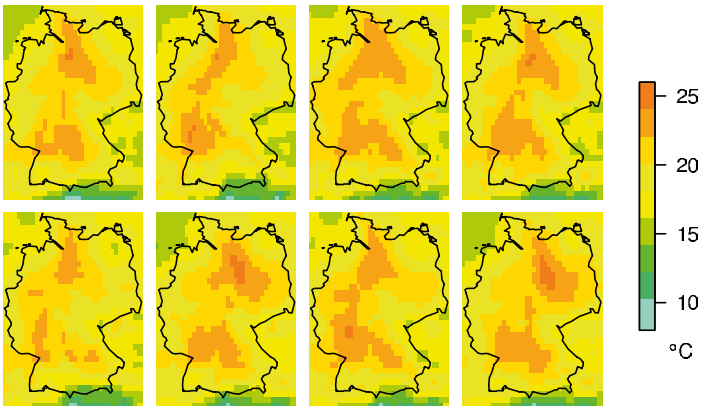


Fig. 5: Ensemble forecast of the surface temperature over Germany valid at 2:00 am on 4 April 2011. The panel shows eight members of the operational ensemble system run by the European Centre for Medium-Range Weather Forecasts (ECMWF).

2.4 Data Mining and Uncertainty Quantification (DMQ)

Advances in sensor technology and high-performance computing enable scientists to collect and generate extremely large data sets, usually measured in terabytes and petabytes. These data sets, obtained by means of observation, experiment, or numerical simulation, are not only very large but also highly complex in their structure. Exploring these data sets and discovering patterns and significant structures in them is a critical and highly challenging task that can only be addressed in an interdisciplinary framework combining mathematical modeling, numerical simulation and optimization, statistics, high-performance computing, and scientific visualization.

Besides the size and complexity of these data, quality is another crucial issue in guaranteeing reliable insights into the physical processes under consideration. The associated demands on the quality and reliability of experiments and numerical simulations necessitate the development of models and methods from mathematics and computer science that are able to quantify uncertainties for large amounts of data. Such uncertainties may derive, for example, from measurement errors, lack of knowledge about model parameters or inaccuracy in data processing. The Data Mining and Uncertainty Quantification group headed by Prof. Dr. Vincent Heuveline started work in May 2013. In this group we make use of stochastic mathematical models, high-performance computing, and hardware-aware computing to quantify the impact of uncertainties in large data sets and/or associated mathematical models and thus help to establish reliable insights in data mining. Currently, the fields of application are medical engineering, biology, and meteorology.

Fortschritte in Sensortechnik und High Performance Computing ermöglichen Wissenschaftlern das Erfassen und Erzeugen extrem großer Datenmengen, die vorwiegend in Tera- und Petabyte gemessen werden. Sie werden durch Beobachtung, Experimente und numerische Simulation generiert und sind nicht nur umfangreich, sondern weisen auch eine hochkomplexe Struktur auf. Die Erforschung dieser Datenmengen und das Entdecken von Mustern und signifikanten Strukturen ist eine entscheidende und in hohem Maße herausfordernde Aufgabe, der man nur in einem interdisziplinären Rahmen gerecht werden kann, d.h. durch Verbindung von mathematischer Modellierung, numerischer Simulation und Optimierung, Statistik, High Performance Computing und wissenschaftlicher Visualisierung.

Neben der Größe und Komplexität der Daten spielt deren Qualität eine entscheidende Rolle, um zuverlässige Einblicke in die betrachteten physikalischen Prozesse zu garantieren. Der damit verbundene Anspruch an Qualität und Zuverlässigkeit der Experimente und numerischen Simulationen erfordert die Entwicklung von Modellen und Methoden aus Mathematik und Informatik, die Unsicherheiten für große Datenmengen quantifizieren können. Solche Unsicherheiten können z.B. durch Messfehler, mangelnde Kenntnisse über Modellparameter oder Predictive Analytics von großen Datenmengen entstehen.

Die DMQ-Gruppe unter Prof. Dr. Vincent Heuveline nahm ihre Arbeit im Mai 2013 auf. Unsere Forschungsgruppe nutzt stochastische mathematische Modelle, High Performance Computing und Hardware-Aware Computing, um die Auswirkungen von Unsicherheiten bei großen Datensätzen und/oder zugehörigen mathematischen Modellen im Hinblick auf zuverlässige Einblicke in Data Mining zu quantifizieren. Derzeit sind Medizintechnik, Biologie und Meteorologie typische Einsatzfelder.

WHAT IS UNCERTAINTY QUANTIFICATION (UQ) AND WHY IS IT NECESSARY?

„Measure everything that is measurable and make everything measurable that is not measurable.“

Back in the 16th century, Galileo Galilei recognized the crucial role of mathematical models for representing uncertainties bedeviling a broad understanding of nature and its associated processes. This being the case, the relatively young research area Uncertainty Quantification has emerged. Its goal is to develop theoretical approaches and numerical methods for the reduction and quantification of uncertainties in complex systems, thus enabling validated numerical simulation of many problems arising in the natural and engineering sciences.

Usually, uncertainty is understood as an individual expression of incomplete knowledge or as an individual perception that future events are not fully determined. This poses a number of philosophical questions.

- Are uncertainties characteristics of the real world?
- Do they develop in human perception as a result of information deficiency?

Benjamin Constant argues that

„Uncertainty is an element in all human things. A person that would be free of all uncertainties would stop being a thinking character.“

Fig. 6: Galileo Galilei. (Source: Wikimedia Commons)

The DMQ group in 2013 (f.l.t.r.) Chen Song, Michael Bromberger, Vincent Heuveline, Samo Jordan, Maximilian Hoecker, Michael Schick

Group Leader
Prof. Dr. Vincent Heuveline

Staff Members
Maximilian Hoecker (from July 2013)
Dr. Michael Schick (from May 2013)
Dr. Samo Jordan (Aug. – Dec. 2013)

Scholarship Holders
Michael Bromberger (HITS Scholarship, from Nov. 2013)
Chen Song (HITS Scholarship, from Nov. 2013)

Especially in the engineering sciences, the term „uncertainty“ has been established as a synonym for randomness or stochastic. Accordingly, two classes of uncertainty are normally distinguished.

1) **Aleatoric uncertainties** is the term used for intrinsic variabilities arising from the stochastic character of model input parameters.
2) **Epistemic uncertainties** reflect incomplete knowledge about parameter values, which may be caused, for example, by incomplete measurements or specific model assumptions.

Uncertainties can be modeled in many different ways: convex sets, fuzzy-set approaches, probabilistic, and sto-

chastic concepts (e.g. Kolmogorov, Koopmans, Bayes). It is important to stress that the quantification of uncertainties requires additional data or assumptions. Quantifying and reducing uncertainties by numerical and statistical methods usually make for better prognoses, for example by determining more accurate confidence intervals.

NEW METHODS FOR UQ

Uncertainty Quantification has applications in many research areas in the natural and engineering sciences, for example in climate research, modeling of biological processes, reactive flows, computation of optimal power flows in networks, and numerical simulation of models in medical engineering.

However, complexity increases significantly when we compare purely deterministic approaches with their stochastic counterparts. In particular, the „curse of dimensionality“ (exponential growth in system size) causes headaches for many researchers. This makes the development of efficient numerical methods for uncertainty quantification a crucial and essential task in the closely related areas of applied mathematics, computer science, engineering, and natural science for both academia and industry.

The representation of uncertainties is often carried out by employing stochastic models included in some underlying physical model. Such a physical model accepts uncertainty as a random input, and many different solutions exist for computing the correspondingly uncertain model outputs. Popular methods are based on multiple evaluations (samples) of random model input and solving the corresponding deterministic systems, e.g. by using the Monte Carlo method. These results are later combined to obtain information about stochastic features in model output, such as expected values or variance. In this context, a so-called response surface or surrogate model is normally used to model the stochastic dependency of model output with respect to uncertainty. To this end, Polynomial Chaos has become a widespread model, interpolating the random dimension through orthogonal polynomials enabling us to obtain information about all possible realizations of random output without the need to re-compute it for all possible input realizations. Many numerical approaches exist for determining such Polynomial Chaos representations. These can be divided into two major classes: intrusive and non-intrusive. Intrusive

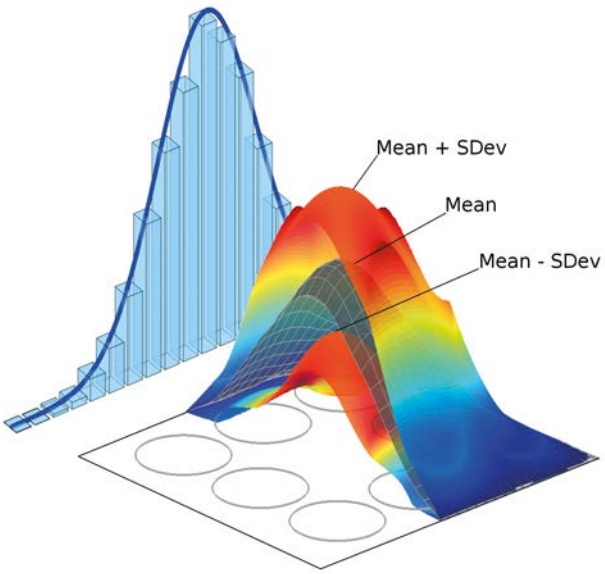


Fig. 7: Symbolic illustration of an example of Uncertainty Quantification. A Gaussian probability distribution is used to parameterize the uncertainty influencing the outcome of a solution (membrane) to a physical model. The uncertainty in model parameters is introduced within the plotted circles in the computational domain. The variation around the mean is plotted by adding and subtracting the standard deviation term.

methods require intensive re-programming of existing software solutions. This frequently needs to be done when employing so-called Galerkin projection methods. Non-intrusive methods allow for a re-use of existing software solutions that merely require a specific realization of the input values to compute deterministic sample solutions with respect to these values. Examples include the Monte Carlo method, cubature methods, and sparse-grid collocation approaches. Both classes, intrusive and non-intrusive, have their specific advantages and disadvantages. Accordingly, a decision about optimal methods will crucially hinge on the kind of application envisaged.

Many physical processes can be modeled by partial differential equations representing the physical model. The consideration of uncertainties results in stochastic counterpart equations, so-called stochastic partial differential equations (SPDE). Since discretization of an SPDE typically results in huge systems with many random variables, an efficient model for representing the stochastic variables is crucial. In this context, methods such as low-rank approximations, which arise in the area of tensor product representations, are a popular tool for model reduction. Karhunen-Loève decomposition and Norbert Wiener's idea of white noise analysis are two other significant examples based on singular value decomposition of covariance functions. From the computational viewpoint, a tensor product formulation helps significantly in reducing the amount of storage required, making it an elementary tool for the reduction of dimensionality in stochastic problems.

Especially for industrial applications it is important for the memory requirements of stochastic data and the numerical cost of using stochastic algorithms to remain as low as possible. One single deterministic numerical simulation, e.g. computing unsteady fluid flow, can create data in the range of hundreds of megabytes. These data form a huge data matrix that should never be stored explicitly. Instead, it seems to be more appropriate to approximate the data matrix by some low-rank approximation that can be re-used during numerical simulation and in post-processing steps.

If additional knowledge about the model output is available, for example by performing experiments or by using expert knowledge, the so-called Bayes' Update method can be used to correct the probability distributions of the random input. In this context, the famous theorem of Bayes plays a crucial role. It links the probability distribution of a prior hypothesis with its later distribution by taking into account the likelihood of the data observed or computed under the assumption that the hypothesis is true. This approach permits completion of an uncertainty quantification cycle, which uses stochastic models as a starting point for the computation of uncertainty propagation and an update correction scheme for adjusting model assumptions to additional information obtained from experiments or expert knowledge.

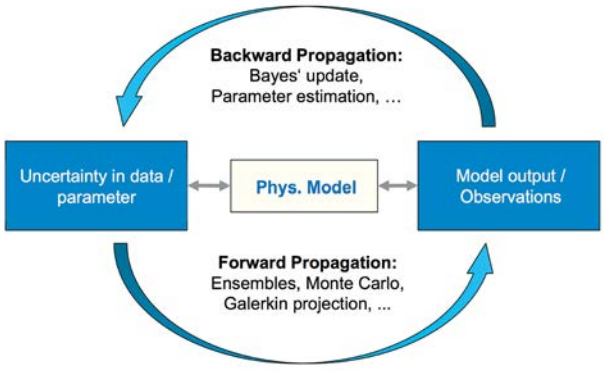


Fig. 8: Uncertainty Quantification: a simplified cycle.

DATA MINING: QUID?

Today's scientific and industrial applications produce and process unstructured multivariate data at considerable speed (Gbyte/sec). These data arise not only from experiments but also from computer simulations of complex phenomena and/or processes arising in fields such as medicine and meteorology. The data obtained are usually massive and complex. They are measured in terabytes and have both spatial and temporal components. As a result, a central issue in this context is related to the ability to explore, analyze, and understand the data obtained.

"Can one acquire valuable new knowledge from the big data collected?"

As a matter of fact, it has become impractical to manually explore and investigate such large data sets. Their size and complexity makes data mining a highly challenging task for many applications calling for innovative hardware and software solutions. A further issue is related to the intrinsic uncertainty associated with the sampled data. This characteristic can be a major impediment, whether the data are obtained by means of experimentation or computer simulation. The associated question, which may be crucial in fields of application like medicine, is:

"Can one derive reliable knowledge from collected data that may be uncertain? "

One of the main goals of the DMQ research group is to develop techniques and new methods enabling us to provide at least partial answers to this fundamental question in the context of complex applications such as medicine. Currently, we are concentrating on the development of adequate algorithms via exploitation of the existing compute power of supercomputers.

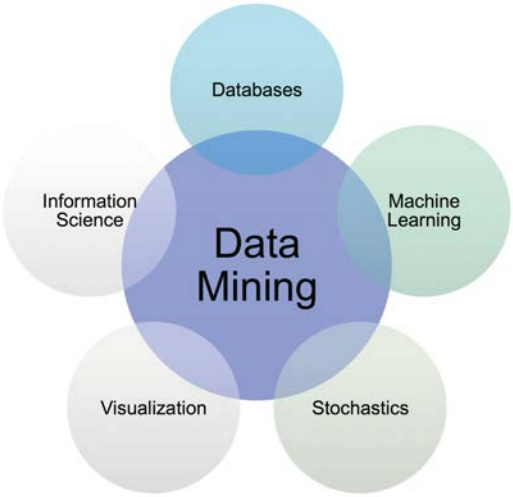


Fig. 9: Bordering fields of research in Data Mining.

The multi-step approach we are considering includes

- processing the raw data to identify objects of interest, assuming there is uncertainty in the data
- detecting patterns among the objects processed
- displaying those patterns for validation by the scientists.

Based on this data-mining process with the assumption of uncertainty in the data, the following scenarios are then considered:

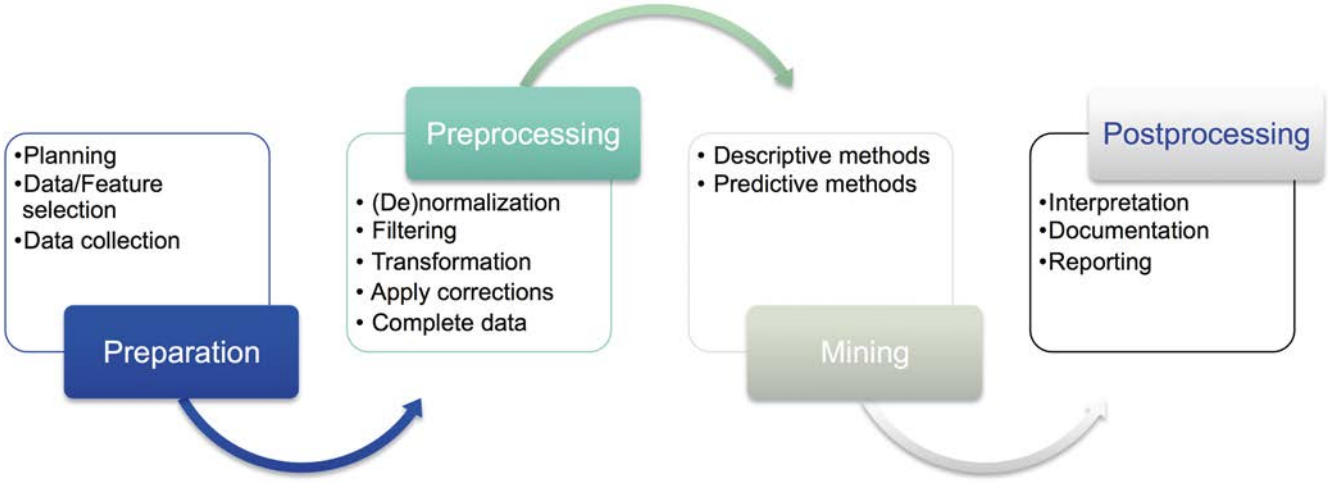
- prediction and forecast, including interval of confidence
- description of the phenomenology investigated (e.g. exploring the reasons)
- verification assuming stochastic effects (e.g. confirming hypothesis)
- exception detection and identification of unusual or exceptional events.

Currently, DMQ is investigating the broad lines of these different scenarios for application in medicine.

NEW PERSPECTIVES FOR COMPUTATIONAL FLUID DYNAMICS BASED ON UQ

At DMQ, Michael Schick is primarily working on the development of numerical methods for the simulation of fluid-flow problems involving uncertain parameters. Such uncertainties may arise, for example, when uncertain boundary conditions or system parameters such as forcing and viscosity are not precisely known. Here Polynomial Chaos is used to express the functional dependency of model outputs such as velocity and pressure variables on these parameters. In his main research, Michael Schick employs the Galerkin projection method to obtain high accuracies in computing the uncertain flow profiles.

Fig. 10: Generic full data-mining process.



However, the discretization of spatial and temporal computational domains along with the discretization of the stochastic components yields a system that exhibits up to billions of strongly coupled degrees of freedom. This makes it very important to build a bridge between the development of numerical methods and high-performance computing so that efficient use of parallel and heterogeneous computing platforms can be exploited to the fullest. This research field is fast-growing, and little is currently known about the optimal methods to employ.

DATA MINING AND UNCERTAINTY QUANTIFICATION IN MEDICAL ENGINEERING

DMQ is strongly involved in research activities associated with the idea of surgery guided by machine cognition, i.e. by a technical cognitive system that not only executes programmed tasks but also interprets a given situation and acts accordingly, supporting the surgeon in this way.

We analyze and develop numerical schemes used in and developed for the overall cognitive system. These are ap-

plied in modeling the behavior of soft tissue, including large deformations such as those induced by incision, and hand-in-hand development of an aortic silicon phantom and computer model. Aspects for consideration include the coupling of experimental data with data taken both from knowledge bases and from numerical simulations in order to yield patient-specific numerical simulations. Data mining techniques play a key role in this context. Also, the idea of inverse problem formulation leads to the identification of unknown material parameters and thus facilitates iterative improvement of forward-backward simulations. The visualization of the simulation results on stereoscopic displays supports surgeons in assessing when and how to perform an operation.

In analyzing risks and specifying clinical treatment scenarios on the basis of simulation results, we consider simulation quality using UQ methods to guarantee reliable information. Due to the medical application requirements (high accuracy and real-time simulations), our simulations are optimized with respect to the underlying hardware. Finally, prompted by the necessity of integrating our software into the clinical operation treatment workflow, we are working on a corresponding software interface in the context of the Medical Simulation Markup Language (MSML) project.

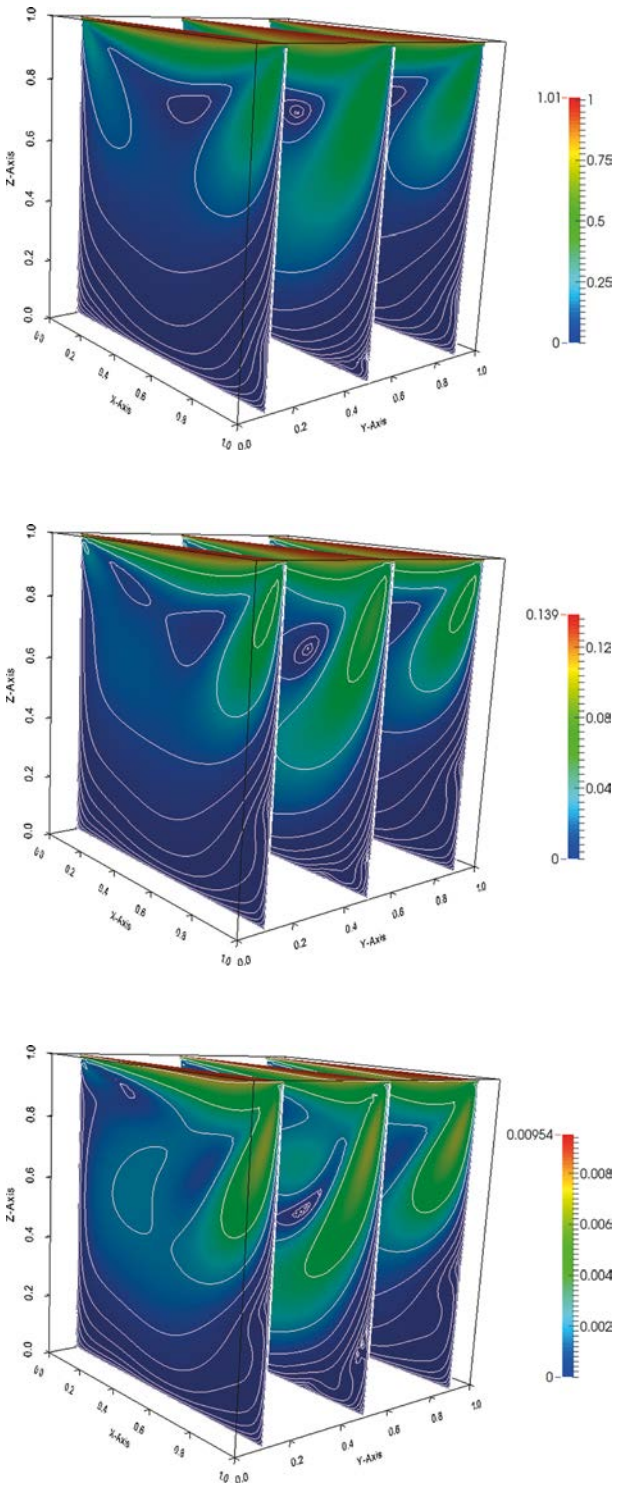


Fig. 11: Flow profile in a 3D domain (lid-driven cavity). Uncertain parameter: viscosity. Plot shows the expected value of the flow profile and its stochastic variations within the 3D domain. Top: Mean flow profile. Middle: Standard deviation. Bottom: Higher-order deviation.

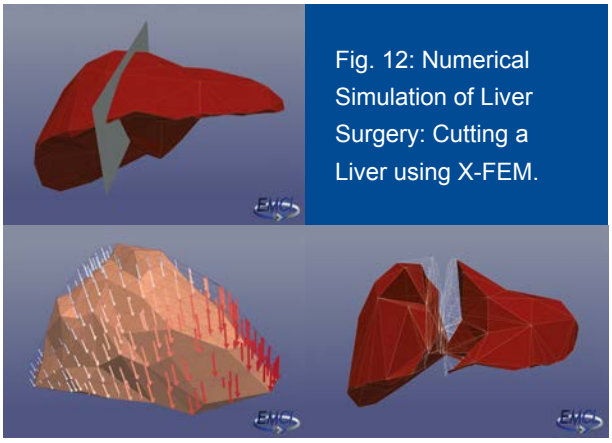


Fig. 12: Numerical Simulation of Liver Surgery: Cutting a Liver using X-FEM.

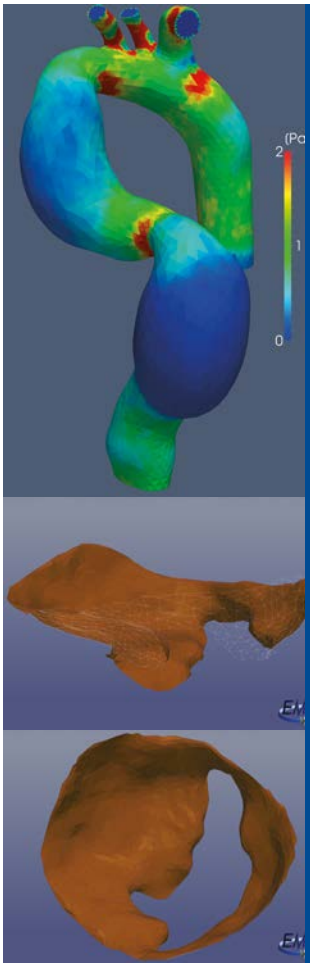


Fig. 13:
Top: Patient-specific wall shear stress analysis of an aneurysmatic aorta.
Before each widening of the vessel, the tissue is exposed to increased stresses. By contrast, hardly any forces act on the aortic wall within an aneurysma.

Middle & Bottom: Mitral valve opening and closing; simulation of annulus ring implantation surgery.

FUTURE OF DATA MINING AND UNCERTAINTY QUANTIFICATION

The solution of real-world problems with due consideration of stochastic uncertainties can only be achieved by steadily increasing computing power resources. Accordingly, high-performance computing will play a crucial role in the future development of numerical methods and algorithms both for data mining and uncertainty quantification. Efficient use of accelerating hardware such as graphic processing units (GPU) or specific co-processors is a sine qua non condition for such problem formulations. Technologies like these open up new and unprecedented vistas for the integration of stochastic effects into numerical simulations.

In his *Meditationes de Prima Philosophia,* René Descartes refers to uncertainty as his worst enemy. Minimizing uncertainty is a basic need for enlightened human beings. Seen thus, uncertainty quantification in the context of data mining provides an exciting bridge for the exchanges between mathematicians, computer scientists, engineers, meteorologists, surgeons, and philosophers.

2.5 Molecular Biomechanics (MBM)

For thousands of years, scientists have been grappling with the question of where we actually come from. Proteins are the most significant players in living organisms, and their structural evolution can give us indications of how life has changed throughout time at the molecular level. However, the further back one goes, the less information there is about life. At present, the sequence data of around a thousand genomes is available, but the information on the structures of proteins transcribed from those genomes is much scarcer. Also, proteins degrade quickly. Protein fossils do not exist, and evolutionary analysis will always be based exclusively on today's protein repertoire.

The major interest of the Molecular Biomechanics group is to decipher how proteins have been designed to specifically respond to mechanical forces in the cellular environment or as a biomaterial. We use Molecular Dynamics simulations, Force Distribution Analysis, Finite Element Analysis, and other computational techniques to study protein dynamics and mechanics on different length and time scales. More recently, we have combined phylogenomics and computational biophysics to take an evolutionary perspective on the mechanical function of protein. This involves mapping protein structures onto phylogenetic trees to reveal trends in protein folding or mechanical stability. The knowledge thus gleaned provides insights into the differences between present-day proteins and their ancestors a few billion years ago. Our aim is to provide a conclusive answer to the question of how mechanical forces influence living organisms at the level of individual molecules.

Seit Jahrtausenden haben Wissenschaftler versucht, die Frage nach unserem Ursprung zu beantworten. Proteine sind die Hauptakteure lebender Organismen. Ihre strukturelle Entwicklung kann Aufschluss darüber geben, wie sich das Leben im Laufe der Zeit auf molekularer Ebene verändert hat. Je weiter man jedoch in der Zeit zurückreist, desto weniger Informationen gibt es über das Leben. Heutzutage stehen Sequenzdaten von rund tausend Genomen zur Verfügung. Informationen über die Proteinstrukturen, die sich aus diesen Genomen ergeben, sind jedoch deutlich seltener. Dazu kommt die schnelle Abbauzeit von Proteinen, so dass keine Proteinfossilien existieren und Evolutionsanalysen lediglich auf dem heutigen Proteinrepertoire basieren.

Das Hauptinteresse der Molecular Biomechanics-Gruppe ist es zu entschlüsseln, wie Proteine aufgebaut sind, um in der zellulären Umgebung oder als Biomaterial gezielt auf mechanische Kräfte reagieren zu können. Wir nutzen dafür Molekulardynamiksimulationen, Kraftverteilungsanalyse, Finite-Elemente-Analyse und andere Rechenverfahren, um Proteindynamik und –mechanik in unterschiedlichen Längen- und Zeitskalen zu untersuchen. Wir haben kürzlich Phylogenomik und computergestützte Biophysik miteinander kombiniert, um die mechanische Proteinfunktion aus einer evolutionären Perspektive aus zu betrachten.

Dafür haben wir Proteinstrukturen auf Stammbäumen abgebildet, um Trends bei Proteinfaltung oder mechanischer Stabilität zu bestimmen. Unsere Erkenntnisse erlauben Einblicke in die Unterschiede zwischen heutigen Proteinen und deren Vorfahren vor ein paar Milliarden Jahren. Wir wollen herausfinden, wie mechanische Kräfte auf der Ebene einzelner Moleküle lebende Organismen beeinflussen.

EVOLUTION OF PROTEIN STRUCTURE

The current catalog of protein structures exhibits a fascinating diversity of functions and shapes. This variety in shape and function is the result of billions of years of evolution. Evolution is known to have selected protein structures, but the mechanisms that govern these selections have yet to be determined. To answer questions like these, we have been exploring physical factors that may have had an impact on the selection of protein structures. In the present project we have been studying one important aspect of proteins: folding. Proteins are folded into their native state after or during their formation, but they can also fold and unfold as they function. Errors in this mechanism can lead to misfolding. Misfolded proteins may form aggregates that can potentially affect cell integrity and lead to diseases like Alzheimer's or Creutzfeld-Jakob. The project aims to understand how proteins have evolved into well-folded structures in order to protect the cell from harmful aggregation.

In this endeavor we have been working with Prof. Gustavo Caetano-Anolles at the University of Illinois at Urbana-Champaign, who has developed a method for exploring the evolution of protein structures. The appearance of domains obeys a molecular clock, and this was used to determine the appearance of proteins, the dynamics of domain organization in proteins, and the first appearance of free oxygen on earth. In the next stage, we evaluated protein foldability using a computational topology-based technique.

Finally we mapped foldability onto the evolutionary timeline. Here we observed the following trend: Through most of protein evolution, folding speed increased (Figure 14) from Archean to multicellular organisms. However, ~1.5 billion years ago, more complex structures emerged and caused a biological 'big bang.' This led to the development of slower-folding protein structures. Our hypothesis is that among those complex structures new mechanisms emerged to assist folding.

The MBM group in 2013 (f.l.t.r.): Eduardo Cruz-Chu, Camilo Aponte-Santamaria, Maxime Louet, Johannes Wagner, Frauke Gräter, Agnieszka Bronowska, Beifei Zhou, Davide Mercadante, Jing Zhou.

Group Leader
Dr. Frauke Gräter

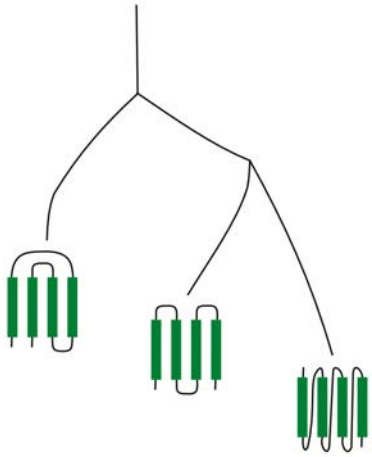
Staff Members
Dr. Camilo Aponte-Santamaria
Ilona Baldus (until April 2013)
Sandeep Patil
Dr. Maxime Louet (from Feb. 2013)
Dr. Eduardo Cruz-Chú (from March 2013)
Dr. Ion Bogdan Costescu
(until July 2013, now ITS group leader)

Scholarship Holders
Dr. Cedric Debes
Dr. Ulf Hensen
Dr. Davide Mercadante (from March 2013)
Christian Seifert (until May 2013)
Johannes Wagner
Jing Zhou

Visiting Scientists
Dr. Agnieszka Bronowska
Fizza Mughal (Dec. 2013)
Dr. Shijun Xiao (until May 2013)
Katra Kolšek (from Sept. 2013)
Dr. Richard Henchman (from Jan. until Sept. 2013)
Beifei Zhou

Student
Sang Paik

Fig.14: Protein topologies favoring short-range inter-aminoacid contacts may be the result of evolutionary foldability optimization and would thus probably have made a late appearance in evolution.



CLASSIFICATION OF PROTEIN DOMAINS

For our analysis, we considered 92,000 proteins and 989 genomes, a task that can only be tackled with computational methods. The group headed by Prof. Gustavo Caetano-Anolles at the Evolutionary Bioinformatics Laboratory at Urbana-Champaign had originally estimated the age of most known structural domains from the Structural Classification Database (SCOP). For the purposes of the study, Dr. Minglei Wang and his lab staff applied an algorithm to the reconstruction of evolutionary protein structure history based on the abundance of these structures in genomes. Subsequently, we applied a mathematical model to predict the folding rate of proteins. Individual folding behavior differed in speed, ranging from nanoseconds to minutes. No experimental methodologies would be able to capture the folding time-scales for so many proteins. Only computational methods enable us to work on such high numbers of protein structures. In the study we evaluated the folding speed of single proteins using structures previously determined in experiments. To fold a protein, a number of molecular interactions have to take place that are determined by contacts between amino-acids in the polypeptide chain. If these contacts are far apart from each other, folding takes longer than if they lie close to each other. With the so-called Size-Modified Contact Order (SMCO) it is possible to predict how fast these contacts will take place and thus how fast the protein will fold, regardless of length. As we go back in time, the foldability of protein decreases, suggesting that over time nature has improved protein folding.

Our analysis revealed that the folding of proteins was gradually optimized from archaea to multicellular organisms between 3.8 and 1.5 billion years ago. The amino-acid chains that proteins are made up of have also become shorter in the course of evolution. This was another factor contributing to the increase in folding speed, as has been shown in the study. When eukaryotes - i.e. organisms with a cell nucleus - emerged, protein folding became somewhat less crucial. To satisfy the need for fast-folding proteins, nature has developed a complex machinery for the prevention and repair of misfolded proteins. The so-called chaperones are an example of this. It seems that nature is prepared to accept a certain degree of disorder if this helps develop structures that could not have evolved otherwise. According to our findings, this happened around 1.5 billion years ago. At that time, there was a reverse trend in both the foldability and the length of the domains. But above all, there occurred an almost explosive increase in the number of domain architectures and rearrangements in multi-domain proteins, which was mainly triggered by increased rates of domain fusion and fission. While one-domain proteins had previously been dominant, multi-domain proteins now took over. Prof. Gustavo Caetano-Anollés therefore refers to the period 1.5 billion years ago as the ‘big bang’ in protein evolution.

In future analyses of protein evolution it may be possible to extend our purview to related questions, namely whether proteins have become more stable or more flexible in the billion-year history of their evolution.

A CRASH TEST FOR CHAPERONE PROTEINS

Proteins are complex molecular machines performing their functions under tight regulation. An intriguing example is the chaperone protein Hsp90. Chaperones are helper proteins – they assist other proteins in finding the most stable structure natural to them. As such, they are essential for the very survival of most cells. This is especially true of cancer cells, the undisputed world champions in producing proteins. They have to cope with highly crowded cell interiors where the folding of a protein into its native functional state is extraordinarily challenging. Chaperones like Hsp90 are thus of primary importance for the survival and proliferation of cancer cells and accordingly for cancer growth. This is why these proteins are primary targets for anti-cancer drugs.

Designing strategies to block Hsp90 function in cancer cells would be greatly facilitated if we could decipher the way this fascinating molecular machine actually functions. Experiments have revealed that Hsp90 undergoes a transition between an open and a closed state. This is an essential feature of its functioning and regulation. Opening and closure transitions are controlled by the binding of a small nucleotide molecule, ATP, which is then transformed into ADP and finally released again.

But how is the binding of these small molecules in one very restricted protein region coupled to the large-scale dynamic transitions undergone by the protein from an open to a closed state and vice versa? How can the protein 'sense' the binding of the molecule and communicate this information to distant regions of the protein scaffold?

Might it be possible to reveal the distribution of stress in the protein by perturbing it with a small molecule, much as we are able to analyze the stress field in a crash test?

Experiments on Hsp90 to date have not been able to reveal dynamic information from the protein at high time resolution. The most advanced experiments can however provide very detailed information on spatial resolution (fluorescence resonance energy transfer or hydrogen/deuterium exchange.

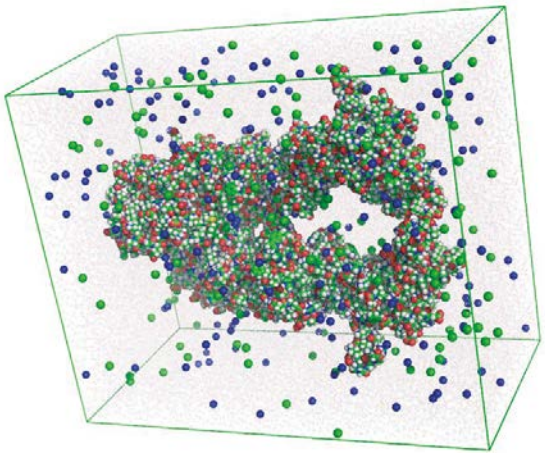


Fig. 15: Graphic representation of the simulation system. The protein HtpG (red/white, balls) in the middle of a triclinic box filled with water (opaque red, lines), Sodium (blue, balls) and chlorine (green, balls) ions representing the physiological salt.

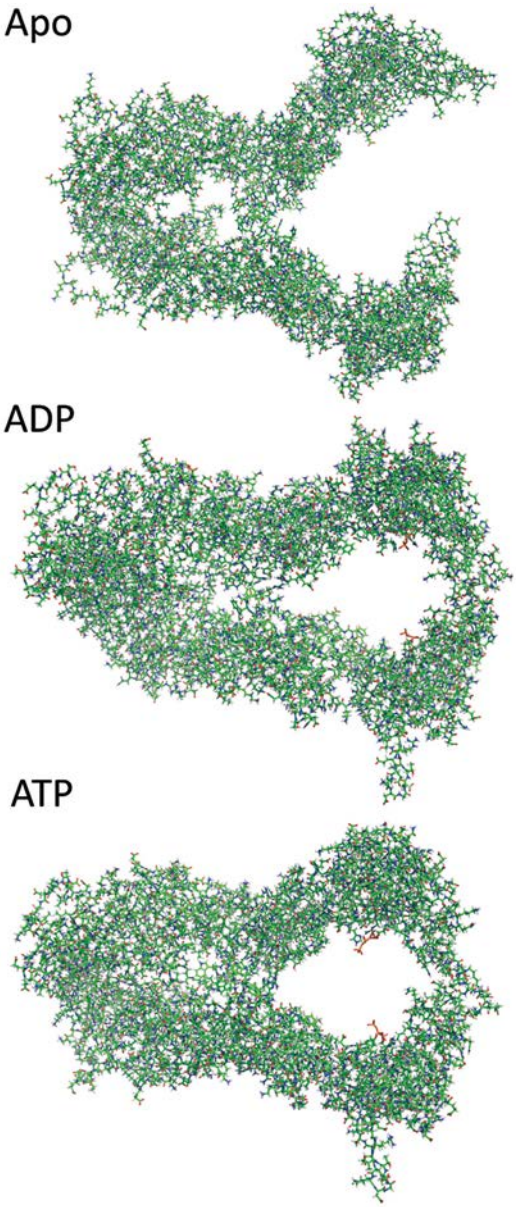


Fig. 16: Line representation of HtpG after a molecular dynamics simulation of 20 ns, the ADP bound state (middle) and the ATP bound state (bottom) as a reference [3].

VALIDATION BY EXPERIMENT: OPENING AND CLOSURE IN HSP90

These complex calculations are based on a whole range of parameters and assumptions. To ensure that the results of these calculations are useful, the basic findings they produce have been compared to experimental results that were not used to build the simulated model in the first place. To this end, we have compared the opening and closing of the protein in our simulation with data from fluorescence resonance energy transfer.]. The results of the simulations are in good agreement with the experimental results.

Sophisticated experiments suggest that these motions happen on timescales of seconds, which however presents an upper limit due to the limited time resolution at hand.. Our dynamic trajectories at femtosecond resolution, however, showed that motion in parts of the protein can be much faster than expected (Figure 16). More specifically, we have shown that in Hsp90 the large domain movement required for activation of the protein can happen on a timescale of nanoseconds.

GOING BEYOND EXPERIMENTS: HSP90 IN THE CRASH TEST

How is this large opening and closing motion involved in activating the protein linked to the binding of the small ATP molecule? To answer this question, the results of the simulation have been analyzed using a method invented in our group: force distribution analysis. As in a crash test, this method can reveal the internal molecular stresses obtaining within the protein scaffold. Obviously, we did not calculate the stress distribution in Hsp90 in terms of the high forces released by a car hitting a wall but in terms of the forces involved when the small ligand ATP is bound to the protein. We were able to demonstrate that the effect

of the ligand is not merely restricted to the vicinity of the binding niche. Instead, force is very selectively channeled through a large helical structure in the protein from the ATP binding area to a part of the functional section of the protein that is responsible for the opening and closing of the protein (Figure 17). In this way, our high-resolution technique based on high-performance computing was able to detect the network of forces responsible for signal transmission in a protein on an atomistic scale. On this basis, we were able to propose biochemical mutations for the protein to validate the model we had suggested for the protein mechanism.

ONGOING RESEARCH/OUTLOOK

Our results in the intriguing case of molecular chaperones demonstrate that by means of novel force distribution analysis it is possible to decipher communication pathways in complex molecules like proteins that help re-engineer functional mechanisms in protein molecular machines. Applications of this approach to other proteins and protein complexes is promising, given that the computational power available will steadily increase in future, a prerequisite for the investigation of the convergence of molecular stresses like the ones discussed here.

But how is the binding of these small molecules in one very restricted protein region coupled to large-scale dynamic transitions of the protein from an open to a closed state and vice versa? How can the protein ‘sense’ the binding of the molecule and communicate this information to distant regions of the protein scaffold? Could one perhaps reveal the distribution of stress in the protein after perturbing it by a small molecule, much in the same way as the stress field is analyzed during a crash test?

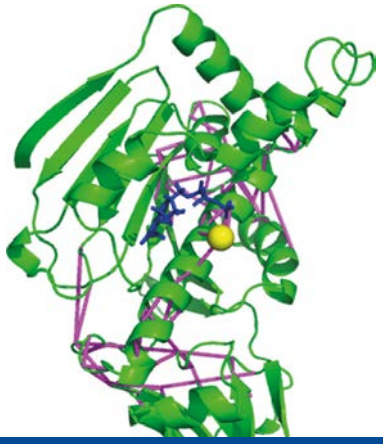


Fig. 17: Graphic representation of the result of a force distribution analysis. The two domains (upper and lower parts of the Figure) of the protein (green) communicate along a pathway (magenta). The communication interface between both domains is shown in a stick representation.

KISSPEPTIN RECEPTOR GPR54 AS A TARGET FOR CANCER RESEARCH

Kisspeptins are are C-terminally amidated peptides that are expression products of the KiSS-1 gene. They have been identified as binders of the G protein-coupled receptor GPR54.

Kisspeptins play a fundamental role in the sexual differentiation of the brain, the metabolic regulation of fertility essential for human puberty, and the maintenance of adult reproduction. Recently, kisspeptins and their cognate receptor, GPR54, have emerged as indispensable factors for cancer metastasis. This makes GPR54 an important drug target. However, the molecular details of kisspeptin-mediated activation of GPR54 remain elusive. This situation is further complicated by the fact that the atomistic structure of GPR54 is still unknown.

With our co-workers from the group headed by Prof. M. Auer (University of Edinburgh) we combined several molecular modeling techniques, such as homology modeling, molecular dynamics (MD) simulations, molecular mechanics (MM), molecular docking calculations, synthetic chemistry, and several experimental biology techniques in order to obtain an atomistic 3D model of GPR54 (Figure 18), deduce the binding site, predict the key features of kisspeptin-GPR54 interactions, and synthesize and test ligands predicted as GPR54 binders to validate the model. This approach would enable us to design new drug-like compounds binding GPR54 with high affinity.

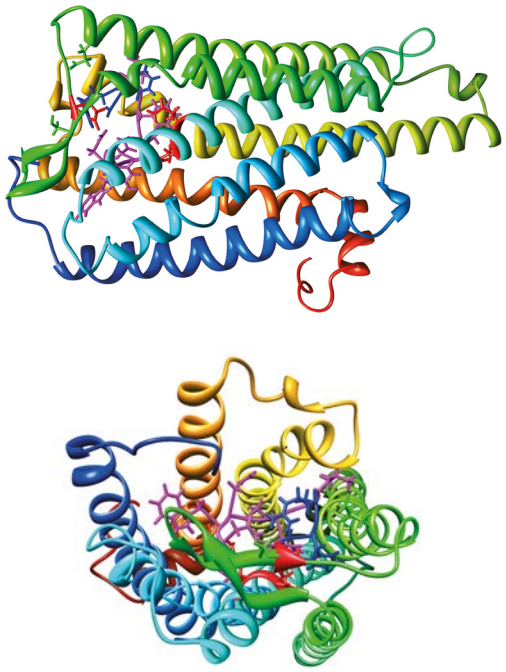


Fig. 18: The atomistic model of human GPR54 developed with kisspeptin KP10 docked at the binding site. (Top) the view along the transmembrane helices, (bottom) the view from the top of the membrane (extracellular site).

We have tested the ability of the sequential docking procedure refined by MD simulations to reproduce experimental binding energies of 19 different ligand-GPR54 complexes for which the structure was previously unknown. More specifically, we looked at the factors influencing binding energetics, including the active conformation of the peptide at the binding site, the number of favorable ligand-receptor contacts, the dynamics of the unbound ligand, and the loss of the ligand's conformational entropy upon binding. In the absence of experimental data on the structure of the receptor and peptide ligands, and with very limited and indirect information about the binding site (Figure 18), we concluded that the procedure was able to select the native-like complexes based on their binding energetics. The success of this procedure represents an important finding for molecular pharmacology and rational drug discovery.

The most important requirement of the molecular docking calculation is its ability to distinguish the real binding modes from 'decoys' (nonspecific or energetically unfa-vorable binding modes). In the absence of structural data for GPR54 and its ligands, the binding energies were used to validate the procedure. In all cases tested, the sequential docking procedure resulted in the correct estimation of the binding strength of the peptides tested and distinguished binders from non-binders. This boosts confidence in the methodology applied and paves the way for efficient design of new ligands for the GPR54 receptor with major therapeutic potential.

PEPTOIDS FOR ALLOSTERIC INHIBITION OF THYMIDYLATE SYNTHASE

Thymidylate synthase is an enzyme upregulated in many types of cancer. Conventional thymidylate synthase inhibitors target its active site, competing with either the co-factor (e.g. methotrexate) or substrate (e.g. 5-fluorouracil)

binding. These compounds are commonly used in cancer chemotherapy. However, many clinically challenging cancers, such as ovarian cancer and cancer of the colon, are resistant to these conventional thymidylate synthase inhibitors. Accordingly, the search is on for novel strategies enabling us to develop thymidylate synthase inhibitors with an alternative inhibition mechanism.

Human thymidylate synthase performs its biological function as a homodimer, with each monomer adopting active or inactive conformation. The dimerization is a prerequisite for the catalytic activity. This makes the dimerization interface an attractive target for drug development.

In 2011, Cardinale et al. (Cardinale D, Guaitoli G, Tondi D, et al., Proc Natl Acad Sci U S A. 2011 Aug 23;108(34):E542-9.) published a study in which a group of peptides binding to the dimerization interface of human thymidylate synthase were developed. The binding mechanism was studied by a combination of experimental and computational techniques, including X-ray crystallography, molecular dynamics (MD) simulations, molecular docking, circular dichroism and fluorescence spectroscopy, kinetic analysis, and isothermal titration calorimetry (ITC). More recently, the structures of three other compounds (one peptide and two organic molecules) targeting the dimerization interface of thymidylate synthase were also published.

In collaboration with the group headed by Prof. Rode (Necki Institute of Experimental Biology, Warsaw, Poland), we used these data (PDB codes 3N5E, 4FGT, 4G6W, and 4G2O) as a guide for the rational design of peptoid scaffolds (Figure 19). The advantages of peptoids over peptides lie in synthetic feasibility, resistance to protease digestion, and resistance to unfolding/solvent effects. Another goal we pursued was to miniaturize the scaffold, thus increasing bioavailability in the prospective inhibitors. We used a combined computational approach consisting of virtual screening, molecular docking, quantum chemical (QM) calculations, and molecular dynamics (MD) simulations.

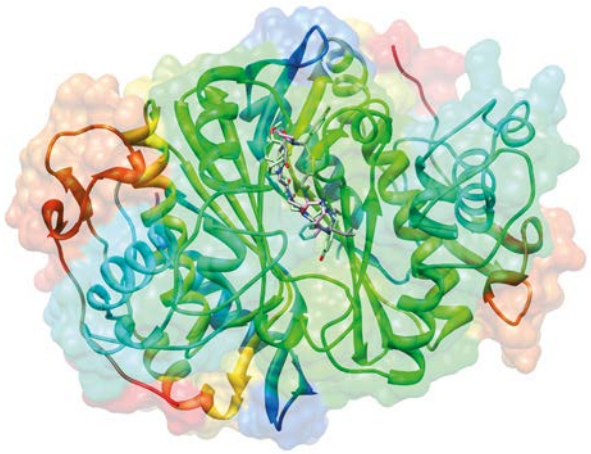


Fig. 19: The dimer of human thymidylate synthase with peptoid inhibitors docked at the dimerization interface.

The resulting peptoid inhibitors (Figure 19) were around half the size of the peptide we used as a guide. They are smaller than 500 daltons (Da), thus potentially fulfilling the classic Lipinski “rule of five,” In silico virtual screening, followed by MD simulations and MM-PBSA, indicated that their binding affinity was higher than that of the guide compound. This stage in our work is soon to be followed by compound synthesis and experimental testing of the binding to human thymidylate synthase.

2.6 Molecular and Cellular Modeling (MCM)

In the MCM group we are primarily interested in understanding how biomolecules interact. What determines the specificity and selectivity of a drug-receptor interaction? How can proteins assemble to form a complex, and what shape can the complex take? How is the assembly of a complex influenced by the crowded environment of a cell? What makes some binding processes quick and others slow? How do the motions of proteins affect their binding properties?
These questions are illustrative of the types of problem we address in our projects via the development and application of computational approaches to the study of biomolecular structure, dynamics, interactions, and reactions. We take an interdisciplinary approach, entailing collaboration with experimentalists and concerted use of computational approaches based on physics and bio-/chemo-informatics. The broad spectrum of techniques employed ranges from interactive, web-based visualization tools to atomicdetail molecular simulations.

This report describes the results achieved this year in three selected projects. They demonstrate the types of method we develop to model macromolecular interactions and their application to problems in biology, biotechnology, and drug design. The projects center on (i) simulating how a protein binds to the ribosome, (ii) the specificity of allosteric regulation of metabolic enzymes in lactic acid bacteria, and (iii) computational methodology for detecting transient pockets in proteins for drug design.

Die MCM-Gruppe ist primär daran interessiert, die Wechselwirkungen zwischen Biomolekülen zu verstehen. Was bestimmt die spezifische und selektive Wirkung beim Zusammenspiel von Wirkstoff und Rezeptor? Wie werden Proteinkomplexe gebildet und welche Formen können sie annehmen? Welche Wirkung hat die beengte Zellumgebung auf die Bildung eines Proteinkomplexes? Warum verlaufen einige Bindungsprozesse schnell und andere langsam? Welche Auswirkungen haben Proteinbewegungen auf ihre Bindungseigenschaften? Diese Fragen versuchen wir in unseren Projekten durch die Entwicklung und Anwendung rechnerischer Methoden zur Untersuchung biomolekularer Strukturen, Dynamik, Wechselwirkungen und Verhaltensweisen zu beantworten. In enger Zusammenarbeit mit Experimentatoren verwenden wir in interdisziplinären Ansätzen rechnerische Methoden aus den Bereichen der Physik-, Bio- und Chemoinformatik. Das Spektrum unserer Methoden reicht dabei von web-basierten Visualisierungswerkzeugen bis hin zu Molekularsimulationen auf atomarer Ebene.
Die Ergebnisse unserer diesjährigen Arbeit präsentieren wir in drei ausgewählten Projekten. Sie demonstrieren einerseits die Methoden, die wir entwickeln, um makromolekulare Interaktionen zu modellieren, und andererseits ihre Anwendungen in Biologie, Biotechnologie und Medikamentenforschung. Die Projekte beschäftigen sich mit (i) der Simulation eines Proteins, das sich an ein Ribosom bindet, (ii) der Spezifität der allosterischen Regulation von Stoffwechselenzymen in Milchsäurebakterien und (iii) Berechnungsmethoden zum Entdecken transienter Bindungstaschen in Proteinen für die Arzneimittelherstellung.

GENERAL NEWS SECTION

2013 saw a number of arrivals and departures among our group members. Priyanka Banerjee successfully completed her project on peptidomimetic design, obtaining a Master's degree in Life Science Informatics from the University of Bonn and has gone on to doctoral studies in Berlin. Prajwal Nandekar joined us from the National Institute of Pharmaceutical Education and Research (NIPER), India, as a doctoral student with a DAAD sandwich scholarship. Dr. Julia Romanowska obtained an EMBO postdoctoral scholarship to do research in the group. Prof. Zaheer-ul-Haq Qasmi from the University of Karachi, Pakistan, came for a six-month visit on a Georg Forster Fellowship from the Alexander von Humboldt Foundation.

Two major projects reached completion this year. Some of the results of SysMOLAB2, an international project on the comparative systems biology of lactic acid bacteria, are outlined in the second project described below. The third section reports on work from our project in the Biotech Cluster Rhein-Neckar (BioRN) to develop a computational methodology for the identification of transient binding pockets on proteins for structure-based drug design.

The MCM group participates in the Human Brain Project, an FET Flagship Project of the European Commission that started in October, The Human Brain Project involves over 130 institutions and is planned to run for ten years. Our role in the initial ramp-up phase is to contribute to the molecular simulation component of the "brain simulation platform."

We have made three software releases this year. TRAPP is a new toolbox for the identification and analysis of TRAnsient Pockets in Proteins and is described below. SDA7 Is a completely restructured version of our Simulation of Diffusional Association software. It is parallelized and allows for the simulation of bigger systems than was previously possible. LigDig is a new web application for investigating protein-ligand interactions. It is designed, in particular, to assist non-experts in bio- and chemo-

The MCM group in 2013 (f.l.t.r.): Xiaofeng Yu, Julia Romanowska, Daria Kokh, Musa Özboyaci, Michael Martinez, Mehmet Öztürk, Rebecca Wade, Mykhaylo Berynskyy, Stefan Henrich, Stefan Richter, Ghulam Mustafa, Zaheer-ul-Haq Qasmi, Prajwal Nandekar, Antonia Stank

Group Leader
Prof. Dr. Rebecca Wade

Staff members
Dr. Jonathan Fuller
Dr. Stefan Henrich
Dr. Daria Kokh
Dr. Michael Martinez
Dr. Stefan Richter
Antonia Stank (from April 2013)

HITS predoctoral scholars
Mykhaylo Berynskyy
Musa Özboyaci
Mehmet Öztürk
Xiaofeng Yu

Visiting scientists
Prof. Zaheer-ul-Haq Qasmi (July-Dec. 2013)
Ghulam Mustafa
Prajwal Nandekar (from Oct. 2013)
Dr. Julia Romanowska

Students
Priyanka Banerjee (until March 2013)
Eduard Bopp

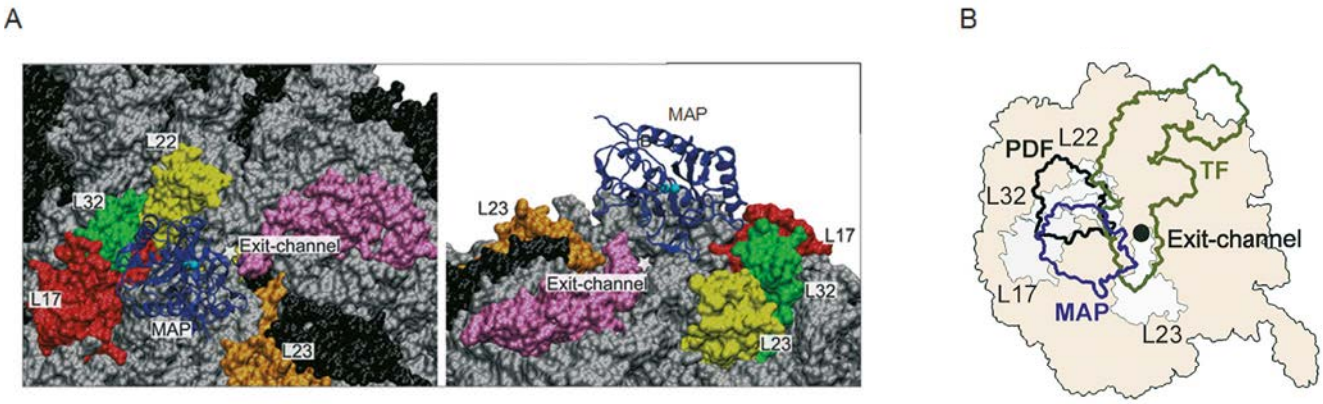
Interns
Laura Armbruster (July-Sept. 2013)
Elena Sizikova (June-Aug. 2013)

informatics with ligand-centric queries about binding properties, e.g. in the context of investigating the effect of a particular metabolite on a biochemical network.
A highlight this year was the third Biological Diffusion and Brownian Dynamics Brainstorm, BDBDB3, which took place on October 7-9 in the Studio, Villa Bosch (see Chapter 5.1.4.)

DYNAMIC ENZYME DOCKING TO THE RIBOSOME COORDINATES N-TERMINAL PROCESSING WITH POLYPEPTIDE FOLDING

Newly synthesized polypeptides go through various co-translational maturation stages, including N-terminal enzymatic processing, chaperone-assisted folding, and membrane targeting. The spatial and temporal coordination of these stages is however unclear. Bernd Bukau, Günter Kramer, and colleagues at ZMBH (Heidelberg University) and the German Cancer Research Center (DKFZ) have studied experimentally how two N-terminal processing enzymes, peptide deformylase (PDF) and methionine aminopeptidase (MAP), associate with ribosomes and process the nascent polypeptide chain. MAP competes with PDF, the first enzyme to act on nascent chains, in binding sites at the ribosomal tunnel exit. While the binding site of PDF on the ribosome has been determined experimentally, it has not been possible to obtain a structure for the MAP-ribosome complex by experimentation. Accordingly, we performed a computational docking of MAP to the ribosome using our SDA software. MAP binds to the ribosome through a positively charged loop that is crucial for nascent-chain processing and cell viability. The positive patch on MAP can make favo¬rable interactions with the negatively charged rRNA, so there are many energetically favourable positions at which MAP

Fig.20: Model of the MAP-ribosome complex obtained by computational docking. MAP is shown in cartoon representation (dark blue, cobalt ions in cyan), rRNA is in light gray, some of the ribosomal proteins are shown in color, and the exit tunnel is indicated by a star. Top and side views are shown on the left. The footprints of PDF (black), MAP (blue), and trigger factor (TF, dark green) on the ribosome are shown on the right. These were determined from crystallographic data (PDF, TF) and computational docking (MAP). [Sandikci, 2013].



can contact the ribosome. Docking was therefore performed subject to constraints from cross-linking and mutagenesis experiments. We identified three docking modes consistent with the experimental data that showed MAP docking with its active site oriented toward the ribosome tunnel exit in a position overlapping with the PDF binding site. The most energetically favorable orientation is shown in Figure 20. Both PDF and MAP display fast ribosome association kinetics, which appear to be crucial for ensuring the sequential processing of the nascent chains after their emergence from the ribosome and for the temporal separation of nascent-chain processing from later maturation events, including chaperone recruitment and folding [Sandikci, 2013].

ORGANISM-ADAPTED SPECIFICITY OF THE ALLOSTERIC REGULATION OF CENTRAL METABOLIC ENZYMES IN FOUR LACTIC ACID BACTERIA

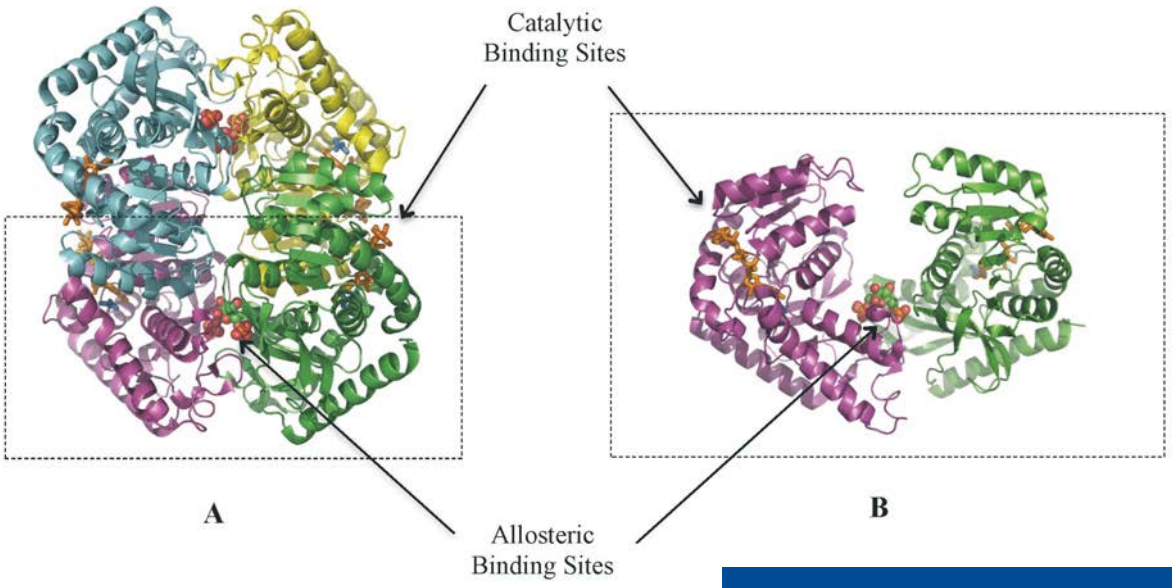
Some lactic acid bacteria (LAB) are antibiotic-resistant pathogens causing severe disease. Others are healthy probiotics used in the food industry. What makes an LAB a friend or a foe, and how do they adapt to survive in such different environments? We have addressed this problem by focusing on the enzymes lactate dehydrogenase and pyruvate kinase, which both play a central role in the metabolism of lactic acid bacteria. These enzymes need to react quickly to changes in the environment, so their activity is strictly regulated. We used computational techniques to predict the cellular substances, called allosteric modulators, that are responsible for quick and effective activation or inhibition of these enzymes and to explore the effects of environmental conditions on allosteric regulation. We modeled the three-dimensional structures of the enzymes from four different bacteria and computed interactions with known and putative modulators using a range of computational techniques including our PIPSA approach.

For lactate dehydrogenase, we found that the activating effects of allosteric compounds could be understood from analysis of the electrostatic properties in the allosteric binding site. The effects of phosphate (Pi), which can have both activating and inhibitory effects, were related to the relative binding energies in the catalytic and allosteric binding sites, Overall, the results of our calculations and the experiments performed by Bernd Kriekemeyer, Tomas Fiedler, and colleagues in Rostock and Hans Westerhoff and Hanan Messiha in Manchester revealed a subtle interplay between the effects of Pi, fructose-1,6-bisphosphate, and pH that results in different regulatory effects on the lactate dehydrogenases of different LABs [Feldman-Salit et al., 2013], see Figure 21.

In the case of pyruvate kinase, we also used enzyme kinetic modeling (in cooperation with Ursula Kummer (Bioquant, Heidelberg University)) to simulate the dynamic behavior of pyruvate kinase activity We predicted activators by considering the cellular concentrations of metabolites in the different organisms, along with the binding properties of the pyruvate kinases from the different lactic acid bacteria. We found that different lactic acid bacteria have different preferences for activators and that the level of activator specificity is related to the environment in which the bacteria live. [Veith, 2013], see Figure 22.

Both these studies show how enzymes with high sequence similarity can display subtle but significant differences in allosteric regulation in different organisms that need to function in different environments.

This work was carried out as part of the SYSMO-LAB2 project in the Systems Biology of Microorganisms (SysMO) Network with support from the Federal Ministry of Education and Research (BMBF).



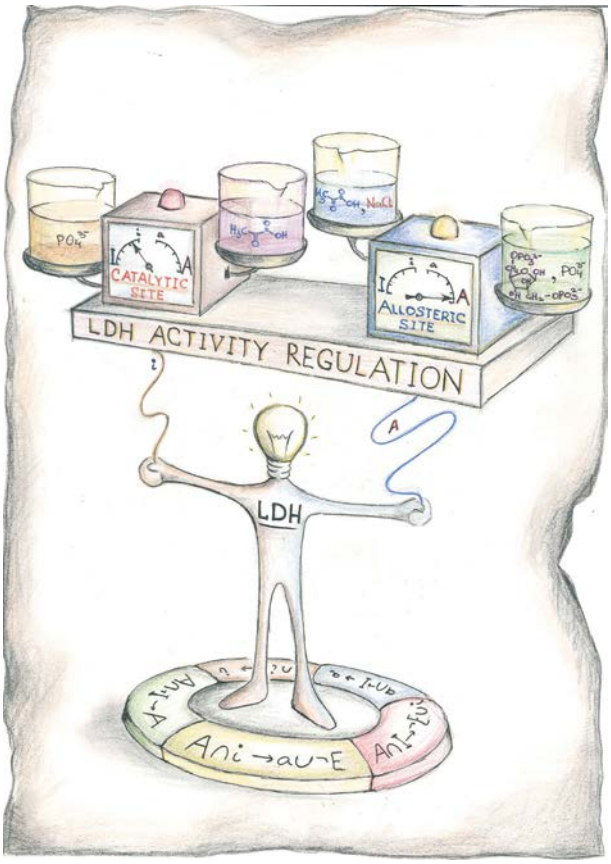


Fig. 21: Model for the regulatory mechanisms of lactate dehydrogenases proposed in [Feldman-Salit et al., 2013]. A combination of experimental and computational approaches reveal a subtle interplay between the effects of phosphate, fructose-1,6-bisphosphate (FBP), NaCl, and pyruvate, resulting in different regulatory effects on the enzymes in lactic acid bacteria. Left: The crystal structure of the lactate dehydrogenase from B. stearothermophilus with the catalytic and allosteric binding sites indicated. The homotetrameric quaternary structure and the homodimeric structure used as a template for modeling the LAB lactate dehydrogenase structures are shown in A and B, respectively. FBP molecules are shown in sphere representation in the allosteric binding sites. NAD (orange) and oxamate (a pyruvate analog, blue) are bound in the catalytic binding sites. Bottom: The catalytic and allosteric sites are illustrated as weight scales indicating inhibitory (i/I) and/or activating (a/A) effects.

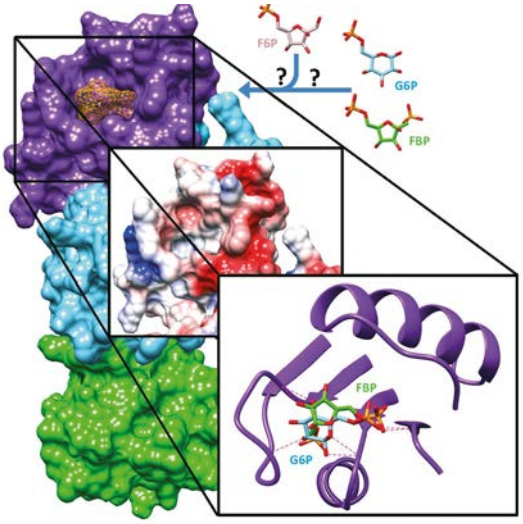


Fig. 22: A structural model of the pyruvate kinase monomer from the lactic acid bacterium Streptococcus pyogenes. Binding of potential activators, like fructose 1,6-bisphosphate (FBP), glucose 6-phosphate (G6P), and fructose 6-phosphate (F6P) to the allosteric site of pyruvate kinase was studied with different computational tools. [Veith, 2013].

TRAPP: A TOOL FOR ANALYSIS OF TRANSIENT BINDING POCKETS IN PROTEINS

Most drugs that target disease-related proteins bind in a concave pocket where many favorable contacts can be made with the target protein. Shape complementarity of a compound and a binding pocket is therefore often the first requirement for selecting a drug candidate. It is common to use a single static structure of a target protein for the detection of a binding pocket. This structure is usually determined by x-ray crystallography. However, proteins are highly flexible, so restricting analysis to one or a few experimentally determined structures may mean that transient, druggable pockets are not detected, implying that potentially druggable proteins are discarded. In order to facilitate the identification of drug targets, we have developed TRAPP as a software platform for exploring binding site dynamics and identifying and characterizing transient pockets. TRAPP has been designed for analysis of the dynamics of the region of a protein where a ligand (natural or non-natural) is known to bind and for which the aim is to identify transient pockets or sub-pockets that can be exploited in ligand design or optimization. The TRAPP platform allows for a range of methods to be used to generate protein motion trajectories of protein structure ensembles. These may encompass conformational changes ranging from local side-chain fluctuations to global backbone motions. TRAPP then performs accurate grid-based calculations of the shape and physico-chemical characteristics of a binding pocket for each structure and detects the conserved and transient regions of the pocket in an ensemble of protein conformations [Kokh, 2013]. It also provides tools for tracing the opening of a particular sub-pocket and residues that contribute to the binding site. TRAPP thus paves the way for an assessment of the druggability of a disease-related target protein that takes its flexibility into account, see Fig. 23 (following page).

TRAPP has been developed as part of a project in the BMBF Biotech Cluster Rhein Neckar (BioRN). This project was carried out in collaboration with Friedrich Rippmann and Paul Czodrowski at Global Computational Chemistry, Merck Serono, Darmstadt.

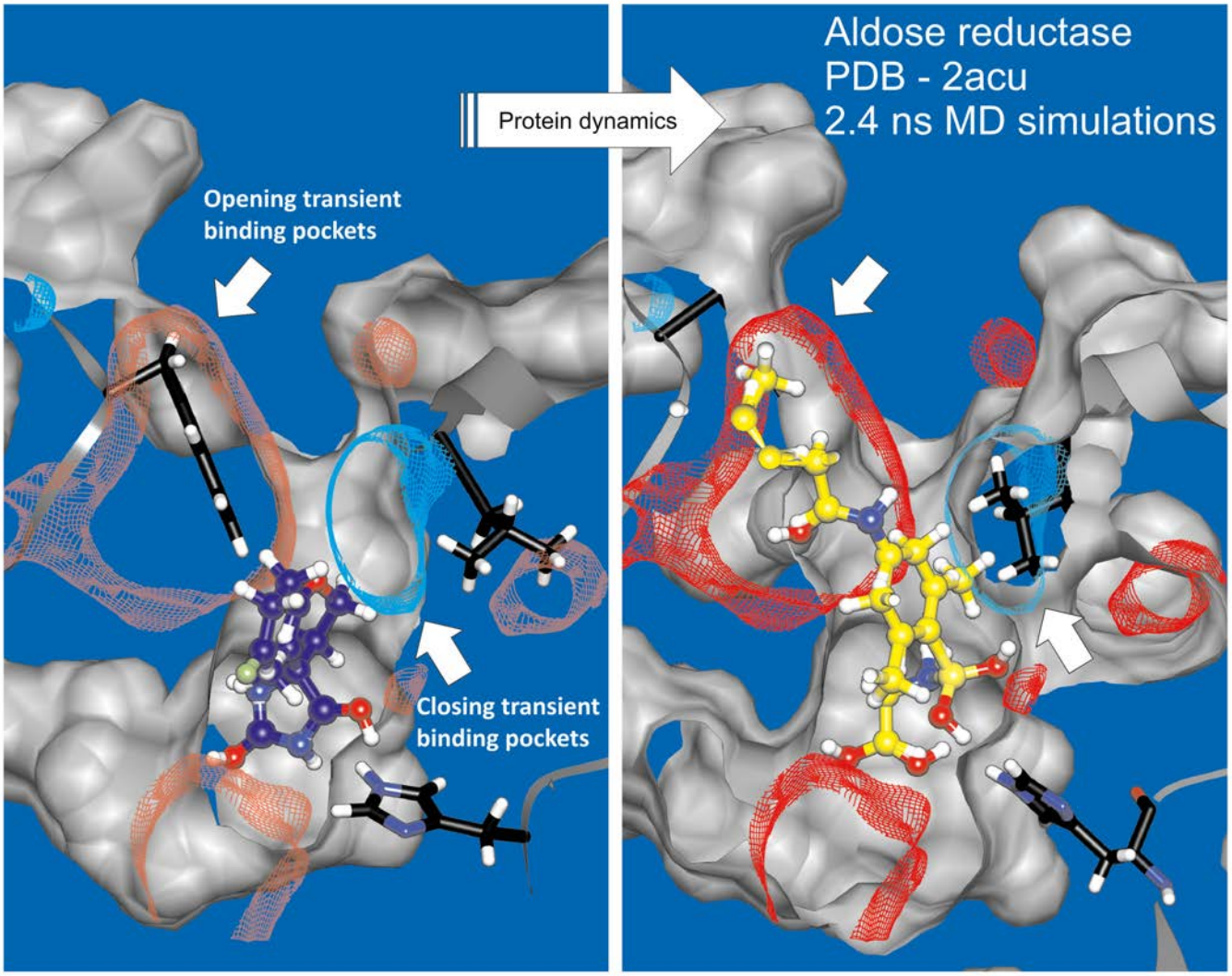


Fig. 23: Example of the application of TRAPP to identify and characterize pockets occurring transiently during a molecular dynamics simulation of an enzyme. A binding pocket of aldose reductase is shown by the grey surface and three residues in stick representation (carbon atoms in black). Transient binding regions are shown by red and cyan meshes (appearing and disappearing, respectively). Left : A reference crystal structure. Right: A snapshot after 2.4 ns of simulation. The ligands shown in the binding site are taken from co-crystallized protein structures superimposed with the reference and snapshot structures. A large transient sub-pocket opens (upper arrow) due to a loop motion and flipping of a tyrosine residue, making space for the ligand shown in the right-hand panel to extend into. [Kokh, 2013].

The Natural Language Processing (NLP) group develops methods, algorithms, and tools for the automatic analysis of natural language. The group focuses on discourse processing and related applications like automatic summarization.

In 2013 the NLP group was very successful, with papers accepted at all important NLP conferences. We also participated again with great success in the shared task of the Text Analysis Conference. Despite strong international and industrial competition, our team headed by Angela Fahrni ranked among the top 25% for mono- and cross-lingual entity linking.

In April 2013, Jie Cai successfully defended her PhD thesis. She is the first PhD student to graduate from the Computational Linguistics Department at the Heidelberg University. Congratulations, Cai! She left HITS at the end of 2012 to join Microsoft in Beijing. Our postdoc Camille Guinaudeau left HITS at the end of August 2013 for a position as maître de conférences (assistant professor) at the Université de Paris Sud. We have two new PhD students in our group: Mohsen Mesgar, who is working on modeling local coherence, and Alex Judea, who is working on concept disambiguation.

In 2013 Katja Markert again visited us for five months on an extension of her Humboldt fellowship. Together with PhD student Yufang Hou, she has been working on bridging resolution. We also had several other visitors, including Nafise Moosavi (Sharif University of Technology, Tehran, Iran), Gordana Ilic Holen (University of Oslo), and Minso Ko (KAIST, Daejeon, Korea).

Die Natural Language Processing-Gruppe (NLP) widmet sich der automatischen Verarbeitung natürlicher Sprache. Der Arbeitsschwerpunkt der Gruppe liegt auf dem Text- oder Diskursverstehen und darauf aufbauenden Anwendungen wie etwa der automatischen Zusammenfassung.

Das Jahr 2013 war wissenschaftlich sehr erfolgreich für die NLP-Gruppe. Wir schafften es, auf allen wichtigen NLP-Konferenzen zu publizieren. Darüber hinaus nahmen wir wieder erfolgreich an einem Wettbewerb im Rahmen der Text Analysis Conference teil. Das von Angela Fahrni geleitete Team erreichte eine Platzierung unter den Top 25% im Bereich mono- und cross-linguales Entity Linking.

Jie Cai verteidigte ihre Dissertation erfolgreich im April 2013. Damit ist sie die erste Doktorandin, die jemals an der Universität Heidelberg in der Computerlinguistik promoviert wurde. Herzlichen Glückwunsch, Cai! Sie hat uns schon Ende 2012 verlassen und arbeitet jetzt bei Microsoft in Peking. Unser PostDoc Camille Guinaudeau verließ uns Ende August 2013, um eine Stelle als Maître de Conférences (Assistant Professor) an der Université de Paris Sud anzutreten. Wir konnten aber auch zwei neue Doktoranden begrüßen: Mohsen Mesgar wird über das Modellieren lokaler Kohärenz arbeiten, Alex Judea über Konzeptdisambiguierung.

2013 konnten wir nochmals Katja Markert als Humboldt-Stipendiatin begrüßen. Sie arbeitet zusammen mit Yufang Hou erfolgreich am Problem der Bridgingauflösung. Daneben verbrachten Nafise Moosavi (Sharif University of Technology, Teheran, Iran), Gordana Ilic Holen (Universität Oslo) und Minso Ko (KAIST, Daejeon, Korea) längere oder kürzere Gastaufenthalte in unserer Gruppe.

CONCEPT AND ENTITY DISAMBIGUATION AND CLUSTERING

Concept and entity disambiguation means linking common and proper nouns in texts to their corresponding entries in a concept and entity inventory. We use Wikipedia as our inventory and consider each article to be a concept or entity. This task involves two challenges: (1) ambiguity (a noun can refer to several mentions) and (2) variety (different nouns can refer to the same concept or entity). Linking common and proper nouns in a text to concepts and entities is a step on the way toward automatic text understanding and can serve as a preprocessing stage in, say, a summarization or information retrieval system. In 2013 we refined our previous joint approach to concept disambiguation and clustering and integrated it into a larger framework.

DISCOURSE-AWARE CONCEPT AND ENTITY DISAMBIGUATION

Here is a short text example in which all common and proper nouns to be disambiguated are highlighted.

*"I have to review a **paper**," the **supervisor** moaned from the door. "Please don't disturb me until I am done with the **review**." His **student** nodded, went to the **cafeteria**, sat down into the sun and read yesterday's **paper**.*

This example illustrates that the interpretation of different nouns is determined by different notions of context. Some nouns depend more on a local sentence-level context, and here the global text-level context is in fact misleading for their disambiguation (e.g. paper in read yesterday's paper). Some nouns are influenced more strongly by the global context (review in I am done with the review). In these cases, the local context is not discriminative. Some other nouns depend on both the global and the local context (paper in review a paper). The context relevant for disambiguating a noun depends on how it is embedded in the discourse. It is not determined by the surface form of a noun.

These aspects have been neglected by previous disambiguation approaches. We propose a novel approach that disambiguates nouns differently depending on the way they are embedded in the discourse. We distinguish three different types of discourse embedding and model them as latent variables using Markov Logic Networks. The use of latent variables enables us to learn and predict the cohesive scope and the disambiguation of a noun at the same time. Another advantage is that learning the discourse-embedding types does not need its own annotated data but is guided by the annotations available for the target prediction task, i.e. the disambiguation. The results on various commonly used data sets indicate that our novel discourse-aware approach consistently improves disambiguation results.

To compare our system to other state-of-the-art approaches, we participated this year for the third time in the shared task of the Text Analysis Conference (http://www.nist.gov/tac/)" on English monolingual and Spanish and Chinese cross-lingual entity linking and clustering [Fahrni et al., 2013]. The results show that our system is competitive across all languages without language-specific tuning.

JOINT CONCEPT AND ENTITY DISAMBIGUATION AND COREFERENCE RESOLUTION

Another aspect of text understanding (and another important research interest in our group) is coreference resolution, i.e., determining which expressions in texts denote the same entities in the world. Consider the following two sentences:

Two Russian punk rockers *freed from jail –* ***The band members****, jailed for an anti-Putin protest, were released under a new amnesty law.*

Here, the expressions in bold print refer to the same entities in the world. We believe that the two tasks concept disambiguation and coreference resolution can be mutually supportive. Concept disambiguation provides access to world knowledge for coreference resolution by linking textual expressions to encyclopedias like Wikipedia. If we know that punk rock is a kind of music and that bands have something to do with music, we can infer that the expressions in bold print probably denote the same thing. On the other hand, we are in a better position to disambiguate textual expressions if we know other textual expressions denoting the same thing. For example, our meanings inventory has no entry for band members. If we know that punk rockers and band members are the same thing, we can infer from this that musical ensemble should be the right meaning for band members as opposed to other kinds of members, e.g., parliament members.

COREFERENCE RESOLUTION

This year, we have introduced two different unsupervised models for coreference resolution. The first model is based on our previous graph-based coreference resolution model, the second is a new probabilistic model based on the statistical analysis of feature distributions among entities.

In the graph-based approach, a document to be processed is modeled as a graph in which the nodes are the mentions and the edges are relations between mentions that are either indicative of a coreference relation (like string matching) or of non-coreference (like number disagreement). In this graph, sets of coreferring mentions are obtained via clustering. One central aspect of this

The NLP group in 2013 (v.l.n.r.): Michael Strube, Sebastian Martschat, Nafise Moosavi, Daraksha Parveen, Nicolas Bellm, Camille Guinaudeau, Katja Markert, Angela Fahrni, Youfang Hou, Thierry Göckel

Group Leader
Prof. Dr. Michael Strube

Staff members
Angela Fahrni

HITS Scholarship holders
Dr. Camille Guinaudeau (until Aug. 2013)
Alex Judea (from Oct. 2013)
Sebastian Martschat
Mohsen Mesgar (from Sept. 2013)
Daraksha Parveen

Visiting scientists
Gordana Ilic Holen (Nov. 2013)
Yufang Hou (Promotionskolleg Scholarship)
Minsu Ko (Aug.-Sept. 2013)
Dr. Katja Markert (AvH Scholarship, April-Aug. 2013)
Nafise Moosavi (from February 2013)
Caecilia Zirn

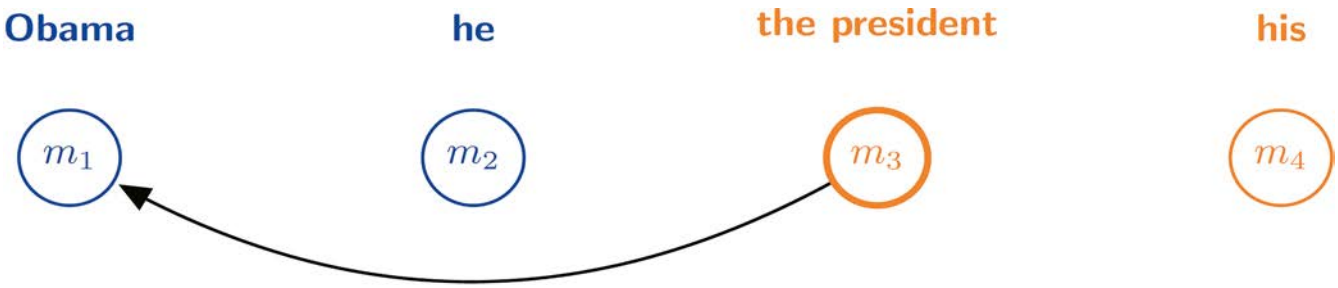
Students
Nicolas Bellm
Samuel Broscheit (until August 2013)
Thierry Göckel
Benjamin Heinzerling (from July 2013)
Hans-Martin Ramsl (from Aug. 2013)

model is assigning weights to edges: given an edge of a specific type (e.g. string matching) between two mentions, what weight should be assigned to the edge? Last year we adapted the weighting scheme from the hypergraph coreference model previously developed in the NLP group, which assigned weights to edges by computing from training data the fraction of edges connecting coreferent mentions in the graphs. This year we devised a different weighting scheme that does not depend on training data at all: edges built from relations indicated for coreference are given weight 1, all other edges are given weight minus infinity. This unsupervised model performs as well as its supervised variant and obtains state-of-the-art results on benchmark datasets [Martschat 2013].

In the second model, our probabilistic modeling of coreference resolution is based on the local and global distribution of features among coreferring and non-coreferring mentions.

Our resolution method is an incremental mention-entity method. The inference method processes all mentions from the beginning of the text to the end, and for each mention it combines the local and global distributions of features in order to decide which entity this mention should be assigned to.

Fig. 24: Coreference resolution error analysis.



The global distribution of features among entities is captured by a measure that helps to score features based on their association with coreference relations. In this way, a feature that figures mostly between coreferring mentions will get a higher score than a feature that is mostly found among non-coreferring mentions.

Similarly, the local distribution measure takes care of the distribution of features among a specific mention and its corresponding candidate entities.

Accordingly, the inference method needs global feature distribution to construct the entities, but to compute global feature distribution; we need to have the set of entities. Estimation of entities using the inference algorithm and estimation of global feature distribution based on estimated entities correspond to the E and M steps of the Expectation Maximization (EM) algorithm. Our method estimates the entities and the global distribution of features iteratively. It continues to estimate these two sets of unknown parameters until the parameters converge to steady values.

Although totally unsupervised, our method performs slightly better than its corresponding rule-based system, in which the importance of features is given beforehand based on linguistic assumptions. Since this method is independent of annotated data, input features, and therefore input language, it is suitable for application to new domains or languages without annotated data (so called low-resource languages).

Another focus of our work was error analysis. We have devised a novel,linguistically informed method for analyzing the links between coreferring mentions not recovered by a coreference resolution system. Suppose the system misses the link from the president to the partial entity consisting of Obama and he. Our method scores the link from the president to Obama as missing because the coreferential relation between these two mentions can in general be more easily recovered than the link from the president to he (Figure 24).

We applied this method to the analysis of our graph-based model. Here we focused on investigating missing links between common nouns and proper names. The analysis revealed that most of the missing links can be attributed to a lack of world knowledge, as in the Obama – the president example. In order to obtain such knowledge, we investigated the use of large knowledge bases built on Wikipedia (such as YAGO and Freebase). From such knowledge bases we can extract the fact that Barack Obama is a president. To make this knowledge accessible to our model, we preprocess the documents with our concept and entity disambiguation system, which links mentions to Wikipedia, and then extract the corresponding relations between the mentions from the Wikipedia-based knowledge base.

However, the coverage of the knowledge bases is still too low to recover more than a small fraction of the missing links. Right now we are investigating relation extraction from very large text collections such as newspaper archives or the web in order to extend the knowledge bases.

BRIDGING ANAPHORA RECOGNITION USING CASCADING COLLECTIVE CLASSIFICATION

Alongside coreference resolution, **bridging anaphora** is essential for the understanding of entity coherence. In the example below, the phrases a resident, the stairs and the lobby are **bridging anaphora** relating to the antecedent one building. Without bridging resolution, entity coherence between the first and second coordinated clause in Example 1 cannot be established. This is a problem both for coherence theories such as the centering model (where bridging figures as an indirect realization of previous entities) and for applications relying on entity coherence modeling, such as readability assessment or sentence ordering.

One building was upgraded to read status while people were taking things out, and **a resident** called up **the stairs** to his girlfriend, telling her to keep sending things down to **the lobby**.

Bridging resolution can be divided into two tasks: (1) bridging anaphora recognition and (2) determining the correct antecedent among several candidates. In the example, we first recognize that the phrases a resident, the stairs and the lobby are bridging anaphors (task 1) and then select their antecedent one building (task 2) [Hou et al. 2013a]. Although those two tasks need to be combined to perform bridging resolution completely, bridging recognition on its own can also be valuable for applications. For example, bridging anaphora are an information status category which, like other such category, has an influence on prosody without any knowledge of the antecedent.

Bridging anaphora recognition is usually handled as part of the information status classification task. Each mention in a text gets assigned to one information status class

describing its accessibility for the reader at a given point in a text, bridging being one possibility. However, bridging anaphora recognition is a difficult task, and in previous work we had very low results to report on this information status class. This is due to the variety of the phenomenon, involving insufficiently clear surface features for recognition.

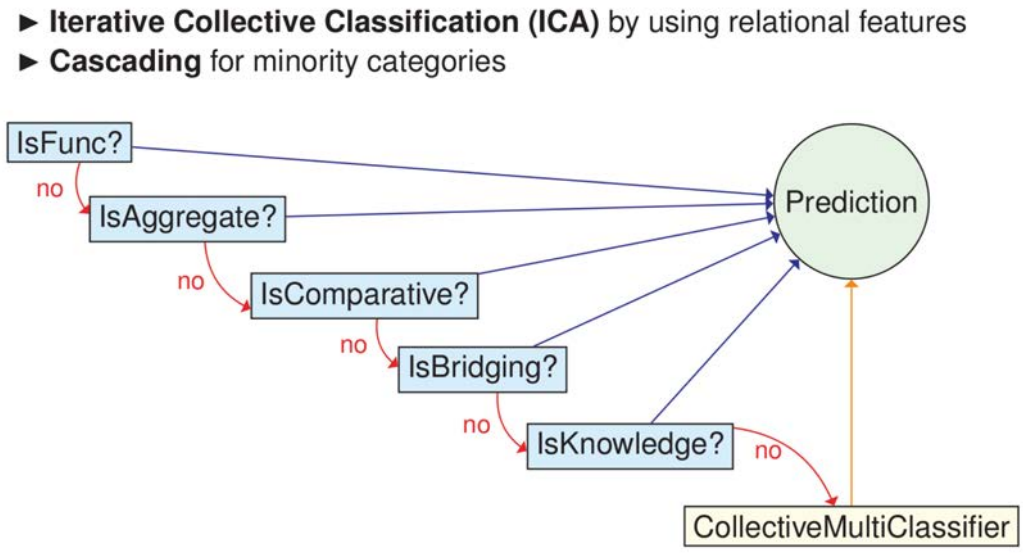
This year we came back to this task. Based on linguistic intuition and corpus research, we developed novel discourse structure, lexico-semantic, and genericity features. In addition, bridging makes for between 5% and 20% of definite descriptions and around 6% of all NPs, making it less frequent than many other information status categories, and recognition of minority categories is known to be more difficult. We therefore developed a cascaded classification algorithm to address this problem (see Fig. 25). We substantially improved bridging anaphora recognition without impairing performance in other information status classes [Hou et al., 2013b].

In future, we plan to combine bridging anaphora recognition and antecedent selection to perform full bridging anaphora resolution. We also plan to study how bridging anaphora can be applied to weak discourse relations to establish local coherence.

EXTENSIVE EVALUATION OF THE GRAPH-BASED LOCAL COHERENCE MODEL

Coreference and bridging relations are linguistic devices for establishing local coherence. Accordingly, both core-

Fig. 25: Iterative Collective Classification for Bridging Anaphora Recognition.



ference and bridging resolution are tightly connected with local coherence and can serve as input for a model dealing with that phenomenon. In 2012, we thus developed a graph-based local coherence model (Guinaudeau 2013a) inspired by one of the most popular local coherence models in the literature: the entity grid model. Our model has two main advantages over the entity grid model. First, as the graph used for document representation contains information about entity transitions, our model does not need a learning phase. Second, as it relies only on graph centrality, our model does not suffer from the computational complexity and data sparsity problems of the entity grid model.

In 2013 we focused on an extensive evaluation of the model. To this end, we performed a sentence ordering task, where the model has to distinguish between documents where the sentences are presented in the correct order and documents with rearranged sentences. Then, as the first task uses “artificial” documents (the ones with rearranged sentences), we also worked on two other tasks that involve “real” documents: summary coherence rating and readability assessment.

The sentence ordering task involves comparing a document to a random permutation of its sentences. Here our system associates local coherence values with the original document and its permutation, the output of our system being considered as correct if the
score for the original document is higher than the score of its permutation. This task was performed on the test part of the CoNLL 2012 shared task.

The objective of the summary coherence rating task is to order a pair of summaries according to local coherence. For this experiment, pairs to be ordered are composed of summaries extracted from the Document Understanding Conference (DUC 2003) corpus. Summaries provided either by humans or by automatic systems were judged by seven human annotators and associated with a coherence score. 80 pairs were then created, each of them made up of two summaries of the same document where the score of one of the summaries is significantly higher than the score of the other.

Finally the readability assessment task involves evaluating how difficult a document is to read. We perform this task on a dataset comprising 107 articles about capital cities, each of them in two versions, one for adults, from the Encyclopedia Britannica, and one for children, from the Britannica Elementary. To estimate the complexity of a document, our model computes the local coherence score for each article in the two categories. The article associated with the higher score is considered to be the more readable as it is more coherent and needs less interpretation on the reader’s part than a document associated with a lower local coherence score.

Fig. 26: Accuracy values for sentence ordering, summary coherence rating, and readability assessment tasks.

Sentence Ordering	Summary Coherence Rating	Readability	Assessment
B&L	87.7	83.3	50.9
E&C	91.5	-	-
Graph-based	88.9 (88.9)	70.0 (80.0)	76.6 (76.6)

The evaluation of our model is presented in Fig. 26. In this table, the first two lines represent the accuracy values – i.e. number of correct decisions divided by the number of comparisons – obtained by two state-of-the-art algorithms: the entity grid model proposed by Barzilay and Lapata (2008) (B&L) and its extended version proposed by Elsner and Charniak (2011) (E&C). The results obtained by our graph-based model accounting for syntactic information and distance between sentences are presented in the last line of the table.

The extensive evaluation performed shows that our model is robust among tasks and gives reasonable results for all tasks with the same parameter settings. Moreover, our model can be optimized and obtains results comparable with entity grid-based methods when optimal settings are used for each task (values in parentheses).

TOPIC-BASED MULTI-DOCUMENT SUMMARIZATION

The aim of automatic summarization is to shorten a document without any loss of information in the summary. Common approaches to this task select the most important sentences and arrange them in the order in which they appear in the original document. This is more difficult when summarizing multiple documents. Here the input is a set of documents dealing with the same topic, and the task is to select the most important sentences from all documents and arrange them in a coherent way. A topic (or query) can be used to represent the reader‘s specific interest. Hence multi-document summarization should measure the importance of sentences with respect to the given topic.

In our work we build on the graph-based local coherence model presented above. The input documents are presented as a bipartite graph, where one node set consists of sentences in different documents and the other node set consists of the entities mentioned in the documents. We apply the HITS (Hyperlink-induced Topic Search) algorithm. where one of our node sets corresponds to the hubs and the other to the authorities of the HITS algorithm. We give entities an initial rank. If an entity is present both in a document and in the topic, its initial rank will be heightened. Then the HITS algorithm performs repeated updates on hub and authority scores until convergence is achieved. The output is a ranked list of sentences in descending order of importance with respect to the topic. We choose the top-ranked sentences for the summary and carry on until a predetermined summary length has been reached.

Since we do not check whether the summary is coherent, our future work will include the development of a robust technique for measuring the coherence of the multi-document summary (building on the graph-based local coherence model described above). Nor do we control for redundancy in the summary, a task which is however indispensable if we want to produce brief summaries with a high information content.

The Scientific Computing Group (SCO) focuses on developing algorithms, computer architectures, and high-performance computing solutions for bioinformatics.

We mainly focus on

- computational molecular phylogenetics
- large-scale evolutionary biology data analyses
- supercomputing
- quantifying biodiversity
- next-generation sequence-data analysis

Secondary research interests include, but are not limited to,

- emerging parallel architectures (FPGAs, GPUs, Xeon PHI)
- discrete algorithms on trees
- population genetics

Here we outline SCO's current research activities. Our research is situated at the interface(s) between computer science, electrical engineering, biology, and bioinformatics. The overall goal is to devise new methods, algorithms, computer architectures, and freely available/accessible tools for molecular data analysis and make them available to evolutionary biologists. In other words, our overarching goal is to support research. One aim of evolutionary biology is to infer evolutionary relationships between species and the properties of individuals within populations of the same species. In modern biology, evolution is a widely accepted fact and can nowadays be analyzed, observed, and tracked at the DNA level. A famous dictum widely quoted in this context comes from evolutionary biologist Theodosius Dobzhansky: "Nothing in biology makes sense except in the light of evolution."

Die Gruppe wissenschaftliches Rechnen (SCO) beschäftigt sich mit Algorithmen, Hardware-Architekturen und dem Hochleistungsrechnen für die Bioinformatik.

Unsere Hauptforschungsgebiete sind:

- Rechnerbasierte molekulare Stammbaumrekonstruktion
- Analyse großer evolutionsbiologischer Datensätze
- Hochleistungsrechnen
- Quantifizierung von Biodiversität
- Analyse von Sequenzdaten der nächsten Generation

Sekundäre Forschungsgebiete sind unter anderem:

- Neue parallele Rechnerarchitekturen (FPGAs, GPUs, Xeon PHI)
- Diskrete Algorithmen auf Bäumen
- Methoden der Populationsgenetik

Im Folgenden beschreiben wir unsere Forschungsaktivitäten. Unsere Forschung setzt an der Schnittstelle zwischen Informatik, Elektrotechnik, Biologie und Bioinformatik an. Unser Ziel ist es, Evolutionsbiologen neue Methoden, Algorithmen, Computerarchitekturen und frei zugängliche Werkzeuge für die Analyse molekularer Daten zur Verfügung zu stellen. Unser grundlegendes Ziel ist es, Forschung zu unterstützen. Die Evolutionsbiologie versucht die evolutionären Zusammenhänge zwischen Spezies sowie die Eigenschaften von Populationen innerhalb einer Spezies zu berechnen. In der modernen Biologie ist die Evolution eine weithin akzeptierte Tatsache und kann heute anhand von DNA analysiert, beobachtet und verfolgt werden. Ein berühmtes Zitat in diesem Zusammenhang stammt von Theodosius Dobzhansky: „Nichts in der Biologie ergibt Sinn, wenn es nicht im Licht der Evolution betrachtet wird".

WHAT HAPPENED IN OUR GROUP IN 2013?

The following is an account of the important events in our group in 2013.

In summer 2013, Alexis finished teaching the first one-year course module entitled “Introduction to Bioinformatics for Computer Scientists” at the Karlsruhe Institute of Technology (KIT). This course consists of a winter lecture on bioinformatics and a summer seminar on hot topics in bioinformatics.
The course was well received, and Alexis obtained a very positive teaching evaluation from the students.

In this context, we also have three computer science students from KIT working with us at present. David Dao is doing his Bachelor thesis here, Sebastian Mendes has been working with us as student programmer for GPU programming, and Patrick Flick is doing his Master‘s thesis with us in collaboration with colleagues from Scandinavia.

In March, Alexis also visited the Department of Ecology and Evolutionary Biology at the University of Arizona at Tucson for the first time after being appointed adjunct professor at Tucson.

Another highlight was Alexis’ appointment as member of the scientific advisory board of the recently established computational biology institute in Montpellier, France.

The year was also very important for our former PhD student Simon Berger. He defended his thesis successfully at the Karlsruhe Institute of Technology in January 2013. Simon started working in the field of automotive software development in Munich right after finishing his PhD.

Our former PhD student Nikos Alachiotis, who graduated in 2012 and then did his military service in Greece, will soon be moving to the US to start a postdoc at Carnegie Mellon University.

We were also delighted to hear that our former postdoc Pavlos Pavlidis has recently been offered a tenure-track researcher position in bioinformatics at the Institute of Computer Science of the Foundation for Research and Technology Hellas in Heraklion, Crete.

At the end of 2013 we also had to say goodbye to our former PhD student Fernando Izquierdo-Carrasco, who has now finished writing his PhD thesis, which he will defend at the Karlsruhe Institute of Technology in summer 2014. Starting in February 2014, he will be working as a bioinformatician in industry.

Our postdoc Solon Pissis, whom we shared with Pam and Doug Soltis at the University of Florida, left us in summer 2013 and is now a lecturer in computer science at King‘s College London.

Our PhD student Andre Aberer spent three summer months in Stockholm working with Fred Ronquist, one of the ‚gurus‘ of Bayesian phylogenetic inference.

We were also happy to welcome our new PhD student Alexey Kozlov, the first PhD candidate we have recruited from KIT.

In 2013 we hosted a visiting PhD student within the auspices of our Visiting PhD Student program. Paschalia Kapli, a biologist from the University of Crete who works on the evolution of lizards actually visited us twice. She spent time at our lab up to March 2013 and then came back again for another 6 months in September 2013.

Another highlight was running the summer school on computational molecular evolution for the 5th time on the Wellcome Trust campus in Hinxton, UK. Alexis was a co-organizer of this event. We received 80 applications for

the 40 places available. Our postdoc Tomas and our former PhD student Fernando also participated at the summer school as teaching assistants.

OVERVIEW

The term “evolutionary bioinformatics” is used to refer to computer-based methods for reconstructing evolutionary trees from DNA or from, say, protein or morphological data. The term also refers to the design of programs for estimating statistical properties of populations, that is, for disentangling evolutionary events taking place within a single species.

The very first evolutionary trees were inferred manually by comparing the morphological characteristics (traits) of the species under study. Nowadays, in the age of the molecular data avalanche, the manual reconstruction of trees is no longer feasible. This is why evolutionary biologists now rely on computers for phylogenetic and population-genetic analyses. In fact, following the introduction of so-called short-read sequencing machines (machines used in the wet-lab by biologists to extract DNA data from organisms) that can generate over 10,000,000 short DNA fragments (containing between 30 and 400 DNA characters), the community as a whole is facing novel and exciting challenges. One of the key problems to be tackled is the fact that the amount of molecular data available in public databases is growing at a significantly faster pace than the computers capable of analyzing the data can keep up with. In addition, the cost of sequencing a genome is decreasing at a faster pace than the cost of computation. Accordingly, as computer scientists we are facing a scalability challenge, that is, we are constantly trying to catch up with the data avalanche and make molecular data analysis tools more scalable with respect to dataset sizes. At the same time, we also want to implement more complex and hence more realistic and compute-intensive models of evolution.

Another difficulty is that next-generation sequencing

The SCO group in 2013 (f.l.t.r.): David Dao, Alexey Kozlov, Diego Darriba, Tomáš Flouri, Jiajie Zhang, Paschalia Kapli, Kassian Kobert, Alexandros Stamatakis, Andre Aberer

Group Leader
Prof. Dr. Alexandros Stamatakis

Staff members
Andre Aberer
Tomáš Flouri
Fernando Izquierdo-Carrasco (until Dec. 2013)
Dr. Simon Berger (until Jan. 2013)

Scholarship holders
Kassian Kobert (HITS Scholarship)
Jiajie Zhang (HITS Scholarship)
Alexey Kozlov (HITS Scholarship, from June 2013)
Visiting scientists
Paschalia Kapli (until March 2013 & from Sept. 2013)
Dr. Solon Pissis (until June 2013)

Students
David Dao (from June 2013)
Patrick Flick (from June 2013)
Sebastian Mendez (April – Dec. 2013)

technology is changing rapidly. Accordingly, the output of those machines with respect to the length and quality of the sequences they can generate is also changing all the time. This requires the continuous development of new algorithms and tools for filtering, puzzling together, and analyzing these molecular data.

Yet another big challenge is reconstructing the tree of life based on the entire genome sequence data of each living organism on earth.

Phylogenetic trees (evolutionary histories of species) are important in many domains of biological and medical research. The programs for tree reconstruction developed in our lab can be deployed to infer evolutionary relationships among viruses, bacteria, green plants, fungi, mammals, etc. In other words, they are applicable to all types of species. In combination with geographical and climate data, evolutionary trees can be used, for instance, to disentangle the geographical origin of the H1N5 viral outbreak, determine the correlation between the frequency of speciation events (species diversity) and climatic changes in the past, or analyze microbial diversity in the human gut. For conservation projects, trees can also be deployed to determine endangered species that need to be protected, based on the number of non-endangered close relatives they have.

Studies of population-genetic data, i.e., genetic material from a large number of individuals of the same species (a human population, for instance) can be used to identify mutations leading to specific types of cancer or other serious diseases.

As we have seen, one key challenge for computer science is to scale existing analytical methods to the huge new datasets produced by next-generation sequencing methods. As we are involved in a number of large-scale empirical data-analysis projects, we face these challenges every day.

In the One Thousand Insect Transcriptome project (1KITE, www.1kite.org), for example, we intend to disentangle the evolution of insects by using 1,000 insect transcriptome sequences. To analyze these data we need to use a large supercomputer such as the SuperMUC system in Munich. The transcriptome is the fraction of the DNA that is translated into RNA in each cell and encodes important functions with respect to processes and development in that cell.

Another major challenge is the so-called multi-core revolution in parallel computing architectures. Throughout the 1980s and 1990s, computers became faster and faster as a result of increases in clock frequencies (the clock frequency of a processor essentially represents the number of arithmetic operations that can be executed per second). We have now reached a point where sheer physical limitations dictate that clock frequencies cannot be increased beyond approximately 4GHz (4,000,000,000 instructions per second). Accordingly, the computer industry has started producing systems with more than one processor, so-called multi-core processors, so that further speed improvements (e.g. for analyzing larger and more complex phylogenetic trees) can be achieved. This represents a significant paradigm shift, because in order to exploit the available computational resources (cores) in a new-generation processor, programs now need to be executed in parallel. In other words, programs need to be re-written so that they can perform their computational steps simultaneously on several processing cores.

It is a non-trivial task to transform a serial/sequential program into a parallel program, so the transition to multi-core architectures is a genuine revolution. Put briefly, the task of parallelizing a code so that it can be executed simultaneously requires human intuition. In addition, each new generation of processors makes more cores available to the application programmer. As a consequence, programming environments are becoming increasingly complex. Current hardware is also becoming hybrid, i.e. a collection of processors with distinct characteristics for different types of computations are integrated on the same chip or system. Evidently, this development makes the process of developing efficient parallel codes even more complex and error-prone. A substantial amount of manpower needs to be invested to efficiently exploit the

capabilities of modern hardware for molecular evolutionary analyses.
In the following, we briefly outline some research highlights in 2013.

THE PHYLOGENETIC LIKELIHOOD LIBRARY

A project most lab members are directly involved in is the development of the so-called Phylogenetic Likelihood Library. The project is headed by postdoc Tomas Flouri, but most lab members have contributed code or are using the library. The goal of this project is to design a scalable and well-documented library allowing for rapid development of likelihood-based phylogenetic tools (Bayesian and Maximum Likelihood).
The Library as such has not been published yet, but we have already released and made available the source code and API (Application Programmer Interface) documentation. The Library is available at http://www.libpll.org/

The Library implements highly tuned functions for calculating likelihood scores on trees, optimizing branch lengths and model parameters. It supports common data types such as DNA and protein data. Furthermore, it can be run sequentially using 128-bit SSE, 256-bit AVX, 256-bit AVX2 (including fused multiply-add operations), and vector intrinsics and also in parallel, using either a PThreads parallelization for multi-core systems or a MPI (Message Passing Interface) parallelization for cluster and supercomputer systems.

It also supports GPUs (Graphics Processing Units), but the results so far have been rather disappointing with respect to performance (see [Izquierdo-Carrasco 2013]). This is mainly due to the fact that the x86 version of the Library is highly efficient on standard processors because we have manually inserted and tuned the code using vector intrinsics.
We have also deployed the Library to substantially accelerate a code called DPPDiv for the Bayesian inference of divergence times. This code was not developed in our lab. It took only about a month to integrate the Library functions with the DPPDiv code, and the results were promising. By integrating the Library, we obtained a sequential speedup over the native likelihood function implementation in DPPDiv of up to a factor of 7.8. By using the PThreads version of the library that is entirely transparent to the application program (DPPDiv), we obtained a total acceleration of factor 350 in the best case on a 48-core multi-core system (see Figure below). For details, please see [Darriba 2013].

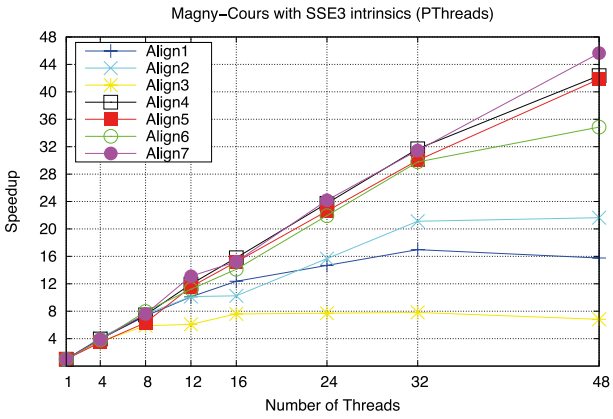


Fig. 27: Speedup plot for DPPDiv on a 48-core system using the SSE3-based and PThreads parallelized version of the Phylogenetic Likelihood Library for different alignment sizes. With increasing input (alignment) size, the speedups become almost linear.

DELIMITING SPECIES

A rather tricky problem that we have been grappling with in 2013 is that of species delimitation. This problem centers around finding an answer to the following question: Given a set of sequences from the same gene or genes, how many species does such a sample contain? Or, put differently, how many sequences belong to different individuals representing the same species? The problem as such is difficult, because there is still huge unsettled controversy in biology as to what a species actually is.
We decided to adopt the so-called phylogenetic species concept and developed two closely related methods for delimiting species based on their sequences.

The first is a stand-alone method which, given a rooted phylogenetic tree for the sequences under consideration, determines the branches on which the species boundaries should be placed. This means that, below these boundaries, we actually have individuals from the same population. This is outlined in the figure below. The example tree contains 4 Species C, D, E, F, where species D and F are each represented by two individuals d1, d2 and f1, f2, respectively.

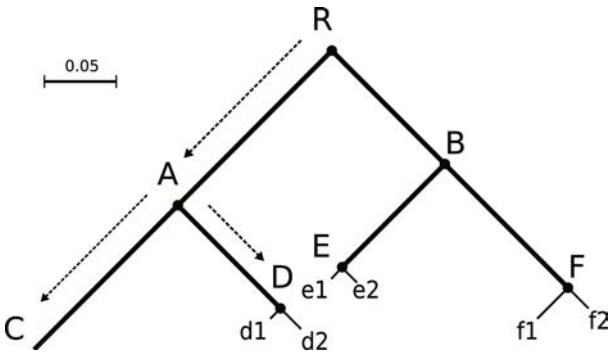


Fig. 28. Example of a species delimitation (limits at nodes C, D, E, and F) for a given rooted phylogenetic tree.

The actual delimitation software deploys a statistical technique which we have termed PTP: Poisson Tree Process (see [Zhang 2013]).
One important advantage over existing methods for species delimitation is that the input tree for PTP does not need to be what we call ultrametric (an ultrametric tree is a tree that has paths of equal length from the root to all leaves). A phylogenetic tree obtained via likelihood-based phylogenetic inference is typically not ultrametric, and generating an ultrametric tree from such a non-ultrametric tree is a complicated, time-consuming, and error-prone process that can be circumvented by PTP.

The second method we have developed is mainly for conducting species delimitations on huge datasets as obtained for instance by next-generation sequencing experiments. It combines the evolutionary placement algorithm (EPA) with the PTP method. Here we already have a given phylogenetic reference tree and a reference dataset that encapsulate a predefined (empirical) species concept. That is, we assume that each species is represented by a single sequence and a corresponding leaf in the tree. Initially, we place the sequences from the sample into that reference phylogeny and then execute PTP separately and independently for each branch of the reference tree that contains sequences from our sample.

Overall, our method, especially when used in conjunction with the EPA (see above), outperforms other current methods for species delimitation and so-called OTU (Operational Taxonomic Units) clustering approaches.
Like all our codes, the software is freely available as open-source code, and we have also developed a web-service that implements PTP and other methods for species delimitation. A screenshot has been included below.

HOW PLAUSIBLE ARE LARGE TREES?

In the past we have designed tools such as RAxML that allow for the reconstruction of very large phylogenetic trees, which contain tens of thousands of sequences that are also called taxa. One unresolved issue is how one can assess the plausibility of such large trees. In other words, when dealing with phylogenies comprising 50,000 or more taxa, we need to be able to decide whether a tree makes ‚biological sense.'

The key issue here is that such large trees cannot be scrutinized visually by biological experts because we lack appropriate visualization tools and because inspecting a tree comprising more than 50,000 species would most probably lead to insanity.
Thus we need to provide automated methods for assessing the plausibility of such large trees. One way to do this is to compare such a large tree with a set of substantially smaller and hence probably more accurate subtrees from some database comprising only 100-500 taxa (more accurate because the tree search space is much smaller). Of course, the taxa in the small reference trees need to form a proper subset of the taxa in the large tree we want to assess. Fortunately, such a curated database containing small reference trees does exist (STBase http://stbase.org/), containing 1 billion reference trees that can be queried.

Given the prolegomena, we now simply need to devise a method for comparing the small trees with the large tree whose plausibility we intend to assess. This can be done by the following procedure: Initially, we need to extract the partial tree (induced subtree) from the large tree that covers the taxon set of the small reference tree. Once this is done, we can then compare the topologies of the two trees that now have an exactly identical taxon set by computing some topological distance measure. We can then simply average the topological distances between all small trees and the corresponding induced subtrees from the large tree to obtain a notion of plausibility or dissimilarity.

Algorithmically, the main challenge was to develop a method that enables us to rapidly extract the induced subtree from the huge phylogeny for a given taxon set as defined by the small reference tree. Our postdoc Tomas and our bachelor student David worked on this algorith-

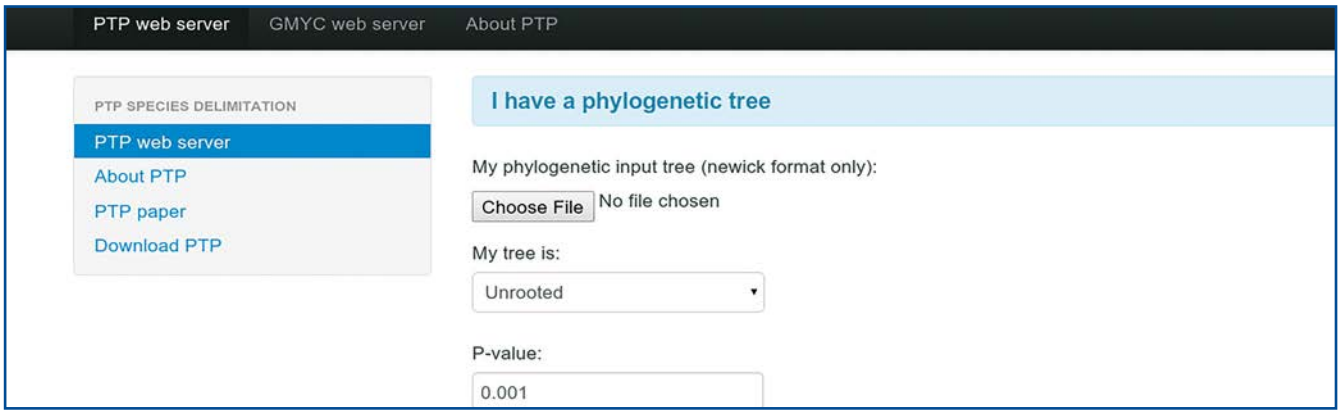


Fig. 29: Screenshot of the species delimitation web-service offering the PTP and the GMYC (General Mixed Yule Coalescent) methods.

mic problem, and Alexis integrated the algorithm into the RAxML tool.

The novel, efficient algorithm for subtree extraction turned out to be 5 orders of magnitude faster than a naïve algorithm we used for initial testing (see book chapter preprint at http://sco.h-its.org/exelixis/pubs/Exelixis-RR-DR-2013-6.pdf).

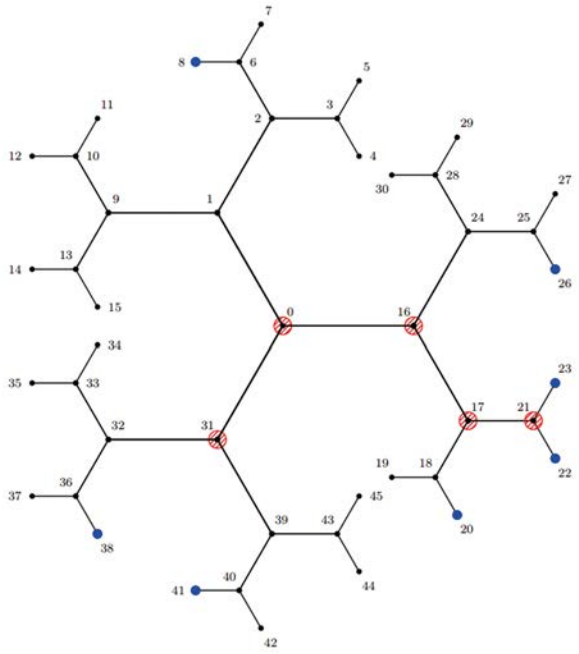


Fig. 30: A graph from our book chapter explaining the fast procedure for extracting an induced subtree from the huge tree topology.

Furthermore, the initial results with respect to average topological distance obtained for a huge real-world phylogeny were quite promising.

For a huge tree with 55,000 taxa, we obtained an average topological distance of 31.8% between the large tree and the reference trees from STBase (see above). Because of the substantially larger tree search space for the 55,000-taxon tree, we consider this average topological distance of approximately 32% to be rather low. For a tree with 2000 taxa there are about 3.00 × 106328 possible unrooted tree topologies, whereas for 55,000 taxa there exist approximately 2.94 × 10253380 possible unrooted tree topologies. In other words, the tree search space of the 55K taxon tree is about 10247052 times larger than for the 2,000-taxon tree. Taking into account further that different procedures were used to automatically construct the corresponding alignments and that the trees have also partly been constructed from different genes, an average error of around 30% appears to be low.

In 2014 we intend to further develop the method and test it more systematically with empirical biological data.

INTRODUCING EXAML

In 2013 Alexis and Andre also released the new code for conducting large-scale maximum likelihood analyses on supercomputers. This new code is called ExaML (Exascale Maximum Likelihood) and is available under GNU GPL via Alexis' github repository at: https://github.com/stamatak/ExaML
Andre presented the corresponding paper [Stamatakis 2013] at the International Parallel and Distributed Processing Symposium (IPDPS) in Boston.

While the ExaML search algorithm and likelihood function implementation are the same as in RAxML, ExaML implements a radically new parallelization scheme that greatly reduces parallel communication and synchronization overhead. We have abandoned the earlier master-worker approach from RAxML and RAxML-Light and now use an approach where each MPI process executes a consistent copy of the search algorithm. This has enab-

led us to dramatically reduce communication overhead, in particular for large, partitioned multi-gene datasets, where each gene evolves under a distinct statistical model of sequence evolution.

In comparison with the predecessor (RAxML-Light), this radical change in the parallelization strategy has led to a performance improvement of up to a factor of 3.2, while executing exactly the same search algorithm on 192 cores using the HITS cluster.

For an exascale scalability workshop at the Munich supercomputing center that Andre attended, we showed that ExaML scales well even beyond 4,000 cores. Furthermore, we identified ExaML performance problems with respect to the I/O of the input alignment, which currently takes too long during the startup phase of the program. We intend to address this issue in 2014.

ExaML has also served as a basis for the ExaBayes (Exasacle Bayesian inference) code that Andre and Kassian are currently working on. It deploys the same efficient

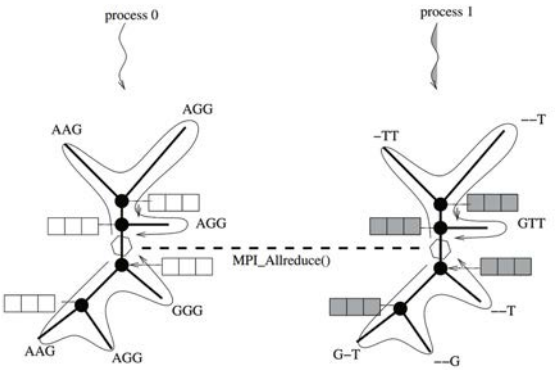


Fig. 31: Schematic outline of the new parallelization approach in ExaML.

parallelization scheme. A pleasant side-effect of this new parallelization approach is that it also substantially reduces code complexity in comparison with the old master-worker approach. We are already regularly using ExaML to infer trees on the Munich supercomputer in the framework of the 1000 Insect Transcriptome Evolution project.

SCALABLE & PARALLEL CODES FOR POPULATION GENETICS

In 2013 we also continued our work on developing and making available fast and efficient programs for analyzing population genetic data.

Alexis implemented a new model of evolution in RAxML that accommodates the so-called ascertainment bias. Without going into further details, this model allows for proper analysis of DNA datasets that entirely consist of so-called SNP sites (Single Nucleotide Polymorphisms) as typically obtained from individuals of the same population.

In early 2013, our former lab members Nikos and Pavlos finished work [Pavlidis 2013] on a more efficient, more scalable, and numerically more stable version of the widely-used SweepFinder tool for the detection of positive election that we call SweeD. The sequential version of SweeD is up to 22 times faster than SweepFinder and, more importantly, is able to analyze thousands of sequences. We have also developed parallel implementation of SweeD for multi-core processors.

In addition, we have implemented a checkpointing mechanism that enables us both to deploy SweeD on cluster systems with queue execution-time restrictions and to resume long-running analyses after processor failures. We have also integrated new statistical models into SweeD. SweeD validation was undertaken via simulations and the use of data from the 1000 Human Genomes project.

Finally, Andre finished and published his work on a tool for forward-in-time simulation of evolutionary events in populations. In population genetics, simulation is a fun-

damental tool for analyzing how basic evolutionary forces such as natural selection, recombination, and mutation shape the genetic landscape of a population. Forward simulation represents the most powerful and at the same time most compute-intensive approach to simulating the genetic material of a population.
To this end, we have developed a tool called AnA-FiTS, a highly optimized forward simulation software that is up to two or even three orders of magnitude faster than other state-of-the-art software.
These substantial performance improvements will now enable researchers to conduct forward simulations at the chromosome and genome level.

FUNDING

Our postdoc Tomas Flouri is funded by the DFG (German Research Foundation).

Andre Aberer, Jiajie Zhang, Kassian Kobert, and Alexey Kozlov are funded directly by HITS. Our visiting PhD student Paschalia Kapli was funded partly by HITS and partly by the EU and Greek national funds.

The summer school in the UK was funded by the Wellcome Trust.

OUTLOOK

We have assembled a strong team of bioinformaticians, computer scientists, biologists, and mathematicians to address the challenges that lie ahead. We intend to make use of this broad and diverse reservoir of knowledge to improve statistical models of evolution, scale up evolutionary bioinformatics algorithms, analyze and exploit emerging parallel computer architectures, and infer large phylogenies of insects in the framework of the 1KITE project. One of the key challenges will be to design even more scalable parallel codes for inferring large phylogenies on supercomputers under complex statistical models of evolution. In 2014 we intend to release and further extend our new ExaBayes code for Exascale Bayesian analysis of molecular data on supercomputers. We also intend to develop novel and scalable Bayesian tools and methods for divergence time estimation. Finally, we plan to enhance the parallel efficiency of ExaML by improving I/O performance and considering fault tolerance issues.

Our mission is to improve data storage and the search for life science data, making storage, search, and processing simple to use for domain experts who are not computer scientists. We believe that much can be learned from running actual systems and serving their users, who can then tell us what is important for them.

We employ computer science methods, notably human-computer interaction and information retrieval, as well as domain knowledge. In the past 15 years, the group has specialized in the life sciences, a domain where the data integration tasks are daunting. Here we are mainly interested in data at the boundary between “big” data and “small” data.

The strength of the group is its bench scientists (Golebiewski, Ilkavets, Kania, Krebs, Rey, Weidemann, Wittig), who have all worked at the bench in biological, biochemical, and chemical labs. This facilitates development, as developers can closely monitor the acceptance of changes through contact with domain experts (An, Bittkowski, Nguyen, Savora, Shi, Zander).

Unser Ziel ist, die Datenspeicherung und die Suche nach Life-Science Daten zu verbessern, so dass Domänenexperten, die keine Informatiker sind, Daten einfach speichern, suchen und verarbeiten können. Wir glauben, dass man durch die Ausführung realer Systeme und die Arbeit für richtige Kunden viel lernen kann. So erfahren wir, worauf Nutzer Wert legen.

Wir arbeiten mit Methoden aus der Informatik, vor allem aus den Bereichen Mensch-Computer-Interaktion und Information Retrieval, sowie mit Domänenwissen. In den vergangenen 15 Jahren hat sich die Gruppe auf Life Sciences spezialisiert, ein Fachgebiet in dem die Integration von Daten eine enorme Herausforderung darstellt. Wir sind dabei vor allem an Daten an der Grenze zwischen „big“ data und „small“ data interessiert.

Die Stärke der Gruppe sind ihre Wissenschaftler mit Laborerfahrung (Golebiewski, Ilkavets, Kania, Krebs, Rey, Weidemann, Wittig), die durch ihre frühere Arbeit in Biologie-, Biochemie- und Chemielaboren einen großen Praxisbezug haben. Dies erleichtert die Entwicklungsaufgaben, da die Software-Entwickler einen direkten Einblick haben, wie die Domänenexperten auf Veränderungen reagieren (An, Bittkowski, Nguyen, Savora, Shi, Zander).

NEW CHALLENGES

In addition to the continuation of long-term projects, the past year saw the inception of two new projects complementing the group's work between big and small, curated and non-curated data. One is the "Virtual Liver Network (VLN) Portal," a site that breaks down the many interesting properties of the liver into bite-sized articles. The other is the OperationsExplorer (we are still looking for another name), a site that not only facilitates the browsing of health data licensed from from the Statistisches Bundesamt (German Federal Statistics Office) for data journalists but is also fun to use for laypersons.

What do these projects have in common? Both are logical extensions of what we've been doing all along. The VLN portal is important in enabling the VLN to showcase the findings stored in VLN SEEK, the database and cataloguing system that we are building and running as part of VLN data management. The portal attempts to be more user-friendly than a scientific database can be. In the process, we can also learn how to improve the usability of systems like SEEK.

The OperationsExplorer also reaches out to a wide audience. Currently, it is the most visual of the tools we are building and maintaining. The problems involved in its visualization and in interactive search are well-defined. So in a technical sense, OperationsExplorer enables us to try out new things that will later be useful for SABIO-RK and SEEK. At the same time, the OperationsExplorer has already had appreciable impact; we feel that it is a tool the target community has been waiting for.

This complements our long-term projects SABIO-RK (curated kinetic data) and SEEK-based data management. In the past year, SABIO-RK's user interface has been further stabilized, and import functionality for SBML files has been etablished and improved. Our plans now turn more towards improved curation support. After years of adding many new features, SEEK is also in a stabilization phase, emphasizing incremental feature improvement over big changes. In both SEEK projects, the social aspects of scientific data management and sharing are just as much part of the equation as they have always been.

DATA MANAGEMENT FOR SYSTEMS BIOLOGY

Systems Biology is the study of dynamic processes in living systems. Modern technologies make it possible to analyze the molecular inventory of cells and the complex interplay of these components with a degree of depth and comprehensiveness that has never been attainable before. Research in systems biology sets out to draw a holistic picture of dynamic processes of life on all the different biological scales and taking into account all the different inventories, from the genome through the proteome, the transcriptome, and the metabolome, all the way up to the intricate interactions between all the individual components that generate life functions in all their variety. The data collected at all these different levels and for all the different components are the basis for complex computer models that help to better unravel the biological processes going on in cells, tissues, organs, and finally in the highly complex interplay of all these factors in the entire organism. Mathematical models of systems can be created by integrating heterogeneous biological knowledge about parameters changing in their respective contexts as a result of different temporal and environmental conditions. These models can subsequently be used to explore and predict the functioning of the systems. The modeling process is dependent on the ability to access and integrate heterogeneous data from databases and other published sources as well as directly from ongoing experimental investigations.

The data in systems biology projects are obtained from a wide variety of sources, and the methods employed are just as variegated. They have to be structured, combined, and integrated in order to make them comparable and to

construct simulatable computer models based on these data. Together with its collaboration partners, the SDBV group develops and maintains complex data management platforms that help to structure, exchange, integrate, and publish their experimental data, models, workflows, and additional information pertaining to them. The group's main focus lies in its responsibility for the data management support provided for two large-scale and distributed systems biology research networks: the German Virtual Liver Network (http://www.virtual-liver.de/) and the transnational European research network SysMO (Systems biology of MicroOrganisms). To this end, two separate versions of the web-based data management platform SEEK (http://www.seek4science.org) serve as a central hub and a main access point to the data waiting to be tapped in each of these systems biology consortium. The SEEK system was initially developed as the major SysMO assets catalogue (http://www.sysmo-db.org/seek) by the SysMO-DB team, a joint unit consisting of members of the SDBV group at HITS and of colleagues from the University of Manchester in the UK. SEEK contains information about who has what data, models, protocols, and expertise, and where those assets are to be found. It provides an access control layer enabling researchers to restrict access to collaborators, colleagues, or other individuals until they are ready to share with the whole consortium or the wider community.

The Virtual Liver Network (VLN) is a major national research initiative in systems biology funded by the German Federal Ministry for Education and Research. According to its mission statement it aims at "developing a dynamic mathematical model that represents, rather than fully replicates, human liver physiology, morphology, and function, integrating qualitative and quantitative data from all levels of organization, from sub-cellular levels to the whole organ. The main focus is on delivering a true multi-scale representation of liver physiology that helps in understanding the dynamics of liver functions in normal and diseased states." The network is made up of 70 research

The SDBV group in 2013 (f.l.t.r.): Ivan Savora, Iryna Ilkavets, Andreas Weidemann, Martin Golebiewski, Quyen Ngyuen, Meik Bittkowski, Maja Rey, Lihua An, Wolfgang Müller, Renate Kania, Ulrike Wittig, Lei Shi

Group Leader
Priv.-Doz. Dr. Wolfgang Müller

Staff members
Lihua An
Meik Bittkowski
Martin Golebiewski
Dr. Iryna Ilkavets (from Dec. 2013)
Renate Kania
Dr. Olga Krebs
Quyen Nguyen
Dr. Maja Rey (from Feb. 2013)
Ivan Savora
Lei Shi
Dr. Andreas Weidemann
Dr. Ulrike Wittig

Student
Jill Zander

groups distributed across Germany and involves about 250 researchers from both experimental and theoretical science.

The VLN project and data hub (http://seek.virtual-liver.de/) was first implemented in 2010 and has been constantly improved and extended to include new features providing better support for users. The most important features are the manifold possibilities for structuring and cross-relating corresponding information, such as raw experimental data and the standardized workflows (so called Standard Operating Procedures) leading to those results, including the specimens and samples used in the experiments. The computer models, publications and scientific presentations resulting from those experiments can also be cross-related to the original data. At the same time, the SEEK systems help VLN researchers in their collaborative ventures, not only by supporting them with resources for exchanging data and information, but also by offering them yellow pages that give details of all participating institutions and scientists and an outline of their specific expertise and the projects they are involved in. Also, internal meetings and international or national conferences or workshops can be announced as events and corresponding material (in the form of posters or presentations) can be uploaded to SEEK and linked to those events. In this way, SEEK provides substantial support for collaboration and project management in this complex and wide-ranging research network.

The focus in 2013 was dedicated to increasing user acceptance of the system. This included extensive development work to accelerate the system's performance, especially the loading speed of the webpages, while browsing or querying SEEK. The major focus during the past year, however, was on the social side, i.e. on extensive user training, which was conducted by all-hands and local hands-on data management sessions and attendance at most meetings of the Virtual Liver Network to provide data management consulting for the VLN scientists.

The most valuable user contact comes about by way of a unique group of so-called PALs, 'front-line' experts working on the project and acting both as data management advocates and multipliers. These PALs (Project Area Liaisons) are an invaluable support for the data management team at HITS, collecting and collating all the different requirements voiced by users throughout the whole Virtual Liver network.

The portal (see next section) complements the content in SEEK, making it interesting for a broader range of potential users.

VLN PORTAL – A NEW KNOWLEDGE DATABASE POPULATING SHOWCASING VLN ACHIEVEMENTS IN THE FIELD OF LIVER SCIENCE

Over the past three years, the Virtual Liver Network (VLN) has produced lots of high-quality data and results and published hundreds of interesting publications in scientific journals. Making the highlights of these scientific findings available for the non-scientific general public and to exp-

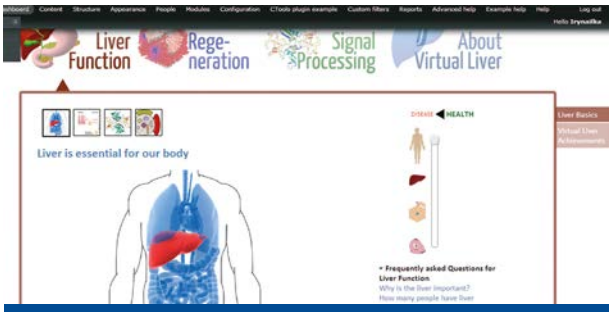


Fig. 32: Homepage of the portal. The general portal solution implemented using Drupal CMS. Navigation though the portal is provided by "facets" and thumbnails.

lain these research highlights in the wider context of the human liver and its manifold functions, we have started an initiative to create an appealing, shop-window-style public website, the “VLN Portal” (hereafter referred to as ‘the portal’). The design of the portal was developed and implemented in a Drupal content management system (CMS) to facilitate creation and editing of the content: http://portal-vl.h-its.org/.

The ultimate goal of the portal is to generate an easy-to-read and interactive website for the scientifically interested general public. This website is designed to describe the functions of the liver and how these functions can be impaired by various diseases. At the same time, the portal is meant to showcase the most important and interesting outcomes of VLN in an intelligible way. Each content page at the portal consists of a brief story with a title, images, and a text of between 50 and 200 words. The stories divide into two major groups: “basic knowledge” (general textbook knowledge) and “VLN achievements” (results from the work of scientists in the Virtual Liver Network). They are targeted at people above high-school graduation level with a general interest in scientific topics.
The portal has been structured into so-called facets:

CHAPTERS: “Liver function,” “Regeneration,” and “Signal processing”
SCALES: “Organism,” “Organ,” “Lobule,” and “Cell”
STATES: “Health” and “Disease”
SOURCES: “Basic knowledge” and “Virtual Liver achievements”

These facets can be easily accessed by different navigation elements provided at portal level: a scale slider and switches for states or sources as well as navigation elements for the general chapters and clickable thumbnails to highlight different aspects.
Several VLN members from different groups have contributed content describing their work for a broader public audience. Altogether, we have generated more than 40 content pages for the following topics: basic knowledge about the liver, bile flow, metabolism, bile acids, hepatocyte polarity, glucose metabolism, pharmacokinetics, and HGF. To guide text contributors to the portal within the Virtual Liver Network, guidelines have been compiled for the selection and the processing of the material. To enhance both the usability and the visibility of the portal for its intended audience, tests using an professional eye-tracking system have been initiated and will help to better suit the portal to users’ needs in future. In this way, the portal will ultimately function on the basis of a user-centered philosophy.
Using feedback from discussions with our project partners within VLN and with potential external users of the portal, we take account of the details reflecting the priorities and requirements of the audience. The portal is an attempt to help people to acquire valuable knowledge and familiarize themselves with hitherto unknown aspects of the liver both in a healthy and diseased state.

EXCEMPLIFY

In the last reporting period, we described the implementation of Excemplify, a web-based application developed to support experimentalists in facilitating the sharing of their data in a standardized manner via a corporate database. This improves the exchange of data between different experimentalists, makes data searchable, comparable, and exchangeable, and finally facilitates coordinated long-term storage both of the data and of corresponding descriptions of the exact conditions under which the data were generated.
This DFG-funded project was initiated with the laboratory for Systems Biology of Signal Transduction at the German Cancer Research Center (Prof. Klingmüller, DKFZ). The data are of different types, e.g. immunoblot images, Excel sheets for the description of the experimental design including cell lines and antibodies used, and as the record of numerical quantitative data. Apart from the data

storage capabilities of Excemplify, experimentalists are thus freed from time-consuming data-handling procedures and error-prone data entry. As an additional part of the data handling facilities of Excemplify, a web-based user interface has been developed for the administration of antibody repositories, replacing the Excel sheets used previously. Additionally, Excemplify was further developed, tested, and refined in collaboration with the group for Systems Biology of Signal Transduction, which finally resulted in the release of an initial production version of Excemplify at the DKFZ.

Excemplify is a tool designed for use in single workgroups. Excemplify links to SEEK for data upload and sharing beyond the confines of single work groups.

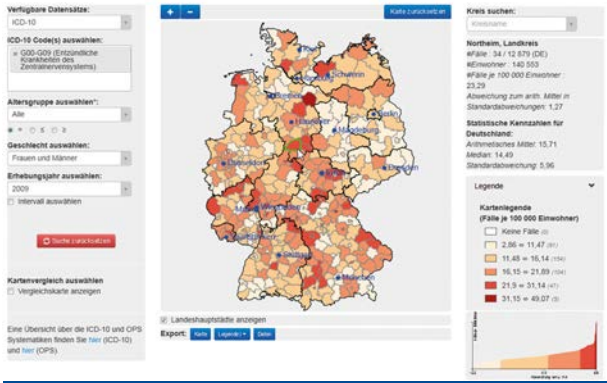


Fig. 33: The OperationsExplorer.

OPERATIONSEXPLORER

Given the huge amount of statistical data about the situation of medical care in Germany, how can one possibly find the really interesting pieces of information in it? How can one find those few nuggets of information that are so surprising that a curious person could not resist asking further questions about the data?

If the curious person asking those questions is a renowned science journalist like Volker Stollorz, those questions will be prompted by the hope that there may be a narrative, a story to be told about the underlying pattern that explains the startling data.

During his stay at HITS as first Journalist in Residence, Volker Stollorz asked us for help in interviewing big datasets. More specifically, he asked for a tool with which he could mine the sizable datasets of all diagnoses made and all operations performed and remedial procedures resorted to in all German hospitals in the course of one year. These datasets are not big data. But they are too big to be handled by Excel alone.

We were quickly able to devise a preliminary solution for Volker Stollorz' requirements. It was shown under the title "HITS Krankenatlas" at HITS Open Day and at Explore Science 2013. We then obtained funding from the Robert Bosch Foundation to expand the "Krankenatlas" into a more general solution: the OperationsExplorer.

The OperationsExplorer (see Figure 33) is an interactive single-page web application that enables the user to search for any diagnosis (ICD10) or operation/procedure (OPS) code. The search can be restricted to certain age groups, to certain years, or to a single gender. The results returned encompass the number of incidents for each code in each district of Germany and related to each district's normalized population. These values are clustered into color-encoded classes that are then used to color a map of all German districts. This map is the main resource for representing the search results to the user. But users can also export the data used to color the map directly into an Excel spreadsheet for further custom analysis.

By hovering over any district, users can view additional information about the number of incidents for that district, including additional statistical information about the total distribution of incidents in Germany. There is as yet no test for statistical significance available to the user, but there is coarse outlier detection based on interquartile

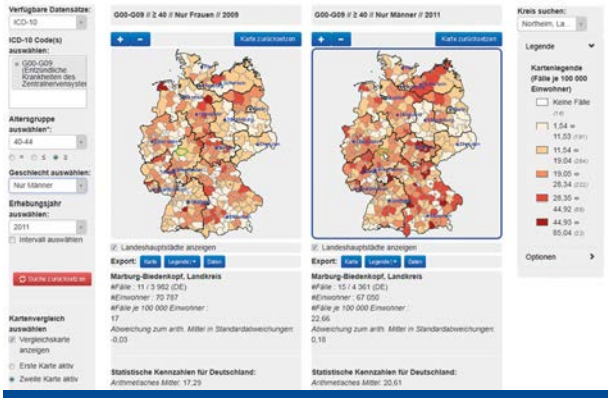


Fig. 34: The Operations Explorer displaying two maps side by side, each map representing the results for a single search.

ranges that helps users to distinguish outliers from normal values that have merely chanced into the color code of the highest class (and hence just seem to be unusual). A common-use case for data journalists like Volker Stollorz is to examine the comparative distributions of such incidents in order to get a grip on the peculiarities of the data. For that purpose, we have developed the map comparison mode of the OperationsExplorer (see Fig. 34): two maps are displayed side by side, each map representing the results for a single search.

Development of the OperationsExplorer was supported by invaluable feedback from a highly motivated group of beta testers recruited by Volker Stollorz. In November 2013, we hosted a user meeting at HITS to specify the agenda for the second half of the project. Our warmest thanks to all the beta testers!

SABIO-RK DATABASE: NEW FEATURES

SABIO-RK (http://sabio.h-its.org/) is a manually curated database containing data about biochemical reactions and their kinetic properties. Reactions are described by details about reaction participants, catalysts, and their biological sources. These data are mainly based on information reported in the scientific literature that is manually extracted and stored in a structured format. In addition, SABIO-RK also offers direct automatic data upload from lab experiments. SABIO-RK users have made an increasingly large number of requests pertaining to database content and greater flexibility for data access and standard export. Accordingly, we have improved the accessibility and usability of the database.

Beside the ongoing expansion of database content for metabolic reactions, signalling reactions or events can now also be inserted in greater detail, including extra information on this special type of reaction. For signalling reactions, it is necessary to include details on proteins (e.g. UniProtKB_AC, subunit composition, modifications) participating either as substrates/products or as reaction modifiers (e.g. inhibitors, activators, cofactors). Modification types like “phosphorylation” and the modification position within the protein can be defined. Signalling reactions/events and modification types are annotated and linked to Gene Ontology. Modification, parameters, and

Fig. 35: The beta testers at HITS: Science journalists at the user meeting with Volker Stollorz (third left) in November 2013.

kinetic-law types have now also been annotated and newly linked to Systems Biology Ontology.

As new content, the search results can now be filtered for transport reactions. In the past, the detailed information about chemical compounds already included IUPAC International Chemical Identifier (InChI) strings; from now on, InChI strings also can be used as a search option.

Beside existing cross-references from the external pathway database KEGG (Kyoto Encyclopedia of Genes and Genomes) and the chemical compound database ChEBI (Chemical Entities of Biological Interest) to SABIO-RK in 2013, UniProtKB has now started to link up with SABIO-RK on the basis of protein annotations. Currently, 1,313 reactions in KEGG, 6,654 chemical compounds in ChEBI, and 3,279 database entries in UniProtKB refer to corresponding SABIO-RK database entries, enabling external database users to obtain detailed kinetic information on whatever interests them.

There have been many user requests for the submission of experimental results directly from the laboratory or kinetic data from simulation experiments, so we are currently (beta) testing the upload of data in SBML (Systems Biology Markup Language) format to our web-based input interface. Up to now, the input interface has been used by biological experts to curate inserted literature data before inserting the data into the final SABIO-RK database. Using the same interface for automatic data upload via SBML format will accelerate the submission process and improve the flexibility of the system.

Since spreadsheet programs like Microsoft Excel are among the favourite data processing tools of our users, we have decided to provide a flexible way of exporting database search results to XLS or TSV files. As of now, users can tailor their own custom-made export format by selecting properties of the entries in the result set the user wants to export. (Fig. 36)

Fig. 36: Export of search results in table format represented as a merged SABIO-RK screenshot of the result page with selected entries for export (A) and entry point to the additional overview export page for all selected entries (B) with a screenshot of the alterable table view of results (C).

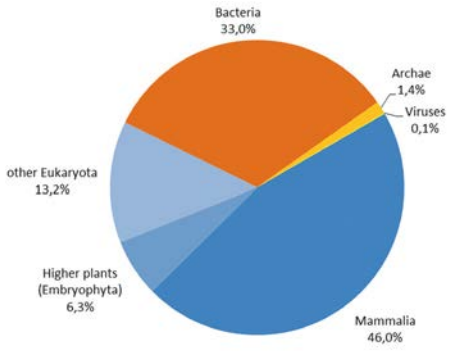


Fig. 37: Taxonomic distribution of organisms in SABIO-RK.

As of December 2013, the database contains more than 46,400 different entries related to 4,483 publications from about 220 different journals. Kinetic data are available for more than 6,100 biochemical reactions catalysed by 1,380 enzymes (represented as EC numbers) in 787 different organisms and 282 tissues or cell types. The list of kinetic parameters comprises more than 33,150 velocity constants (e.g. Vmax, kcat, rate constants), about 38,100 Km or S_half values, and more than 9,600 inhibition constants (Ki and IC50).

ISABEL ROJAS (1968-2013)

For us, 2013 was marked by the death of Isabel Rojas in July. Isabel Rojas founded our group in 1999. The group‘s name reflects her conviction that scientific databases need data exploration support if they are to be useful. From its inception, the group worked both on the modeling of biological data and on visualization schemes for graphs arising in the modeling of biological systems. In 2005 work started on SABIO-RK, which is still a central element in our research efforts. Its successful work made the group a sought-after collaboration partner and opened up the opportunity (2008) to do project data management in systems biology projects, the beginnings of what was later to become SysMO-SEEK and then SEEK. Gradually databases became more important than visualization in the group‘s work, today we see the „V“ in the group‘s name as standing for „visual exploration“.

Health reasons forced Isabel to withdraw from group leadership. Wolfgang Müller arrived as deputy leader in 2008 and took over the group one year later. Isabel continued to consult with the group, discussing changes and improvements to SABIO and participating in devising new projects like our entry in the Executable Paper grand challenge in 2011 that took us to the finals in Singapore that same year.

Isabel was an enthusiastic person full of ideas, and she had the confidence required to turn those ideas into reality. At the same time she was a team player, recognizing other people’s ideas and helping them grow. We will miss her as a leader, as a colleague, and as a friend.

In her memory, we will continue to dedicate our efforts to the same goal we set out to achieve back in 1999: simplifying the handling of complex scientific data.

2.10 Theoretical Astrophysics (TAP)

The Theoretical Astrophysics group at HITS seeks to understand the physics of cosmic structure formation over the last 13.5 billion years, from briefly after the Big Bang until today.

The group is especially interested in how galaxies and stars form and aims to constrain the properties of dark matter and dark energy, the two enigmatic matter and energy components that dominate today's cosmos. A prominent role in this work is played by numerical simulations on a variety of scales, both of the collisionless and the hydrodynamic type. To this end, group leader Volker Springel and his team members develop novel numerical schemes that can be used efficiently on very large supercomputers with the goal of exploiting them in full to link the initial conditions of the universe with its complex evolved state today. The simulation models are indispensable for the interpretation of observational data and comparison of those data with theoretical models.

Recently, the TAP group developed the novel moving-mesh code AREPO, which offers significant advantages over previous simulation techniques used in cosmic structure formation. It combines the high accuracy of traditional Eulerian hydrodynamics with the automatic adaptivity and Galilean invariance of smoothed particle hydrodynamics. The application and further improvement of this new method has been one of the priorities in the group's recent research.

Die Gruppe „Theoretische Astrophysik" am HITS versucht, die Physik der kosmischen Strukturentstehung während der letzten 13.5 Milliarden Jahre, vom Urknall bis heute, zu verstehen.

Das besondere Interesse der Gruppe gilt der Entstehung von Galaxien und Sternen, sowie einer Bestimmung der Eigenschaften der Dunklen Materie und der Dunklen Energie, jenen rätselhaften Komponenten die den heutigen Kosmos dominieren. Eine besonders wichtige Rolle in dieser Arbeit spielen numerische Simulationen auf verschiedenen Skalen. Zu diesem Zweck entwickeln Gruppenleiter Volker Springel und seine Teammitglieder neue numerische Verfahren, die effizient auf sehr großen Supercomputern eingesetzt werden können, mit dem Ziel, deren volle Kapazität für eine Verknüpfung der Anfangsbedingungen des Universums mit seinem heutigen komplexen Zustand auszunutzen. Die Simulationen sind für die Interpretation von Beobachtungen und deren Vergleich mit theoretischen Modellen unverzichtbar.

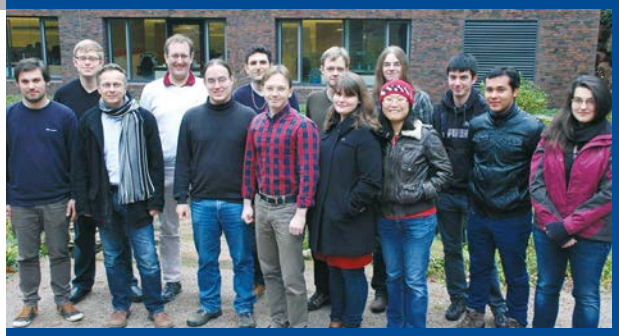
Die TAP Gruppe hat den neuen Code AREPO für bewegte Gitter entwickelt, der gegenüber früheren Simulationstechniken erhebliche Vorteile aufweist. Er verbindet die hohe Genauigkeit traditioneller Eulerscher Hydrodynamik mit der automatischen Adaptivität und Galilei-Invarianz der teilchenbasierten SPH Methode. Die Anwendung und weitere Verbesserung dieser neuen Methode bildet einen Schwerpunkt der aktuellen Forschung der Gruppe.

FORMING MILKY-WAY-LIKE GALAXIES

Due to the intrinsically multi-scale and multi-physics nature of the problem, modeling the formation and evolution of galaxies in their cosmological context is a very challenging task for numerical astrophysics. Unlike dark matter, stars and gas that make up galaxies are subject not only to gravity but also to a series of physical processes that act on much smaller spatial and temporal scales than those relevant for the galaxy as a whole. Nevertheless, these processes play a fundamental role in shaping the global features of galaxies. The complexity of this problem is further increased by the fact that any faithful representation of galactic evolution must include the full cosmological context. Indeed, in the standard cosmological framework, interactions between galaxies are the norm, and they can profoundly change the evolutionary histories of these systems. Fig. 38 shows an example of such an interaction in cosmological simulations [Marinacci et al. 2014].

These challenges culminate in studies of the formation and evolution of galaxies similar to the Milky Way (the so-called late-type galaxies). For decades, even obtaining objects featuring extended star-forming discs - the distinctive morphological feature of late-type galaxies - remained an elusive goal. Only recently, thanks to a combination of a better understanding of the conditions that lead to disc formation and a more sophisticated treatment of baryonic physics, cosmological simulations have started to produce disc galaxies with properties that are in reasonable agreement with observational features.

During 2013 the TAP group was particularly active in this research area. Much of the work focused on a suite of cosmological hydrodynamic simulations of eight haloes selected for total mass within the mass range estimated for the Milky Way. The simulations were designed with the aim of following the formation and the evolution of the galaxies inside these haloes so as to make progress in



The TAP group in 2013 (f.l.t.r.): Federico Marinacci, Christopher Hayward, Volker Springel, Christoph Pfrommer, Rüdiger Pakmor, Kevin Schaal, Denis Yurin, Andreas Bauer, Christine Simpson, Christian Arnold, Dandan Xu, Dominik Steinhauser, Juan Carlos Basto Pineda, Jolanta Krzyszkowska

Group Leader
Prof. Dr. Volker Springel

Staff members
Andreas Bauer
Dr. Fabio Fontanot (until Aug. 2013)
Dr. Christopher Hayward
Dr. Federico Marinacci
Dr. Rüdiger Pakmor
Dr. Christoph Pfrommer
Dr. Ewald Puchwein (until Sept. 2013)
Kevin Schaal (from May 2013)
Dr. Christine Simpson (from Sept. 2013)
Denis Yurin
Dr. Dandan Xu (from June 2013)

Visiting scientists
Juan Carlos Basto Pineda (from April 2013)
Martin Spare (from June - Nov. 2013)
Dominik Steinhauser (from Nov.- Dec. 2013)

Students
Christian Arnold (from March 2013)
Jolanta Krzyszkowska (from Oct. 2013)
Vitali Pauz (until Sept. 2013)

producing realistic late-type discs. All the runs were performed with our moving-mesh code AREPO, using the so-called zoom-in technique that permits high-resolution tracking of the evolution of the central halo while the rest of the cosmological volume is sampled with a resolution that degrades with increasing distance from the region of interest. This saves computer time without sacrificing accuracy. It was the first time we employed the AREPO code in combination with a novel comprehensive model for baryonic physics, including all the most relevant physical processes, such as metal cooling, a self-consistent treatment of stellar evolution and the associated mass and metal return to the gaseous phase, galactic winds, and black-hole feedback.

Our predicted galaxy morphologies are presented in Fig. 39. The images show edge-on and face-on views of stellar density projections for all the systems at the present time. They show that in most of the haloes (apart from Aq-E) our simulations are able to produce galaxies with extended stellar disc components. The colors of the images are directly related to the photometry of the stellar populations comprising the galaxies, giving a visual impression of stellar ages. Old stellar populations appear in red, while young populations are blue. The presence of a blue component in the vast majority of the systems indicates that their discs are still actively star-forming, another important feature common to all late-type galaxies. Other key properties of the simulated galaxies, such as masses, luminosities, sizes, and kinematics, agree very well with observed scaling relations.

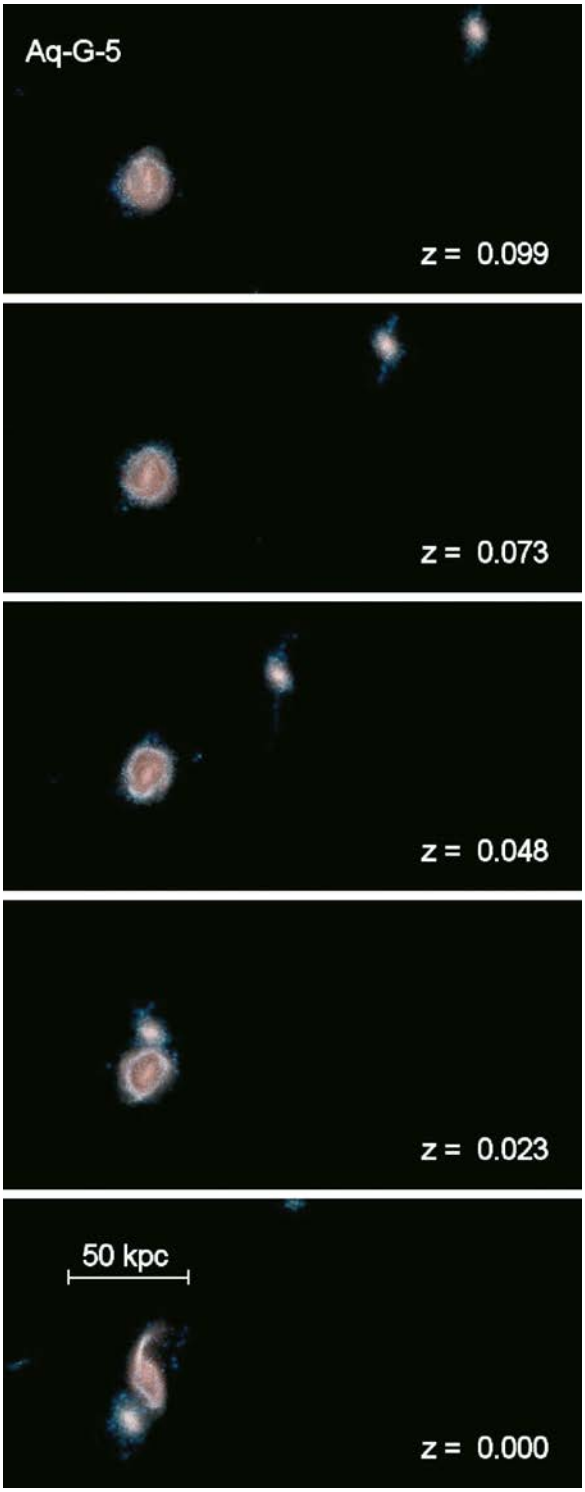
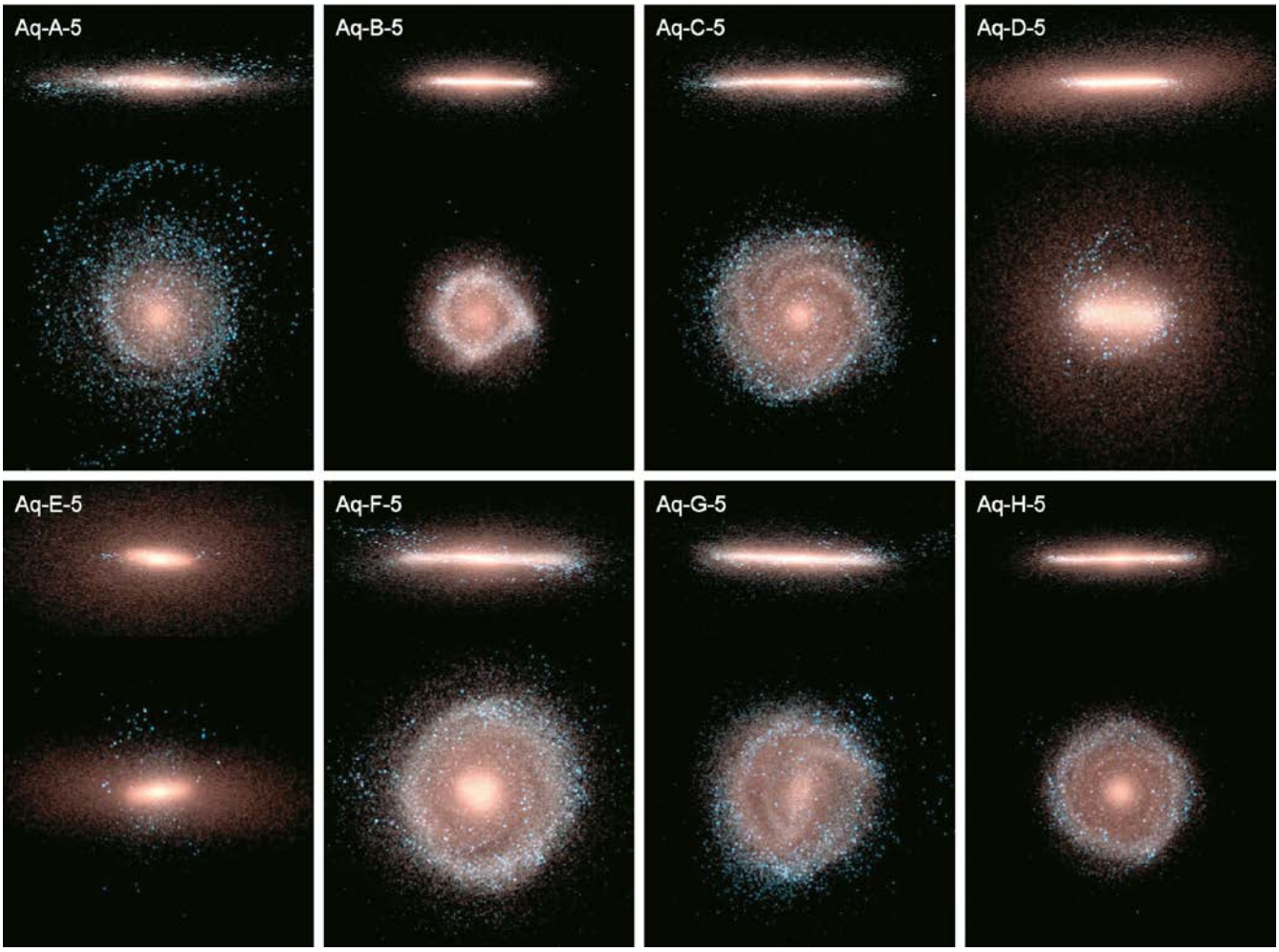


Fig. 38: An example of how interactions between galaxies shape their evolution. The figure shows the time evolution of an encounter between the central galaxy of the Aq-G halo and a massive satellite. Note how the stellar disc of the central galaxy is distorted by the encounter.

In the near future we plan to extend this simulation set by including additional physical phenomena (for example magnetic fields) and by increasing the maximum resolution to an unprecedented level in order to better resolve the interstellar medium. This strategy will enable us to achieve a more complete and faithful understanding of the physical processes involved in galaxy formation and to make a decisive step forward in advancing our theoretical understanding of the Universe.

Fig. 39: Stellar density projection maps for all the systems at the final time of our simulations. Each panel features an edge-on (top) and a face-on (bottom) view of the stellar component of the resulting galaxy. Colors in the images correspond to the actual emission in different bands of the stellar populations within the galactic disc.



THE ORIGIN OF COSMIC EXPLOSIONS

Supernovae are very luminous transient events. For a few weeks they rival their host galaxy in brightness and release about the same energy in light as a billion suns before they fade away. Human records of supernovae extend as far back as 185 AD. Today, we know that they are gigantic explosions of stars that release a vast amount of energy on a timescale of only a few seconds, turning a star into a huge explosion and producing most of the heavy elements in the Universe.

One particular class of supernovae, so-called type Ia supernovae, is of particular interest for cosmology because their absolute brightness can be inferred from the time they take to become fainter. This means that we can use brightness measurements to calculate the distance to the supernova and its host galaxy very accurately. In 1999, two independent teams of astronomers using this method showed that the expansion of the Universe is accelerating.

We know that type Ia supernovae are thermonuclear explosions of white dwarf stars made from carbon and oxygen which are inert remnants of stars slightly more massive than the Sun. However, we still do not know what causes them to explode. Since they are inert objects, the explosion must be caused by some interaction with another star, usually assumed to be in a binary system with the white dwarf.

It was thought for a long time that for theoretical reasons a second white dwarf eventually merging with the primary dwarf after a long inspiral phase due to the emission of gravitational waves could be excluded as the companion star (it was thought that the system never ignited nuclear fusion). However, this scenario has been revived in recent years, both by new theoretical results and by new observational data, notably by the nearest SN Ia (SN 2011fe) discovered in recent history. This supernova places strong constraints on the nature of the companion star, excluding almost all alternatives to a system of two white dwarfs as progenitors of the explosion.

Back in 2012 we were able to show that a merger of two white dwarfs ignited during the merger itself was fully able to reproduce observational data of the kind generated by type Ia supernovae. However, it was not clear whether the conditions in the simulation during the merger are actually sufficient to ignite nuclear fusion. Now we were able to repeat this simulation with the new moving-mesh code AREPO, which enables us to conduct the simulation more accurately and at much higher resolution. In particular, we were able for the first time to resolve the thin helium shells on top of carbon-oxygen white dwarfs that they retained after all other material was burned to carbon and oxygen. Since helium has a much smaller charge than carbon and oxygen, the temperature needed to overcome their repulsion and to ignite helium is significantly lower. In our simulation, we were able to show that the temperature in the helium shell skyrockets due to the interaction with material accreted from the secondary white dwarf to the primary white dwarf, such that nuclear fusion in the helium starts. Following nuclear burning in the code (see Fig. 40), we showed that helium burning propagates supersonically as a detonation around the primary white dwarf on its surface. During burning, the detonation sends a shockwave to the core from its current position. Since the shockwave propagates through the core at a slightly slower velocity than the detonation around it, the shockwave converges on a single point in the outer parts of the core opposite to the initial ignition point (see Fig. 41).

This convergence then increases the temperature and density in a small part of the carbon-oxygen core to such an extent that the repulsion of carbon and oxygen can be overcome, and they also start a nuclear fusion process. This second detonation in the core will then burn all the

core completely, mostly to heavy elements like iron, and release so much energy that the white dwarf becomes unbound and explodes as a type Ia supernova. Thus we were able to show that merging white dwarf binaries of a sufficient mass should naturally be expected to ignite their helium shells first, and that the burning of the helium shell will cause a converging shock in the core of the primary white dwarf that will ignite the core and lead to a type Ia supernova.

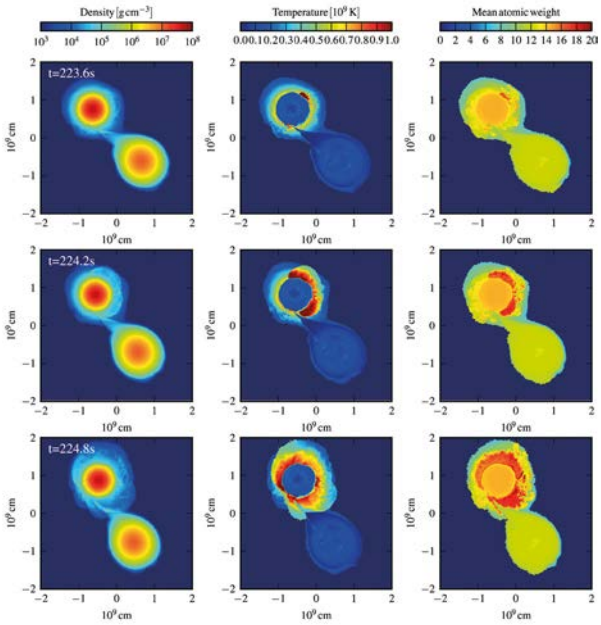


Fig. 40: Time evolution of the density, temperature, and mean atomic weight (columns from left to right) for our simulation of the merger of a 1.1 with a 0.9 Msun white dwarf. The first row shows the simulation about 4 minutes after mass transfer from the secondary to the binary has started. On the top right part of the surface of the primary white dwarf, the helium has become so hot that it ignites nuclear fusion. The nuclear burning front then propagates as a detonation around the primary white dwarf, burning all the helium on the surface.

THE COOLING FLOW PROBLEM IN GALAXY CLUSTERS

Galaxy clusters are the largest gravitationally collapsed objects in the Universe and have formed in very recent cosmic history. Most of the ordinary matter in clusters is

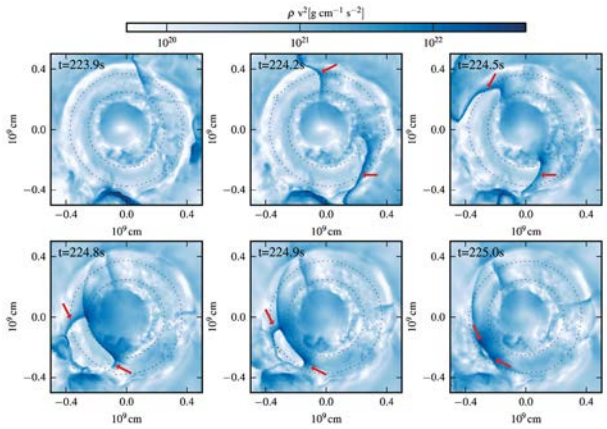


Fig. 41: Propagation of the shock in the primary white dwarf caused by the detonation burning the helium on its surface shown by the kinetic energy density. The shock propagates through the center of the white dwarf and is continuously sent from the surface to the core by the detonation. It converges opposite to the point of ignition but still within the core of the white dwarf. At the convergence point, it will increase the temperature and density of the carbon-oxygen mixture to such an extent that it triggers nuclear burning that leads to the explosion of the whole white dwarf, the type Ia supernova.

composed of hot gas that emits in X-rays, thereby cooling the gas. Cooling gas becomes denser, increasing the cooling rate further. This is a run-away process that should result in many stars and cold gas at the cluster centers. Astonishingly, the predicted amount of cold material has not been observed. This is the famous „cooling flow problem." Over the last decades, it has become clear that cluster centers host super-massive black holes of sizes exceeding billions of solar masses. When a small amount of gas cools, it falls to the center and accretes onto the super-massive black hole. However, a fraction of it gets ejected in the form of powerful outflows of very energetic particles (electrons and protons). These push the ambient, X-ray-emitting gas away to create spectacular, radio-emitting lobes - potentially representing the signposts of a self-regulated feedback loop. While the total available energy in these lobes is more than enough to offset the cooling, it is far from clear whether this is the long-sought-after solution to the „cooling flow problem" and if so, how exactly this heating process works. Moreover, the gas displays a central temperature floor that previously suggested models have failed to explain.

New observations of the closest galaxy cluster, which is situated in the constellation of Virgo, have enabled us for the first time to put forward a comprehensive model for the physical heating mechanism involved in this long-standing problem [Pfrommer 2013]. Low-frequency radio emission traces out an aged population of energetic electrons, thus enabling a glimpse at the distant past of the feedback cycle. However, observations by the European low-frequency radio interferometer LOFAR revealed the absence of fossil energetic electrons in the radio halo surrounding the center of the Virgo cluster (Fig. 42). This puzzle can be resolved by accounting for the release of these energetic electrons from the radio lobes and subsequent mixing with the dense ambient intracluster gas. As a result, the energetic electrons thermalize on a timescale similar to the radio halo age of 40 million years, hiding an aged electron population from LOFAR's „radio eyes".

However, this picture also implies the release of energetic protons from the lobes that will inevitably interact hadronically with the ambient gas to produce an observable gamma-ray signal. To our great surprise, such a signal has been detected toward Virgo by the gamma-ray observatories Fermi and H.E.S.S. The signal shows spectral characteristics that match our expectations from the ra-

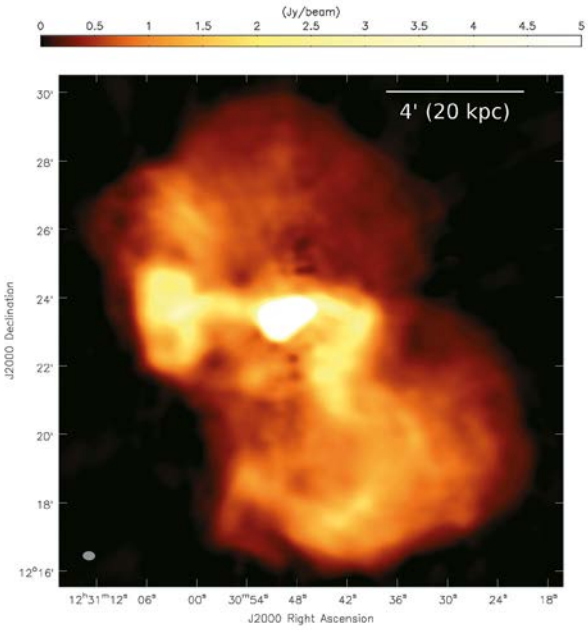


Fig. 42: The LOFAR radio image of the center of the Virgo cluster [de Gasperin et al. 2012] shows the energetic outflow from the center, the bright radio lobes, and the fainter diffuse radio halo surrounding them. The radio halo emission fills the entire cooling region and exactly resembles the radio image at high frequencies, which implies the absence of fossil electrons. Moreover, it indicates that energetic particles are distributed in a volume-filling manner, which is a necessary condition if they are to heat the cooling cluster gas.

dio results, enabling us to count the number of energetic protons responsible for the observed gamma-ray signal.

These energetic protons are bound to move along magnetic fields and to excite magnetic waves. Dissipation of these waves heats the surrounding thermal plasma at a rate that scales with the amount of energetic protons. It turns out that the amount of energetic protons required to explain the gamma-ray emission yields a heating rate that balances that of average radiative cooling at each radius (see Fig. 43). However, the resulting global thermal equilibrium is locally unstable, and its thermal instability permits the formation of the observed multi-phase medium. Provided that this heating process balances cooling during the emergent „cooling flow," the collapse of the

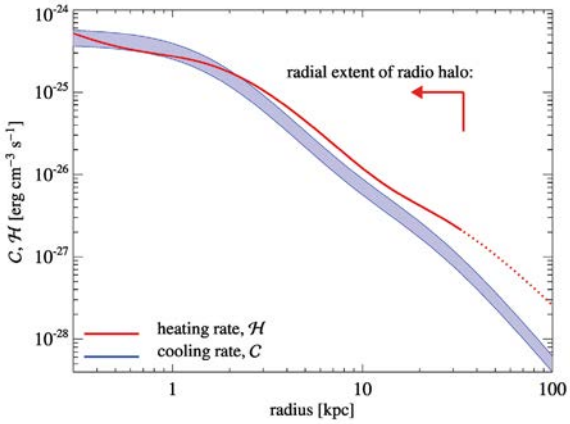


Fig. 43: Cooling vs. heating rates in the Virgo cluster as a function of cluster-centric radius. The radiative cooling rates (blue) are globally balanced at each radius by the heating due to energetic protons (red), suggesting a physical solution to the „cooling flow problem." Note that the heating rates become very uncertain outside the boundary of the radio halo (dotted red)

majority of the gas is halted at a temperature floor that is in accordance with the X-ray data. We show that both the existence of a temperature floor and the similar radial scaling of the heating and cooling rates are generic results of the heating model described. This model makes several predictions that will be tested with future observations. If successful, this will have profound implications for our understanding of the evolution of galaxy clusters and star formation in the central cluster galaxies.

FOLLOWING THE UNIVERSE IN MODIFIED GRAVITY THEORIES

Roughly 15 years ago, astrophysicists discovered that the expansion of the Universe is accelerating. This finding, referred to earlier, was completely unexpected as it contradicts the theoretical notion that gravity is always attractive and should slow down the expansion over time. In 2011, the Nobel Prize for physics was awarded for this discovery. The accelerating expansion of the Universe has since been firmly established by observations. But we still do not know why this should be the case.

Models proposed to explain accelerating expansion can be divided into two main categories. In the first case, accelerating expansion is obtained within the framework of Einstein's general relativity theory by proposing the existence of a mysterious form of energy that permeates the Universe and has a strong negative pressure. The density of this „dark energy" would have to exceed the rest of the mass energy density of all known contents of the Universe in order to explain the observation of accelerating expansion.

In the second scenario, accelerating expansion is not caused by otherwise invisible „dark energy" but is a consequence of a change in the laws of gravity, i.e. in general relativity theory. In this case, no additional form of energy

is needed. However, Einstein's field equations describing gravity need to be altered. This is quite a delicate issue. General relativity has been confirmed with a high degree of accuracy by experiments on Earth and within our solar system. Any changes to it would thus have to be effected in such a way that the numerous successes of the theory on solar-system and smaller scales are preserved, while at the same time yielding accelerating expansion of the Universe on larger scales. Recent theoretical studies indicate that this is indeed possible. Models of this type have been dubbed „screened modified gravity," as they hide the effects of modified gravity in comparably dense environments such as the solar system. To confirm or rule out such models, detailed theoretical predictions of what a universe with such modified gravity would look like are required, so that a comparison with observations can be carried out. Computer simulations of the formation of

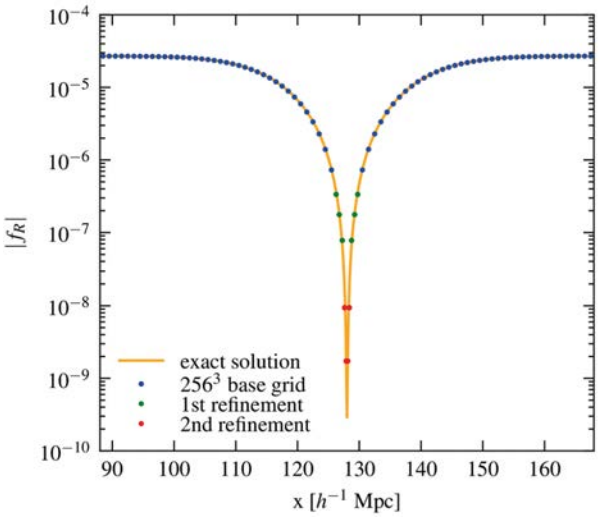


Fig. 44: The accuracy of the non-linear partial differential equation solver is tested for a 1D density peak. For this simple configuration, the analytic solution is known (orange line). The circles indicate the numerical solution that our code computes on an adaptively refined mesh. Blue corresponds to the base grid of the mesh, while green and red indicate the solutions obtained at the first and second grid-refinement levels. The analytic and numerical solutions are in excellent agreement.

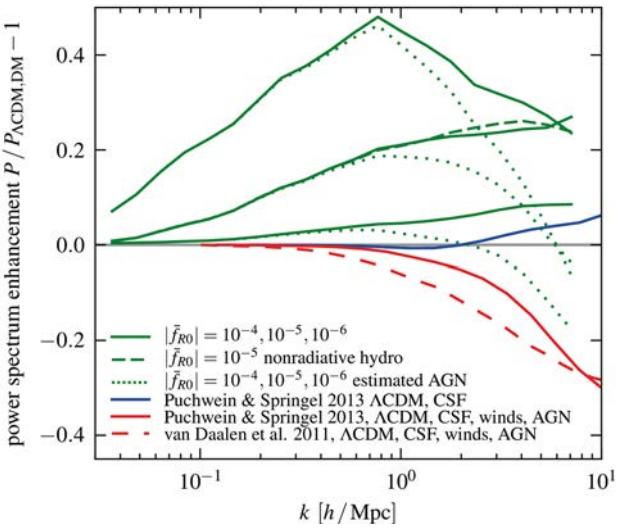


Fig. 45: The effects of modified gravity and baryonic physics on the matter distribution as a function of scale, i.e. on the power spectrum of the density field. Green solid lines indicate the relative change of the matter power spectrum due to modified gravity for three different models. The blue line shows how radiative cooling and star formation affect the matter distribution compared to a simulation that ignores gas physics. The simulations indicated by the red lines also account for energy injection due to accretion onto supermassive black holes. It is important to account for both at the same time in order to predict observable signatures of modified gravity theories.

structure in the universe are the tool of choice for this. When studying the formation of cosmic structures like galaxies in general relativity, it is normally adequate to operate within the limits of weak gravitational fields. In this case, Einstein‘s field equations can be replaced by Newton‘s much simpler law of gravity. Thus, when performing computer simulations of cosmic structure formation with standard gravity, it is sufficient to solve a relatively simple, linear partial differential equation.

By contrast, the equations of gravity are highly non-linear in screened modified gravity, even within the limits of weak fields. This makes numerical simulations of such models much more challenging. To be able to perform such studies, we have written a new numerical code that enables us to efficiently solve the partial differential equations arising in such models. Our new code solves the non-linear equations by an iterative relaxation technique on a mesh that adaptively refines on the forming structures. It employs a multi-grid approach, meaning that the large-scale features of the solution are quickly found on a relatively coarse grid, while the computationally more extensive calculations on high-resolution grids are only needed to recover the small-scale features. The code is highly parallel so that the power of modern supercomputers can be used.

Our new code „MG-GADGET“ has been written as a module for the widely used GADGET code. This enables us to simultaneously follow both modified gravity and a wide range of additional baryonic astrophysics included in the latter code, such as the radiative cooling of gas or the formation of stars. We have already used MG-GADGET for a number of studies on how modified gravity affects the properties of galaxy populations and galaxy clusters, as well as on the degeneracies that exist between the effects of modified gravity and uncertainties in baryonic and neutrino physics (see Fig. 44 and Fig. 45).

3 Centralized Services

3.1 Administrative Services

In 2013 the activities of the HITS Admin group can best be summarized under the heading “consolidation.” This was the main focus of the year’s work with respect to personnel, organization, structure, and a number of technical issues.

The installation of four new research groups and one technical group (ITS) made up a significant part of the workload for the HITS administration in 2013. For each new group, a budget had to be defined that meets the group‘s specific requirements. Furthermore, office space had to be allocated, and all the necessary technical resources had to be made available. For all groups personnel had to be hired or transferred from other institutions. In a number of cases, projects and other activities had to be relocated from the new group leaders’ previous organizations, so it was necessary to make sure that all the responsibilities were assigned correctly, projects properly re-adjusted, etc. Moving third-party projects across organizational boundaries proved to be a particularly complicated and time-consuming undertaking. But looking back, we can confirm that the procedures for resource planning, budget allocation etc. developed during the last couple of years proved to be perfectly adequate to those tasks, so the new groups had what one might call a running start and were fully integrated into the institute within a very short time.

One major project of a completely different type involved the administration group and the ITS group (as well as some other companies belonging to the KT group). We had to take stock of all the software licenses from a specific large software vendor - as part of a routine inspection initiated by that vendor. This was the first real-life test of our inventory system (except for the annual inventory control routines), and we are happy to report that there was an almost perfect match between the software installed on a large variety of devices and the licenses registered with the vendor. The outcome convinced even those members of the institute who have been heard to say that inventory control and all the procedures related to it are an unnecessary administrative impediment.

This year the administration group completed the three projects it had been conducting over the previous two years. The outcomes were very different.

The online application system is now fully operational. During the beta-test phase we only encountered minor problems, and the only adjustments needed had to do with clearly specifying the different documentation requirements for different types of application (scholarship holders, scientific staff member, postdoc, etc.). The system is now being used regularly for all kinds of job openings, and the gains in terms of greater effectiveness and reliability for the application processes are considerable. In addition, it is now much easier to assure adherence to the restrictions imposed by privacy law than with traditional paper-based procedures. Thanks to the modular structure of the system, it can easily be adapted to new and/or changing requirements and scaled to any number of applications HITS may have to deal with in the future.

The second project had to do with handling travel-expense claims. The key idea is to free the staff members from the necessity of entering all the details of their trips into an online system for further processing, because this requires some familiarity with that system and even then can be quite time-consuming. Instead, people only need to collect all the tickets, receipts etc. pertaining to their trips; the rest is done by an external service provider specializ-ing in administering business trips. The new scheme has now been in use for about two years and is considered to be a great improvement over the previous procedure.

Our third project was aimed at installing a web-based system for project planning, controlling, and reporting, with a special focus on the reporting needs of the various types of third-party projects (funded, for example, by the German Research Foundation (DFG), the Federal Ministry of Education and Research (BMBF), and the EU). We collaborated closely with other research organizations and

system developers and invested well over one person-year in testing and evaluating various approaches. In the end, we decided to suspend this activity for the time being, because it turned out that none of the solutions available fulfilled the essential requirements in all areas. The gaps that would have to be filled by resorting to other tools were so legion that, all things considered, we would not have achieved any simplification of processes or reductions in the workload. We will, however, be watching this space very closely and starting another round of evaluations as soon as we have identified new solutions that may potentially fill the bill in this respect.

We started out by saying that the big word for the reporting year was consolidation. This did not only happen by way of "natural transition," it was also forced on us by an unscheduled event. The manager of the administration group, Bärbel Mack-Reuter, left HITS in 2013. She had established, structured, and directed the group and its precursors more or less since the very first days of EML Research. She defined the internal procedures, the workflows, and assignments of responsibility, and her primary goal – which she firmly emphasized in many different situations – was to build an administrational service group that supports the scientists at HITS so they can focus on what they are really good at: research. The fact that we were able to handle all the routine operations plus the additional challenges in 2013 in an effective way despite the absence of the group leader proves that everything had been extremely well designed and all the group members were perfectly prepared to live up to their respective responsibilities. HITS is deeply indebted to Bärbel Mack-Reuter for the great work she contributed towards the overall goal of setting up a first-class research institute.

f.l.t.r.: Ingrid Kräling, Rebekka Riehl, Kerstin Nicolai (1st row), Andreas Reuter, Christina Bölk-Krosta, Benedicta Frech, Stefanie Szymorek, Christina Blach (2nd row).

Group Leader
Bärbel Mack-Reuter (until Sept 2013)
Prof. Dr. Andreas Reuter (acting, since Oct 2013)

Group members
Christina Blach I Office
Christina Bölk-Krosta I Controlling
Benedicta Frech I Office
Kornelia Gorisch I Assistant to Managing Director (until Feb 2013)
Ingrid Kräling I Controlling
Kerstin Nicolai I Controlling
Rebekka Riehl I Human Resources and Assistant to Managing Director (since May 2013)
Stefanie Szymorek I Human Resources

3.2 IT Infrastructure and Network

In July 2013, the IT Services (ITS) group took over responsibility for all IT-related issues from the Scientific Computing group. It strives to provide a coherent set of IT services centered on the needs of the other groups at HITS. The main focus is the management of the HITS computer center and network, plus administration and maintenance of all HITS computer systems. Another important aspect is user support and education, ranging from solving day-to-day computer-related problems to organizing internal workshops on the efficient usage of the HPC cluster or introducing new technologies. A scientific software developer and a web applications developer will be joining ITS in 2014, thus widening the range of services we can offer to the other groups.

From the outset we have had to tackle the difficult challenge of providing the expected services to an increasing number of scientific groups and HITS members. This has been reflected in the growing amount of issues reported in the ticket system or in face-to-face discussions. It has also manifested itself in the increasing number of computer systems to be maintained: the automation of simple or frequently performed administration tasks has become even more important than before. The additional workstations, laptops, and VoIP phones have required reorganization of the research group networks and an increase in the number of network ports. To improve performance and add new security features, the firewall protecting the HITS network has also undergone hardware replacement and a software update.

As in many academic and commercial institutions worldwide, the data storage requirements of the scientific groups at HITS have increased significantly over the past few years, especially in 2013. To cope with the demand, we have significantly extended the capacity of the RAID systems serving the workstations. We have also introduced two FhGFS-based storage systems designed to hold data produced on our HPC cluster, supplementing the existing Panasas distributed storage system. FhGFS is a high-performance parallel file system from the Fraunhofer Competence Center for High Performance Computing in Kaiserslautern, Germany. With its outstanding scalability and flexibility, it has enabled to supply close to 1 petabyte of very fast storage to the HPC cluster users, while at the same time providing the potential for simple and cost-effective expansion in the future. Due to its compatibility with POSIX standards, it was also easy to integrate into the existing cluster infrastructure and could immediately be used for the available applications. We have also taken advantage of the possibility of exporting FhGFS via NFS, making it directly accessible to the workstations and significantly reducing the need to copy data to and from the cluster.

To stay in line with the increase in main storage capacity, we have extended the backup system. In the first few months of 2014, there will be further extension and reorganization to decrease the time required for copying the data and to improve the efficiency of deduplication.

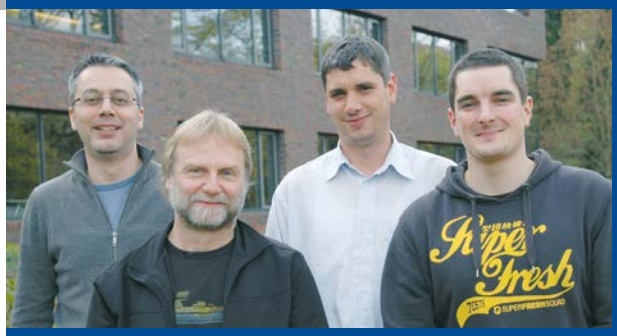
Long-term storage of scientific data provides access to research results beyond the limits of individual projects. Following a test phase in the first part of 2013, HITS has started to use the archiving facilities at the Karlsruhe Institute of Technology. A new 10 Gb/s connection to the German National Research and Education Network (DFN) has been installed and optimized for large transfers resulting from archiving operations and restoration of data.

Most of the scientific data at HITS is generated on the HPC cluster, for which the initial design focused on CPU-intensive calculations. Alongside access to the very fast FhGFS storage systems, the cluster was expanded in 2013 with nodes containing large amounts of internal memory and based on the latest “Ivy Bridge” CPU technology from Intel. These measures have not only enhanced overall cluster performance but have also catered specifically to the needs of the new groups joining

the institute. Many bioinformatics calculations operate on large in-memory data sets, while data-mining jobs require access to vast amounts of data both in-memory and on external storage. We have also installed Apache Hadoop, a popular framework for scalable distributed computing, which will be used by several scientific groups for data-intensive research.

Most of the central network services are hosted in a virtual datacenter environment based on VMWare ESX and DataCore storage, which we operate in a high-availability setup together with the other institutions located on our campus. Part of the infrastructure is located in the HITS building, the other part in Villa Bosch. During 2013, the physical servers, storage shelves, and Fibre Channel switches were upgraded, significantly improving performance and enabling us to manage the virtual infrastructure more effectively.

The central systems managing user accounts, authentication, e-mail, and various other network services have been updated to the latest release of the Microsoft Windows Server (2012R2), increasing performance and security. This update also provides the basis for further improvements related to data backup, e-mail, and user authentication planned for 2014. To cope with the increase in the number of HITS members and support organizational changes, the e-mail structure is now hierarchical, providing greater clarity and facilitating management. HITS passed a comprehensive license audit by Microsoft in the first part of 2013, which it passed successfully. The audit took into consideration the whole range of Microsoft products, including those installed in virtual environments. As the scientists at HITS have strong connections and, in some cases, appointments with other educational institutions that benefit from an academic licensing scheme, clean separation between the license types is called for. The audit led to some improvements in the way the licenses for Microsoft products are managed internally, so that we can respond more efficiently to any similar audits in the future.



The ITS group in 2013 (f.l.t.r.): Ion Bogdan Costescu, Norbert Rabes, Christian Goll, Andreas Ulrich

Group Leader
Dr. Ion Bogdan Costescu

Staff Members
Dr. Christian Goll | System Administrator
Norbert Rabes | System Administrator
Andreas Ulrich | System Administrator

HITS is one of the organizers of the Heidelberg Laureate Forum (HLF), a yearly event assembling both young and renowned researchers in the fields of mathematics and computer science. We provided technical support throughout the preparatory stages and during the 1st HLF in 2013. In particular, ITS participated in the recording and live-streaming of the sessions, thus sharing the presentations and discussions outside the physical limits of the conference venue.

As we have seen, plans for next year are already afoot, and others can be expected to materialize in the natural course of events. The IT industry never sleeps! Finally, and most importantly, we will be facing new challenges in 2014 as we try to meet the demands associated with ongoing progress in the computational and data-driven research that the HITS scientists are very much a part of.

The year 2013 was marked by new experiences for HITS communications, too: the second leg of "Journalist in Residence," four new research groups, and new, additional communication tasks.

An important project for HITS is the "Journalist in Residence" program. It is addressed to science journalists and offers them a paid sojourn at HITS. During their stay they can learn more about data-driven science and get to know researchers and research topics in more detail and without pressure from the "daily grind." TV journalist Pia Grzesiak was the second journalist to rise to the bait, staying at the institute from April to July. In this period, she produced two films about topics related to data-driven research: research on blood coagulation, to which Frauke Gräter (MBM) has made important contributions, and the calculation of phylogenetic trees in developmental biology, a field in which Alexandros Stamatakis (SCO) has been working very successfully. Both TV contributions were broadcast in the German scientific TV show "nano." Like her predecessor Volker Stollorz, Pia Grzesiak also held a lecture for the general public that attracted a lot of interest. Under the title "Tabloidization of a Mass Medium," she gave a carefully researched assessment of science journalism on television.

Volker Stollorz' stay at HITS led to a collaboration with Wolfgang Müller and the SDBV group. In their joint project "OperationsExplorer," they developed a data tool for medical journalists. The project was funded by the Robert Bosch Foundation in the framework of its "Innovation in Science Journalism" program. Volker Stollorz and Meik Bittkowski (SDBV) presented a prototype of the tool at "Explore Science" (see Chapter 5.4) and at "Wissenswerte," the annual conference of German-speaking science journalists in Bremen (see Fig. 47).

In summer, HITS showcased the "Journalist in Residence" program for the first time in an international setting, presenting it at the World Conference of Science Journalists (WCSJ) in Helsinki at the end of June. The response was positive, with 36 journalists from 22 countries subsequently applying for a stay at HITS. In December 2013, the jury, consisting of scientists and chief editors of scientific journals, chose Michele Catanzaro as the next "Journalist in Residence." The Barcelona-based science journalist works freelance for media in Italy, Spain, Mexi-

Fig. 46: Science through the camera lens: Pia Grzesiak (left) with Dirk Weiler, director of photography, while shooting the movie with Frauke Gräter.

Fig. 47: Volker Stollorz presenting the OperationsExplorer at "Wissenswerte," the annual conference of German-speaking science journalists in Bremen, in late November 2013.

co, and Great Britain. He will be starting his sabbatical at HITS in the latter half of 2014.
At HITS he will get to know at least ten research groups, including the four new ones that started work in 2013 (see chapter 2). They increase the variety of research topics at HITS by covering new mathematical and astronomical aspects, as well as issues in developmental biology. The groups deal with interesting and forward-looking topics from data-driven science like regenerative medicine, technological improvements in operating rooms, and the probable accuracy of weather forecasts.
Also in summer, HITS intensified its social media activities. In addition to the existing Youtube channel, the institute now shares information about research news with the public on Twitter and Facebook. The invitation for applications to qualify for the "Journalist in Residence" program, for example, was mainly distributed via the social media; Twitter in particular is intensively used by journalists.

To cope with these additional communication tasks, Isabel Hartmann joined the communications department in November 2013. She had previously worked as an intern at HITS and helped establish and enhance HITS presence in the social media.
This year again, research by the "HITSters" spawned scientific publications that received an enthusiastic response from professionals and the general public. Frauke Gräter (MBM) and her colleague Gustavo Caetano-Anollés (University of Illinois at Urbana-Champaign) found out that the speed of protein folding plays a key role in the evolutionary history of proteins. In this connection, they analyzed 100,000 proteins and 1,000 genomes. The "protein origami" story attracted international interest and was taken up by media in Europe, the U.S., and Japan.
At HITS Open House Day on June 8, the general public had an opportunity to find out more about research at HITS with simulations, films, lectures, and hands-on activities for children (see Chapter 5.3).

The HITS Communications team in 2013 (f.l.t.r.): Juliane Repp, Peter Saueressig, Isabel Hartmann

Head of Communications
Dr. Peter Saueressig

Members
Isabel Hartmann I Public Relations Assistant (since November 2013, after an internship from May to October)
Katsiaryna Sazonava M.A. I Intern (until April 2013)
Natalie Achims BA I Student (until February 2013)
Juliane Repp BA I Student (since June 2013)

Communications

5 Events

5.1 Conferences, Workshops & Courses

5.1.1 INTERNATIONAL BIOCURATION CONFERENCE

April 7-10, 2013 Cambridge, UK

Organiz ng Committee:
Alex Bateman (European Bioinformatics Institute, UK), Claire O'Donovan (European Bioinformatics Institute, UK), Mike Cherry (Stanford University, USA), Jen Harrow (Wellcome Trust Sanger Institute, UK), Valerie Wood (University of Cambridge, UK), Elspeth Bruford (European Bioinformatics Institute, UK), Renate Kania (Heidelberg Institute for Theoretical Studies, Germany), Raja Mazumder (George Washington University, USA), Sandra Orchard, (European Bioinformatics Institute, UK), Monica C. Munoz-Torres (Lawrence Berkeley National Laboratory, USA), Susan Tweedie (European Bioinformatics Institute, UK), Francis Ouellette (Ontario Institute for Cancer Research, Canada), Kimmen Sjölander (University of California, Berkeley, USA), Peter McQuilton (University of Cambridge, UK)

The 6th International Biocuration Conference brought together over 300 scientists to exchange news about their work and discuss issues relevant to the International Society for Biocuration's (ISB) mission. The conference was organized around seven sessions and nine workshops. Two poster sessions were held with over 70 posters at each event. The Best Poster Awards went to Scott Cain (gold) for his GMOD poster, Jerven Bolleman (silver) for the Catching inconsistencies in UniProtKB/Siss-Prot with the semantic web, and Siew-Yit Yong (bronze) for her poster about the InterPro web site - refreshed, revamped, refined. Ewan Birney, Fiona Brinkman, and Nobel Prize winner Sir Richard Cotton gave stimulating plenary lectures. Among the topics covered were annotation issues, data integration, ontologies, and data mining. The conference series is an essential part of the ISB's goal to support exchanges among members of the biocuration community. The meeting provided a forum for biocurators and developers of biological databases to discuss their work, promote collaboration, and foster a sense of community in this very active and quick-growing research area. Participants from academia, government, and industry interested in the methods and tools employed in the curation of biological data were encouraged to attend. The conference included four plenary talks by established scientists who have influenced the field of biocuration.

HITS Research Associate **Renate Kania** was among the conference organizers. She is one of the founders of the International Society of Biocurators (ISB) and a member of the ISB Executive Committee.

5.1.2 ADVANCED COURSE IN COMPUTATIONAL MOLECULAR EVOLUTION

April 29 – May 10, 2013, Hinxton, Cambridge /UK

The molecular evolutionary analysis of sequence data is crucial to effectively tackle the challenges of the increasing amount of biological sequence data. Accordingly, it is essential to provide researchers with the theoretical knowledge and practical skills they need to do so. To this end, the 5th summer school on computational molecular evolution was organized on the Wellcome Trust campus in Hinxton, UK. **Alexis Stamatakis**, group leader of the HITS research group Scientific Computing, was one of four organizers of the event, which took place from 29 April to 10 May 2013. The course dealt with the basic principles of sequence data analysis and cutting-edge methodologies, thus making it an important event for experts and beginners alike. The multi-faceted program covered topics like uses and interpretations of molecular phylo-

genies, sequence alignments, and genomics resources, Markov models of sequence evolution, phylogeny reconstruction, hypothesis testing in molecular phylogenetics and evolution, coalescent models, and inference from population data. Forty researchers participated in the summer school. SCO postdoc **Tomáš Flouri** and PhD student **Fernando Izquierdo-Carrasco** also took part in the summer school as teaching assistants.

5.1.3 AREPO WORKSHOP

September 25-27, 2013, Heidelberg

The international AREPO workshop was held at the Heidelberg Academy of Science. Volker Springel from the Theoretical Astrophysics group and his team organized the event in order to bring together current users and developers of the AREPO moving-mesh code, as well as people interested in analysis projects of the very large cosmological simulations carried out by the group and its collaborators, in particular Prof. Lars Hernquist's group at the Harvard-Smithsonian Center for Astrophysics. AREPO is a powerful hydrodynamical simulation code for cosmic structure formation whose development is led by the HITS Theoretical Astrophysics group. The workshop focused on scientific applications of this software and also on the question how to further advance the numerical algorithms and the physical modelling capabilities of the code.

More than thirty scientists, including several HITS Astrophysicists, held short presentations for about fifteen to twenty minutes, interleaved with very lively discussion sessions. The topics of these talks ranged from current simulation results on galaxy formation over small-scale physics applications to future code development directions. The workshop proved highly successful in initiating a large number of new scientific projects as well as in discussing and establishing new collaboration guidelines for the AREPO code.

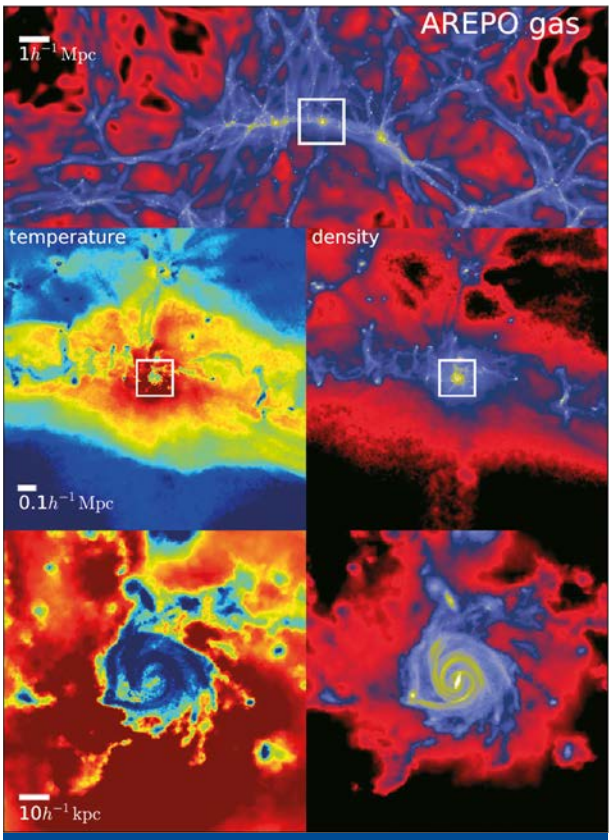


Fig. 48: A simulation generated with the AREPO code. Picture: Vogelsberger et al., 2013, MNRAS, 436, 3031.

5.1.4 BIOLOGICAL DIFFUSION AND BROWNIAN DYNAMICS BRAINSTORM 3

7-9 October, 2013, Heidelberg/Germany

Studio Villa Bosch, Heidelberg with live video-conferencing sessions every evening (Heidelberg)/morning (UC San Diego)
Workshop organizers: Franziska Matthäus (Bioquant, Heidelberg University), Rommie Amaro (UC San Diego), Rebecca Wade, Daria Kokh, Stefan Richter, Jon Fuller, and Xiaofeng Yu (all MCM/HITS).

Following on from the two previous BDBDB meetings, BDBDB3 provided a forum for intensive discussion about the state of the art in Brownian Dynamics simulations of biological macromolecules and related methodologies. The topics covered included measurements of macromolecular diffusion, calculation of binding-rate constants, membrane diffusion and protein association, nano-particles and surface-protein interactions, macromolecular crowding and confinement, hydrodynamic interactions, macromolecular dynamics and flexibility, and approaches to multi-scale simulation. From 7-9 October the workshop in Heidelberg brought together about 50 theoreticians and experimentalists from around the world plus a further 10 scientists in San Diego participating via live video-conferencing sessions.

Among the sponsors were the BIOMS Heidelberg, the National Biomedical Computation Resource (NCBR), San Diego, and HITS.

Fig. 49: The participants of BDBDB3 during a coffee break in the Villa Bosch garden.

Dr. Francisco Kitaura
Leibniz Institute for Astrophysics Potsdam (AIP)
January 28, 2013: Unveiling the Initial Conditions and Structure of the Universe

Dr. Kai Polsterer
Astronomisches Institut der Ruhr-Universität Bochum
February 21, 2013: Machine Learning in Astronomy

Dr. Lisa Kaltenegger
Max-Planck-Institut für Astronomie, Heidelberg
April 22, 2013: Super-Earths and Life - an interdisciplinary puzzle

Dr. Christian Kirches
IWR Universität Heidelberg
May 27, 2013: New numerical and algorithmic approaches to mixed-integer combinatorial problems in industrial process control

Pia Grześiak
Freie Wissenschaftsjournalistin, Journalist in Residence am HITS
June 3, 2013:
Wissenschaft im Fernsehen – Boulevardisierung eines Bildungsmediums?

Prof. Dr. G. Ulrich Nienhaus
KIT Institut für Angewandte Physik
July 15, 2013: Optical Nanoscopy of Biomolecular Structure and Dynamics

Prof. Dr. Wilfred F. van Gunsteren
ETH Zürich. Lab. für Physikalische Chemie
October 21, 2013: Calculation and measurement of protein stability: methodological issues and structural and thermodynamic aspects

Prof. Dr. Thomas Scheibel
Fakultät für Angewandte Naturwissenschaften, Universität Bayreuth
November 18, 2013: Engineering, processing and applications of structural proteins: about spider silk, mussel byssus and lacewing eggstalks

5.2 Colloquia

Dr. Paul Maragakis
D.E. Shaw Research, New York
December 2, 2013:
Early steps in targeting “undruggable” proteins using long molecular dynamics simulations

Debate: The Standard Cosmological Model

The Standard Cosmological Model was the subject of a panel discussion on June 19, 2013 in Peterskirche, Heidelberg. **Alexander Unżcker, Volker Springel, Wolfgang Kundt,** and **Joachim Wambsganß** exchanged views on Dark Matter, Dark Energy and the question of how stable the foundations of the Standard Model actually are. Moderator was **Ulf von Rauchhaupt** (FAZ). HITS and the Klaus Tschira Foundation were the joint organizers of the event. The discussion was recorded on video and can be viewed on the HITS Youtube channel (www.youtube.com/TheHITSters).

On 8 June 2013 HITS once again opened its doors to the public. Many visitors from Heidelberg and beyond seized the opportunity to get an inside view of the institute and its research on this sunny Saturday. Seven research groups presented their work ranging from galaxies, molecules, and genomes to computational linguistics, and databases. In addition to meeting the HITS researchers, visitors could also listen to two talks by Volker Springel, leader of the Theoretical Astrophysics group. In his first talk about "Astrophysics with the Supercomputer," he explained how important supercomputers are for researchers who want to explore the Universe and enlarged on the scientific knowledge that has been discovered with this technology. His second talk called "How Astronomers Look Back in Time" was targeted specially at the younger members of the audience. Both talks were very well received and went on much longer than scheduled due to the large number of questions from the audience.

The Open Day event was planned to be fun for the whole family. Some of the HITS scientists improvised hands-on areas for children to not only entertain them but also show them how fascinating science can be. MBM member Dr. Agnieszka Bronowska created an outdoor station demonstrating what non-Newtonian fluids are and how they behave under pressure. The scientists used a mixture made of corn starch, water, and food coloring which is usually fluid but starts hardening under pressure. Another hands-on station devised by Dr. Stefan Richter of the MCM group showed how soapy water finds the shortest distance between a number of points. Both events astonished children and their parents equally.

During the whole day there were half-hourly guided tours through the HITS building and the HITS computer center, which were both crowded to capacity. For the rest of the time, visitors could talk to the scientists, enjoy the scenic environment, and have some pizza fresh from the oven or some cake.

Impressions from the Open House day on 8 June, 2013.

5.4 Explore Science

This year, HITS participated once again in Explore Science, the hands-on science event organized by the Klaus Tschira Foundation in Mannheim's Luisenpark. The topic for 2013 was "Fascination Earth!", and the event took place from June 26 to 30. Together with scientists from Heidelberg University's Institute of Computer Science, the HITSters manned three hands-on exhibits demonstrating how spatial context is made easier to grasp by visualization via digital maps.

Two of the hands-on exhibits were created by the database specialists of the SDBV group. The station called "HITS Cards" was a puzzle demonstrating how Google Maps is able to display so many maps of different scales in a short time. The scientists explained that the Maps Service uses a tile system that only loads the data required by the user. The tiles were represented by a self-made puzzle showing maps of the world in different degrees of resolution. The children could test their knowledge of geography and additionally find out about all the data used to create digital maps. After that, they could create their own personal map on a computer showing the Luisenpark from above and mark the spots they wanted to visit next.

The other exhibit created by SDBV was devoted to the "OperationsExplorer" project they developed in conjunction with former HITS Journalist in Residence Volker Stollorz. The children discovered that maps not only display geographical content but also information about things like diseases. Where do get people get sick? What diseases do they have? Are men or women more seriously affected? What is the average age for measles? The database created by the SDBV group uses real statistical data to display the incidence of illness on a map. The visitors were keen to know more about the map and asked a lot of questions.

Along much the same lines, Prof. Michael Gertz and his students from the Heidelberg University developed an interactive map showing the web activities of social media networks. Gertz and his team had collected and analyzed data from the social network Twitter, and with reference to a map of San Francisco bay, they were able to show who tweeted about what in that area. This project was especially popular among the teenagers, who frequently use various social media platforms and enjoyed filtering the map for different keywords like Alcatraz, beach, or restaurant.

Impressions from "Explore Science" 2013.

In 2013 the Heidelberg Laureate Forum first saw the light of day. The Forum is to be established as an annual meeting bringing together the winners of the most prestigious scientific awards in Mathematics (Abel Prize, Fields Medal), Computer Science (Turing Award), and Theoretical Computer Science (Nevanlinna Prize) with a selected group of highly talented young researchers. The Forum has been initiated by the Klaus Tschira Foundation and the Heidelberg Institute for Theoretical Studies. It is modeled on the annual Lindau Meeting for Nobel Laureates established more than 60 years ago.

The first Heidelberg Laureate Forum was held from September 23 to 27 in various places in and around Heidelberg. Thirty-eight laureates and 200 young researchers had accepted the invitation to the first meeting of this kind. During their five days in Heidelberg, they were treated to a full-scale program including not only panel discussions and talks given by the laureates but also group excursions and dinners, so that the young researchers and laureates had ample opportunity to get together and exchange ideas. The participants went on outings to the residence in Schwetzingen and Heidelberg Castle, they enjoyed a boat trip along the Neckar and even had their own Oktoberfest at the University cafeteria.

On one of the days, the young researchers were invited to visit various scientific institutes in Heidelberg. HITS welcomed 20 of them and also Turing Award (1999) winner Fred Brooks, who is renowned for landmark contributions to computer architecture, operating systems, and software engineering. At HITS some of our scientists held presentations about their work and engaged in lively discussions with the young researchers.

The next Heidelberg Laureate Forum will take place from 21 to 26 September 2014.

At the opening ceremony of the Heidelberg Laureate Forum in Germany´s oldest university (picture: HLF).

Fred Brooks during the discussion with the young researchers and HITSters in the Alan Turing room at HITS.

Group picture: The young researchers at the end of their short stay at HITS.

6 Cooperation

As the number of research groups increases, cooperation with other scientific organizations is becoming more intense and more diverse. In the following, we briefly describe some of them, covering both institutional and project-based collaborations. The list is not meant to be complete, but it does attempt to convey the scope and depth of our joint activities with partners in the scientific community.

Heidelberg University

Starting with the framework agreement between the University and EML Research signed in 2007, areas of common interest have expanded in many directions. There are the jointly appointed professors Andreas Reuter, Volker Springel, and Rebecca Wade; Michael Strube holds an honorary professorship. All of them (plus some other HITSters) regularly teach courses at the University, and they participate in a variety of projects and research programs: Collaborative Research Center (SFB/Transregio) "The Dark Universe"; interdisciplinary research program "Modeling and Simulation in the Bio-Sciences" (BIOMS); doctoral program "Semantic Processing"; Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences, and many others.

Through the associated HITS research group headed by Prof. Vincent Heuveline, who is scientific director of the University Computing Center, we expect to develop a strategic cooperation venture with the University aimed at building an infrastructure for scientific computing.

Karlsruhe Institute of Technology (KIT)

Following the collaboration contract between HITS and KIT, the second joint professorship was established in 2013. Prof. Tilmann Gneiting accepted the offer of a full professorship in Computational Statistics. The chair is part of the Faculty of Mathematics at KIT and is combined with a HITS research group.

Further activities have been prepared throughout the year and are now ready for implementation in 2014.

European Southern Observatory (ESO) Garching

The research group "Theoretical Astrophysics" (TAP) maintains close collaborative links with the European Southern Observatory. The Gadget and AREPO codes, developed and supported by Volker Springel, are used by various groups at ESO. Volker Springel and a leading supernova expert at ESO, Dr. Bruno Leibundgut, are members of the DFG-funded Transregio Collaborative Research Center (TRR33) "The Dark Universe." These collaborative ventures are perfect examples of the (indispensable) exchange between theoreticians on the one hand and observers gathering data on the other.

This long-standing cooperation has also been instrumental in creating a joint project between the Klaus Tschira Foundation and ESO for building a planetarium and outreach center at the ESO headquarters in Garching.

MPI Dresden

The new junior research group "Computational Biology" (CBI) was established in close cooperation with Prof. Eugene (Gene) Myers at the Max Planck Institute for Molecular Biology and Genetics (MPI-CBG) in Dresden. Gene Myers is one of the pioneers in the field of algorithms for sequence analysis; his BLAST code is still the most widely used tool in that area. When HITS decided to buy two sequencing machines using a new technological approach, it was agreed that those machines would be operated and supported by Gene Myers' lab in Dresden. The interesting thing about the new machines is that the sequence genome fragments are about two orders of magnitude longer that what was possible before. The downside is an error rate that is about one order of magnitude higher than that of "standard" sequencers. In principle, the new machines will facilitate so-called de novo assembly, but because of the higher error rates new methods for genome assembly are required. Developing and optimizing algorithms for that purpose is the main task of the CBI group headed by Dr. Siegfried Schloissnig. Gene Myers is supporting the group as mentor, contributing his vast fund of experience in the field of genome assembly.

KAIST

HITS has a formal exchange agreement with the Korea Advanced Institute of Science and Technology (KAIST). The partner on our side is the research group “Computational Linguistics” (NLP) headed by Prof. Michael Strube; the partner at KAIST is the Division of Web Science and Technology headed by Prof. Sung-Hyon Myaeng. The contract defines a framework that enables scientists from both sides to spend time in each other’s labs working on topics of mutual interest in computational linguistics and related areas.

University of Illinois at Urbana-Champaign

The research group “Molecular Biomechanics” (MBM) headed by Prof. Frauke Gräter is involved in many international collaborations. One of them, with the Institute for Genomic Biology (University of Illinois at Urbana-Champaign) under Prof. Gustavo Caetano-Anollés, is particularly intensive – and particularly ambitious. The goal is to understand, from first principles, how life began, how the genetic code came into being, and how the various evolutionary mechanisms influence and control each other.

7 Publications

[Aberer et al. 2013] Andre J. Aberer and Alexandros Stamatakis. Rapid forward-in-time simulation at the chromosome and genome level. BMC Bioinformatics, 14(1):216, 2013.

[Ackermann et al. 2013] The Fermig-LAT Collaboration: M. Ackermann, M. Ajello, A. Albert, A. Allafort, W. B. Atwood, L. Baldini, J. Ballet, G. Barbiellini, D. Bastieri, K. Bechtol, R. Bellazzini, E. D. Bloom, E. Bonamente, E. Bottacini, T. J. Brandt, J. Bregeon, M. Brigida, P. Bruel, R. Buehler, S. Buson, G. A. Caliandro, R. A. Cameron, P. A. Caraveo, E. Cavazzuti, R. C. G. Chaves, J. Chiang, G. Chiaro, S. Ciprini, R. Claus, J. Cohen-Tanugi, J. Conrad, F. D'Ammando, A. de Angelis, F. de Palma, C. D. Dermer, S. W. Digel, P. S. Drell, A. Drlica-Wagner, C. Favuzzi, A. Franckowiak, S. Funk, P. Fusco, F. Gargano, D. Gasparrini, S. Germani, N. Giglietto, F. Giordano, M. Giroletti, G. Godfrey, G. A. Gomez-Vargas, I. A. Grenier, S. Guiriec, M. Gustafsson, D. Hadasch, M. Hayashida, J. Hewitt, R. E. Hughes, T. E. Jeltema, G. Johannesson, A. S. Johnson, T. Kamae, J. Kataoka, J. Kn□odlseder, M. Kuss, J. Lande, S. Larsson, L. Latronico, M. Llena Garde, F. Longo, F. Loparco, M. N. Lovellette, P. Lubrano, M. Mayer, M. N. Mazziotta, J. E. McEnery, P. F. Michelson, W. Mitthumsiri, T. Mizuno, M. E. Monzani, A. Morselli, I. V. Moskalenko, S. Murgia, R. Nemmen, E. Nuss, T. Ohsugi, M. Orienti, E. Orlando, J. F. Ormes, J. S. Perkins, M. Pesce-Rollins, F. Piron, G. Pivato, S. Raino, R. Rando, M. Razzano, S. Razzaque, A. Reimer, O. Reimer, J. Ruan, M. Sanchez-Conde, A. Schulz, C. Sgro, E. J. Siskind, G. Spandre, P. Spinelli, E. Storm, A. W. Strong, D. J. Suson, H. Takahashi, J. G. Thayer, J. B. Thayer, D. J. Thompson, L. Tibaldo, M. Tinivella, D. F. Torres, E. Troja, Y. Uchiyama, T. L. Usher, J. Vandenbroucke, G. Vianello, V. Vitale, B. L. Winer, K. S. Wood, S. Zimmer, C. Pfrommer, and A. Pinzke. Search for cosmic-ray induced gamma-ray emission in Galaxy Clusters. ArXiv e-prints, 2013. eprint: 1308.5654.

[Alachiotis et al. 2013] Nikolaos Alachiotis, Simon Berger, Tomáš Flouri, Solon P. Pissis, and Alexandros Stamatakis. libgapmis: extending short-read alignments. BMC Bioinformatics, 14(Suppl 11):S4, 2013.

[Alachiotis et al. 2013] Nikolaos Alachiotis, Emmanouella Vogiatzi, Pavlos Pavlidis, and Alexandros Stamatakis. Chromatogate: A tool for detecting base mis-calls in multiple sequence alignments by semi-automatic chromatogram inspection. Computational and Structural Biotechnology Journal, 6, 2013.

[Angulo et al. 2013] R. E. Angulo, S. D. M. White, V. Springel, and B. Henriques. Galaxy formation on the largest scales: The impact of astrophysics on the BAO peak. ArXiv e-prints, 2013. eprint: 1311.7100.

[Arnold et al. 2013] C. Arnold, E. Puchwein, and V. Springel. Scaling relations and mass bias in hydrodynamical f(R) gravity simulations of galaxy clusters. ArXiv e-prints, 2013. eprint: 1311.5560.

[Balbo et al. 2013] Jessica Balbo, Paolo Mereghetti, Dirk-Peter Herten and Rebecca C. Wade. The shape of protein crowders is a major determinant of protein diffusion. Biophys. J., (2013) 104, pages 1576-1584.

[Baldi et al. 2013] M. Baldi, F. Villaescusa-Navarro, M. Viel, E. Puchwein, V. Springel, and L. Moscardini. Cosmic Degeneracies I: Joint N-body Simulations of Modied Gravity and Massive Neutrinos. ArXiv e-prints, 2013. eprint: 1311.2588.

[Barton et al. 2013] Carl Barton, Tomáš Flouri, Costas S. Iliopoulos, and Solon P. Pissis. Gapsmis: flexible sequence alignment with a bounded number of gaps. In Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, page 402. ACM, 2013.

[Barton et al. 2013] Carl Barton, Mathieu Giraud, Costas S. Iliopoulos, Thierry Lecroq, Laurent Mouchard, and Solon P. Pissis. Querying highly similar sequences. International journal of computational biology and drug design, 6(1), pages 119-130, 2013.

[Barton et al. 2013] Carl Barton, Costas S. Iliopoulos, Inbok Lee, Laurent Mouchard, Kunsoo Park, and Solon P. Pissis. Extending alignments with k-mismatches and i-gaps. Theoretical Computer Science, 2013.

[Barton et al. 2013] Carl Barton, Costas S. Iliopoulos, and Solon P. Pissis. Circular string matching revisited. Invited contributions, page 200, 2013.

[Battaglia et al. 2013] N. Battaglia, J. R. Bond, C. Pfrommer, and J. L. Sievers. On the Cluster Physics of Sunyaev-Zel'dovich and X-Ray Surveys. III. Measurement Biases and Cosmological Evolution of Gas and Stellar Mass Fractions. Astrophysical Journal, 777:123, 2013.

[Berynskyy 2013] Mykhaylo Berynskyy and Rebecca C. Wade. Treating conformational flexibility in protein-protein docking. Curr. Phys. Chem., (2013) 3, pages 27-35.

[Bird et al. 2013] S. Bird, M. Vogelsberger, D. Sijacki, M. Zaldarriaga, V. Springel, and L. Hernquist. Moving mesh cosmology: properties of neutral hydrogen in absorption. Monthly Notices of the Royal Astronomical Society, 429, pages 3341-3352, 2013.

[Brahmkshatriya et al. 2013] P. S. Brahmkshatriya, P. Dobeš, J. Fanfrlik, J. Rezáç, K. Paruch, A. K. Bronowska, M. Lepšík and P. Hobza. Quantum mechanical scoring: structural and energetic insights into cyclin-dependent kinase 2 inhibition by pyrazolo[1,5-a]pyrimidines.

Curr Comput Aided Drug Des, 9(1), pages 118-129, 2013.

[Bringmann et al. 2013] T. Bringmann and C. Pfrommer. Bringmann and Pfrommer.
Reply. Physical Review Letters, 111(19):199002, 2013.

[Broderick et al. 2013a] A. E. Broderick, C. Pfrommer, E. Puchwein, and P. Chang. Implications of Plasma Beam Instabilities for the Statistics of the Fermi Hard Gamma-ray Blazars and the Origin of the Extragalactic Gamma-Ray Background. ArXiv e-prints, 2013a. eprint: 1308.0340.

[Broderick et al. 2013b] A. E. Broderick, C. Pfrommer, E. Puchwein, and P. Chang. Lower Limits upon the Anisotropy of the Extragalactic Gamma-Ray Background implied by the 2FGL and 1FHL Catalogs. ArXiv e-prints, 2013b. eprint: 1308.0015.

[Büchel et al. 2013] Finja Büchel, Nicolas Rodriguez, Neil Swainston, Clemens Wrzodek, Tobias Czauderna, Roland Keller, Florian Mittag, Michael Schubert, Mihai Glont, Martin Golebiewski, Martijn van Iersel, Sarah Keating, Matthias Rall, Michael Wybrow, Henning Hermjakob, Michael Hucka, Douglas. B. Kell, Wolfgang Müller, Pedro Mendes, Andreas Zell, Claudine Chaouiya, Julio Saez-Rodriguez, Falk Schreiber, Camille Laibe, Andreas Dräger, and Nicolas Le Novère (2013). Path2Models: large-scale generation of computational models from biochemical pathway maps. BMC Systems Biology, 7(1):116, November 2013.

[Costescu 2013] I. B. Costescu and F. Gräter. Time-resolved force distribution analysis.
BMC Biophys (2013) 6(5), pages 1-5.

[Crochemore et al. 2013] Maxime Crochemore, Costas S. Iliopoulos, Tomasz Kociumaka, Marcin Kubica, Alessio Langiu, Solon P. Pissis, Jakub Radoszewski, Wojciech Rytter, and Tomasz Waleń. Order-preserving incomplete suffix trees and order-preserving indexes. In String Processing and Information Retrieval, pages 84-95. Springer, 2013.

[Dao et al. 2013] David Dao, Tomáš Flouri, and Alexandros Stamatakis. Automated plausibility analysis of large phylogenies. John Wiley&Sons, 2013.

[Darriba et al. 2013] D. Darriba, A. Aberer, T. Flouri, T.A. Heath, F. Izquierdo-Carrasco, and A. Stamatakis. Boosting the performance of bayesian divergence time estimation with the phylogenetic likelihood library. In Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing Workshops PhD Forum (IPDPSW), pages 539-548, 2013.

[Debès et al. 2013] C. Debès, M. Wang, G. Caetano-Anollés and F. Gräter. Evolutionary optimization of protein folding. PLoS Comput. Biol (2013) 9(1), pages 1-9.

[Eskevich et al. 2013] Maria Eskevich, Gareth J.F. Jones, Shu Chen, Robin Aly, Roeland Ordelman, Danish Nadeem, Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot, Tom De Nies, Pedro Debevere, Rik Van de Walle, Petra Galuščáková, Pavel Pecina, and Martha Larson. Multimedia information seeking through search and hyperlinking. In Proceedings of the ACM International Conference on Multimedia Retrieval, Dallas, Tex., 16-19 April 2013, pages 287-294, 2013.

[Fahrni et al. 2013] Angela Fahrni, Thierry Göckel, and Michael Strube. HITS' monolingual and cross-lingual entity linking system at TAC 2012: A joint approach. In Proceedings of the Text Analysis Conference, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 5-6 November 2012, 2013.

[Fang et al. 2013] Hai Fang, Matt E. Oates, Ralph B. Pethica, Jenny M. Greenwood, Adam J. Sardar, Owen J.L. Rackham, Philip C.J. Donoghue, Alexandros Stamatakis, David A. de Lima Morais, and Julian Gough. A daily-updated tree of (sequenced) life as a reference for genome research. Scientific Reports, 3, 2013.

[Feldman-Salit et al. 2013] Anna Feldman-Salit, Silvio Hering, Hanan L. Messhia, Nadine Veith, Vlad Cojocaru, Antje Sieg, Hans V. Westerhoff, Bernd Kriekemeyer, Rebecca C. Wade and Tomas Fiedler. Regulation of the activity of lactate dehydrogenases from four lactic acid bacteria. J. Biol. Chem., 288:21295-306, 2013.

[Ferrari et al. 2013] Stefania Ferrari, Marco Ingrami, Fabrizia Soragni, Rebecca C. Wade and M. Paola Costi. Ligand-based discovery of n-(1,3-dioxo-1h,3h-benzo[de] isochromen-5-yl)-carboxamide and sulfonamide derivatives as thymidylate synthase inhibitors.

Bioorg Med Chem Lett., (2013) 23, pages 663-668.

[Ferreras et al. 2013] I. Ferreras, R. Sharples, J. S. Dunlop, A. Pasquali, F. La Barbera, A. Vazdekis, S. Khochfar, M. Cropper, A. Cimatti, M. Cirasuolo, R. Bower, J. Brinchmann, B. Burningham, M. Cappellari, S. Charlot, C. J. Conselice, E. Daddi, E. K. Grebel, R. Ivison, M. J. Jarvis, D. Kawata, R. C. Kennicutt, T. Kitching, O. Lahav, R. Maiolino, M. J. Page, R. F. Peletier, A. Pontzen, J. Silk, V. Springel, M. Sullivan, I. Trujillo, and G. Wright. Chronos: A NIR spectroscopic galaxy survey. From the formation of galaxies to the peak of activity. ArXiv e-prints, 2013. eprint: 1306.6333.

[Flouri et al. 2013a] Tomáš Flouri, Jan Janoušek, Bořlvoj Melichar, and Solon P. Pissis. Tree template matching in ranked ordered trees by pushdown automata. Journal of Discrete, 17, pages 15-23, Journal of Discrete Algorithms, Volume 17, December, 2012
Pages 15-23.

[Flouri et al. 2013b] Tomáš Flouri, Costas S. Iliopoulos, Tomasz Kociumaka, Solon P. Pissis, Simon J. Puglisi, W.F. Smyth, and Wojciech Tyczyński. Enhanced string covering. Theoretical Computer Science, 506, pages 102-114, 2013.

[Fontanot et al. 2013a] F. Fontanot, E. Puchwein, V. Springel, and D. Bianchi. Semi-analytic galaxy formation in f(R)-gravity cosmologies. Monthly Notices of the Royal Astronomical Society, 436, pages 2672-2679, 2013.

[Fontanot et al. 2013b] F. Fontanot, G. De Lucia, A. J. Benson, P. Monaco, and M. Boylan-Kolchin. A Research Note on the Implementation of Star Formation and Stellar Feedback in Semi-Analytic Models. ArXiv e-prints, 2013. eprint: 1301.4220.

[Fraternali et al. 2013] F. Fraternali, A. Marasco, F. Marinacci, and J. Binney. Ionized Absorbers as Evidence for Supernova-driven Cooling of the Lower Galactic Corona. Astrophysical Journal, 764:L21, 2013.

[Fuller et al. 2013] Jonathan C. Fuller, Pierre Khoueiry, Holger Dinkel, Kristoffer Forslund, Alexandros Stamatakis, Joseph Barry, Aidan Budd, Theodoros G Soldatos, Katja Linssen, Abdul Mateen Rajput. Biggest challenges in bioinformatics. EMBO reports, (2013) 14(4), pages 302-304.

[Garg et al. 2013] Divita Garg, Alexander V. Beribisky, Glauco Ponterini, Alessio Ligabue, Gaetano Marverti, Andrea Martello, M. Paola Costi, Michael Sattler and Rebecca C. Wade. Translational repression of thymidylate synthase by targeting its mRNA. Nucleic Acids Res., (2013) doi:10.1093/nar/gkt098: 1-12.

[Genel et al. 2013] S. Genel, M. Vogelsberger, D. Nelson, D. Sijacki, V. Springel, and L. Hernquist. Following the flow: tracer particles in astrophysical fluid simulations. Monthly Notices of the Royal Astronomical Society, 435, pages 1426-1442, 2013.

[Gombos et al. 2013] Linda Gombos, Annett Neuner, Mykhaylo Berynskyy, Luca L. Fava, Rebecca C. Wade, Carsten Sachse and Elmar Schiebel. GTP regulates the microtubule nucleation activity of gamma-tubulin. Nat Cell Biol., 15, pages 1317-1327, 2013.

[Greif et al. 2013] T. H. Greif, V. Springel, and V. Bromm. On the operation of the chemothermal instability in primordial star-forming clouds. Monthly Notices of the Royal Astronomical Society, 434, pages 3408-3422, 2013.

[Guinaudeau 2013] Camille Guinaudeau and Michael Strube. Graph-based local coherence modeling. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, pages 93-103, 2013.

[Hajian et al. 2013] A. Hajian, N. Battaglia, D. N. Spergel, J. R. Bond, C. Pfrommer, and J. L. Sievers. Measuring the thermal Sunyaev-Zel'dovich effect through the cross correlation of Planck and WMAP maps with ROSAT galaxy cluster catalogs. Journal of Cosmology and Astroparticle Physics, 11:064, 2013.

[Hass et al. 2013a] M. R. Haas, J. Schaye, C. M. Booth, C. Dalla Vecchia, V. Springel, T. Theuns, and R. P. C. Wiersma. Physical properties of simulated galaxy populations at z = 2 - II. Effects of cosmology, reionization and ISM physics. Monthly Notices of the Royal Astronomical Society, 435, pages 2955-2967, 2013.

[Hass et al. 2013b] M. R. Haas, J. Schaye, C. M. Booth, C. Dalla Vecchia, V. Springel, T. Theuns, and R. P. C. Wiersma. Physical properties of simulated galaxy populations at z = 2 - I. Effect of metal-line cooling and feedback from star formation and AGN. Monthly Notices of the Royal Astronomical Society, 435, pages 2931-2954, 2013.

[Hauser et al. 2013] Jörg Hauser, Kassian Kobert, Fernando Izquierdo-Carrasco, Karen Meusemann, Bernhard Misof, Michael Gertz, and Alexandros Stamatakis. Heuristic algorithms for the protein model assignment problem. In Bioinformatics Research and Applications, pages 137-148, Springer Berlin Heidelberg, 2013.

[Hayward 2013] C. C. Hayward. The star formation rate and stellar mass limits for submillimetre galaxies implied by recent interferometric observations. Monthly Notices of the Royal Astronomical Society, 432:L85, 2013.

[Hayward et al. 2013a] C. C. Hayward, P. S. Behroozi, R. S. Somerville, J. R. Primack, J. Moreno, and R. H. Wechsler. Spatially unassociated galaxies contribute significantly to the blended submillimetre galaxy population: predictions for follow-up observations of ALMA sources. Monthly Notices of the Royal Astronomical Society, 434, pages 2572-2581, 2013.

[Hayward et al. 2013b] C. C. Hayward, D. Narayanan, D. Keres, P. Jonsson, P. F. Hopkins, T. J. Cox, and L. Hernquist. Submillimetre galaxies in a hierarchical universe: number counts, redshift distribution and implications for the IMF. Monthly Notices of the Royal Astronomical Society, 428, pages 2529-2547, 2013.

[Hayward et al. 2013c] C. C. Hayward, P. Torrey, V. Springel, L. Hernquist, and M. Vogelsberger. Galaxy mergers on a moving mesh: a comparison with smoothed-particle hydrodynamics. ArXiv e-prints, 2013. eprint: 1309.2942.

[Henriques et al. 2013] B. M. B. Henriques, S. D. M. White, P. A. Thomas, R. E. Angulo, Q. Guo, G. Lemson, and V. Springel. Simulations of the galaxy population constrained by observations from z = 3 to the present day: implications for galactic winds and the fate of their ejecta. Monthly Notices of the Royal Astronomical Society, 431, pages 3373-3395, 2013.

[Hoecker et al. 2013] M. Hoecker and M. Kunze. An on-demand scaling stereoscopic 3D video streaming service in the cloud. Journal of Cloud Computing: Advances, Systems and Applications, 2:14, 2013.

[Hopkins et al. 2013a] P. F. Hopkins, T. J. Cox, L. Hernquist, D. Narayanan, C. C. Hayward, and N. Murray. Star formation in galaxy mergers with realistic models of stellar feedback and the interstellar medium. Monthly Notices of the Royal Astronomical Society, 430, pages 1901-1927, 2013.

[Hopkins et al. 2013b] P. F. Hopkins, D. Keres, N. Murray, L. Hernquist, D. Narayanan, and C. C. Hayward. Resolving the generation of starburst winds in Galaxy mergers. Monthly Notices of the Royal Astronomical Society, 433, pages 78-97, 2013.

[Hou et al. 2013a] Yufang Hou, Katja Markert, and Michael Strube. Global inference for bridging anaphora resolution. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, pages 907-917, 2013.

[Hou et al. 2013b] Yufang Hou, Katja Markert, and Michael Strube. Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash., pages 814-820, 2013.

[Izquierdo-Carrasco et al. 2013] Fernando Izquierdo-Carrasco, Nikolaos Alachiotis, Simon Berger, Tomas Flouri, Solon P. Pissis, and Alexandros Stamatakis. A generic vectorization scheme and a gpu kernel for the phylogenetic likelihood library. In Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing Workshops and PhD Forum, IPDPSW ,13, pages 530-538, Washington, DC, USA, 2013. IEEE Computer Society, 2013.

[Kiefer et al. 2013] Markus Kiefer, Roswitha Schmickl, Dmitry A German, Terezie Mandáková, Martin A. Lysak, Ihsan A. Al-Shehbaz, Andreas Franzke, Klaus Mummen-

hoff, Alexandros Stamatakis, and Marcus A. Koch. Brassibase: Introduction to a novel knowledge database on brassicaceae evolution. Plant and Cell Physiology, page pct158, 2013.

[Knebe et al. 2013] A. Knebe, F. R. Pearce, H. Lux, Y. Ascasibar, P. Behroozi, J. Casado, C. C. Moran, J. Diemand, K. Dolag, R. Dominguez-Tenreiro, P. Elahi, B. Falck, S. Gottlöber, J. Han, A. Klypin, Z. Lukic, M. Maciejewski, C. K. McBride, M. E.Merchan, S. I. Muldrew, M. Neyrinck, J. Onions, S. Planelles, D. Potter, V. Quilis, Y. Rasera, P. M. Ricker, F. Roy, A. N. Ruiz, M. A. Sgro, V. Springel, J. Stadel, P. M. Sutter, D. Tweed, and M. Zemp. Structure finding in cosmological simulations: the state of affairs. Monthly Notices of the Royal Astronomical Society, 435, pages 1618-1658, 2013.

[Kociumaka et al. 2013] Tomasz Kociumaka, Solon P. Pissis, Jakub Radoszewski, Wojciech Rytter, and Tomasz Waleń. Fast algorithm for partial covers in words. In Johannes Fischer and Peter Sanders, editors, Combinatorial Pattern Matching, volume 7922 of Lecture Notes in Computer Science, pages 177-188. Springer Berlin Heidelberg, 2013.

[Kokh et al. 2013] Daria B. Kokh, Stefan Richter, Stefan Henrich, Paul Czodrowski, Friedrich Rippmann and Rebecca C. Wade. TRAPP: a tool for analysis of transient binding pockets in proteins. J. Chem Inf Model, (2013) 53, pages 1235-1252, 2013

[Kolar et al. 2013] M. Kolar, P. Hobza and A. K. Bronowska. Plugging the explicit sigma-holes in molecular docking. Chem Commun (Camb), 1;49(10):981-3, 2013.

[Kromer et al. 2013a] M. Kromer, M. Fink, V. Stanishev, S. Taubenberger, F. Ciaraldi-Schoolman, R. Pakmor, F. K. Röpke, A. J. Ruiter, I. R. Seitenzahl, S. A. Sim, G. Blanc, N. Elias-Rosa, and W. Hillebrandt. 3D deflagration simulations leaving bound remnants: a model for 2002cx-like Type Ia supernovae. Monthly Notices of the Royal Astronomical Society, 429, pages 2287-2297, 2013.

[Kromer et al. 2013b] M. Kromer, R. Pakmor, S. Taubenberger, G. Pignata, M. Fink, F. K. Roepke, I. R. Seitenzahl, S. A. Sim, and W. Hillebrandt. SN 2010lpa Type Ia Supernova from a Violent Merger of Two Carbon-Oxygen White Dwarfs. Astrophysical Journal, 778:L18, 2013.

[Lanz et al. 2013] L. Lanz, A. Zezas, N. Brassington, H. A. Smith, M. L. N. Ashby, E. da Cunha, G. G. Fazio, C. C. Hayward, L. Hernquist, and P. Jonsson. Global Star Formation Rates and Dust Emission over the Galaxy Interaction Sequence. Astrophysical Journal, 768:90, 2013.

[Li et al. 2013] W. Li, S. A. Edwards, L. Lu, T. Kubar, Patil S. P., H. Grubmüller, G. Groenhof and F. Gräter Force distribution analysis of mechanochemically reactive dimethylcyclobutene.
ChemPhysChem , 14(12), pages 2687-2697, 2013.

[Liu et al. 2013a] Z.-W. Liu, R. Pakmor, F. K. Röpke, P. Edelmann, W. Hillebrandt, W. E. Kerzendorf, B. Wang, and Z. W. Han. Rotation of surviving companion stars after type Ia supernova explosions in the WD+MS scenario. Astronomy & Astrophysics, 554:A109, 2013.

[Liu et al. 2013b] Z.-W. Liu, M. Kromer, M. Fink, R. Pakmor, F. K. Röpke, X. F. Chen, B. Wang, and Z. W. Han. Predicting the Amount of Hydrogen Stripped by the SN Explosion for SN 2002cx-like SNe Ia. Astrophysical Journal, 778:121, 2013.

[Liu et al. 2013c] Z.-W. Liu, R. Pakmor, I. R. Seitenzahl, W. Hillebrandt, M. Kromer, F. K. Röpke, P. Edelmann, S. Taubenberger, K. Maeda, B. Wang, and Z. W. Han. The Impact of Type Ia Supernova Explosions on Helium Companions in the Chandrasekhar-mass Explosion Scenario. Astrophysical Journal, 774:37, 2013.

[Louet et al. 2013] M. Louet, E. Karakas, A. Perret, D. Perahia, J. Martinez and N. Floquet.
Conformational restriction of G-proteins Coupled Receptors (GPCRs) upon complexation to G-proteins: A putative activation mode of GPCRs? FEBS Lett, 587, pages 2656-2661, 2013.

[Ludlow et al. 2013a] A. D. Ludlow, J. F. Navarro, R. E. Angulo, M. Boylan-Kolchin, V. Springel, C. Frenk, and S. D. M. White. The Mass-Concentration-Redshift Relation of Cold Dark Matter Halos. ArXiv e-prints, 2013. eprint: 1312.0945.

[Ludlow et al. 2013b] A. D. Ludlow, J. F. Navarro, M. Boylan-Kolchin, P. E. Bett, R. E. Angulo, M. Li, S. D. M. White, C. Frenk, and V. Springel. The mass profile and accretion history of cold dark matter haloes. Monthly Notices of the Royal Astronomical Society, 432, pages 1103-1113, 2013.

[Marasco et al. 2013] A. Marasco, F. Marinacci, and F. Fraternali. On the origin of the warm-hot absorbers in the Milky Way's halo. Monthly Notices of the Royal Astronomical Society, 433, pages 1634-1647, 2013.

[Marinacci et al. 2013] F. Marinacci, R. Pakmor, and V. Springel. The formation of disc galaxies in high-resolution moving-mesh cosmological simulations. Monthly Notices of the Royal Astronomical Society, 437, pages 1750-1775, 2013.

[Martschat 2013] Sebastian Martschat. Multigraph clustering for unsupervised coreference resolution. In Proceedings of the Student Research Workshop at the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, pages 81-88, 2013.

[Mercadante et al. 2013] D. Mercadante, L. D. Melton, G. B. Jameson, M. A. K. Williams and Alfonso De Simone. Substrate Dynamics in Enzyme Action: Rotations of Monosaccharide Subunits in the Binding Groove are Essential for Pectin Methylesterase Processivity.
Biophys J, 104(8), pages 1731-1739, 2013.

[Mereghetti 2013] Paolo Mereghetti and Rebecca C. Wade. Brownian dynamics simulation of protein diffusion in crowded environments. AIP Conf. Proc., 1518, pages 511-516, 2013.

[Munoz et al. 2013] D. J. Munoz, V. Springel, R. Marcus, M. Vogelsberger, and L. Hernquist. Multidimensional, compressible viscous flow on a moving Voronoi mesh. Monthly Notices of the Royal Astronomical Society, 428, pages 254-279, 2013.

[Nastase 2013] Vivi Nastase and Michael Strube. Transforming Wikipedia into a large scale multilingual concept network. Artificial Intelligence, 194, pages 62-85, 2013.

[Nelson et al. 2013] D. Nelson, M. Vogelsberger, S. Genel, D. Sijacki, D. Keres, V. Springel, and L. Hernquist. Moving-mesh cosmology: tracing cosmological gas accretion. Monthly Notices of the Royal Astronomical Society, 429, pages 3353-3370, 2013.

[Pakmor et al. 2013a] R. Pakmor and V. Springel. Simulations of magnetic fields in isolated disc galaxies. Monthly Notices of the Royal Astronomical Society, 432, pages 176-193, 2013.

[Pakmor et al. 2013b] R. Pakmor, M. Kromer, S. Taubenberger, and V. Springel. Helium-ignited Violent Mergers as a Unified Model for Normal and Rapidly Declining Type Ia Supernovae. Astrophysical Journal, 770:L8, 2013.

[Pakmor et al. 2013c] R. Pakmor, F. Marinacci, and V. Springel. Magnetic fields in cosmological simulations of disk galaxies. ArXiv e-prints, 2013b. eprint: 1312.2620.

[Pandey et al. 2013] B. Pandey, S. D. M. White, V. Springel, and R. E. Angulo. Exploring the non-linear density field in the Millennium Simulations with tessellations - I. The probability distribution function. Monthly Notices of the Royal Astronomical Society, 435, pages 2968-2981, 2013.

[Patil et al. 2013] Sandeep P. Patil, Bernd Markert, and Frauke Gräter. Refining a Bottom-up Computational Approach for Spider Silk Fibre Mechanics. Proceedings of the 3rd GAMM Seminar on Continuum Biomechanics Report No. II-21, pages 75–88, 2013.

[Pavlidis et al. 2013] Pavlos Pavlidis, Daniel Živković, Alexandros Stamatakis, and Nikolaos Alachiotis. Sweed: Likelihood-based detection of selective sweeps in thousands of genomes. Molecular Biology and Evolution, pages 2224-2234, 2013.

[Pfrommer et al. 2013] C. Pfrommer, A. E. Broderick, P. Chang, E. Puchwein, and V. Springel. The physics and cosmology of TeV blazars in a nutshell. ArXiv e-prints, 2013. eprint: 1308.6284.

[Pfrommer et al. 2013] C. Pfrommer. Toward a Comprehensive Model for Feedback by Active Galactic Nuclei: New Insights from M87 Observations by LOFAR, Fermi, and H.E.S.S. Astrophysical Journal, 779: 10, 2013.

[Pinzke et al. 2013] A. Pinzke, S. P. Oh, and C. Pfrommer. Giant radio relics in galaxy clusters: reacceleration of fossil relativistic electrons? Monthly Notices of the Royal Astronomical Society, 435, pages 1061-1082, 2013.

[Pissis et al. 2013] Solon P. Pissis, Christian Goll, Alexandros Stamatakis, and Pavlos Pavlidis. Accelerating string matching on MIC architecture for motif extraction. In 10th International Conference on Parallel Processing and Applied Mathematics, Lecture Notes in Computer Science. Springer, Berlin Heidelberg, 2013.

[Pissis et al. 2013] Solon P. Pissis, Alexandros Stamatakis, and Pavlos Pavlidis. Motexach: A word-based hpc tool for motif extraction. In Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, p. 13, ACM, 2013.

[Pointecouteau et al. 2013] E. Pointecouteau, T. H. Reiprich, C. Adami, M. Arnaud, V. Bi, S. Borgani, K. Borm, H. Bourdin, M. Brueggen, E. Bulbul, N. Clerc, J. H. Croston, K. Dolag, S. Ettori, A. Finoguenov, J. Kaastra, L. Lovisari, B. Maughan, P. Mazzotta, F. Pacaud, J. de Plaa, G. W. Pratt, M. Ramos-Ceja, E. Rasia, J. Sanders, Y.-Y. Zhang, S. Allen, H. Boehringer, G. Brunetti, D. Elbaz, R. Fassbender, H. Hoekstra, H. Hildebrandt, G. Lamer, D. Marrone, J. Mohr, S. Molendi, J. Nevalainen, T. Ohashi, N. Ota, M. Pierre, K. Romer, S. Schindler, T. Schrabback, A. Schwope, R. Smith, V. Springel, and A. von der Linden. The Hot and Energetic Universe: The evolution of galaxy groups and clusters. ArXiv e-prints, 2013. eprint: 1306.2319.

[Puchwein et al. 2013] E. Puchwein, M. Baldi, and V. Springel. Modified-Gravity-GADGET: a new code for cosmological hydrodynamical simulations of modified gravity models. Monthly Notices of the Royal Astronomical Society, 436, pages 348-360, 2013.

[Puchwein and Springel 2013] E. Puchwein and V. Springel. Shaping the galaxy stellar mass function with supernova- and AGN-driven winds. Monthly Notices of the Royal Astronomical Society, 428, pages 2966-2979, 2013.

[Ruiter et al. 2013] A. J. Ruiter, S. A. Sim, R. Pakmor, M. Kromer, I. R. Seitenzahl, K. Belczynski, M. Fink, M. Herzog, W. Hillebrandt, F. K. Röpke, and S. Taubenberger. On the brightness distribution of Type Ia supernovae from

violent white dwarf mergers. Monthly Notices of the Royal Astronomical Society, 429, pages 1425-1436, 2013.

[Sales et al. 2013] L. V. Sales, F. Marinacci, V. Springel, and M. Petkova. Stellar feedback by radiation pressure and photoionization. ArXiv e-prints, 2013. eprint: 1310.7572.

[Sandikci et al. 2013] Arzu Sandikci, Felix Gloge, Michael Martinez, Matthias P. Mayer, Rebecca C. Wade, Bernd Bukau and Guenter Kramer. Dynamic enzyme docking to the ribosome coordinates N-terminal processing with polypeptide folding. Nat Struct Mol Biol., 20, pages 843-850, 2013.

[Schefzik et al. 2013] R. Schefzik, T.L. Thorarinsdottir and T. Gneiting. Uncertainty quantification in complex simulation models using ensemble copula coupling. Statistical Science, 28, pages 616-640, 2013.

[Schloissnig et al. 2013] S. Schloissnig, M. Arumugam, S. Sunagawa, M. Mitreva, J. Tap, A. Zhu, A. Waller, D.R. Mende, J.R. Kultima, J. Martin, K. Kota, S.R. Sunyaev, G.M. Weinstock, and P. Bork. Genomic variation landscape of the human gut microbiome. Nature, 493, pages 45-50, 2013.

[Seifert 2013] C. Seifert and F. Gräter. Protein mechanics: How force regulates molecular function. Biochim. Biophys. Acta - General Subjects, 1830(10), pages 4762-4768, 2013.

[Seitenzahl et al. 2013a] I. R. Seitenzahl, G. Cescutti, F. K. Röpke, A. J. Ruiter, and R. Pakmor. Solar abundance of manganese: a case for near Chandrasekhar-mass Type Ia supernova progenitors. Astronomy & Astrophysics, 559:L5, 2013.

[Seitenzahl et al. 2013b] I. R. Seitenzahl, F. Ciaraldi-Schoolmann, F. K. Röpke, M. Fink, W. Hillebrandt, M. Kromer, R. Pakmor, A. J. Ruiter, S. A. Sim, and S. Taubenberger. Three-dimensional delayed-detonation models with nucleosynthesis for Type Ia supernovae. Monthly Notices of the Royal Astronomical Society, 429, pages 1156-1172, 2013.

[Shi et al. 2013] Lei Shi, Lenneke Jong, Ulrike Wittg, Philippe Lucarelli, Markus Stepath, Stephanie Mueller, Lorenza Alice D`Alessandro, Klingmüller Ursula, and Müller Wolfgang (2013). Excemplify: A Flexible Template Based Solution, Parsing and Managing Data in Spreadsheets for Experimentalists. Journal of Integrative Bioinformatics, 10(2):220, 2013.

[Sim et al. 2013] S. A. Sim, I. R. Seitenzahl, M. Kromer, F. Ciaraldi-Schoolmann, F. K. Röpke, M. Fink, W. Hillebrandt, R. Pakmor, A. J. Ruiter, and S. Taubenberger. Synthetic light curves and spectra for three-dimensional delayed-detonation models of Type Ia supernovae. Monthly Notices of the Royal Astronomical Society, 436, pages 333-347, 2013.

[Snyder et al. 2013] G. F. Snyder, C. C. Hayward, A. Sajina, P. Jonsson, T. J. Cox, L. Hernquist, P. F. Hopkins, and L. Yan. Modeling Mid-infrared Diagnostics of Obscured Quasars and Starbursts, Astrophysical Journal, 768:168, 2013.

[Stamatakis et al. 2013] Alexandros Stamatakis and Andre J. Aberer. Novel parallelization schemes for large-scale likelihood-based phylogenetic inference. Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing, pages 1195-1204, 2013.

[Starkenburg et al. 2013] E. Starkenburg, A. Helmi, G. De Lucia, Y.-S. Li, J. F. Navarro, A. S. Font, C. S. Frenk, V. Springel, C. A. Vera-Ciro, and S. D. M. White. The satellites of the Milky Way - insights from semianalytic model-

ling in a LCDM cosmology. Monthly Notices of the Royal Astronomical Society, 429, pages 725-743, 2013.

[Summa et al. 2013] A. Summa, A. Ulyanov, M. Kromer, S. Boyer, F. K. Röpke, S. A. Sim, I. R. Seitenzahl, M. Fink, K. Mannheim, R. Pakmor, F. Ciaraldi-Schoolmann, R. Diehl, K. Maeda, and W. Hillebrandt. Gamma-ray diagnostics of Type Ia supernovae. Predictions of observables from three-dimensional modeling, Astronomy & Astrophysics, 554:A67, 2013.

[Sunagawa et al. 2013] Shinichi Sunagawa, Daniel R. Mende, Georg Zeller, Fernando Izquierdo-Carrasco, Simon A. Berger, Jens Roat Kultima, Luis Pedro Coelho, Manimozhiyan Arumugam, Julien Tap, Henrik Bjørn Nielsen, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nature methods, 10(12), pages 1196-1199, 2013.

[Tauberberger et al. 2013a] S. Taubenberger, M. Kromer, S. Hachinger, P. A. Mazzali, S. Benetti, P. E. Nugent, R. A. Scalzo, R. Pakmor, V. Stanishev, J. Spyromilio, F. Bufano, S. A. Sim, B. Leibundgut, and W. Hillebrandt. `Super-Chandrasekhar' Type Ia Supernovae at nebular epochs. Monthly Notices of the Royal Astronomical Society, 432, pages 3117-3130, 2013.

[Tauberberger et al. 2013b] S. Taubenberger, M. Kromer, R. Pakmor, G. Pignata, K. Maeda, S. Hachinger, B. Leibundgut, and W. Hillebrandt. [O I] 6300, 6364 in the Nebular Spectrum of a Subluminous Type Ia Supernova. Astrophysical Journal, 775:L43, 2013.

[Thorarinsdottir et al. 2013] T.L. Thorarinsdottir, T. Gneiting and N. Gissibl. Using proper divergence functions to evaluate climate models. SIAM/ASA Journal on Uncertainty Quantification I, pages 522-534, 2013.

[Veith et al. 2013] Nadine Veith, Anna Feldman-Salit, Vlad Cojocaru, Stefan Henrich, Ursula Kummer and Rebecca C. Wade. Organism-adapted specificity of the allosteric regulation of pyruvate kinase in lactic acid bacteria. PLoS Comput Biol., 9:e1003159, 2013.

[Vogelsberger et al. 2013] M. Vogelsberger, S. Genel, D. Sijacki, P. Torrey, V. Springel, and L. Hernquist. A model for cosmological simulations of galaxy formation physics. Monthly Notices of the Royal Astronomical Society, 436, pages 3031-3067, 2013.

[Wacker et al. 2013] S. J. Wacker, C. Aponte-Santamaria, P. Kjellbom, S. Nielsen, B. L. de
Groot, and M. Rützler. The identification of novel, high anity AQP9 inhibitors in an intracellular binding site. Mol Membr Biol 30, pages 246-260, 2013.

[Wang et al. 2013] B. Wang, S. Xiao, S. A. Edwards and F. Gräter. Isopeptide bonds mechanically stabilize spy0128 in bacterial pili. Biophys. J 104(9), pages 2051-2057, 2013.

[Wilman et al. 2013] D. J. Wilman, F. Fontanot, G. De Lucia, P. Erwin, and P. Monaco. The hierarchical origins of observed galaxy morphology. Monthly Notices of the Royal Astronomical Society, 433, pages 2986-3004, 2013.

[Wittig et al. 2013] Ulrike Wittig, Maja Rey, Renate Kania, Meik Bittkowski, Lei Shi, Martin Golebiewski, Andreas Weidemann, Wolfgang Müller, and Isabel Rojas (2013). Challenges for an enzymatic reaction kinetics database. FEBS Journal, 281(2), pages 572-582, 2013.

[Wolstencroft et al. 2013] Katherine Wolstencroft, Stuart Owen, Olga Krebs, Wolfgang Müller, Quyen Nguyen, Jacky L. Snoep, and Carole Goble (2013). Semantic Data and Models Sharing in Systems Biology: The Just Enough Results Model and the SEEK Platform. The Semantic Web-ISWC, 8219, pages 212-227, 2013.

[Xiao et al. 2013] Se. Xiao, Sh. Xiao and F. Gräter. Dissecting the structural determinants for the difference in mechanical stability of silk and amyloid beta-sheet stacks. Phys. Chem. Chem. Phys 15(22), pages 8765-8771, 2013.

[Young et al. 2013] H.Young, S.A. Edwards and F.Gräter. How fast does a signal propagate through proteins. PLoS One 8(6), pages 1-5, 2013.

[Yu et al. 2013] Xiaofeng Yu, Vlad Cojocaru and Rebecca C. Wade. Conformational diversity and ligand tunnels of mammalian cytochrome P450s. Biotechnology and Applied Biochemistry, 60, pages 134-145, 2013.

[Zandanel et al. 2013] F. Zandanel, P. Colin, S. Lombardi, M. Doro, D. Eisenacher, D. Hildebrand, F. Prada, for the MAGIC Collaboration, C. Pfrommer, and A. Pinzke. MAGIC Gamma-ray Observations of the Perseus Galaxy Cluster. ArXiv e-prints, 2013. eprint: 1308.0492.

[Zhang et al. 2013] Jiajie Zhang, Paschalia Kapli, Pavlos Pavlidis, and Alexandros Stamatakis. A general species delimitation method with applications to phylogenetic placements. Bioinformatics, 29(22), pages 2869-2876, 2013.

[Zhang et al. 2013] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis. Pear: A fast and accurate illumina paired-end read merger. Bioinformatics, published online: October 18, 2013. Print: Bioinformatics (2014) 30 (5): pages 614-620.

DEGREES

[Banerjee 2013] Banerjee, Priyanka: “An algorithmic approach to peptidomimetics”, Master’s Thesis, Life Science Informatics, University of Bonn and HITS: Stefan Henrich & Rebecca Wade, 2013.

[Berger 2013] Berger, Simon: “Phylogeny-Aware Placement and Alignment Methods for Short Reads”, Ph.D. Thesis, Computer Science, Karlsruhe Institute of Technology, Karlsruhe and HITS: Alexandros Stamatakis, 2013.

[Cai 2013] Jie Cai: “Coreference Resolution via Hypergraph Partitioning”, Ph.D. Thesis, Department of Computational Linguistics, Heidelberg University and HITS: Michael Strube, 2013.

[Costescu 2013] Bogdan Costescu: “Time-Resolved Force Distribution Analysis for Molecular Communication”, Ph.D. Thesis, Physics, Heidelberg University and HITS: Prof. Dr. Robert B. Russell and Frauke Gräter (2013).

[Debes 2013] Cedric Debes: “Physical constraints on protein structure evolution”, Ph.D. Thesis, Physics, Heidelberg University and HITS: Dr Frauke Gräter (2013).

[Krzyszkowska 2013] Jolanta Krzyszkowska: “Chemical Enrichment in Gadget2 and Arepo Simulated Galaxy Clusters”, Bachelor Thesis, Physics, Heidelberg University and HITS: Volker Springel, 2013.

[Pauz 2013] Vitali Pauz: “Anisotropic Thermal Conduction and the Cooling Flow Problem in Galaxy Clusters”, Master Thesis, Physics, Heidelberg University and HITS: Volker Springel, 2013.

COURSES

Camilo Aponte-Santamaría
Course on “Simulation of the dynamics of biomolecules using GROMACS“.
Antioquia University, Medellín (Colombia). June 3-13 2013.

Meik Bittkowski, Volker Stollorz
User Meeting OperationsExplorer, HITS, Heidelberg, November 4, 2013.

Agnieszka Bronowska
Lecture course on “Projektowanie molekularne w chemii leków i nanotechnologii (Molecular design in medicinal chemistry and nanotechnology)”, Faculty of Chemistry, University of Warsaw, Poland, Fall semester 2013/2014.

Eduardo R. Cruz Chu
Workshop in “Computational Biophysics”. Laboratory of Research and Development (LID). Peruvian University Cayetano Heredia (UPCH). Lima, Peru. Feb, 2013.

Course on “General Chemistry. Alpha Cycle for Freshmen College Students”. Peruvian University Cayetano Heredia (UPCH). Lima, Peru. Feb, 2013.

Jonathan Fuller
(with Grainne Kerr, Holger Dinkel, Matthew Betts, Heidelberg University)
Practical course on “Introduction to the Linux Command Line for NGS Analyses”, Participants from across Heidelberg, Heidelberg University, 20 September 20, 2013.

Martin Golebiewski, Lihua An
„Meet the Virtual Liver Data Management” (local user meeting with hands-on training)
Dr. Margarete Fischer-Bosch Institute of Clinical Pharmacology (IKP), Stuttgart, March 21, 2013.

Martin Golebiewski, Meik Bittkowski
"Virtual Liver LPS Showcase Meeting" (local user meeting with hands-on training)
Universitätsklinikum Mannheim, Germany, July 4, 2013.

Martin Golebiewski, Meik Bittkowski, Ivan Savora
Data management session (half-day) with all-hands hands-on training
Virtual Liver Retreat, Hünfeld, Germany, October 28-30, 2013.

Frauke Gräter
Lecture course "Biomolecular simulations change over from molecular dynamics to continuum mechanics", Summer School-Multiscale Modeling and Simulation : 3rd Summer School of the IMPRS, Magdeburg, Germany, 3th Sept, 2013.

Sandeep Patil
Tutorial on "Biomolecular simulations change over from molecular dynamics to continuum mechanics". Summer School-Multiscale Modeling and Simulation : 3rd Summer School of the IMPRS, Magdeburg, Germany, 3th Sept, 2013.

Christoph Pfrommer
Experimental Physics II, tutoring classes, Department of Physics and Astronomy, Heidelberg University, April - July 2013.

Volker Springel
Experimental Physics II, tutoring classes, Department of Physics and Astronomy, Heidelberg University, April - July 2013.

Alexandros Stamatakis, Fernando Izquierdo-Carrasco, Tomas Flouri
Course on "Computational Molecular Evolution", Wellcome Trust, Hinxton, UK, 29 April – 10 May 2013.

Rebecca Wade, Stefan Richter, Musa Ozboyaci, Stefan Henrich, Jonathan Fuller (with Damien Devos, COS, Heidelberg University)
"Grundkurs Bioinformatik" practical course, BSc Biosciences, Heidelberg University, January 14-18, 2013.

Johannes Wagner
Seminar on the „Physics of the Genome", Heidelberg University, Heidelberg. Winter semester 2012-2013.

„PreCourse in Mathematics for physicists", Heidelberg University, Heidelberg. Sep 23 - Oct 11 2013.

Andreas Weidemann
Modelling and Simulation of Quantitative Biological Models. Online tools, model databases and data management: SYCAMORE.
Tutorial at COMBINE workshop, Copenhagen (Denmark) September 4, 2013.

LECTURES

Tilmann Gneiting
Seminar Statistical Forecasting, Karlsruhe Institute of Technology (KIT), winter semester 2013/14.

Martin Golebiewski
"SEEK and Find: The Virtual Liver Data Management", Virtual Liver Retreat 2013, Hünfeld , Germany, October 30, 2013.

Christoph Pfrommer
Cosmology, Department of Physics and Astronomy, Heidelberg University, October 2013 - February 2014.

Ewald Puchwein
Advanced Cosmology, Department of Physics and Astronomy, Heidelberg University, April 2013 - July 2014.

Andreas Reuter
Course on “Wissenschaftliches Arbeiten“ (in German), Heidelberg University, Heidelberg, winter semester 2013/2014.

Stefan Richter
Technical Introduction for the Bioinformatics course, Heidelberg University, Bioquant, January 2013.

Ivan Savora, Meik Bittkowski
“News from VLN SEEK - Usability improvement and new features” and “The Virtual Liver PORTAL”, Virtual Liver Retreat 2013, Hünfeld , Germany, October 28-30, 2013.

Volker Springel
Fundamentals of Simulation Methods, Department of Physics and Astronomy, Heidelberg University, October 2013 - February 2014.

High perfomance computing and numerical modeling, 43rd Saas-Fee Winter School, Switzerland, March 2013.

Alexandros Stamatakis
Introduction to Bioinformatics for Computer Scientists, Institute of Theoretical Informatics, Karlsruhe Institute of Technology (October 2012 – February 2013)
Hot Topics in Bioinformatics, Institute of Theoretical Informatics, Karlsruhe Institute of Technology, April 2013 – July 2013.

Michael Strube
Seminar: “Discourse Processing”, Department of Computational Linguistics, Heidelberg University (October 2012 – February 2013)
PhD Colloquium, Department of Computational Linguistics, Heidelberg University, October 2012 - February 2013.

Blockseminar: “Extracting Knowledge from and Linking to Wikipedia”, Computer Science Department, Korea Advanced Institute for Science and Technology, Daejeon, Korea, May 2013.

PhD Colloquium, Department of Computational Linguistics, Heidelberg University, April 2013 - July 2013.

Rebecca Wade
Ringvorlesung “Structure and Dynamics of biological macromolecules”, lecture on “Electrostatics, solvation and protein interactions”, BSc Biology, Heidelberg University, July 2, 2013.

Ringvorlesung „Biophysik“,“Receptor-Ligand Interactions: Structure and Dynamics”, Bachelor Molecular Biotechnology, Heidelberg University, November 28, 2013.

9 Miscellaneous

9.1 Guest Speaker Activities

Agnieszka Bronowska

"Small changes, big impact: halogen bonds, peptidomimetics, and structure-based design.", Faculty of Chemistry, University of Edinburgh, Scotland, April, 2013.

"Molecular modelling of GPR54 and its interactions with kisspeptins: towards the discovery of novel peptidomimetic ligands of GPR54.", Faculty of Biology, University of Edinburgh, Scotland, April 2013.

Habilitation colloquium: "Oddziaływania lek-białko w ujęciu termodynamicznym: entropia konformacyjna, efekty rozpuszczalnikowe i kompensacja entalpowo-entropowa." Faculty of Chemistry, University of Warsaw, Poland, October 2013.

Martin Golebiewski

"Assembling the jigsaw puzzle of liver data: The Virtual Liver Data Management", Seminar presentation at the Center for Cell Analysis and Modeling - University of Connecticut Health Center, Farmington, Connecticut (USA) May 16, 2013.

Frauke Gräter

"Evolution of protein foldability", Biophysical Society meeting, Philadelphia, USA, 2-7 Feb, 2013.

"Protein function from force distribution analysis", Instruct, EBML, Heidelberg, Germany, 22 May, 2013.

"How proteins are designed for mechanical function: from unfolding kinetics to force sensors", Gordon Research Conference on Proteins, Holderness, USA, 17-21, Jun, 2013.

"Protein mechanics from molecular simulations: myomesin and titin", EMBO workshop on Z-disc protein assemblies, Hamburg, Germany, 14-17, Oct, 2013.

Christopher Hayward

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", Massachusetts Institute of Technology, Boston, USA, November 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", Tufts University, Medford, USA, November 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", University of Massachusetts, Amherst, USA, November 2013.

"Galaxy mergers on a moving mesh",
Harvard University, Cambridge, USA, November 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", University of Chicago, Chicago, USA, November 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", Northwestern University, Chicago, USA, November 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", University of Arizona, Tucson, USA, November 2013.

"The heterogeneity of the submillimeter galaxy population", Infrared Processing and Analysis Center, Pasadena, CA, October 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical

technique", Carnegie Observatories, Pasadena, USA, October 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", California Institute of Technology, Pasadena, USA, October 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", University of California, Santa Cruz, USA, October 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", University of California, San Diego, USA, October 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", University of California, Berkeley, USA, October 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", University of Michigan, Ann Arbor, USA, October 2013.

"Advances in galaxy-formation simulations: calculating mock observables & using a more-accurate numerical technique", Max Planck Institute for Astronomy, Heidelberg, Germany, October 2013.

"Galaxy mergers on a moving mesh", Max Planck Institute for Astronomy, Heidelberg, Germany, September 2013.

"The heterogeneity of the submillimeter galaxy population", Institut d'astrophysique de Paris, Paris, France, May 2013.

"The heterogeneity of the submillimeter galaxy population", Max Planck Institute for Astronomy, Heidelberg, Germany, April 2013.

"Physically modeling high-redshift starbursts and obscured AGN", Max-Planck-Institut für Astrophysik, Garching, Germany, March 2013.

"Physically modeling high-redshift starbursts and obscured AGN", Max-Planck-Institut für extraterrestrische Physik, Garching, Germany, March 2013.

Vincent Heuveline

"Mathematische Modellierung und Optimierung zukünftiger Energienetze", Internationales Wissenschaftsforum Heidelberg (IWR), 7-8 May 2013.

"Big Data Technology Descriptions & Uses Cases", Chair of the Session, ISC Cloud and Big Data, Heidelberg, Germany, 25 – 26, September 2013.

"3D in the era of Supercomputer", Festival Beyond 3D, Karlsruhe, Germany, 3-6 October 2013

"Partial Differential Equations with Random Coefficients", WIAS Workshop Berlin, Germany, October 28 –– 30, 2013.

Vincent Heuveline, F. Nobile, A. Chernov

"Numerical Methods for Uncertainty Quantification", Organization of Workshop at Hausdorff Center for Mathematics, Bonn, Germany, 13-17 May 2013.

Maxime Louet

"G-proteins activation and CAP allostery explored by molecular mechanics",
École Normale Supérieur Cachan, France, October, 16, 2013

Federico Marinacci

"Forming disc galaxies on a moving mesh: zoom-in simulations of Milky-Way sized haloes", Osservatorio Astronomico di Trieste, Trieste, Italy, November 2013.

"Forming disc galaxies on a moving mesh: zoom-in simulations of Milky-Way sized haloes", Osservatorio Astronomico di Trieste, Trieste, Italy, October 2013.

"The formation of disc galaxies in moving-mesh cosmological simulations", SFB seminar series, Heidelberg, Germany, July 2013.

"The formation of disc galaxies in moving-mesh cosmological simulations",
Max Planck Institute for Astronomy, Heidelberg, Germany, July 2013.

Davide Mercadante

"Brownian ratchets to break the wall. Inception of plant infection on the nanoscale.",
Université de Picardie Jules Verne, France, May, 25, 2013.

Rüdiger Pakmor

"Arepo - operA", Astrophysical Seminar, University of Toronto, Toronto, Canada, March 2013.

Christoph Pfrommer

"High-Energy Astrophysics meets Cosmology", Current Topics in Astronomy and Astrophysics, Department of Physics and Astronomy, Heidelberg University, February 2013.

"The Physics and Cosmology of TeV Blazars", GRAPPA Institute, University of Amsterdam, The Netherlands, September 2013.

"Schwarze Löcher im Universum", Vortragsreihe "Faszination Astronomie" am Haus der Astronomie, Heidelberg, October 2013.

"Cosmic rays in galaxy formation: a solution to the faint and bright-end of the population?" Galaxy Theory Meeting at the MPIA, Heidelberg, December 2013.

Solon Pissis

"Finding subtree repeats to slash the time for phylogenetic analyses", Theo Murphy International scientific meeting on Storage and Indexing of massive data, The Royal Society at Chicheley Hall, home of the Kavli Royal Society International Centre, Buckinghamshire, UK, February 2013.

Andreas Reuter

Talk at International Research Seminar. "25 Years of Research Cooperation in the Fields of Computational Science and Parallel Computing", in Donezk, Ukraine, 4 November 2013.

Volker Springel

"Supercomputer Simulations of Cosmic Structure Formation", Physics Colloquium, University of Amsterdam, Amsterdam, March 2013.

"Astrophysik mit dem Supercomputer", Haus der Astronomie, Heidelberg, January 2013.

"Wie Astronomen zurück in die Zeit blicken", Kinderuni, Heidelberg University, March 2013.

"Ist das kosmologische Standardmodell tragfähig?", Podiumsdiskussion, Peterskirche Heidelberg, June 2013.

"Forming the Milky Way on a Supercomputer", Astrophysical Colloquium, Laboratoire d'Astrophysique de Marseille, Marseille, France, June 2013.

"Die dunkle Seite des Universums", Planetarium Münster, Münster, January 2013.

"Forming the Milky Way Galaxy on a Supercomputer", Physics Colloquium, University of Nottingham, Nottingham, UK, May 2013.

"Forming the Milky Way Galaxy on a Supercomputer", Physics Colloquium, University of Tübingen, Tübingen, July 2013.

"Forming the Milky Way Galaxy on a Supercomputer", Astrophysics Colloquium, Stockholm University, Stockholm, Sweden, December 2013.

Alexandros Stamatakis

"Using Supercomputers to build Evolutionary Trees from DNA data", Panhellenic Electrical and Computer Engineering Students Conference, Athens, Greece, April 2013.

"ExaML and ExaBayes: Maximum Likelihood and Bayesian Phylogenetic Inference on Exascale Supercomputers", Arizona State University, Phoenix, USA, March 2013.

"What's up at the Exelixis Lab?", Museo Nacional de Ciencias Naturales, Madrid, Spain, January 2013.

Denis Yurin

"The stability of stellar disks in cosmological environments", IMPRS Seminar, Heidelberg University, Heidelberg, July 2012.

TALKS

Andre Aberer

"Novel Parallelization Schemes for Large-Scale Likelihood-based Phylogenetic Inference", IPDPS 2013, Boston, USA, May 2013.

Andreas Bauer

"Studying reionization using GPUs", Mind the gap: from microphysics to large-scale structure in the Universe, University of Cambridge, Cambridge, UK, July 8 - 12, 2013.

"Subsonic turbulence", Virgo Consortium Meeting, Durham, UK, December 12 - 14, 2012.

Mykhaylo Berynskyy

"Protein-Protein interactions: chaperones", BDBDB3, Heidelberg, October 8,2013.

Meik Bittkowski

"OperationsExplorer: Daten, Technologien, Tests", OnlineCamp "Data Checking", Wissenswerte 2013, Bremen (Germany) November 2013.

"Parsing of BioSample spreadsheets & Introduction to the Content Management System behind the VLN Portal", VLN Retreat and VLN PALS Meeting, Hünfeld (Germany) October 2013.

Eduardo R. Cruz Chu

"Introductory Talk about the Chemistry as a Career", Alpha Cycle for Freshmen College Students. Peruvian University Cayetano Heredia (UPCH). Lima, Peru. 18, Jan, 2013.

"Computer Modeling of Polymers with Applications in Biology and Nanotechnology",
Laboratory of Research and Development (LID). Peruvi-

an University Cayetano Heredia (UPCH). Lima, Peru. 25, Jan, 2013.

"Computer Modeling of Polymers with Applications in Biology and Nanotechnology", Peruvian Chemical Society. Lima, Peru. 31, Jan, 2013.

Eduardo R. Cruz-Chu (shared with Rudiger Pakmor)
"Cosmological versus Molecular Simulation Methods: I. Efficient Algorithms for Particle-Cased Simulations", HITS, Heidelberg, Germany. 11 Nov. 2013.

Eduardo R. Cruz-Chu (shared with Dr. Volker Springel and Dr. Michael Martinez)
"Cosmological versus Molecular Simulation Methods: II. Efficient Parallelization", HITS, Heidelberg, Germany. 25, Nov. 2013.

Tomas Flouri
"An optimal algorithm for computing all subtree repeats in trees", IWOCA 2013, Rouen, France, July 2013.

"A generic Vectorization Scheme and a GPU kernel for the Phylogenetic Likelihood Library", IPDPS 2013, Boston, USA, May 2013.

Fabio Fontanot
"Low Mass Galaxies as Tracers of Cosmic History", Lorentz Workshop "What Regulates Galaxy Evolution?", Leiden, The Netherlands, Apr. 22 - 17, 2013.

"Comparing Semi-Analytical Models for Star Formation and Stellar Feedback",
Mind the gap: from microphysics to large-scale structure in the Universe, University of Cambridge, Cambridge, UK, July 8 - 12, 2013.

Jonathan Fuller
"How can I find out which proteins my small-molecule inhibitor is really inhibiting?",
DKFZ-ZMBH Alliance 'Pizza' Seminar Series, Heidelberg, Germany, April 10, 2013.

Martin Golebiewski
"Virtual Liver SEEK: Planned Features and Additions" and "SABIO-RK - Reaction Kinetics Database", HARMONY 2013: The Hackathon on Resources for Modeling in Biology, Farmington, Connecticut (USA) May 20-23, 2013.

"Standardization of Systems Biology Data & Models", All-Bio workshop on Standard Operation Procedures (SOP), Copenhagen (Denmark) September 3, 2013.

"Standardization in distributed research networks: The Virtual Liver Experience", COMBINE 2013, Paris (France) September 16-20, 2013.

Christopher Hayward
"Calculating synthetic UV-mm SEDs of simulated galaxy merger", RT13: Dust Radiative Transfer 2013 - Codes & Benchmarks, Grenoble,
France, October 9 - 11, 2013.

"The surprising complexity of the submillimetre galaxy population", Galaxy Evolution Over Five Decades, Cambridge, UK, September 3 - 6, 2013.

"Galaxy mergers on a moving mesh", Virgo Consortium Workshop, Garching, Germany, June 10 - 12, 2013.

"Theoretical models of submillimeter galaxies", Aspen Summer Workshop: The Obscured Universe, Aspen, USA, May 26 - June 7, 2013.

Fernando Izquierdo-Carrasco
"Heuristic Algorithms for the Protein Model Assignment Problem", ISBRA 2013, Charlotte, USA, May 2013.

Olga Krebs

"SEEK and JERM: Sharing Data and Models in Systems Biology"
International German/Russian Workshop on Integrative Biological Pathway Analysis and Simulation, Gatersleben (Germany) March 20, 2013.

"SEEK: A COMMUNITY BASED APPROACH FOR SHARING AND EXCHANGING DATA AND MODELS IN SYSTEMS BIOLOGY"
17. International Pushchino School Conference of YOUNG SCIENTISTS "Biology - The Science of the XXI Century", Pushchino (Russia) April 22-26, 2013.

"SEEK: An integration platform for systems biology projects" German-Russian Forum Biotechnology, Rostock (Germany) June 4-5 2013.

"JERM templates: overview & recent development"
SysMO PALs Meeting/ Combine 2013, Paris (France) September 16–20, 2013.

"RightField: Embedding ontology annotation in spreadsheets"
5th Virtual Liver Data Management & PALs Meeting, Hünfeld (Germany) October 30-31, 2013.

Federico Marinacci

"Moving-mesh cosmological simulations of disk galaxy formation", Exascale Computing in Astrophysics, Ascona, Switzerland, Sep 8 - 13, 2013.

"Moving-mesh cosmological simulations of disk galaxy formation", Mind the Gap: from microphysics to large-scale structure in the Universe, Cambridge, UK, July 8 - 12, 2013.

"Moving-mesh cosmological simulations of disk galaxy formation", The Physical Link between Galaxies and their Halos, Garching bei Muenchen, Germany, June 24 - 28, 2013.

"Moving-mesh cosmological simulations of disk galaxy formation", Virgo Consortium Workshop, Garching bei Muenchen, Germany, June 10 -12, 2013.

Wolfgang Müller & Martin Golebiewski

"The PORTAL into Virtual Liver Data", Mid-term Evaluation Meeting Virtual Liver, Berlin (Germany) January 30-31, 2013.

Musa Oeb oyaci

"Adsorption of 3H-BLIP on a gold surface". Workshop on Computer Simulation and Theory of Macromolecules, Hünfeld, Germany, April 26-27, 2013.

"Adsorption of 3H-BLIP on a gold surface". Biological Diffusion and Brownian Dynamics Brainstorm, Heidelberg, Germany, October 7-9, 2013.

Mehmet Ali Öz ürk

"Mechanism of pioneer transcription factor (FoxA1) complex formation with androgen receptor", Third Biological Diffusion and Brownian Dynamics Brainstorm: BDBDB3, Heidelberg, Germany, October, 7-9, 2013.

Rüdiger Pakmor

"Magnetic fields in simulated disk galaxies", Workshop on magnetic fields in galaxies, Bonn, Germany, March 6 - 8, 2013.

"Magnetic fields in simulated disk galaxies", Virgo Consortium Meeting, Garching, Germany, June 10 - 12, 2013.

"Magnetic fields in simulated disk galaxies", The Physical Link between Galaxies and their Halos, Garching, Germany, June 24 - 28, 2013.

"Challenges in modelling Type Ia Supernovae", Exascale Computing in Astrophysics, Ascona, Schweiz, September 8 - 13, 2013.

"White Dwarf Mergers", Observational Signatures of Type Ia Supernova Progenitors II, Leiden, Neitherlands, September 23 - 27, 2013.

"SN Ia explosion models and synthetic observables", SN Ia mini-workshop, Garching, Germany, October 17, 2013

Sandeep Patil

"Multiscale modeling of spider silk fiber mechanics", 3rd International Conference on Material Modelling (ICMM), Warsaw, Poland." 8-11, Sept, 2013.

Christoph Pfrommer

"Introduction to extragalactic sources of very high-energy photons", Rencontres de Moriond: Very High Energy Phenomena in the Universe,
La Thuile, Italy, Mar 9 - 16, 2013.

"Galaxy Clusters as Laboratories for Astroparticle Physics", SnowCLUSTER 2013, Utah, USA, Mar 24 - 29, 2013.

"LOFAR's role in unveiling the physics of galactic winds and AGN feedback",
LOFAR Magnetism Key Science Project Workshop, Sardinia, Italy, May 13 - 15, 2013.

"Blazar heating: physical mechanism and cosmological consequences", ENIGMA workshop, MPIA Heidelberg, Germany, Jun 18, 2013.

"Non-thermal Emission from Galaxy Clusters: Cosmic Rays and Dark Matter",
The Anisotropic Universe: from Microwaves to Ultrahigh Energies, GRAPPA Institute, University of Amsterdam, The Netherlands, Sep 25 - 27, 2013.

Solon Pissis

"GapsMis: flexible sequence alignment with a bounded number of gaps", ACM BCB 2013, Washington DC, USA, September 2013.

"MoTeX: an HPC word-based tool for MoTif eXtraction", ACM BCB 2013, Washington DC, USA, September 2013.

"Accelerating string matching on MIC architecture for motif extraction", 10th International Conference on Parallel Processing and Applied Mathematics, Warsaw, Poland, September 2013.

Kai Polsterer

Solving Regression Problems with Machine Learning, 4th Annual Astroinformatics Conference: Knowledge from Data, December 2013

Ewald Puchwein

"Simulating modified gravity models", Arepo Workshop, Heidelberg, Germany, September 2013.

"The formation and evolution of central cluster and group galaxies", European Week of Astronomy and Space Science, Turku, Finland, July 2013.

"The Lyman-alpha forest in a TeV-blazar heated Universe", Enigma workshop, Heidelberg, Germany, June 2013.

"Simulating modified gravity models", Cosmological Tests of Gravity, Oxford, UK, March 2013.

Julia Romanowska

"Studying lysozyme adsorption onto a charged surface with multimolecular BD simulations"; "Third Biological Diffusion and Brownian Dynamics Brainstorm: BDBDB3", Heidelberg, Germany, October 7-9, 2013.

"Research overview — Molecular and Cellular Modeling Group @ HITS"; Heidelberg Laureate Forum, HITS, Heidelberg, Germany, September 25, 2013.

Lei Shi

"Excemplify: A Flexible Template Based Solution, Parsing and Managing Data in Spreadsheets for Experimenta-

lists", 9th International Symposium on Integrative Bioinformatics 2013, Gatersleben (Germany) March 18-20, 2013.

Volker Springel

"Simulating the Universe", Statistical and related methods, University of Graz, Graz, Austria, April 10, 2013.

"Simulating the Universe with Particles: Current techniques, results and challenges",
PARTICLES 2013: Particle-based methods, fundamentals and challenges", University of Stuttgart, Stuttgart, September 18 - 20, 2013.

"Physics and resolution challenges in galaxy formation simulations", Mind the gap: from microphysics to large-scale structure in the Universe, University of Cambridge, Cambridge, UK, July 8 - 12, 2013.

"Simulations of Cosmic Structure Formation", Large-scale structure and the first objects, USP Cosmology Conference, Sao Paulo, Brazil, Feb 4 - 7, 2013.

Alexandros Stamatakis

"Software bugs and identification of mislabeled sequences", Symposium on detecting Errors in Phylogenies, BioSyst.EU 2013, Vienna, Austria, February 2013.

Rebecca Wade

"Protein dynamics and binding: From computation to experiment and back", DKFZ-ZMBH Alliance Scientific Advisor Board Meeting, Heidelberg, Germany, November 27, 2013.

Ulrike Wittig

"Publications from a Biocurator's Point of View", Workshop
"Biocuration and scholarly communication cycle: roles and opportunities for biocurators" at the 6th International Biocuration Conference, Cambridge (UK) April 7-10, 2013.

"Publications from SABIO-RK's Point of View", 6th Beilstein Symposium on Experimental Standard Conditions of Enzyme Characterizations (ESCEC), Rüdesheim (Germany) September 16-20, 2013.

Xiaofeng Yu

"Cytychrome P450: an important metabolic enzyme family studied by molecular dynamics simulation". DKFZ-ZMBH alliance retreat, Kloster Schöntal, Germany, February 5-7, 2013

Denis Yurin

"A new iterative method for the construction equilibrium N-body models", MODEST13a, Almaty, Kazakhstan, Aug 19 – 23, 2013.

"An iterative method for constructing N-Body equilibrium models", Virgo Consortium Meeting, Garching, Germany, Jun 10 – 12, 2013.

Beifei Zhou

"Prestress in Protein Disulfide Bonds Tunes Their Stabilities", Biophysical Society 57th Annual Meeting. Philadelphia, USA. 2-6, Feb, 2013.

"Pre-Stress Tunes Stability of Regulatory Protein Disulfide Bonds", Theory workshop, Computer Simulations and Theory of Macromolecules,
Hünfeld, Germany, 26-27 Apr, 2013.

POSTERS

Camilo Aponte-Santamaría, Carsten Baldauf, and Frauke Gräter

"Computer Simulation and Theory of Macromolecules", Theory workshop, Computer Simulations and Theory of Macromolecules,
Hünfeld, Germany, 26-27 Apr, 2013.

Camilo Aponte-Santamaría, Carsten Baldauf, and Frauke Gräter

"Force-Dependent Platelet Binding of the Von Willebrand Factor A1-A2",
Complex from Atomistic Simulations". 57th Biophysical Society meeting, Philadelphia, USA, 2-6 Feb, 2013

Andreas Bauer

"Studying reionization using GPU", Radiative transfer treatments for astrophysical applications, Leiden University, Leiden, Netherlands, May 27 - 31, 2013.

"Studying reionization using GPU", Exascale Computing in Astrophysics, Ascona, Switzerland, September 8 -13, 2013.

Mykhaylo Berynskyy, Antonia Stank, Rebecca C. Wade

"Modeling the interactions and structural dynamics of chaperone proteins". Third Biological Diffusion and Brownian Dynamics Brainstorm, Heidelberg, Germany, October 7-9, 2013.

Bronowska Agnieszka, Kolar, M., Hoba P.

"OLD NEWS, NEW STAR: Halogen bonding for medicinal chemistry and rational drug design." BDBDB3), Heidelberg, Germany, 7-9, Oct, 2013.

Bronowska Agnieszka, Kolar M., Fanfrlik J., Lepsik M., Hoba P.

"The devil is in details: halogen-bonding as a powerful tool for medicinal chemists." RICT2013 Drug Discovery and Selection, Nice, France, 3-5, Jul, 2013.

Cedric Debes, Frauke Gräter

"Evolution of protein mechano-stability", Holderness School, Holderness, USA, 16-21, Jun, 2013

Jonathan Fuller, Martin Golebiewski, Michael Martinez, Wolfgang Müller, Maja Rey, Ulrike Wittig, Rebecca Wade

"F3: Kinetic Data Handling". Virtual Liver Network – Midterm review, Berlin, Germany, January 30-31, 2013.

Jonathan Fuller, Michael Martinez, Stefan Henrich, Stefan Richter, Rebecca Wade

"LigDig: A tool to identify off-target effects and promiscuous binding in Systems Biology studies". 3DSIG, Berlin, Germany, July 19-20, 2013.

"LigDig: A tool to identify off-target effects and promiscuous binding in Systems Biology studies". ISMB/ECCB 2013, Berlin, Germany, July 21-23, 2013.

Jonathan Fuller, Michael Martinez, Rebecca Wade

"A3.2: Cross-talk of signaling pathways and endocytic machinery in heptatocytes: Impact on cell polarity and presence of transporters". Virtual Liver Network – Midterm review, Berlin, Germany, January 30-31, 2013.

Martin Golebiewski, Lihua An, David Shockley, Andreas Weidemann, Elina Wetsch, Wolfgang Müller

"The Virtual Liver Data Management System", Mid-term Evaluation Meeting Virtual Liver, Berlin (Germany) January 30-31, 2013.

Martin Golebiewski, Andreas Weidemann, Lihua An, Meik Bittkowski, Quyen Nguyen, Ivan Savora, David Shockley, Wolfgang Müller

"Assembling the jigsaw puzzle of liver data: The Virtual Liver Data Management", 14th International Conference on Systems Biology (ICSB 2013), Copenhagen (Denmark) August 30 – September 3, 2013.

Martin Golebiewski, Katy Wolstencroft, Olga Krebs, Stuart Owen, Quyen Nguyen, Lihua An, Meik Bittkowski, Ivan Savora, David Shockley, Dawie van Niekerk, Franco du Preez, Andreas Weidemann, Jacky L. Snoep, Wolfgang Müller, Carole Goble

"SEEK & Find: A Platform for Experimentalists and Modellers", 14th International Conference on Systems Biology (ICSB 2013), Copenhagen (Denmark) August 30 – September 3, 2013.

Stefan Henrich, Jonathan Fuller, Michael Martinez, Antonia Stank, Stefan Richter, Rebecca C. Wade

"Computational analysis of interactions between enzymes, reactants and allosteric modifiers". International Conference on the Systems Biology of Human Disease, Heidelberg, Germany, June 12-14, 2013.

Vincent Heuveline, Maximilian Hoecker, Michael Schick

DMQ / HITS Booth at the International Supercomputing Conference (ISC), Leipzig, Germany, 16-20 June 2013.

A. Kadlcikova, D. Palanisamy, J. Fanfrlik, M. Lepsik, A.K. Bronowska, P. Majer, P. Hoba

"Novel Selective CK2 Inhibitors based on Polybrominated Benzimidazols", ISTCP-VIII Congress, Budapest, Hungary, 25-31 Aug, 2013.

Daria Kokh

TRAPP- a tool for detection and analysis of protein cavity dynamics and prediction of transient binding pockets , Gordon Research Conference, "Computer Aided Drug Design" , Mount Snow, VT, USA July 21-26, 2013.

Olga Krebs

"SEEK and JERM: Standard-compliant integration of systems biology data", 9th International Symposium on Integrative Bioinformatics 2013, Gatersleben (Germany) March 18-20, 2013.

Maxime Louet, David Perahia, Jean Martinez, Nicolas Floquet

"GDP release in G-protein deciphered by Molecular Mechanics", Theory workshop, Computer Simulations and Theory of Macromolecules 2013, Hünfeld, Germany,April 26-27, 2013.

Maxime Louet, Wolfram Stacklies, Christian Seifert, Senbo Xiao, Bogdan Costescu, Frauke Gräter

"Force Distribution Analysis for Molecular Dynamics Simulations", 18th conference of the French Graphics and Molecular Modeling Group,
Oléron, France, May, 21-23, 2013.

Maxime Louet, Wolfram Stacklies, Christian Seifert, Senbo Xiao, Bogdan Costescu, Frauke Gräter

"Force Distribution Analysis for Molecular Dynamics", Workshop Defis - Biophysics of large macromolecular assemblies: experiments and simulations, École Normale Supérieur Cachan, France, Oct. 14-15, 2013.

Michael Martinez, Julia Romanowska, Daria Kokh, Stefan Richter, Raż f Gabdoulline, Rebecca Wade

"SDA 7 (SDA flex)", BDBDB3, Heidelberg, Germany, October 7-9 2013.

Davide Mercadante, Frauke Gräter

"Force-extension and energy profiles of Intrinsically Disordered Proteins using MARTINI", Theory workshop, Com-

puter Simulations and Theory of Macromolecules, Hünfeld, Germany, Apr 26-27, 2013.

Ghulam Mustafa, Xiaofeng Yu, Zaheer ul-Haq, Rebecca Wade

"Modeling and Simulation of Cytochrome P450 - Membrane, Substrate and Product Interactions", CCP5 Method in Molecular Simulation Summer School 2013, Manchester University, UK, July 21-30 2013.

Musa Oeb oyaci, Daria Kokh, Rebecca Wade

"Adsorption of 3H-BLIP on a gold surface". Computational Biology: Then and Now, Rehovot, Israel, May 6-9, 2013.

Juan Calos Basto Pineda

"Dark Matter Inferences from Rotation Curve Fitting", Deconstructing Galaxies: Structure and Morphology in the Era of Large Surveys. Santiago de Chile, Chile, November 17-22, 2013.

Kai Polsterer, Fabian Gieseke, Christian Igel, Tomotsugu Goto

"Improving the Performance of Photometric Regression Models via Massive Parallel Feature Selection". 23Rd Annual Astronomical Data Analysis Software and Systems (ADASS) Conference, October 2013.

Stefan Richter, Antonia Stank, Jonathan Fuller, Rebecca C. Wade

"Simple visualisation of functional annotations on protein structures". International Conference on the Systems Biology of Human Disease, Heidelberg, Germany, June 12-14, 2013.

Kevin Schaal

"HAVOC - Here's another Voronoi code", The CosmoComp/CHARM 2013 Spring School: Radiative transfer treatments for astrophysical applications, Leiden, Netherlands, May 27-31, 2013.

Siegfried Schloissnig

"A genome-scale resource for in vivo tag-based protein function exploration in C. elegans.", 19th International C.elegans Meeting, University of California, Los Angeles, June 26-30, 2013.

"C. briggsae genomic fosmid library", 19th International C.elegans Meeting, University of California, Los Angeles, June 26-30, 2013.

Andreas Weidemann

"F2: The Liver Knowledgebase: an integrated environment for liver molecular network representation and analysis", VLN Evaluation Meeting, Berlin (Germany) January 30-31, 2013.

Andreas Weidemann, Stefan Richter, Stefan Henrich, Wolfgang Müller, Rebecca Wade, Ursula Kummer

"SYCAMORE – web-based modeling and simulation", ICSB, Copenhagen (Denmark) August 30- September 3, 2013.

Ulrike Wittig, Renate Kania

"Publications from a Biocurator's Point of View", 6th International Biocuration Conference, Cambridge (UK) April 7-10, 2013.

Ulrike Wittig, Renate Kania, Maja Rey

"Publications from SABIO-RK's Point of View", 6th Beilstein Symposium on Experimental Standard Conditions of Enzyme Characterizations (ESCEC), Rüdesheim (Germany) September 16-20, 2013.

Xiaofeng Yu, Vlad Cojocaru, Ghulam Mustafa, Outi Salo-Ahen, Galina Lepesheva, Rebecca Wade

"The Dynamics of Parasitic CYP51". 18th International Conference on Cytochrome P450, Seattle, Washington, USA, June 18-22, 2013.

Xiaofeng Yu, Jon Fuller, Stefan Henrich and Rebecca Wade

"Protein modeling in the Molecular and Cellular Mode¬ling

(MCM) group at HITS". DKFZ-ZMBH alliance retreat, Kloster Schöntal, Germany, February 5-7, 2013.

Xiaofeng Yu, Michael Martinez, Ghulam Mustafa, Rebecca Wade
"The Association of Cytochrome P450s and Their Redox Partners". Third Biological Diffusion and Brownian Dynamics Brainstorm: BDBDB3, Heidelberg, Germany, October 7-9, 2013.

Jing Zhoung, Frauke Gräter
"Forcal adhesion kinase as cellular sensor", Complex from Atomistic Simulations". 57th Biophysical Society meeting, Philadelphia, USA, Feb 2-6, 2013

DEMOS

Martin Golebiewski & Andreas Weidemann
"Bridging Experiments and Modelling: SABIO-RK - Reaction Kinetics Database", ICSB 2013 tutorial workshop, Copenhagen (Denmark) September 4, 2013.

Olga Krebs
Software demo: SEEK, 9th International Symposium on Integrative Bioinformatics 2013, Gatersleben (Germany) March 18-20, 2013.

Michael Martinez
"SDA 7 (SDA flex)", BDBDB3, Heidelberg (Germany), October 7-9 2013.

Quyen Nguyen
"Seek new and upcoming features", 7th All Hand PALS Meeting, Paris (France) September 19-20, 2013.

Stefan Richter
"Mitmachstation" for kids, HITS Open Day, HITS, June 4, 2013.

Andreas Weidemann
"SYCAMORE", ICSB 2013 tutorial workshop, Copenhagen (Denmark) September 4, 2013.

Ulrike Wittig
"SABIO-RK – Database for Reaction Kinetics"
9th International Symposium on Integrative Bioinformatics 2013, Gatersleben (Germany) March 18-20, 2013.
Katy Wolstencroft & Martin Golebiewski
"SysMO-DB and Virtual Liver SEEK: Sharing and Managing Systems Biology Experiments", ICSB 2013 tutorial workshop, Copenhagen (Denmark) September 4, 2013.

9.3 MEMBERSHIPS

Fabio Fontanot
Member of the EUCLID Satellite Consortium

Tilmann Gneiting
Editor for Physical Science, Computation, Engineering, and the Environment, Annals of Applied Statistics

Guest Editor, Nonlinear Processes in Geophysics

Affiliate Professor, Department of Statistics, University of Washington, Seattle (USA)

Guest faculty member, Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University

Associated faculty member, HGS MathComp Graduate School, Heidelberg University

Faculty member, Research Training Group 1653, Spatial/Temporal Probabilistic Graphical Models and Applications in Image Analysis, Heidelberg University

Faculty member, Research Training Group 1953, Statistical Modeling of Complex Systems and Processes: Advanced Nonparametric Methods, Heidelberg University and Mannheim University

Chair, Committee on the Special Year of the Mathematics of Planet Earth (MPE 2013), Institute of Mathematical Statistics (IMS)

IMS Representative, Committee of Presidents of Statistical Societies (COPSS) Awards Committee

Member of the Dissertation Award Committee, Fachgruppe Stochastik (Probability and Statistics Section of the German Mathematical Society)

Martin Golebiewski
Board member of the COMBINE network (Computational Modeling in Biology network)

Frauke Gräter
Member of BIOMS (Heidelberg Center for Modeling and Simulation in the Biosciences) Steering Committee

Guest faculty member, Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University

Associated faculty member, HGS MathComp Graduate School, Heidelberg University

Faculty member, Hartmut Hoffmann-Berling International Graduate School of Molecular and Cellular Biology (HBIGS), Heidelberg University

Christopher Hayward
Member of the American Astronomical Society

Member of the American Physical Society

Member of the Association for the Advancement of Science

Renate Kania
Member of the International Society for Biocuration (ISB) Executive Committee

Christoph Pfrommer
Associate member of the LOFAR Magnetism Key Science Project

External collaboration member of the MAGIC Cherenkov Telescope Collaboration

External collaboration member of the Fermi Space Telescope Collaboration

Kai Polsterer
Member of the Astronomische Gesellschaft (A.G.)

Member of the International Astrostatistics Association

Member of the Knowledge Discovery in Databases Interest Group of the International Virtual Observatory Alliance

Andreas Reuter
Scientific Member of Max-Planck-Gesellschaft (Max Planck Institute of Computer Science, Saarbrücken)

Member of the Scientific Committee, BIOMS, Heidelberg

Member of the Advisory Board of Fraunhofer Gesellschaft Informations- und Kommunikationstechnik (IuK)

Member of the Heidelberg Club International

Member of the Board of Trustees of the Wissenschaftspressekonferenz, Bonn

Co-editor “Database Series”, Vieweg-Verlag

Member of a research group on scientific computing named “WIR”

Member of Dagstuhl`s Industrial Curatory Board of „Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI)“, Dagstuhl (Leibniz Center for Computer Science)- bis Mai 2015

Member of Schloss Dagstuhl´s Scientific Advisory Board

Chairman of the Supervisory Board of SICOS GmbH, Stuttgart

Member of the Board of Directors at IWR, Heidelberg University

Member of the Program Committee for the International Joint EDBT/ICDT Ph. D. Workshop 2012, March 30, 2012 in Berlin

Member of the search committee for a professorship on “Computational Statistics” at KIT, Karlsruhe

Member of the search committee for a professorship on “High-Energy Astrophysics” at Heidelberg University

Member of the search committee for a professorship on “Scientific Computing” at Heidelberg University

Member of the search committee for a professorship on “Molecular Bio-Mechanics” at Heidelberg University

Member of the advisory committee “E-Science” at MWK, Stuttgart
Member of the advisory committee “IKT 2020” at MWK, Stuttgart

Christine Simpson
Member of the American Astronomical Society

Volker Springel
Member of the Interdisciplinary Center for Scientific Computing (IWR), Heidelberg

External Scientific Member of the Max-Planck-Institute for Astronomy, Heidelberg

Member of the Interdisciplinary Astronomical Union (IAU)

Member of the Cosmological Simulation Working Group (CSWG) of the EUCLID mission of ESA

Member of the Research Council of the Field of Focus “Structure and pattern formation in the material world” at Heidelberg University

Member of the Board of SFB 881 “The Milky Way System”

Alexandros Stamatakis
Member of the council of the Society of Systematic Biologists

Member of the steering committee of the Munich Supercomputing System HLRB at LRZ (Leibniz RechenZentrum)

Member of the scientific advisory board of the Computational Biology Institute in Montpellier, France

Michael Strube
Editorial Board: Dialogue & Discourse Journal, The Journal of Data Semantics

Rebecca Wade
Associate Editor: Journal of Molecular Recognition

Section Editor: BMC Biophysics

Associate Editor: PLOS Computational Biology

Editorial Board: BBA General Subjects; Journal of Computer-aided Molecular Design; Biopolymers; Current Chemical Biology; Protein Engineering, Design and Selection; Computational Biology and Chemistry: Advances and Ap-

plications; Open Access Bioinformatics

Member of Scientific Advisory Council of the Leibniz-Institut für Molekulare Pharmakologie (FMP), Berlin-Buch

Member: BIOMS Steering Committee, Heidelberg

Member at Heidelberg University of: CellNetworks Cluster of Excellence, HBIGS (Hartmut Hoffmann-Berling International Graduate School of Molecular and Cellular Biology) faculty, HGS MathComp faculty, IWR, DKFZ-ZMBH Alliance

9.4 CONTRIBUTIONS TO THE SCIENTIFIC COMMUNITY

REFEREE WORK

Andre Aberer
Bioinformatics

Genetics

Molecular Phylogenetics and Evolution

Camilo Aponte Santamaria
PLoS One - Journal of Structural Biology, DNA and cell biology

Agnieszka Bronowska
Molecular BioSystems
Journal of Materials Chemistry B
Metallomics
Journal of Computer-Aided Molecular Design

Tomas Flouri
Theoretical Computer Science

Journal of Discrete Algorithms

Acta Informatica

Algorithmica

Fabio Fontanot
Monthly Notices of the Royal Astronomical Society

Astrophysical Journal

Jonathan Fuller
PLoS One - Journal of Molecular Recognition, BMC Structural Biology.

Tilmann Gneiting
Journal of Statistical Software

Statistical Science

Frauke Gräter
Biophysical Journal

Journal of the American Chemical Society

Journal for Physical Chemistry B

Nature Chemistry

Proceedings of the National Academy of Sciences

German Research Society (DFG)

Danish Council for Independent Research

Netherlands Organization for Scientific Research

PRACE

Christopher Hayward
Monthly Notices of the Royal Astronomical Society

Astrophysical Journal

Gemini Observatory Canadian Time Allocation Committee

Polish National Science Centre

Fernando Iguerdo-Carrasco
Systematic Biology

Daria Kokh
PLOS Computational Biology

Rüdiger Pakmor
Astrophysical Journal

Astrophysical Journal Letters

Christoph Pfrommer
Astrophysical Journal

Astrophysical Journal Letters

Monthly Notices of the Royal Astronomical Society

Marsden Fund of New Zealand

Physical Review Letters

Physical Review D

Solon Pissis
Journal of Theoretical Computer Science

Philoshophical Transactions A

ACM Journal of Experimental Algorithmics

PLOS Currents

Journal of Discrete Algorithms

Algorithms

International Journal on Artificial Intelligence Tools

Andreas Reuter
Deutsche Forschungsgemeinschaft; Fonds zur Förderung der wissenschaftlichen Forschung (Österreich)

Julia Romanowska
Biophysical Journal

Siegfried Schloissnig
Bioinformatics

RECOMB2014 (18th Annual International Conference on Research in Computational Molecular Biology)

Volker Springel
Astrophysical Journal

Monthly Notices of the Royal Astronomical Society

Nature

Physical Review Letters

Astronomische Nachrichten

Agence nationale de la recherche, France

Max-Planck Society

German Science Foundation (DFG)

Royal Society of the UK

European Research Council

German-Israel Science Foundation

PRACE

Jülich Supercomputer Center

Alexandros Stamatakis
PLOS One

Systematic Biology

Genome Biology

PLOS Computational Biology

Future Generation Computer Systems

Computing – Archives for Scientific Computing

Michael Strube
Computational Linguistics Journal

Journal of Artificial Intelligence Research

Journal of the American Society for Information Science and Technology

Journal of Machine Learning Research

Journal of Natural Language Engineering

Transactions of the Association for Computational Linguistics

PROGRAM COMMITTEE MEMBERSHIPS

Tomas Flouri
String Processing and Information Retrieval Symposium (SPIRE 2013), Jerusalem, Israel, October 2013

Conference on Implementation and Application of Automata (CIAA 2013), Halifax, Canada, July 2013

International Colloquium on Automata, Languages and Programming (ICALP 2013), Riga, Latvia, July 2013

Jonathan Fuller
Virtual Liver Network Junior Researchers Retreat 2013, Hünfeld, Germany, October 28-30, 2013
Heidelberg Unseminars in Bioinformatics, Heidelberg, Germany, 2013

Renate Kania
Member of the Organization Committee for the Conference of the International Society for Biocuration, Cambridge, (UK), 7-10 April 2013.

Solon Pissis
Conference on Advances in Databases and Information Systems (ADBIS 2013) Genoa, Italy, September 2013

International Workshop on Data Mining in Bioinformatics (BIOKDD 2013), Chicago, USA, August 2013
International Workshop on Combinatorial Algorithms (IWOCA 2013) Rouen, France, July 2013

International Workshop on High Performance Computational Biology (HiCOMB 2013) Boston, Usa, May 2013

FASTAR/Espresso Workshop 2013, South Africa, November 2013

Andreas Reuter
Member of the Scientific Committee of the 1st Heidelberg Laureate Forum, September 2013

Volker Springel
Mind the gap: from microphysics to large-scale structure in the Universe, University of Cambridge, Cambridge, UK, July 8 - 12, 2013.

Alexandros Stamatakis
BIOIT 2013, Utah, US, December 2013

ICPP 2013, Lyon France, October 2013

Parallel Bio-Computing Workshop (PBC 2013) held in conjunction with PPAM, Warsaw Poland, September 2013

International Supercomputing Conference 2013, Leipzig, Germany, June 2013

Michael Strube
ACL 2013: 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria

IJCNLP 2013: International Joint Conference on Natural Language Processing, Nagoya, Japan

NAACL 2013: 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies
StarSEM 2013: 2nd Joint Conference on Lexical and Computational Semantics

Rebecca Wade
Scientific Advisory Committee, ESF-EMBO 4th International Conference on Molecular Perspectives on Protein-Protein Interactions, Pultusk, Poland, May 25-30, 2013

ORGANIZATION COMMITTEE MEMBERSHIP (CHAIR)

Jonathan Fuller
Virtual Liver Network Junior Researchers Retreat 2013, Hünfeld, Germany, October 28-30, 2013

Heidelberg Unseminars in Bioinformatics, Heidelberg, Germany, 2013

Daria Kokh, Stefan Richter, Jon Fuller , Xiaofeng Yu (with Rommie Amaro (UCSD) and Franziska Matthäus (Bioquant, Heidelberg University) Third Biological Diffusion and Brownian Dynamics Brainstorm: BDBDB3, Heidelberg, Germany, October 7-9, 2013

Andreas Reuter
Scientific Chair of the 1st Heidelberg Laureate Forum, September 2013

Michael Strube
SIGdial 2013: 14th Annual SIGdial Meeting on Discourse and Dialogue, Metz, France, August 23-24 (General Co-Chair)

WORKSHOP ORGANIZATION

Wolfgang Müller
ODLS workshop: Ontologies and Databases in the Life Sciences, GI Jahrestagung 2013, Koblenz.

Volker Springel
AREPO Workshop 2013, Heidelberg Academy of Science, Heidelberg, September 25 - 27, 2013.

9.5 AWARDS

Eduardo R. Cruz Chu (shared with Davide Mercadante and Frauke Gräter)
Grant of Supercomputer Time. Two-million CPU hours on the supercomputer Juropa at the Julich Supercomputing Center for the project "Computational Studies of Nacre". Nov, 2013.

Davide Mercadante (shared with Eduardo R. Cruz Chu and Frauke Gräter)
Grant of Supercomputer Time. 100,000 CPU hours on the supercomputer FERMI@Cineca, Italy, for the project "Probing force-dependency of formin-mediated actin polymerization".

Grant of Supercomputer Time. 100,000 CPU hours on the supercomputer SuperMUC, GAUSS@LRZ, Germany, for the project "Probing force-dependency of formin-mediated actin polymerization".

Grant of Supercomputer Time. Ten-million CPU hours on the supercomputer Juqueen at the Julich Supercomputing Center for the project "Computational Studies of Nacre". Nov. 2013.

Ghulam Mustafa
"Best Poster Prize" during the CCP5 Summer School "Methods in Molecular Simulation" at the University of Manchester, July 20-30 2013.

Andreas Reuter
Honorary Certificate by Donetsk National Technical University, Ukraine, November 2013.

Annual Report

2013

Heidelberg Institute for Theoretical Studies

HITS

Edited by

HITS gGmbH
Schloss-Wolfsbrunnenweg 35
D-69118 Heidelberg
www.h-its.org
@HITStudies
Facebook.com/HITStudies

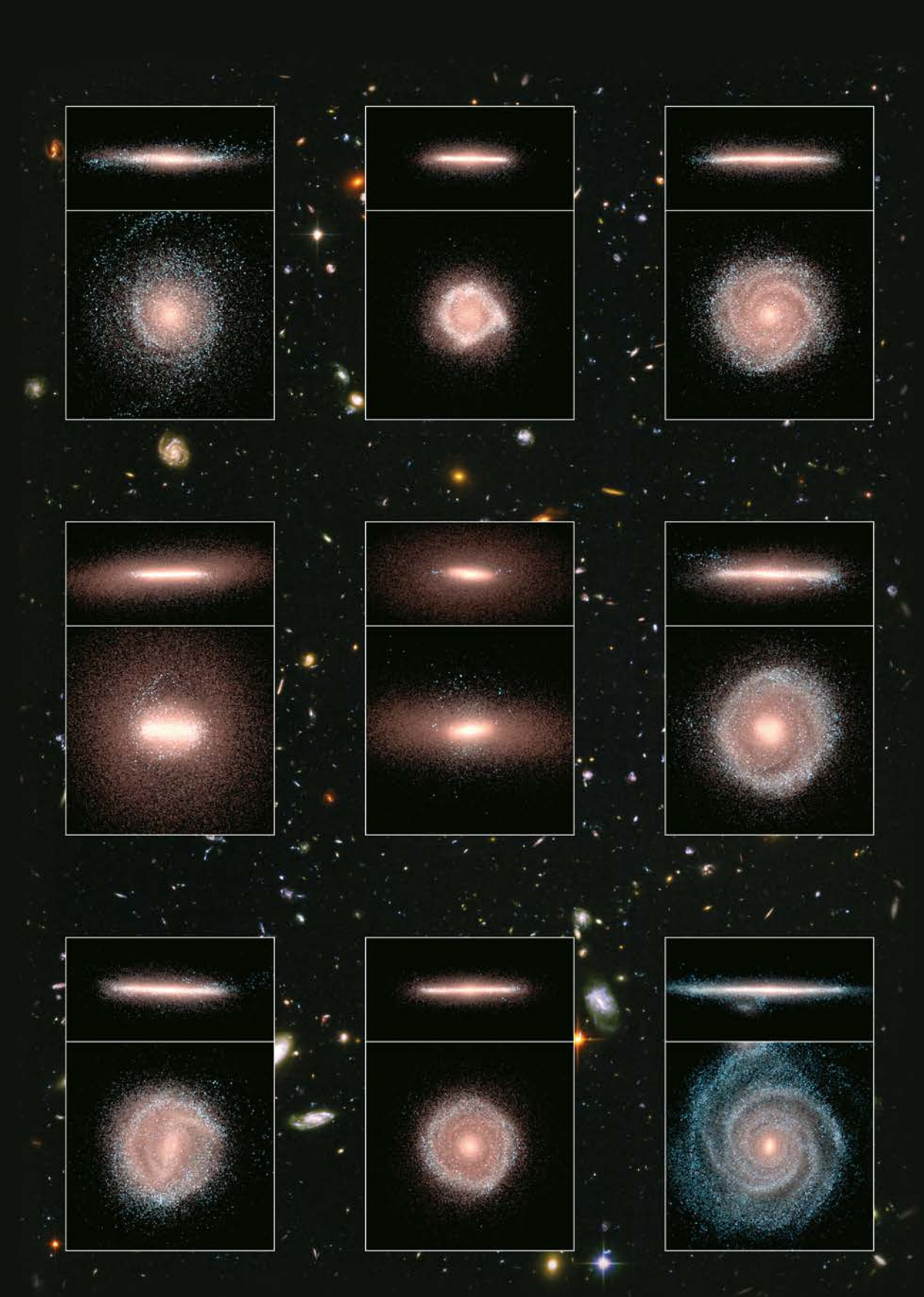
Our e-mail adresses have the following structure:
Firstname.lastname@h-its.org

Contact
Dr.Peter Saueressig
Phone: +49-6221-533 245
Fax: +49-6221-533 298
@HITStudies

Editor
Dr.Peter Saueressig
Public Relations

Pictures
HITS gGmbH,
(unless otherwise indicated)

ISSN 1438-4159



Predicted galaxy morphologies: Maps of the stellar component in different simulations of Milky Way-sized galaxies, in each case in an edge-on (top) and a face-on (bottom) view, overlaid on the Hubble Ultra Deep Field (background image). Colors in the images correspond to the actual emission in different spectral bands of the stellar populations. (see chapter 2.10)

Vorhergesagte Morphologie von Galaxien: Karten der stellaren Komponente in verschiedenen Simulationen von Galaxien der Größe der Milchstraße, jeweils in einer Seitenansicht (oben) und einer Aufsicht (unten). Das Hintergrundbild zeigt das Hubble Ultra Deep Field. Die Farben entsprechen dabei der tatsächlichen Emission in verschiedenen Spektralbereichen der Sternpopulationen. (siehe Kapitel 2.10)