Heidelberg Institute for
Theoretical Studies

HITS

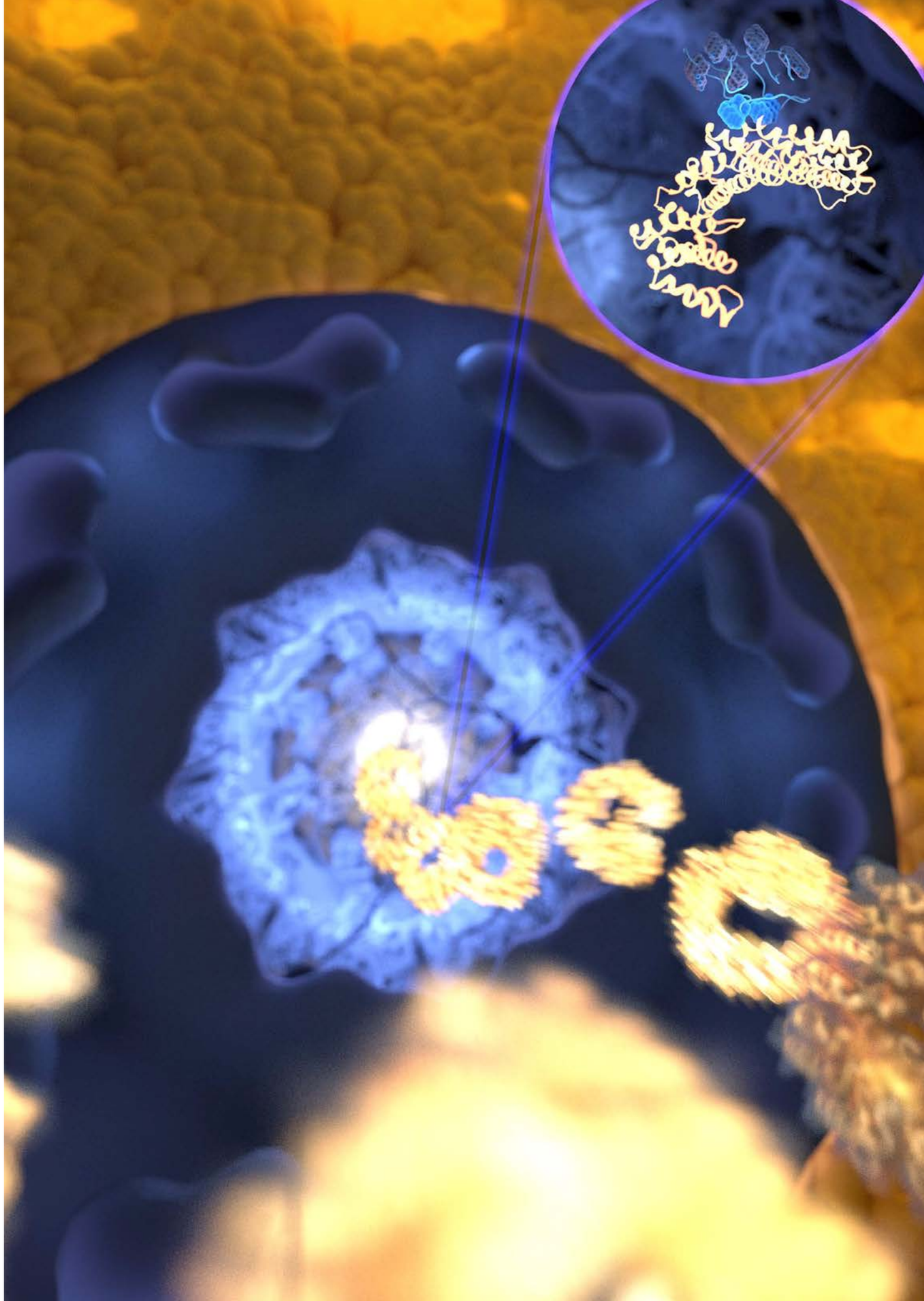
THINK BEYOND THE LIMITS

2015 Annual Report | Jahresbericht

The ultrafast and yet selective binding allows the receptor (gold) to rapidly travel through the pore filled with disordered proteins (blue) into the nucleus, while any unwanted molecules are kept outside *(see chapter 2.6, p. 77).*

Durch die ultraschnelle, aber zugleich gezielte Bindung rast der Rezeptor (goldfarben) durch die mit ungeordneten Proteinen gefüllte Pore in den Zellkern, während unerwünschte Moleküle ferngehalten werden (siehe Kapitel 2.6, S.77).

Image/Bild: Davide Mercadante, Sebastian Sturm, Molecular Biomechanics group.



In Memoriam Klaus Tschira (1940 – 2015)

© Klaus Tschira Stiftung

When in mid-2007 Klaus Tschira came up with the idea of establishing a new institute, he laid down three basic rules:

It was to be called "Heidelberg Institute for Theoretical Studies" – HITS, for short. (His favorite tongue-in-cheek explanation was: "Mannheim has a Pop Academy, we have HITS.")

It was to become an internationally recognized, interdisciplinary research institute covering the natural sciences, mathematics and computer science.

The main aim was not to establish a particular mix of research disciplines, but to do top-level cutting-edge research.

All other decisions were subordinate, simply means for achieving the goal described by the basic rules. We already had "EML Research," a small, interdisciplinary institute with a good reputation and a research agenda fitting in with the above definitions. So the pragmatic solution was to use it as the nucleus of the new institute. The "T" in HITS (referring to the focus on theory) was interpreted as "computational science" due to the fact that in all areas of scientific endeavor, computer-based methods are becoming ever more important. And when deciding on new research groups, the key criterion was never the research topic. The only thing that mattered was to attract top talents to head our research groups.

Klaus Tschira strictly adhered to these principles, and he defended them against all attempts to persuade him otherwise. And it was this clarity and this dependability that were the foundation for the high reputation that HITS very quickly earned in both the scientific community and in political circles – and which is demonstrated by the fact that Heidelberg University and Karlsruhe Institute of Technology have become shareholders of HITS.

It is just as important that, beyond these structural propositions, Klaus Tschira gave the research groups complete freedom to decide on the topics they wanted to investigate and the methods they intended to use. He was always very keen on learning about new projects and new results – he was a regular at the "Scientific Seminars" held every other Monday. With HITS, he created a paradigmatic institution for "unfettered research." There aren't many of them around anymore. Limelight for himself was never Klaus Tschira's motive in doing all this – he was not that kind of person. He was genuinely proud when he could tell people about the three ERC grants that HITS researchers had won, or when he was told that two of the most frequently-cited researchers worldwide are members of our institute. Those are the kind of results he had in mind, he saw them as the rewards for his personal efforts.

All HITSters (the people working at HITS) agree that this is a special place, a place where you can pursue curiosity-driven research and get generous support for doing so. This is Klaus Tschira's legacy, and we all feel the obligation – it is at the same time a very great honor – to preserve this legacy and to build on it, in line with the basic rules he set down back in 2007.

Klaus Tschira occasionally teased us, pointing out that we hadn't won a Nobel Prize yet and that he wouldn't mind at all if we did. Of course, we cannot promise anything in this respect, but we will do our very best. And even if we do not succeed in the foreseeable future, for all of us it is both a privilege and a pleasure to have the chance to try, right here at HITS, in the spirit of Klaus Tschira and in fond remembrance of him.

Andreas Reuter

Als Klaus Tschira Mitte 2007 den Plan zur Einrichtung eines neuen Institutes entwickelte, waren drei Dinge fest vorgegeben:

Der Name sollte „Heidelberger Institut für Theoretische Studien" – kurz: HITS – lauten. (Er hat ihn späterhin gern augenzwinkernd so erklärt: „In Mannheim haben sie die Pop-Akademie, wir haben die HITS.")

Es sollte ein international sichtbares, interdisziplinäres Forschungsinstitut auf den Gebieten der Naturwissenschaften, der Mathematik und der Informatik sein.

Ziel sollte nicht eine irgendwie definierte fachliche Ausgewogenheit sein, sondern Forschung auf höchstem Niveau.

Alle anderen Entscheidungen waren nachrangig und dienten nur dem Zweck, diese drei Prämissen umzusetzen. Da gab es z. B. das „EML Research", ein kleines, multidisziplinäres Institut mit gutem Renommee, dessen fachliche Ausrichtung mit der zweiten Vorgabe kompatibel war, also wurde pragmatisch entschieden, dies zum Ausgangspunkt des neuen Institutes zu machen. Das „T" im Namen (die theoretische Ausrichtung) wurde im Sinne von „Computational Science" interpretiert, der in allen Gebieten der Wissenschaften immer wichtiger werdenden Nutzung von rechnergestützten Methoden bei der Forschung. Und bei der Entscheidung über die Einrichtung neuer Gruppen stand nie die thematische Ausrichtung im Vordergrund; wichtig war und ist allein, ob es gelingt, herausragende Wissenschaftler für die Leitung der Gruppen zu gewinnen. Diese Prinzipien hat Klaus Tschira konsequent befolgt und auch gegen alle möglichen Versuche der Einflussnahme verteidigt. Und diese Klarheit, diese Verlässlichkeit waren entscheidend dafür, dass sich das HITS in Kreisen der Wissenschaft, aber auch in der Politik sehr schnell ein hohes Ansehen erwerben konnte – was nicht zuletzt dadurch zum Ausdruck kommt, dass die Universität Heidelberg und das KIT in Karlsruhe als Mitgesellschafter in die HITS gGmbH eingetreten sind.

Genauso wichtig ist aber die Tatsache, dass er jenseits der strukturellen Vorgaben die Forschungsgruppen völlig frei hat arbeiten und entscheiden lassen. Er hat sich immer sehr für ihre Projekte und Ergebnisse interessiert – so war er zum Beispiel ein regelmäßiger Teilnehmer an den montäglichen „Scientific Seminars" – aber hinsichtlich der Themenstellung oder der Methodenauswahl waren und sind die Gruppen völlig ungebunden. Er hat mit dem HITS eine paradigmatische Einrichtung des „unfettered research" geschaffen, von denen es auf der Welt viel zu wenige (und über die Zeit hin immer weniger) gibt.

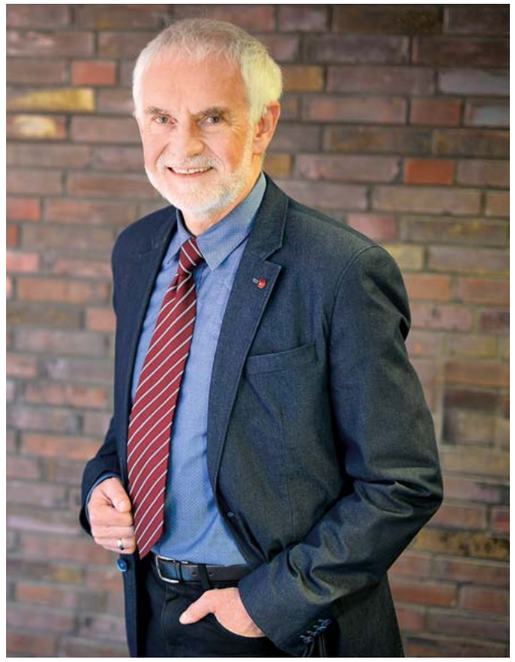
Und Klaus Tschira hat dies nicht getan, um Aufmerksamkeit auf sich zu lenken; das hätte ihm ohnehin nicht gelegen. Dagegen war er aufrichtig stolz, wenn er berichten konnte, dass die Wissenschaftler des HITS schon drei ERC-Grants eingeworben haben oder dass zwei der weltweit meistzitierten Forscher dort arbeiten. Das sind die Arten von Ergebnissen, die er angestrebt hat und die er als Belohnung seiner eigenen Arbeit ansah.

Alle, die am HITS arbeiten, stimmen darin überein, dass dies ein besonderer Platz ist, ein Ort, in dem freie, Neugier-getriebene Forschung möglich ist und großzügig unterstützt wird. Dies ist Klaus Tschiras Vermächtnis, und wir alle fühlen die Verpflichtung – die zugleich eine große Ehre ist –, dieses Vermächtnis zu erhalten und auszubauen, ganz im Sinne der ursprünglichen Vorgaben.

Klaus Tschira hat gelegentlich scherzhaft darauf hingewiesen, dass wir noch keinen Nobelpreis ans HITS geholt hätten, dass ein solcher aber durchaus billigend in Kauf genommen würde. Wir können natürlich in dieser Hinsicht nichts versprechen, aber wir alle werden uns sehr ernsthaft bemühen. Und auch wenn daraus in absehbarer Zeit nichts werden sollte: Es ist für uns alle ein Privileg und ein Vergnügen, es hier, im Institut von Klaus Tschira, versuchen zu dürfen – im Geiste Klaus Tschiras und in Erinnerung an ihn.

Andreas Reuter

Inhalt
Table of Contents



Prof. Dr. Rebecca Wade
(Scientific Director / Institutssprecherin)

Prof. Dr.-Ing. Dr. h.c. Andreas Reuter
(Managing Director / Geschäftsführer)

2015 could be defined as a year of firsts, transformation, and continuity for HITS. In other words, there is much to report and the pages that follow this foreword are full of interesting science and impressive achievements. But first, let's reflect on what defined 2015 for HITS.

Firsts: This is the first year that, following the establishment of the new HITS charter at the end of 2014, HITS has been under the dual leadership of a Managing Director and a Scientific Director. Over the year, we have worked out together how to share and divide the tasks previously carried out by Andreas Reuter and Klaus Tschira as Managing Directors. The new charter has also initiated new procedures, such as the first scientific evaluation by external reviewers coordinated by the Scientific Advisory Board. The first evaluation was of our three biologically oriented research groups, Computational Biology (CBI), Molecular and Cellular Modelling (MCM), and Molecular Biomechanics (MBM). The plan is to review different, thematically related clusters of about three research groups every year. The review process and the annual meetings of the Scientific Advisory Board are intended to assist in ensuring scientific excellence at HITS. Already this year, they have provided valuable impulses.

Transformation and continuity: Klaus Tschira's untimely death in March was obviously a watershed for HITS. But at the same time, HITS' new structure has enabled us to steadfastly continue to follow our agenda of multidisciplinary computational research. Indeed, HITS has continued to expand, its newest addition being the Physics of Stellar Objects group (PSO) headed by Fritz Röpke, who came from the University of Würzburg to join HITS in January 2015. Christoph Pfrommer won an ERC Consolidator grant to set up his own Junior Research group, High-Energy Astrophysics and Cosmology (HAC), which will start work in April 2016. Thus, HITS has a very strong presence in theoretical astrophysics and astronomy with four research groups active in these fields. In June, the Groups and Geometry (GRG) group led by Anna Wienhard was established as an associated research group. Anna, who also holds an ERC Consolidator grant, will retain her primary position as a professor at the Mathematics Department of Heidelberg University, but her affiliation with HITS will enable her to expand her research on symmetry, geometry, and topology in new, interdisciplinary directions. Furthermore, we are in the process of recruiting a Scientific Visualization group in conjunction with Heidelberg University. This will be part of a virtual center for Scientific Visualization with KIT, Heidelberg University, and HITS as the participating organizations.

It is five years since HITS saw the light of day and we moved into the 'new' building that this summer officially was named the "Klaus Tschira HITS Building". The "Mathematikon", a new building for Mathematics and Computer Science built with the support of the Klaus Tschira Stiftung on the Heidelberg University campus, opened at the end of the year. It will be home to the new Scientific Visualization group – and (by mid-2016) to our colleagues from EML, thus freeing up space on the HITS campus for some of HITS' 120+ scientists.

In January 2015 we marked HITS' 5th birthday with a symposium as part of the 20th anniversary celebrations of the Klaus Tschira Stiftung *(see chapter 5.3)*. In November, we held a symposium in honor of Andreas Reuter *(see chapter 5.9)*, who will step down as Managing Director in April 2016 after an overlap period with our new Managing Director, Gesa Schönberger, who started at HITS in January 2016. Andreas Reuter will remain Managing Director of the HITS Stiftung and senior professor at Heidelberg University, moderating a task force dedicated to developing a blueprint for a future IT infrastructure accommodating the needs of the research institutions in the Heidelberg-Mannheim-Karlsruhe region.

We are confident that with all these changes HITS is well-positioned for its consolidation phase, which is expected to start next year.

2015 könnte für das HITS als ein Jahr der Premieren, der Veränderung und der Kontinuität bezeichnet werden. Mit anderen Worten: Es gibt viel zu berichten, und die Seiten, die diesem Vorwort folgen, sind gefüllt mit faszinierender Wissenschaft und beeindruckenden Leistungen. Aber lassen Sie uns zunächst darüber nachdenken, welche Ereignisse das Jahr 2015 am HITS geprägt haben.

***Premieren:** Nach dem Inkrafttreten der neuen Gesellschaftssatzung Ende 2014 steht dem Institut seit diesem Jahr erstmals eine Doppelspitze vor, bestehend aus einem Geschäftsführer und einer Institutssprecherin. Im Laufe des Jahres haben wir gemeinsam festgelegt, wie die Aufgaben verteilt werden sollen, die zuvor von Andreas Reuter und Klaus Tschira als Geschäftsführer bewältigt worden waren. Durch die Satzung wurden auch neue Verfahren ins Leben gerufen, wie etwa die wissenschaftliche Evaluation durch externe Gutachter, die vom Wissenschaftlichen Beirat koordiniert wird. Als erste wurden unsere drei Forschungsgruppen aus dem Bereich Biologie, Computational Biology (CBI), Molecular and Cellular Modeling (MCM) und Molecular Biomechanics (MBM), evaluiert.*

Geplant ist, jedes Jahr unterschiedliche, thematisch verwandte Cluster von etwa drei Forschungsgruppen zu evaluieren. Dies und die jährlichen Sitzungen des Wissenschaftlichen Beirats werden dabei helfen, die wissenschaftliche Exzellenz des HITS zu gewährleisten; sie haben bereits in diesem Jahr wertvolle Impulse gegeben.

***Veränderung und Kontinuität:** Klaus Tschiras plötzlicher Tod im März war eine schwere Zäsur für das HITS. Allerdings ermöglichte uns die neue Struktur des Instituts jedoch, unsere Agenda der multidisziplinären computergestützten Forschung unbeirrt fortzuführen. Tatsächlich expandierte das HITS kontinuierlich, etwa mit der neuen Gruppe Physik Stellarer Objekte (PSO) unter der Leitung von Fritz Röpke, der im Januar 2015 von der Universität Würzburg ans HITS kam.*

Christoph Pfrommer gewann einen ERC Consolidator Grant, um seine eigene Junior-Forschungsgruppe High-Energy Astrophysics and Cosmology (HAC) aufzubauen, die im April 2016 ihre Arbeit aufnehmen wird. In der theoretischen Astrophysik und der Astronomie ist das HITS mit seinen vier Forschungsgruppen in diesem Bereich nunmehr sehr präsent.

Im Juni wurde die Forschungsgruppe „Gruppen und Geometrie“ (GRG) als assoziierte Gruppe unter der Leitung von Anna Wienhard gegründet. Anna, die auch einen ERC Consolidator Grant hat, wird ihre Stelle als Professorin an der Fakultät für Mathematik der Universität Heidelberg behalten, ihre Zugehörigkeit zum HITS wird es ihr allerdings erlauben, ihre Forschung über Symmetrie, Geometrie und Topologie in neue, interdisziplinäre Richtungen zu erweitern.

Darüber hinaus sind wir dabei, gemeinsam mit der Universität Heidelberg eine neue Professur für wissenschaftliche Visualisierung zu besetzen, die Teil eines virtuellen Zentrums für wissenschaftliche Visualisierung sein wird, an dessen Organisation das KIT, die Universität Heidelberg und das HITS beteiligt sein werden.

Fünf Jahre sind inzwischen seit der Gründung des HITS und unserem Umzug in ein neues Gebäude, das in diesem Sommer offiziell „Klaus Tschira HITS Gebäude“ genannt wurde, vergangen. Das „Mathematikon“, ein neues Gebäude für Mathematik und Informatik, das mit der Unterstützung der Klaus Tschira Stiftung auf dem Campus der Universität Heidelberg gebaut wurde, öffnete Ende des Jahres seine Pforten. Es wird die neue Gruppe für wissenschaftliche Visualisierung beherbergen und – ab Mitte 2016 – auch unsere Kollegen vom EML, wodurch auf dem HITS-Gelände mehr Platz für die über 120 HITS-Wissenschaftler geschaffen wird.

Im Januar 2015 begingen wir den fünften Geburtstag des HITS mit einem Symposium als Teil der Feierlichkeiten zum 20. Jubiläum der Klaus Tschira Stiftung (siehe Kapitel 5.3).

Im November veranstalteten wir ein Symposium zu Ehren von Andreas Reuter (siehe Kapitel 5.9). Er wird im April 2016 seine Position als Geschäftsführer niederlegen, nach einer kurzen Übergangsphase mit unserer neuen Geschäftsführerin Gesa Schönberger, die seit Januar 2016 am HITS ist. Andreas Reuter wird Geschäftsführer der HITS Stiftung und Seniorprofessor an der Universität Heidelberg bleiben, um eine Arbeitsgruppe zur Entwicklung des Entwurfs einer IT-Infrastruktur zu leiten, welche den wachsenden Bedürfnissen der Forschungseinrichtungen in der Region Heidelberg-Mannheim-Karlsruhe entspricht.

Wir sind zuversichtlich, dass das HITS durch all diese Veränderungen gut für die Konsolidierungsphase aufgestellt ist, die voraussichtlich im nächsten Jahr beginnen wird.

2 Research

2.1 Astroinformatics (AIN)

2.1 Astroinformatics (AIN)

The Astroinformatics group develops new methods and tools for dealing with the complex, heterogeneous, and large datasets currently available in astronomy.

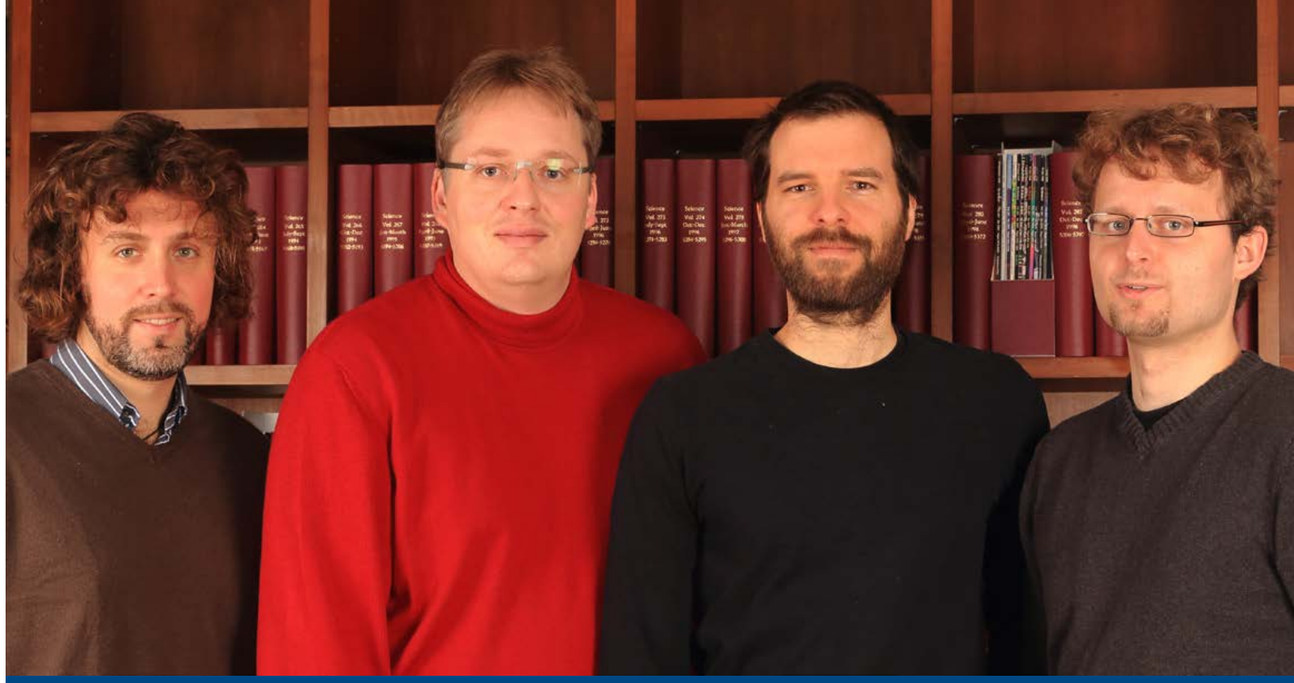
Over the last two decades, computers have revolutionized astronomy. Advances in technology have given rise to new detectors, complex instruments, and innovative telescope designs. These advances enable today's astronomers to observe more objects than ever before and with high spatial/spectral/temporal resolution. In addition, there are new untapped wavelength regimes still to be investigated. Dedicated survey telescopes map the sky and constantly collect data. The task we have set ourselves is to enable scientists to analyze this increasing amount of information in a less biased fashion.

The group is interested in the development of improved methods for time-series analysis and redshift models based on photometric measurements. These are key tools for the analysis of the data in upcoming large survey projects like SKA, Gaia, LSST, and Euclid. Another scientific concern is to develop methods and tools for the extraction and filtering of rare objects (outliers) for detailed follow-up analysis with 8-m-class telescopes. With estimated occurrences of only a few objects per million, manual inspection of the existing catalogs is out of the question. The Astroinformatics group's other interests include the morphological classification of galaxies based on imaging data as well as measuring similarity in high-dimensional data spaces.

Die Astroinformatik Gruppe entwickelt neue Methoden und Werkzeuge, um eine Analyse der heutzutage verfügbaren komplexen, heterogenen und großen 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 wiederholt durchmustert und die so gewonnen Daten werden frei zur Verfügung gestellt. Durch unsere Forschung ermöglichen wir es Wissenschaftlern, diese riesigen Datenmengen durch neue Analysemethoden explorativ und unvoreingenommener zu erschließen und somit effizienter zu nutzen.

Unsere Gruppe beschäftigt sich mit der Zeitreihenanalyse sowie der Entwicklung photometrischer Rotverschiebungsmodelle. Dies wird 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 der Astroinformatik Gruppe.



Group leader

Dr. Kai Polsterer

Staff member

Dr. Nikos Gianniotis

Scholarship holders

Dr. Sven Dennis Kügler *(HITS Scholarship)*
Antonio D'Isanto *(HITS Scholarship, since February 2015)*

Classification of heterogeneous time series in astronomy

The variation in brightness of an astronomical object over time (hereafter called light curve or time-series) is an important source for obtaining knowledge about, and constraining the properties of, the observed object. With the advent of large sky surveys such as the Large Synoptical Sky Survey (www.lsst.org) the incoming data stream will be so immense that the applied methodology has to be reliable and fast at the same time. A huge proportion of the variable objects in the sky has a stellar origin, while their variability is due to a number of physical phenomena. A major proportion of astronomical time-series observed from the ground are subject to numerous effects, such as seasons and weather. The observed data are thus irregularly sampled, which makes the use of standard methodology, such as recurrent neural networks or auto-regressive moving average (ARMA) approaches, quite difficult. Interpolation could potentially convert these irregular time-series into regular ones, but the observed noisy and complicated behavior of these widely varying astronomical sources also makes this solution a real challenge.

In the past, the classification of light curves in astronomy has been mainly performed manually. Domain experts would visually inspect light curves and assign them a label (i.e. class type) they considered plausible on the basis of their subjective experience. Apart from the potential bias involved, it is evident that manual classification is not possible for very large datasets. However, since there is currently no other way of obtaining labels for the observed objects, those subjectively assigned labels are commonly taken as ground truth. Accordingly, reliable automation of this classification task is a major goal in astronomy. Achieving this goal will enable astronomers to cope with the large volumes of data from upcoming survey telescopes.

Current efforts towards automating the classification of light curves are typically variations of the following general scheme: a set of statistical features are defined and extracted from the light curves. The extracted features define a vector which, in conjunction with the subjective labels, is used to train a classifier. Apart from subjectivity in the choice of these features, this method lacks a set of properties that are crucial in astronomy. Firstly, the proposed approach is not capable of handling heteroscedastic noise on the individual photometric data points. Secondly, the sampling may introduce a significant bias in time-dependent features, as the absence or presence of a given data point can drastically change the value of a given feature. Lastly, a given survey can only capture objects up to a certain brightness, and thus, due to variability, an object might not be observable at a given time. The consensus in astronomy is now to assign upper brightness limits to these measurements which cannot be naturally included in a feature-based representation.

In our work, we have introduced a novel way of representing time-series that aims to replace the static features (not time-dependent) that are used in the general classification scheme outlined above [Kügler, 2015a]. We represent each noisy data point by a Gaussian; the mean of the respective Gaussian is the measured value, and the standard deviation is the associated measurement uncertainty (photometric error). Every time-series is represented by a mixture of Gaussian densities that conserves all the encoded information including the commonly used static features. We suggest that this is a simple and natural choice. Unlike the representation of mixtures of Gaussians, static features can be seen as derivatives (such as moments) of this density model and therefore describe only certain properties of it. Moments describe the tails of a distribution but do not necessarily give a good overall description of the underlying distribution. As a consequence, the proposed density-based representation presents an upper boundary to the static information content made available to a given classifier. Instead of measuring distances between extracted feature vectors, as is generally the case in feature-based classifications, the presented scheme measures distances between probability density functions and accounts for the uncertainty of the individual measurements in a natural manner. Once these distances have been calculated, the formed distance matrix is employed in distance-based classification methods. The scheme of the classification algorithm is visualized in *Figure 1.*

Even though the approach presented here is conceptually quite different from feature-based classification approaches, it is interesting to see that the performance achieved by our methods is on a par with existing methods. This shows that the features provide as much information to the classifier as the density-based approach. On the other hand, the results suggest that we can replace an arbitrary feature choice by a more principled approach that accounts for the particularities of astronomical data. Beyond classification performance issues, we have also discovered that inconsistencies in the way data are recorded by the surveys leads to the extraction of biased features, which in turn lead to a spurious boost in classification performance. Since the light curves are obtained from surveys that monitor a given region of the sky in a regular scheme, we would not expect one class of objects to be observed more often than the others. However, this is exactly what has been observed in practice *(see Figure 2).*

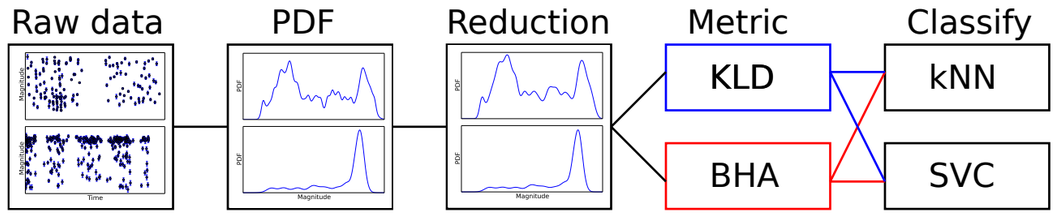


Figure 1: Scheme of the featureless classification of light curves. The raw data are converted to a probability density function between which distances are then measured. These distances are then fed into a classifier.

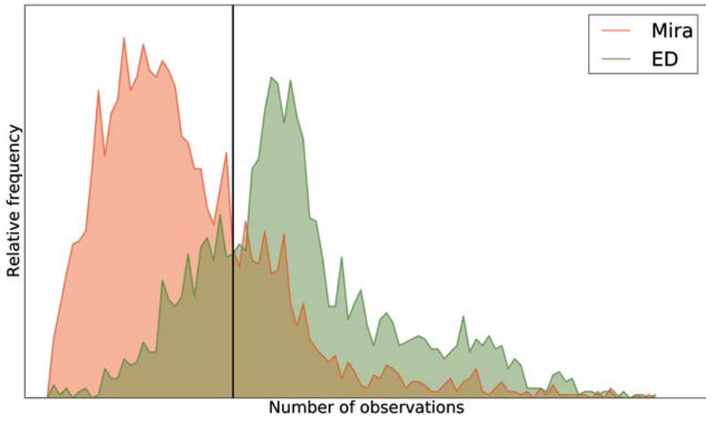


Figure 2: Histogram of number of observations for the given classes in ASAS. The distribution depends heavily on the observed variability type. While the plain number of observations improves classification, it is evident that this is a consequence of the recording strategy and not an inherent physical property.

Naturally, the question arises as to where this bias comes from and how it can boost the performance of certain features. In the survey under consideration, non-detections have not been noted as lower limits in brightness (which, unfortunately, is common in astronomy) but have been discarded entirely. Therefore, strongly varying objects (e.g., MIRA) tend to exceed the lower limit more often and thus seem to be observed less often on average than other classes (e.g., detached eclipsing binaries, ED). As a consequence, features that include a direct measure of the number of observations are strongly biased towards the respective classes and boost the classification performance artificially while neglecting the true physical reason behind the phenomenon.

In summary, a new representation of noisy, inhomogeneous, and heteroscedastic time-series has been introduced, and its performance has been successfully evaluated in classifying astronomical light curves. We have found that in terms of performance the new density-based approach performs just as well as the feature-based approach, but is superior in its ability to respond adequately to the noise of the individual measurements and to missing data points. In our future work, we would like to test this scheme in a transfer-learning setting, where learning is performed on one database of observations and predictions are made on another. This would enable the classifier to learn sequentially from different databases and could thus accommodate the recurring stage of manual classification for each newly published survey.

Exploring homogeneous astronomical time-series

The launch of the Corot and Kepler spacecrafts (https://corot.cnes.fr; http://kepler.nasa.gov) initiated a new era in the study of stellar variability. The unobstructed view of the light from astronomical sources enables continuous monitoring of stellar sources with short cadences and very low photometric error. While the primary goal of both missions was the detection of solid exoplanets, the continuous monitoring of 150,000 variable main-sequence stars allowed a very detailed study of their variability behavior. While astroseismology [Gilliland RL et al. Kepler Asteroseismology Program: Introduction and First Results, PASP (2010) 122:131-143] and exoplanet detection [Lissauer JL et al. Almost All of Kepler's Multiple-planet Candidates Are Planets, ApJ (2012) 750:112] greatly benefit from these observations, many objects still remain unlabeled. Though the labeling of (quasi-) periodic sources is quite reliable [e.g., Benkö JM et al., Flavours of variability: 29 RR Lyrae stars observed with Kepler, MNRAS (2010) 409, 1585-1593; Slawson RW et al., Kepler Eclipsing Binary Stars. II. 2165 Eclipsing Binaries in the Second Data Release, AJ (2011) 142, 160], a large proportion of the objects displaying non-periodic behavior remain unclassified. To investigate the true nature of objects that cannot be explained by known variability mechanisms, alternative approaches are required. Visualization (i.e., dimensionality reduction), clustering, and outlier detection are the most prominent tools from the field of unsupervised learning. Visualization in particular allows for intuitive inspection of the properties of the observed data by projecting them into a lower dimensional space. Data analysts can interpret visualization plots and look for structures and similarities that could not be detected in the original data space. However, it is not possible to directly apply dimensionality reduction to raw sequential data, as the individual measurements are not independent and therefore the time-series cannot be treated like vectorial data. Furthermore, the time-series may differ in length, and their temporal behaviors would not be aligned within the time domain.

In our work, visualization is performed in two steps. First we use the echo state network (www.scholarpedia.org/article/Echo_state_network) to describe time-series as sequences. The ESN is a discrete time-recurrent neural network used to forecast the behavior of time-series. The ESN, like other neural networks, is parametrized by a vector of weights **W.** Training the ESN on a time-series yields an optimized vector of weights which, in this approach, is used as a fixed-length vector representation for regularly sampled time-series. The advantage of this new representation is that it is invariant both to the length of time-series and to the presence of time shifts. We perform visualization on this new representation, not on the original data items. In a second step, we would like to project this high-dimensional representation onto a two-dimensional hyperplane, so that we can visually inspect the data items. In the literature, many algorithms for dimensionality reduction have been proposed, such as principal component analysis or non-linear PCA (also called autoencoder. These reduction algorithms would encode the high-dimensional data in such a way that the decoded part would be as close to the original representation as possible, i.e. is commonly minimized, where **W** is the original, high-dimensional representation. and Θ are free parameters of the reduction function f. However, what is more important for us to study is not how similar the weights vectors of the ESN are, but how well the reconstructed weights can still predict on a given time-series. Thus, the objective function to be optimized changes to which measures the ability of a reconstructed weight vector to predict on a given time-series. A scheme of this visualization approach is shown in *Figure 3*.

In the following, we would like to emphasize that this new objective function is not just mathematically principled but also has direct implications for the meaning of the visualization. To show this, 6,206 Kepler light curves have been visualized using the proposed approach. *Figure 4 (next page)* shows the visualization with over-plotted physical properties of the observed objects. One can clearly see the strong correlation between visualization, temperature, and surface gravity. It seems that the physical properties have a direct implication for the dynamic behavior of the stars. The correlation between different dynamic regimes of giant stars (low-surface gravity) and main-sequence (high-surface-gravity) stars has already been the subject of study [Bastien FA. An observational correlation between stellar brightness variations and surface gravity, Nature (2013) 500:427-430]. It is highly gratifying that this correlation can be recovered without explicitly biasing the algorithm towards finding it. This shows the great potential for knowledge discovery in an unsupervised setting such as visualization.

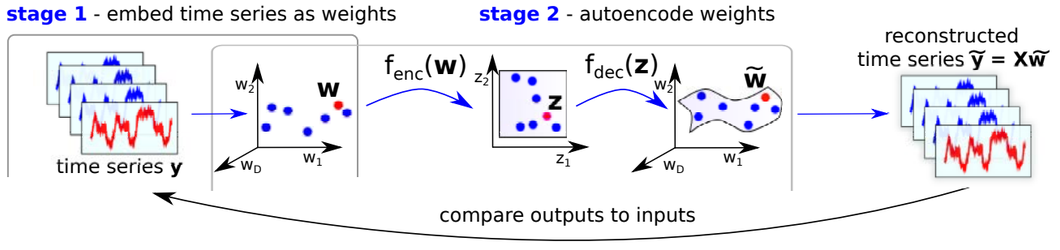


Fig 3: Sketch of proposed method. In the first stage, time-series are cast to readout weights in the weight space. In the second stage, the autoencoder projects readout weights onto latent coordinates residing in a two-dimensional space and reconstructs approximate versions of them. The reconstructed readout weights are then mapped to the time-series space wherethe reconstruction error is measured.

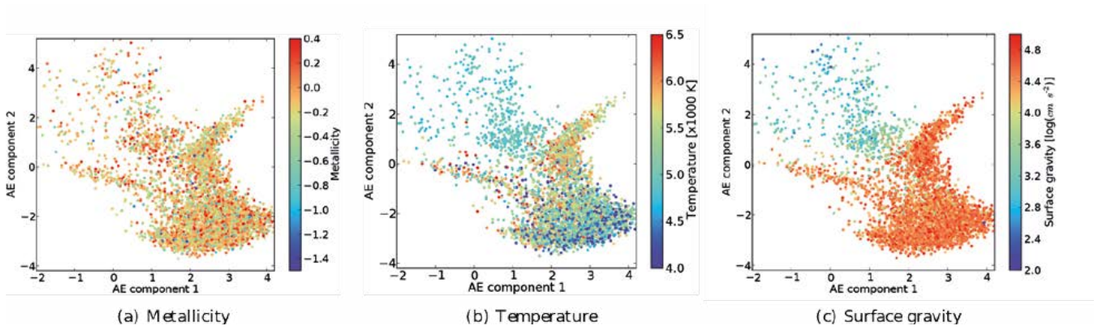


Figure 4.: Correlations between the projection derived by our method and the physical properties of the stars. From left to right, metallicity, temperature, and surface gravity are over-plotted as color gradients. Apparently, the major cause for different variability behaviors is surface gravity, which indicates that giants and main-sequence stars are governed by intrinsically different dynamics.

In conclusion, we have also investigated a new approach to visualizing regularly sampled time-series. As opposed to the visualization algorithms employed in astronomy, the approach presented here does not require any pre-alignment of data and respects the sequential nature of the time-series [Kügler, 2016]. Also, it delivers a shift-invariant vector representation for sequential data of variable length. Compared to the common use of visualization algorithms in astronomy, we do not directly apply dimensionality reduction to the data but to model parameters instead. In future work, we would like to be able to include the uncertainty of the individual data points. Together with a generic regression algorithm like Gaussian processes, this would allow for an extension of this framework to irregular sampled light curves.

Immersive and affordable three-dimensional data representation

The virtual observatory (VO) has become a success story in providing uniform access to a huge amount of data sets. Hidden in those data sets are correlations, distributions, and relations that have to be unveiled. Visual inspection has always been a key tool in understanding these complex structures. It is very common practice to project high-dimensional information down to a two-dimensional plane to obtain a diagnostic plot. Besides expensive stereoscopic visualization environments like hyper-cubes or hyper-walls, stereoscopic displays provide a way of inspecting three-dimensional data spaces.

In the recent past, the rapid development of affordable immersive technologies for entertainment and gaming purposes has been observable. Besides stereoscopic images, three-dimensional movies and computer games have had a revival due to the latest developments in display technologies. Most of the screens sold today support the presentation of stereoscopic content, either with shutter goggles or glasses equipped with polarization filters. These glasses are necessary to present two independent images for both eyes so as to create the impression of depth. There are display solutions available that do not demand the wearing of glasses, but those solutions are more expensive and fail when a larger number of users are present. In 1833, the first stereoscopic viewer was presented by Sir Charles Wheatstone. This was before the invention of photography, so it made use of two hand-made paintings presented side-by-side. By combining this very old idea of stereoscopic images with a smart-phone capable of rendering and presenting stereoscopic images, a virtual data space can be created. With its built-in orientation sensors and accelerometers, the smart-phone is able to measure the direction of view and thus translate motions of your head in the real world into motions of the camera in the virtual environment presented. As most smart-phones are equipped with a camera on the back, simple augmented reality becomes possible by mixing virtual with real-world content.

We developed the Virtual Observatory Virtual Reality (VOVR) software to enable us to use smart-phones with simple stereoscopic viewers and game controllers to explore data from the VO *(Figure 5)*. VOVR is a software tool developed within the scope of the international virtual observatory alliance and thus features the interoperability functions defined there. The basic purpose of VOVR is to add 3D exploring capabilities to already existing VO applications and thus to enable scientists to use immersive data exploration as an integral part of their usual workflow *(Figure 6, next page)*. Our software consists of two applications communicating via a simple socket-based network connection. The central server application is responsible for importing, annotating, and managing the data in order to define a set of plots. Multiple clients can connect to the central server to retrieve the data for plotting and to exchange the positions of the inspecting scientists. All rendering is done on the client side only, so a compact description of the data space is merely exchanged. The communication between the server and the client is based on a plain byte protocol encoding the plot information. The protocol has been kept very simple to facilitate the creation of other client or server implementations and to minimize data transfer. A compression layer will be added in future to improve speed.

Figure. 5.: Using Virtual Observatory Virtual Reality (VOVR) on a smart-phone together with the Zeiss VR-One stereoscopic viewer and an X-Box controller. Navigation is very intuitive and easy to learn, enabling new users to navigate through space after just a few seconds.

The server application is written in Java only and runs on multiple platforms. It is responsible for managing the plots and clients. The graphical user-interface is only available through the server, as user interaction on the client side is limited to data inspection. All connected clients are displayed, and as soon as they report back their position in the virtual space, this information is sent to all other clients. This enables the clients to render avatars for the other clients and thus create an interactive data space. In addition, it is possible to automatically synchronize the position to another client, thus enabling a scientist to guide other observers through the data. As soon as a data set is to be rendered, the data is pre-processed as specified in the plot annotation and transferred to the selected clients. As the pre-processing is done off-line, it just needs to

Figure 7: Altitude information on southern Australia including Tasmania, displayed via the VOVR client.

be done once before transmitting the data to the clients. Currently a TCP socket communication is being used, so broadcasting the data to the clients is not supported. In a future version, data streaming will be considered to allow for the exploration of extremely large data sets. Until these additional features are available, clients are treated sequentially.

The client applications provided for VOVR are based on the Unit3D framework intended for use in game development. This framework makes for efficient rendering on different target hardware platforms. Therefore the client application was developed once and compiled for android mobile devices as well as for non-stereoscopic clients operating on windows, linux, and mac-os. A scene needs to be defined in order to populate the virtual data space. After this is done, the gaming engine is used to navigate through the scene graph, internally making use of spatial indexing structures to accelerate the rendering by clipping unseen objects. Commands transfer the necessary information to the client, which creates the required graphical elements. By choosing a specific plot type, simple point meshes, particle systems, or multiple sprites are added to the graph. Real 3D bodies with light and shadow calculation are currently not supported because this would greatly reduce rendering speed. The current client is capable of rendering 2M *(Figure 7)* simple points or 0,2M sprites smoothly on a Samsung Galaxy S5, which should be enough for most visualization uses. ■

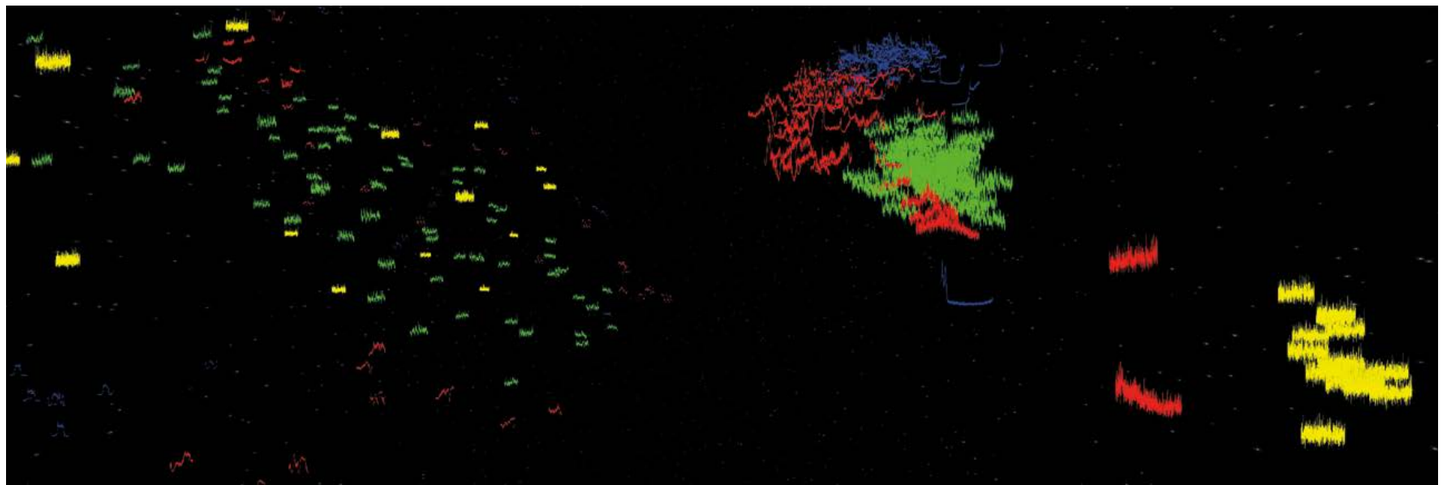


Figure 6: Non-stereoscopic view of the Linux VOVR stand-alone client. The data presented is time-series data from black hole binaries projected with t-SNE (left) and our ESN autoencoder (right) to a three-dimensional space. The different colors represent different classes. Note that each time-series is displayed in a billboard mode, i.e. always facing the user. By navigating around, differences in temporal behavior can be observed as well as the better discriminative power of our approach.

2 Research

2.2 Computational Biology (CBI)

2.2
Computational Biology (CBI)

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 members. Philipp Kämpfer and Philipp Bongartz, two PhD scholarship holders, joined us 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, starteten Juni bzw. August, und Martin Pippel nahm seine Tätigkeit als PostDoc im Juli 2013 auf. Des weiteren hat 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 Informatik, 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 Fragmente, 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.

Group leader

Dr. Siegfried Schloissnig

Staff member

Martin Pippel

Scholarship holders

Philipp Bongartz *(HITS Scholarship)*
Philipp Kämpfer *(HITS Scholarship)*

Introduction

The main focus of the group is on de-novo genome assembly, the reconstruction of a genome's sequence from reads produced by a DNA sequencer. The complexity of the reconstruction is dependent on the genome to be reconstructed (size, ploidy, and repetitiveness) and the characteristics of the sequencer used. Three properties of the sequencer are usually of interest: speed, error rate, and (average) read length. Of these only speed impacts the cost associated with the sequencing. Error rate and read length are functions of the underlying approach employed by the sequencer. The sequencers of the first generation were able to produce reads with lengths in the 100s of base pairs and a relatively low error rate <2%. Unfortunately, this was a very expensive, slow, and only semi-automated process (adding labor costs as a result) and was replaced by NGS (Next Generation Sequencing – the 2nd generation) in the late 1990s, mostly for economic reasons. Cost per base pair dropped significantly in the 2nd generation, while the error rate was maintained at the levels of the previous method, but the average read length dropped by one order of magnitude for the first market-ready models. The disadvantages of this technology became apparent when attempting to assemble anything beyond bacterial genomes. Even though the technology has improved over the years, with read lengths now reaching 200 – 300bp and with error-rates below one percent, the results achieved by applying this technology to genomes of higher organisms were less than satisfactory. The third generation, which has reached maturity in the last couple of years, is usually referred to as single-molecule sequencing. As the name suggests, this basically involves observing the passing or processing of a single strand of DNA through a protein (-complex) and inferring the sequence from the signals generated during this process (e.g. fluorescence or electrical current). For example, the SMRT technology from Pacific Biosystems (PacBio) observes the light emitted when fluorescently labeled nucleotides are incorporated into a DNA strand by a polymerase.

These approaches have pushed read lengths into the tens of kilo-base pairs, but at the price of a high error rate (>10%) due to the stochasticity of the process.

Major changes in technology also drive changes in the software used to deal with the data generated. Initially, the overlap-consensus (OLC) method was employed, which (in the naïve approach) compares each read to every other read, thereby calculating overlaps between them and resolves unique paths in the graph structure induced by all overlaps and producing longer stretches of sequence.

Longer reads improve the chances of correctly reconstructing repetitive regions. This is due to the fact that they potentially span a repetitive element, thus causing fewer junctions in the overlap graph. Obviously, 100s of base pairs were insufficient for most real-world repeat structures, so a hierarchical approach was employed. First the genome was decomposed into overlapping clones (fosmids, cosmids, or BACs) ranging from 20KB up to 300KB that were sequenced and assembled independently, thereby significantly reducing the problem. The assembled clones were then used to assemble the complete genome.

The OLC approach broke down computationally with the advent of NGS, which was able to produce millions of short reads (around 50bp initially) very quickly. Abandoning the overlap-consensus model, assembly moved on to de Bruijn graphs (dBG). dBGs are directed cyclic graphs representing overlaps between k-mers. Each vertex corresponds to a k-mer (a stretch of k-base pairs), and an edge indicates an overlap of length k-1. This rigid structure very quickly allows the creation of a k-mer overlap graph by decomposing all reads into k-mers and adding edges between k-mers shifted by one base (e.g. k = 3 and ACTG would result in vertices ACT and CTG with an edge from the former to the latter). A dBG can be built efficiently, and it implicitly "computes" overlaps between k-mers. Of course, the decomposition of reads into k-mers results in a loss of information, and errors in the reads induce up to k wrong k-mers. Touring the graph (i.e. reading out the unique paths) is coupled with an abundance of heuristics to resolve junctions and deal with graph structures induced by read errors, thus making the different dBG based assemblers produce vastly different result sets from the same input data.

Long reads and high error rates are two properties of single-molecule sequencing that make the reads incompatible (for most applications) with a dBG-based approach. For this reason, we now find the overlap-consensus model making a comeback.

Group leader Dr. Siegfried Schloissnig (left) and the mentor of the Computational Biology group, Prof. Eugene "Gene" Myers.

The research projects outlined below summarize our current work on noisy long-read assembly and hybrid dBG/OLC assembly.

Projects

A new scalable infrastructure for long-read assembly

The long reads from single-molecule sequencers, which can be upwards of 10k in length, are impressive, but with an error rate of 15% the resulting computational complexity of such reads is exuberant. 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 called the Dazzler that has been used to assemble repetitive genomes of up to 32Gb directly from a shotgun, long-read data set currently producible only with the PacBio RS II sequencer.

It is based on the string graph / overlap-consensus paradigm, its most important attributes being: 1) It scales reasonably to gigabase genomes, being one order of magnitude faster than current assemblers for this kind of data due to adaptive repeat masking during the overlap phase. 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. 3) A read correction phase that reduces the error rate to <1% on average.

In collaboration with Gene Myers, we have developed a local alignment finder for noisy reads called Daligner. Daligner works on a read database created from either FASTA or H5 (the PacBio native file format). The database is then partitioned into blocks and all "block against block" comparisons are

computed, producing a set of LAS (local alignments) files in the process. Block partitioning allows for massive parallelization of the overlapping phase at block level. Block comparisons involve k-mer indexing and alignment computation if two sequences share enough k-mers in a diagonal band or the edit distance matrix. Alignments start from the conserved k-mers and terminate when their quality drops below a user-defined threshold.

One of the impediments to scaling the overlapper to large genomes lies in their repetitiveness. If every fragment of the genome were unique, then only true overlaps would be calculated. But even for the simplest bacterial genomes this is only partly the case. Omnipresent repetitiveness (degrees vary) induces partial matches (i.e. local alignments) of the reads that are largely responsible for the overlapper's runtime and storage requirements. In order to alleviate this problem, we have coupled the overlapper with an adaptive repeat finder, which, as soon as enough local alignments have been calculated, starts tagging regions of a read as repeats. This happens parallel to overlapping and results in the continuous exclusion of parts of a read from the overlapping process. This has resulted in savings of one order of magnitude in disk space and runtime. The repeat finder is not specific to a single overlap job. It gathers local alignment statistics from, and distributes repeat interval annotations for, all reads to all overlapper jobs.

Scrubbing detects and corrects read artifacts that would lead to partial alignments, subsequently classified as local alignments (e.g. repeat-induced) and discarded during the assembly process. Artifacts can either be a sudden spike in insertions, lost bases, missed adapters, or polymerase strand jumps. Quality information about a read is derived from the alignments produced by the overlapper. We use this information to a) detect and trim low quality regions, b) repair regions in a read containing large numbers of insertions or add missing sequencing lost through polymerase jumps, and c) split reads at missed adapters and polymerase jumps.

The pipeline was developed with the overlap-correct-overlap paradigm in mind. This means that we overlap the initial raw reads, scrub and correct them, and overlap the corrected reads again. But in the course of this we realized that the correction can actually be skipped and the assembly made directly from the raw reads after scrubbing. This requires a more thorough scrubbing phase that is good enough to guarantee not only a <1% error rate after correction, but also a proper assembly. The switch to uncorrected assembly was motivated by the savings in compute time and the fact that very long reads spanning a repeat can often not be reliably corrected. For the correction to work, there has to be a guarantee that only true overlaps are used, which cannot be achieved for very long reads where the number reliably anchored in unique sequence at both ends drops towards the center of the read, resulting in coverage in the center below what would be needed for correction.

2015 marked the completion of prototypes for all modules in our assembly pipeline.

Uncorrected assembly suffers from the problem of missing overlaps. This is mostly due to low-quality regions in the reads causing either the overlapper to miss overlaps completely or early termination of the alignment when a bad quality stretch is reached, which in its turn makes for problems with transitive reduction of the overlap graph (elimination of as many edges as possible, while still maintaining the same reachability). We investigated transitive inference at this point, but the remaining repeat-induced overlaps proved to be detrimental to this approach. We therefore decided to forgo transitive reduction and directly tour the overlap graph. Reads that exhibit low alignment quality are discarded if better alternatives are available.

By contrast, when assembling from corrected reads, we perform another round of overlapping preceding the correction. This can be done quickly, given the high identity of the reads. It also makes for transitive completeness, allowing use of transitive reduction of the overlap graph and easy touring of the reduced graph.

2015 marked the completion of prototypes for all modules in our assembly pipeline. With a full front-to-end assembler available, we then shifted our focus towards improving algorithmic aspects and the scalability of the separate modules.

Assembly of the Drosophila histone complex

Histones are DNA-packaging proteins that allow for significant DNA-length reduction and play an important role in gene regulation. During cell division not only the chromosomes, but also the number of histone proteins has to be doubled to package the newly created DNA. This is assumed to be the reason for the dozens of copies of the histone genes found within almost every eukaryotic genome. Often these copies are spread out throughout the genome, but sometimes they are clustered in one particular region.

In the Drosophila genome, five histone genes are encoded in one 5kbp sequence. More than a hundred copies of this histone sequence are clustered on chromosome IIR. With previous short-read technology and even with the longer Sanger reads, assembling this cluster has been unthinkable. But with the advent of long-read sequencing machines, the assembly of this highly repetitive region may finally be possible.

Apart from the insights into the evolutionary genesis of a gene cluster that could be derived from a complete assembly, this project is an ideal playground when it comes to trying out ideas for resolving highly repetitive regions.

The longest PacBio reads span five to six copies of the histone sequence. Unfortunately, these copies are highly conserved, as copies of essential genes often tend to be. But even highly conserved genes allow for alternative codons (mutations not affecting the peptide sequence) and the spacers between the five coding regions are able to accumulate a certain number of mutations as well. In order to detect these discriminating variations between different copies of the histone sequence, we calculated a multi-alignment out of all histone sequences that occurred in the PacBio Drosophila Melanogaster dataset.

Based on further analyses of the (multi-) alignments, detection of unique anchor points in the histone complex, clustering, and deep-learning approaches, we assembled the complete complex in one piece.

Sequencing and assembly of the axolotl

The axolotl (Ambystoma mexicanum) belongs to the order of Caudata (tailed amphibians) and is known as the Mexican salamander or Mexican walking fish. Its natural habitat, lakes around Mexico City, is slowly disappearing due to urban expansion, water pollution, and continued draining of lakes and wetlands. Accordingly, the IUCN classified the axolotl in 2006 as "critically endangered" and included it in the "Red List" of endangered species.

An axolotl can reach a length of up to 30cm (12 inches), lives up to 10 years, and comes in various colors, the latter feature is probably the explanation for its popularity as an exotic pet. Due to its ability to regenerate, it is a model organism in regeneration and evolutionary research.

Its high regenerative capabilities enable the adult axolotl to regenerate entire body structures, such as limbs, tail, and jaw.

Another characteristic of axolotl is neotony, i.e. they reach sexual maturity without undergoing metamorphosis, retaining juvenile features during maturation. Besides its regenerative abilities, the size of the genome, estimated at 32gbp base distributed among 14 haploid chromosomes, makes this species both an interesting and challenging task for de-novo genome assembly.

Initial sequencing was based on BAC libraries aimed to construct an expressed sequence tag (EST) library. These approaches indicated that on average the genes of the axolotl are 5x larger than human genes due to increased intron lengths and that the overall genic component of the salamander genome is approximately 2.8gbp.

The genome sequencing was performed by our collaborators at the Max Planck Institute for Molecular Cell Biology and Genetics and the Systems Biology Center in Dresden. The creation of the primary data set required 1.5 years of sequencing and consists of a 28-fold random oversampling of this genome.
The PacBio sequencing technology is under constant development and new sequencing chemistries were adapted as soon as they reached maturity, resulting in an upward trend in the average and maximum read length.

Raw reads were extracted from the PacBio bax.h5 files, which the sequencer produces natively. If the same insert gave multiple reads, which happens when the polymerase reaches the end of the insert and continues on the other strand, we used the read that the sequencer's metadata indicated to be the most stable readout.

Figure 8: The Ambystoma mexicanum, commonly known as the axolotl, possesses extraordinary regenerative abilities and is capable of reconstituting limbs, retina, liver, and even minor regions of the brain. (Picture © kazakovmaksim / Fotolia)

De-novo assembly of a genome of this size could not be performed with the assembly pipeline we had originally implemented. An estimate of the amount of storage necessary for the results of the initial overlapping phase was 1.8PB. To allow for assemblies for complex genomes on cluster environments of a modest size, it was necessary to develop adaptive repeat masking during the overlap phase, as described above.

We created a Dazzler read database and partitioned it into blocks of 400mbp, resulting in a total of 2,282 blocks. In order to enable the adaptive repeat masker to mask highly repetitive elements as quickly as possible, we first calculated the main diagonal of the block vs. block matrix. We then continued block comparison by calculating the remaining diagonals. After the main diagonal, repeat masking already excluded 12% of the bases from further comparisons. Overlapping took 362,180 CPU hours and resulted in 37 billion local alignments, which occupied 2.6TB on disk.

We then filtered the local alignments further, using the final repeat annotation calculated during the overlapping. This is necessary because in some instances the local alignments resulting from the main diagonal block comparisons are unfiltered and, depending on the abundance of a repeat element, it can take a couple of diagonals to detect this. This step decreased the number of local alignments to 6.7 billion (480GB on disk). Overall, we mask 60% of the genome as repetitive, with only repetitive regions >1kbp considered.

Scrubbing of missed repeat-induced local alignments and removal of read artifacts took another 4,000 CPU hours (predominantly I/O). Based on the scrubbed local alignments, we corrected the reads to > 99% identity. The results of this correction were validated via the Tanaka Lab's transcriptome assembly.

After correction of the reads, the resulting coverage remained a respectable 18.1X, which were overlapped, local alignments scrubbed, and then assembled. The assembly graph of this first attempt contained 28 gbp, longest contig was 3mbp and the N50 250 kbp. We are currently investigating whether the repeat structure of the genome is prohibitive to long continuous assembly and whether the jump of one order of magnitude in genome size results in problems with our assembly pipeline.

Sequencing and assembly of Schmidtea mediterranea

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 rule rather than the exception 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.

Schmidtea mediterranea (S.med), is a free-living freshwater flatworm and a member of the invertebrate Plathyhelminthes phylum. The full genome is estimated at 800mbp to 1gbp distributed among four chromosome pairs, and the nucleotide distribution is 70% AT-rich (almost resulting, from a computational point of view, in a three-letter alphabet). S.med has the remarkable capacity to regenerate complete animals from small body parts within a week after amputation. Furthermore, due to a large number of pluripotent stem cells spread throughout its body and being very cheap to keep in a laboratory setting, S.med. has become a model organism in many research areas such as stem cell biology, germ line formation, and regeneration.

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 S.med., which has defied previous assembly attempts for many years now. The assembly is based on reads produced with the current single-molecule real-time (SMRT) sequencing technology from PacBio. Our primary data set consists of a 52-fold random oversampling of this genome.

TS.med. assembly is based on our most recent assembly workflow, which included optimizations for uncorrected assembly. Overlapping required 3,372 CPU hours with adaptive repeat masking and resulted in 61% of the read-mass being annotated as repetitive. Scrubbing and assembly were performed as previously explained and resulted in an initial assembly containing 726mbp, with an N50 of 0.7mbp.

Improving the C. briggsae assembly

The nematode Caenorhabditis briggsae is closely related to the well-studied model organism C.elegans and morphologically almost identical to it. Like C.elegans, the C.briggsae genome (estimated at 108mbp) is divided into 6 chromosomes, contains around 20% repeats, and has a GC content of 37%. A comparative analysis of the genome of these two species can be assumed to provide further insights into the evolution of the nematodes. Using a sequenced fosmids library, Illumina whole-genome sequencing, and PacBio long-reads, we are working on improving the current CB4 assembly of C.briggsae, which still contains many gaps and contigs not integrated into any chromosome. The ultimate goal is a gap-free assembly of all 6 chromosomes by combining the three available data sets and the existing CB4 assembly. Further comparative genomic analyses of the C.elegans and C.briggsae are in the offing, and a manuscript is in preparation.

Other long-term goals include the assembly and a comparative analysis of additional members of the Caenorhabditis genus, more specifically Species 9. PacBio reads for Species 9 have recently been produced, and an initial assembly is already available.

A hybrid-graph approach for short-read assembly

Since the advent of DNA sequencing methods in the late seventies, a variety of new technologies have been developed allowing for cheaper and faster DNA sequencing. Though many consider the assembly of DNA fragments to be a computational problem long since resolved, unordered and fragmented genome assemblies with false joins are widespread, significantly hampering downstream analyses. Next-generation sequencing (NGS) methods have proven to be low-cost and high-throughput through the parallelization of the sequencing process, albeit at the cost of short read-lengths. Short read-lengths and additional shortening by breaking them down into k-mers so as to build an efficient de Bruijn graph are the primary reasons for this, as despite high sequencing coverage, most assembly tools have difficulty producing accurate assemblies with long-range contiguity.

Currently, the reconstruction of genomic information using short-read data utilizes two distinct graph-based approaches, namely the overlap-layout-consensus concept (OVL) and de Bruijn graphs (dBG). The overlap-layout-consensus (OVL) concept is considered superior to the de Bruijn graph (dBG) in that the unit of assembly is a read as opposed to a small k-mer, making the graph and its path structure simpler and easier to disambiguate and resulting in higher contig lengths. However, popular NGS assemblers rely solely on the de Bruijn graph due to its superior runtime efficiency compared to the quadratic nature of the OVL approach.

We have developed a new hybrid graph approach that includes a variety of novel graph algorithms for fast and effective error correction. These eliminate more than 99% of all sequencing errors in linear time while taking advantage of both approaches. We have combined the fast, linear-time construction of a dBG with the higher contig resolution of the OVL approach. This is accomplished by touring the dBG and collecting read information, while simultaneously constructing an overlap graph directly from an expanded dBG containing information on the original sequencing reads prior to their decomposition into k-mers. By then deriving a string graph from the transitively reduced overlap graph, it is possible to reconstruct large unique contigs of the genome. Finally, the paired-end read information is incorporated into the string graph to facilitate scaffolding-like contig ordering and resolution of repetitive sequences. The hybrid graph approach can be used with various insert sizes and sequencing technologies.

Initial tests on several bacterial short-read data sets have shown that the hybrid approach is comparable to the time and space complexity of classical dBG assemblers as well as the high contig resolution of established OVL assemblers.

2 Research

2.3 Computational Statistics (CST)

2.3
Computational Statistics (CST)

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). The group 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 the theory and practice of 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 currently witnessing a transdisciplinary paradigm shift 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 for 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, including the German Weather Service, the German Federal Institute of Hydrology, and the European Centre for Medium-Range Weather Forecasts.

In 2015, the CST group grew to full size, with staff members Sebastian Lerch and Roman Schefzik joining the team.

Die Computational Statistics Gruppe 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 in der Theorie und Praxis von Prognosen. 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 Punktvorhersagen hin zu probabilistischen Vorhersagen. Ziel der CST Gruppe ist es, diese Entwicklungen nachhaltig zu unterstützen, indem sie theoretische Grundlagen für wissenschaftlich fundierte Vorhersagen entwickelt, 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 Kontakte und Kooperationen mit nationalen und internationalen hydrologischen und meteorologischen Organisationen, wie etwa dem Deutschen Wetterdienst, der Bundesanstalt für Gewässerkunde und dem Europäischen Zentrum für mittelfristige Wettervorhersagen.

Im vergangenen Jahr 2015 haben Sebastian Lerch und Roman Schefzik unsere Gruppe verstärkt.

Group leader

Prof. Dr. Tilmann Gneiting

Staff members

Dr. Werner Ehm
Kira Feldmann
Stephan Hemri
Alexander Jordan
Dr. Fabian Krüger
Sebastian Lerch *(from Februar 2015)*
Dr. Evgeni Ovcharov *(until December 2015)*
Dr. Roman Schefzik *(from March 2015)*

Visiting scientist

Prof. Dr. Sándor Baran *(July 2015)*

Students

Stefan Lambert *(November – December 2015)*
Patrick Schmidt

General news

In the year 2015, the CST group grew to full size, as PhD student Sebastian Lerch and postdoc Roman Schefzik joined our team. Our research continues to focus on the theory and practice of forecasting, with support from the Advanced Grant ScienceFore provided by the European Research Council. In this field, we report on two research highlights and an outreach effort at the Explore Science fair. In the area of spatial statistics, we detail progress on flexible probabilistic models for the generation of spherical particles.

An important and very enjoyable aspect of our work is the intense disciplinary and interdisciplinary exchange that we have profited from on many occasions. In 2015, we were happy to welcome guests from all over the world. Sándor Baran of Debrecen University in Hungary visited HITS in July, and Donald and Mercedes Richards of Pennsylvania State University in the United States, both doing sabbaticals at Heidelberg University, were frequent guests of our group, see *Figure 9*. In January, Martin Schlather of Mannheim University gave a computationally oriented short course on Geostatistics with RandomFields at HITS, which attracted PhD students from our project partners at Bonn University, Heidelberg University, and Mannheim University. In July, we hosted a small-scale symposium on the topic of Statistical Post-Processing of Ensemble Forecasts with international participation by statisticians, hydrologists, and meteorologists.

Figure 9: Astronomer Mercedes Richards of Pennsylvania State University in a discussion with CST group members Patrick Schmidt and Sebastian Lerch.

Assessing the accuracy of point forecasts

Comparing the accuracy of two or more forecasting methods is a matter of perennial interest. Can statistical post-processing improve a weather center's physics-based forecasts about tomorrow's temperature? How do economic experts compare to ordinary consumers when it comes to predicting the coming year's inflation rate?

When conducting such forecast comparisons, researchers need to be careful to select appropriate measures of forecast accuracy. For example, when interest is in predicting an expected value, the researcher can choose from among a class of functions identified and described in the ground-breaking research done by Leonard Savage in the 1970s. A performance measure from the Savage class is said to be a consistent scoring function for the expected value, in the sense that truth-telling is every forecaster's optimal strategy. To home in on the idea, suppose Alice asks Bob to supply critical forecasts for her company. Bob's reward depends both on his forecasts and on the respective observations. If the reward is payable in terms of a consistent scoring function, Bob's best strategy is to provide the most honest and careful forecasts he can generate.

However, the choice of a particular measure from the Savage class is often hard to make and justify. This poses a problem, since different choices can lead to different forecast rankings. For example, economic experts may outperform consumers' inflation forecasts in terms of one performance measure, whereas consumers may win in terms of another one. In recent work, Werner Ehm, Tilmann Gneiting, Alexander Jordan, and Fabian Krüger of the CST group have tackled this problem from a novel perspective. Specifically, we have shown that every member of the Savage class admits a representation as a weighted average of very simple and easily interpretable functions, which we call the elementary scoring functions. This finding is not only appealing in mathematical terms, it is also practically useful. It suggests a simple graphical tool, which we call a Murphy diagram, telling us whether a forecast comparison is stable across all performance measures from the Savage class. In a nutshell, a Murphy diagram monitors forecasting performance under the elementary scoring functions. This is easy to implement and avoids the need to select and compute member functions from the Savage class. The term Murphy diagram is in honor of the late Allan Murphy, a meteorologist who made fundamental contributions to the science of forecast evaluation.

To give an example, consider the current quarter probability forecasts of recession in the United States shown in *Figure 10*, where a recession is defined as negative real domestic product growth. The forecasts come either from the Survey of Professional Forecasters (SPF), which depends on experts' knowledge, or from a very simple statistical Probit model. During actual recessions, the SPF panelists tend to assign higher forecast probabilities than the statistical model. In the Murphy diagram in *Figure 11 (next page)*, the graph for the SPF forecast is consistently below the graph for the Probit model, thereby confirming the superiority of the SPF panelists' predictions over the statistical model for current quarter forecasts.

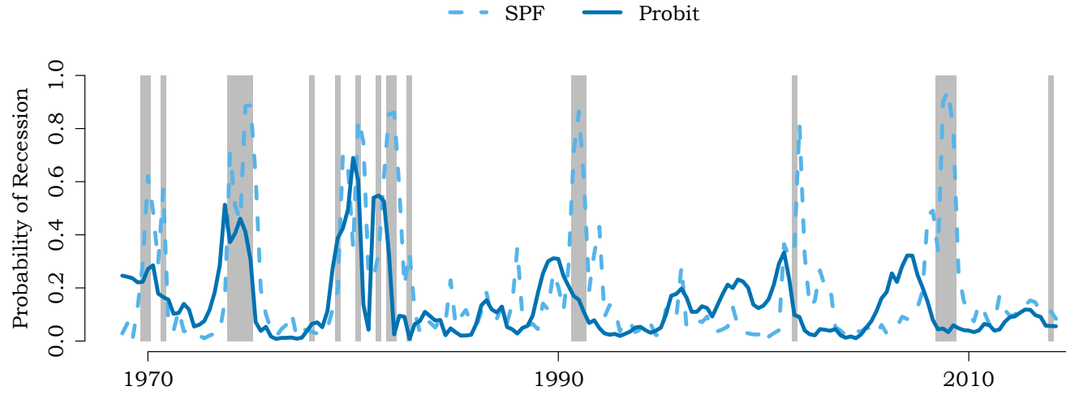


Figure 10: Current quarter probability forecasts of a recession in the United States from the Survey of Professional Forecasters (SPF) and a simple statistical model (Probit). Actual recessionary periods are shaded.

Technical details can be found in a forthcoming paper [Ehm W, Gneiting T, Jordan A, Krüger F. Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings (with discussion and reply). Journal of the Royal Statistical Society Series B (2016)], where analogous results are derived not only for predictions of an expected value, but also in the much more general cases of expectiles and quantiles. We are very pleased to report that the Research Section of the Royal Statistical Society selected our paper as a paper for reading and discussion, a significant token of recognition for a paper's importance and the breadth of its applicability. Group leader Tilmann Gneiting presented the paper at the Ordinary Meeting of the Royal Statistical Society on 9 December 2015 in London. There were invited comments by Chris Ferro of the University of Exeter and Philip Dawid of the University of Cambrige and further comments from the scientific community. The paper will be published in the leading statistical methodology journal, the Journal of the Royal Statistical Society Series B, along with written comments from well over 40 colleagues worldwide.

Basel Framework for banking regulation

The Basel framework prescribes how regulatory bodies request bank's financial risk assessments for monitoring purposes. A natural question then is whether the framework can be designed so that a bank's best strategy is to provide careful and honest assessments of its financial risks to regulators. Currently, there is much debate about a potential revision of the protocol. In this revised form, banks would be required to report a certain feature of their in-house generated predictive distributions, called the expected shortfall or conditional value-at-risk, in lieu of the traditional value-at-risk or quantile measures. However, in a

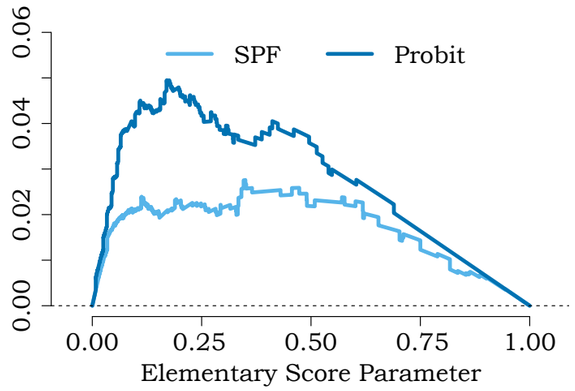


Figure 11: Murphy diagram for the probability forecasts (SPF and Probit) illustrated in Figure 10. The graphs show the mean elementary score for the two forecasts as a function of the elementary score parameter: the smaller the mean score, the better.

technical sense expected shortfall is not elicitable, given that it does not allow for a strictly consistent scoring function. This drawback has provoked extensive debate in the finance sector.

Against this background, recent work by Tobias Fissler and Johanna Ziegel at the University of Berne in Switzerland in collaboration with CST group leader Tilmann Gneiting advocates a change in the practice of banking regulation. In current practice, banks provide risk estimates on a daily basis, and regulators use a so-called backtest to check whether the estimates are correct in some absolute sense. However, procedures of this type are problematic as a basis for decision-making, particularly in view of an anticipated revised standardized or benchmark approach, which banks are supposed to adopt if their internal market risk model fails the backtest. In this situation, the standardized or benchmark approach may fail the backtest, too, and in fact it may be inferior to the internal model. Conversely, the internal model may pass the backtest although a more informative benchmark model is available.

A recent paper [Fissler et al., 2016] that became available in 2015 and was published in the January 2016 issue of Risk, the premier news and analysis magazine for the financial industry, introduces comparative backtests as an alternative.

In the proposed comparative backtest, a bank's internal risk model is held accountable relative to the benchmark approach in terms of both expected shortfall and value-at-risk, rather than being validated in an absolute sense.

This leads to a natural traffic-light approach, as illustrated in *Figure 12.* In a nutshell, the scoring for the benchmark approach takes place via a novel scoring function devised by Fissler and Ziegel that is jointly consistent for expected shortfall and value-at-risk, yielding a certain mean score, S. Using statistical methods, one can then find an interval around S, from a lower limit L to an upper limit U. If the internal model's mean score T is lower than L, the internal risk estimates compare favorably to the benchmark approach, thus passing the backtest. If the internal model's mean score T exceeds U, it has failed the backtest, and it is in the joint interests of all stakeholders for the bank to adopt the benchmark approach. In the yellow zone in between, no decisive ranking can be made, and both banks and regulators are advised to stay alert.

Interestingly, we have here a close analogy to regulatory practice in the health sector, where market authorization for medicinal products hinges on comparative clinical trials and statistical tests. In essence, we argue that in order to maintain our banks' and societies' financial health, the same type of scrutiny ought to be applied.

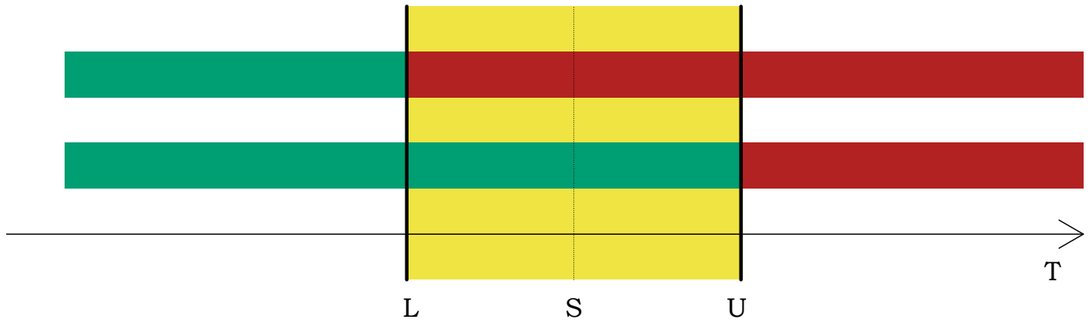


Figure 12: Illustration of the proposed traffic light approach to banking regulation. The limits L and U separating the yellow zone from the green and red zones derive from statistical tests under distinct hypotheses, as indicated by the horizontal bars.

Random Particles of flexible smoothness

Mathematical models for three-dimensional particles have come in for considerable interest in astronomy, biology, geosciences, and material science, among other disciplines. While some particles have a rigid shape, such as crystals and minerals, many real-world objects are star-shaped and vary stochastically. In a joint paper with Linda Hansen of Varde College in Denmark, Thordis Thorarinsdottir of the Norwegian Computing Center, and Donald Richards of Pennsylvania State University in the United States, CST researchers Evgeni Ovcharov and Tilmann Gneiting introduce a flexible framework for modeling three-dimensional star-shaped particles. Mathematically, the particles are represented by a Gaussian random field, which we construct from white noise by means of a kernel smoothing on the sphere.

The roughness or smoothness of a particle can be quantified by a measure called the Hausdorff dimension, which varies between 2 and 3, with the lower limit corresponding to a smooth, differentiable surface, and the upper limit corresponding to a very rough, space-filling surface. The novelty of the approach by Hansen et al. (2015) lies in its flexibility, as it allows for the synthetic generation of random particles of any desired roughness or smoothness. Technically, this is achieved by adapting the shape of the smoothing kernel in the simulation algorithm, as we have rigorously demonstrated using mathematical analysis tools and illustrated in case studies. For example, *Figure 13* shows simulated Gaussian particles with continuous but nowhere differentiable surfaces. *Figure 14* displays synthetic replicates of Venus, the Earth, the Moon, and Mars, with Hausdorff dimensions and other surface properties as reported in the planetary physics literature.

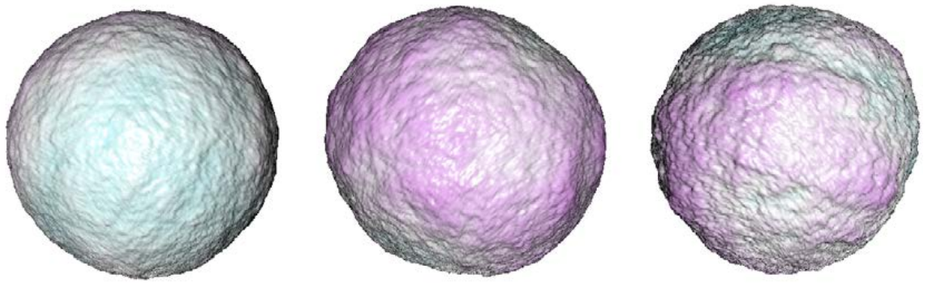


Figure 13: Three-dimensional Gaussian particles with continuous but nowhere differentiable surfaces. The Hausdorff dimension of the particle surfaces is 2.5. (Image provided by Linda Hansen).

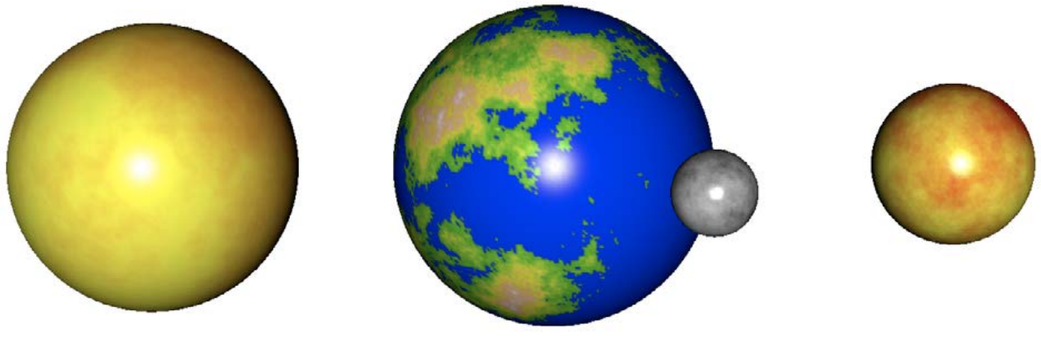


Figure 14: Simulations of Venus, the Earth, the Moon, and Mars in true relative size. (Image provided by Linda Hansen).

Explore Science

In July the CST group contributed to the Explore Science fair under the motto Physics: Pure Motion! at the Luisenpark in nearby Mannheim *(see Section 5.6)*. This outreach event aims to present scientific ideas to students and their families in a playful and entertaining way. With this in mind, members of the CST group put their heads together and designed a game that demonstrates what statistical forecasting is all about: using data and mathematical reasoning to predict the future and make smarter decisions.

These ideas are embodied in a flood forecast game in which the players need to make sure their water basin doesn't get flooded *(Figure 15)*. The players are given toy money, which they can use to build dams of different sizes. The small dam is cheap but shaky; the big dam is expensive but safe. They then throw a dice that determines the amount of water poured into the basin. The goal is to make it through several rounds without getting flooded.

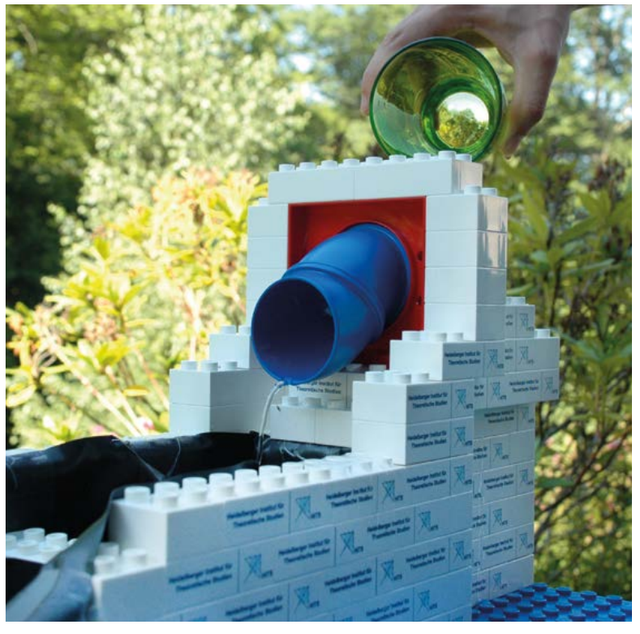


Figure 15: A selfmade flood game using toy bricks.

Figure 16: CST group member Roman Schefzik at Explore Science 2015.

Simple as the game sounds, players' strategies are far from trivial. One central source of complexity is that the game is played over several rounds, with the amount of flooding in one round affecting the outlook for the next round. Furthermore, the optimal solution depends on strategic goals, such as playing it safe versus gambling to make big money. Hence, the game contains integral elements of real-world decision problems of the kind that continue to inspire statisticians in academia and industry.

The students enthusiasm and curiosity made the Explore Science fair a most gratifying experience for the CST members. The children enjoyed the game's gambling character, and many of them put a lot of thought into figuring out the costs and benefits of the dams *(Figure 16)*. The Explore Science fair thus proved to be an excellent opportunity to meet up with the next generation of scientists.

2 Research

2.4 Data Mining and Uncertainty Quantification (DMQ)

2.4
Data Mining and Uncertainty Quantification (DMQ)

In our activities, we are constantly generating and gleaning enormous amounts of data, e.g. via sensors or high-performance computing. In recent years, this process has been dramatically accelerated by technological advances. Of course, extensive data production should enable us to improve knowledge discovery. In fact, however, the increasing size and complexity of the new data sets may lead to the opposite effect, since core information can get lost in huge databases.

The situation gets more involved if we consider the reliability of the derived information, e.g. in numerical simulations of natural processes. Recent advances in computing capability make for much better simulation fidelity. However, this fidelity will not necessarily improve the reliability of knowledge about the process thus simulated. Reliability is achived by quantifying the uncertainties in the simulations, including measurement errors, lack of knowledge about model parameters or inaccuracy in data processing.

In the Data Mining and Uncertainty Quantification group headed by Prof. Dr. Vincent Heuveline, information discovery is addressed by data mining technologies, and reliability considerations are addressed by uncertainty quantification methods. Both fields require a decidedly interdisciplinary approach to mathematical modeling, numerical simulation, hardware-aware computing, high-performance computing, and scientific visualization. Current fields of application are medical engineering, fluid dynamics, energy, astronomy, and meteorology.

Wir sammeln und erzeugen kontinuierlich enorme Datenmengen, zum Beispiel mittels Sensoren oder Hochleistungsrechnern. In den letzten Jahren wurde dies besonders durch technologische Fortschritte angetrieben. Diese Entwicklungen sollten zu einem deutlichen Wissensgewinn führen. Tatsächlich birgt aber die wachsende Komplexität und Größe der Datensätze sogar die Gefahr des Informationsverlustes, da die wesentlichen Informationen nicht mehr einfach erschließbar sind.

Diese Situation wird noch einmal verschärft, wenn wir die Zuverlässigkeit der gewonnenen Informationen, beispielsweise anhand von numerischen Simulationen in den Naturwissenschaften, betrachten. Zwar können wir immer genauere Simulationsergebnisse berechnen, jedoch führt dies nicht zwangsläufig zu zuverlässigeren Informationen über den zu simulierenden Prozess. Eine höhere Zuverlässigkeit können wir dadurch gewinnen, dass wir Unsicherheiten wie zum Beispiel Messfehler, mangelndes Wissen über Parameter und Rechenungenauigkeiten quantifizieren.

In der Gruppe „Data Mining und Uncertainty Quantification“ unter der Leitung von Prof. Dr. Vincent Heuveline wird Wissensgewinn mittels „Data Mining“ – Technologien, sowie Zuverlässigkeitsanalyse mittels Methoden der „Uncertainty Quantification“ durchgeführt. Beide Arbeitsfelder erfordern eine stark interdisziplinäre Arbeit in den Bereichen mathematische Modellierung, numerische Simulation, hardwarenahe Programmierung, Hochleistungsrechnen und wissenschaftliche Visualisierung. Aktuelle Forschungsfelder sind Medizintechnik, Strömungsmechanik, Energieforschung, Astronomie und Meteorologie.

Group leader

Prof. Dr. Vincent Heuveline

Scholarship holders

Chen Song *(HITS Scholarship)*
Michael Bromberger *(HITS Scholarship)*

Staff members

Dr. Michael Schick *(until September 2015)*
Dr. Peter Zaspel *(since October 2015)*
Maximilian Hoecker
Philipp Gerstner

Intern

Eva Vidlickova *(June – September 2015)*

Visiting scientist

Stefanie Speidel

Large high-dimensional clustering

Our aim here is to perform clustering analysis on scientific datasets. Most scientific datasets consist of extremely large numbers of data instances. Data instances are the sum of the measurements of an observed object. To give an example from a non-scientific background, consider a photobook as a dataset. Here, each photo is a data instance, and each pixel of each photo is a measurement. Given the nature of scientific datasets, clustering analysis is faced with a number of challenges.

Today's datasets contain millions of items, which leads to the requirement that analysis methods should be scalable in terms of the number of data instances involved. Some of the currently available clustering algorithms have quadratic or cubic complexities. Assuming that new technology often leads to a growth in data acquisition size by a factor of up to four, the execution of clustering algorithms would last between 16 and 64 times longer. Analysis along these lines is quite simply not feasible.

A dataset comprising millions of items, high-dimensional data, and uncertainties is known as a complex dataset.

Alongside enormous amounts of instances, the high dimensionality of the each instance is another challenging aspect of clustering analysis. High dimensionality is introduced if measurements are not correlated. In that case, each measurement has to be modeled as a separate dimension. Many clustering algorithms are subject to exponential computational complexity dependence in the data dimensionality. Dealing with thousands of dimensions while upholding acceptable output quality in the considered algorithms is often impossible.

Moreover, scientific measurements are subject to uncertainties deriving from aleatoric or epistemic errors in measured objects, measurement devices, systematic errors, and so forth. This uncertainty needs to be incorporated into the analysis, since errors can vary all the way up to a high double-digit percentage of the corresponding measurement. A dataset comprising millions of items, high-dimensional data, and uncertainties is known as a complex dataset.

In 2015, the DMQ group developed a new method for performing clustering analysis on complex datasets in a log-linear time scope called Fractal Similarity Measures. The scientific test case considered was the so-called initial problem statement: unlabeled datasets from the Sloan Digital Sky Survey 3, Data Release 10 (SDSS3 DR10) with a focus on the similarity / dissimilarity relationship of two objects were to be analyzed. Each object is represented by a feature vector of approx. 5,000 dimensions. These vectors display a numeric value for a captured spectrum with uncorrelated noise for all specific wavelengths. The whole dataset consists of 3 million objects with 60 GBytes of raw data in all. As all objects have to be compared with each other, the resulting complexity is $O(n^2)$ in computation and storage.

With regard to the test case mentioned, a naïve full analysis with distance-density clustering algorithms would amount to 542 days of processing time using a single similarity measure on a 128-core computing cluster. The time assumption behind this is 2 ms per comparison, including loading and saving data. The 9×1012 comparisons would effectively produce between 24 TByte and 120 PByte of data depending on the level of detail.

In this research project, our Fractal Similarity Measures approach was able to reduce the processing time to ca. 90 h on the same hardware. In the analysis of this complex dataset, clusters have been created that are visually comparable to clusters created by domain experts.

Uncertainty Quantification (UQ) for a blood pump device

Heart disease is one of the most severe diseases in the world. Around 20% of the patients suffering from heart disease will be affected by heart failure. There are over 2 million heart failure-related deaths each year, and the number is increasing. Although heart transplant operations are already well established, the shortage of heart donations places severe restrictions on heart transplantation frequency. Given this situation, ventricular assist devices (VADs) can play an important role for heart disease patients.

Since the end of the last century, special VADs, i.e. blood pumps are one of the most effective therapeutic instruments for the treatment of cardiac insufficiency. They completely or partially replace the natural function of the heart, thus guaranteeing blood circulation. More specifically, it has been shown that the advantages of the centrifugal pump (the model we are studying) are its ability to work under different conditions and its compact dimensions. Nowadays, the use of numerical simulation (notably by the Finite Element Method (FEM)) is a very common resource for mechanical design. For our blood pump appliance in particular, it can provide important information about the performance of the VAD pertaining to regions of high shear stress, high pressure, vortices, etc.

A vast number of blood pump devices have been successfully implanted for the benefit of heart disease patients. However, there are still various uncertainty parameters that could not be taken into account in the numerical modeling processes. Consequently, Uncertainty Quantification (UQ) is of fundamental importance in enhancing confidence in numerical estimation.

In this project, we enhance the deterministically incompressible Navier-Stokes simulation to probability quantification by using the stochastic Galerkin Projection method in our open source Finite Element library HiFlow3. Different uncertain scenarios are taken into account, such as boundary condition, viscosity, rotation speed, etc. With the help of Polynomial Chaos Expansion we are capable of dealing with physical and stochastic space at the same time. In order to obtain a high-accuracy solution, we need to solve a system with abundant degrees of freedom. We are therefore developing efficient parallel numerical methods (e.g. preconditioning techniques) for the stochastic Galerkin method.

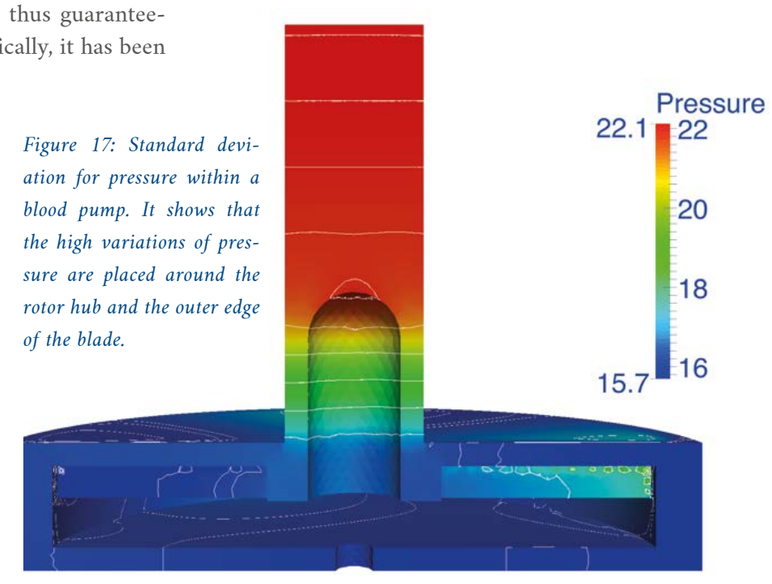


Figure 17: Standard deviation for pressure within a blood pump. It shows that the high variations of pressure are placed around the rotor hub and the outer edge of the blade.

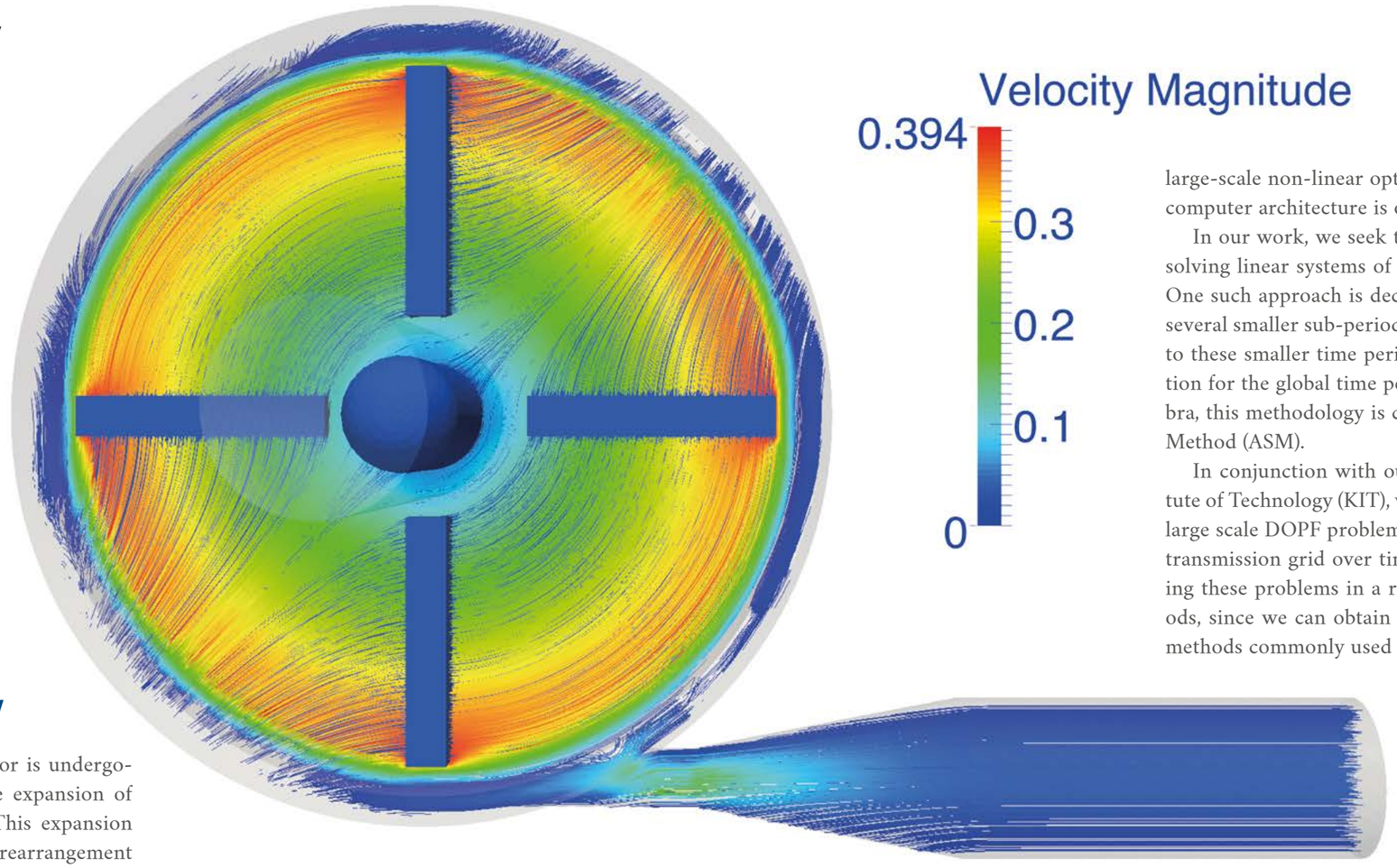


Figure 18: Standard deviation for velocity magnitude within a blood pump. On the surface of the rotor and the connection between housing and outlet, there are high velocity variations. This indicates that those areas need special care (w.r.t. safety) in industrial design.

Numerical methods for dynamic optimal power flow

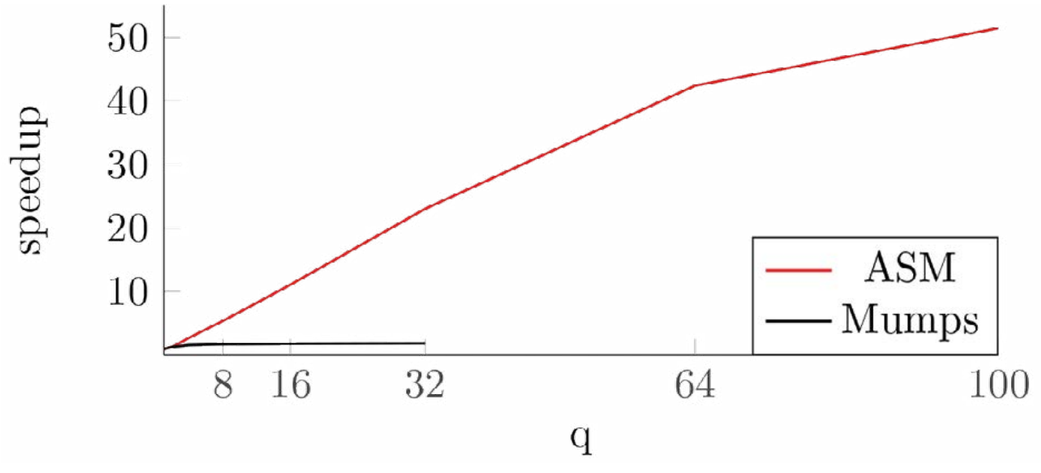
In many countries, the energy sector is undergoing substantial changes due to the expansion of renewable energy sources (RES). This expansion necessitates an extensive structural rearrangement of the power system, with the power grid taking center stage. While today's power grid infrastructure has been designed for centralized and controllable power generation in conventional power plants, RES expansion leads to increasingly uncertain, volatile, and decentralized power generation. To ensure dependable grid operation, methods are needed that can deal with high-resolution regional and temporal input data. Inevitably, this requirement will lead to a target conflict between model complexity and computational intensity on the one hand and model accuracy on the other. Accordingly, the crucial challenge involved is to provide efficient methods for power grid optimization, including accurate consideration of non-linear and non-convex alternating-current (AC) power-flow constraints.

The problem of finding the optimal operating state of a given power grid, also known as Optimal Power Flow (OPF), is stated as the minimization of a cost function with respect to a vector of continuous optimization variables like node voltages, generated active power, and generated reactive power. A feasible vector has to satisfy the AC power-flow equations based on Kirchhoff's circuit law in order to guarantee that the power demand is covered by the power actually generated. In addition, a set of inequalities has to be fulfilled to ensure that no technical restrictions (transmission line limits, etc.) are violated. In dealing with the problem of Dynamic Optimal Power Flow (DOPF), one considers several OPF problems, each corresponding to one specific point in time. Here, the power demand is time-dependent, and one has to take into account additional constraints that couple optimization variables corresponding to different time steps. Among others things, these couplings are introduced by energy storage facilities and limits at the temporal rate of power generation change for conventional power plants. Since DOPF is a large-scale non-linear optimization problem, solving it on a parallel computer architecture is of crucial importance.

In our work, we seek to develop efficient numerical methods for solving linear systems of equations arising in the context of DOPF. One such approach is decomposing the time period of interest into several smaller sub-periods, computing sub-solutions corresponding to these smaller time periods in parallel, and reconstructing a solution for the global time period. In the field of numerical linear algebra, this methodology is commonly known as the Additive Schwarz Method (ASM).

In conjunction with our project partners at the Karlsruhe Institute of Technology (KIT), we use the methods thus developed to solve large scale DOPF problems based on realistic data from the German transmission grid over time periods of several days to weeks. Solving these problems in a reasonable time is possible with our methods, since we can obtain a large parallel speed-up compared to the methods commonly used in DOPF. (*see Figure 19*)

Figure 19: Parallel speed-up for the ASM method developed by our team compared with the state-of-the-art parallel direct solver Mumps in dealing with a large-scale DOPF problem on 2 to 100 cores.



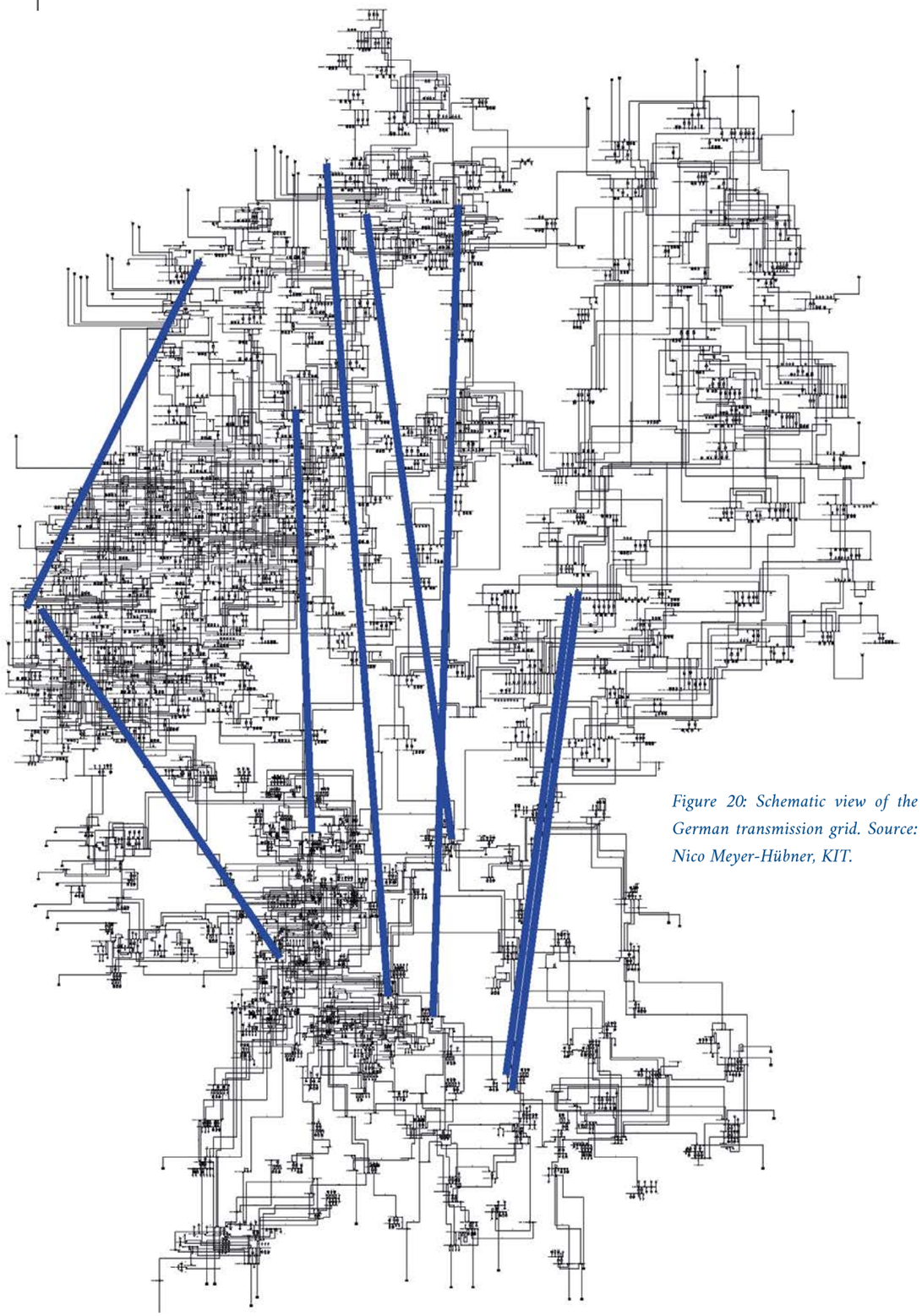


Figure 20: Schematic view of the German transmission grid. Source: Nico Meyer-Hübner, KIT.

Accuracy- and hardware-aware computing

Increasing the frequency of single core processors, an achievement made possible by decreasing line-widths (i.e. transistor gate lengths) while reducing energy consumption at the same level, has resulted in a major performance improvement for all applications. Due to technical limitations, increasing the frequency is no longer feasible. Instead, performance is now increased by using multi-core processors and thus exploiting the inherent parallelism of an application. Moreover, the advent of heterogeneous computing units like Graphics Processing Units makes for a massive speed-up in certain applications. Additionally, reconfigurable architectures like Field-Programmable Gate Arrays are available via PCI-Express cards for the HPC market. In future, they will be closely coupled to host processors, thus enabling specialized hardware support for certain parts of an application.

However, these advances in computer architecture still do not fulfill the requirements of today's scientific and industrial applications. For example, dual-view plane illumination microscopy observes fluorescent probes like living cancer cells and achieves good resolution for the 3D images acquired as well as high acquisition speed and a good sectioning of samples. Such microscopy can produce several hundred gigabytes or even a few terabytes within a short time. Accordingly, we need efficient ways of processing such data. One step in that direction is to accelerate the preprocessing required for images acquired by dual-view plane illumination microscopy. To this end, we have developed a FPGA-based architecture for an efficient convolution of 3D images. We used the FPGA-based convolution as part of the Richardson-Lucy deconvolution algorithm, which enabled us to outperform an Intel Xeon E5-2670 processor, where all 32 hardware threads were used.

However, even such specialized hardware support does not fully meet the needs of scientists, so new ways of improving the processing of huge data volumes have to be found. One way is to consider the accuracy of algorithms, software, and hardware within a computing system as an optimization parameter. For example, certain numerical algorithms benefit from more precise calculations compared to IEEE-754-based 64-bit floating-point operations. The well-known Lanczos method for iteratively calculating eigenvalues is flawed by inexactitude. In-depth evaluation of different software libraries has shown that even for the modified Lanczos method higher accuracy increases the number of correctly calculated eigenvalues for certain matrices *(see Figure 21, next page)*. Providing higher accuracy in this case increases the efficiency measured as execution time per correct eigenvalue.

On the other hand, applications used in machine learning, image processing, and data mining have a degree of tolerance over and against inexact computations, as has been demonstrated in the literature. This aspect can be leveraged to optimize remaining design parameters like latency, throughput, or energy consumption. The research topic approximate computing deals with such tolerably inexact computations. There are existing approaches that can be integrated into software, software libraries, operating systems, and the hardware itself. When integrating approximate computing into a computing system, two things are important: (1) parameters for adapting the accuracy and (2) selecting the right parameters during design time or runtime. For the first of these, it is essential to find approaches that are widely applicable. Accordingly, we investigated the accuracy impact of storing different data types that can be selected statically or dynamically *(see Figure 22)*. Real measurements proved that using data types with lower precision saves energy by up to a factor of 4. In line with these findings, we developed a hardware unit that converts internal data types to lower ones before transferring them to memory. Such a unit has a negligible power and area impact over and against currently embedded processors.

To sum up, considering accuracy as a design parameter is crucial for future computing systems. Avoiding over-precise calculations where they are not needed can improve other design goals. In our future work, we will be turning to the classification of time series. Clustering of time series often requires the comparison of a new object with all known database objects, hence the complexity is $O(n^2)$. Since time series are often inexact or noisy and algorithms are not usually able to compute the correct solution, applying approximate computing approaches for clustering is useful.

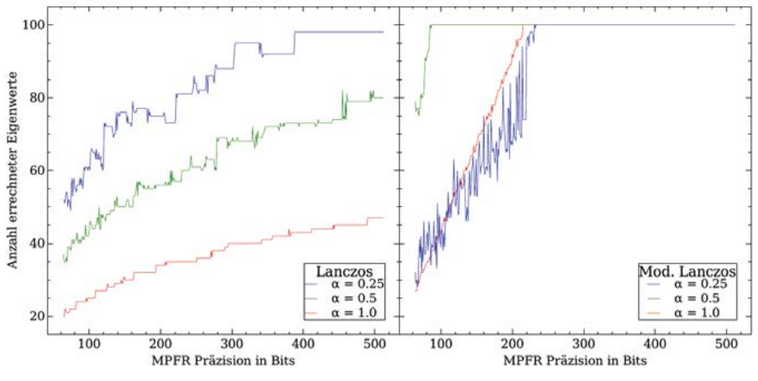


Figure 21: Number of correct eigenvalues calculated by either the unmodified (left) or modified (right) Lanczos method. α is a parameter to specify the distribution of the eigenvalues. Higher accuracy (number of bits for used data type) increases the number of correct calculated eigenvalues. Execution units, thus reducing both energy consumption and execution time.

(a) Input image

(b) Convolution using 64 bit floating point values (Mean Square Error = 0).

(c) Convolution using 16 bit floating-point values (Mean Square Error ≈ 0.0005).

(d) Convolution using 16 bit fixed-point values (Mean Square Error ≈ 5).

Figure 22: Convolution of an image (a) with a 5x5 Gaussian kernel. Two iterations were performed. 64 bit floating-point (b), 16 bit floating-point (c), or 16 bit fixed-point was used as underlying data type to store values. Data types with lower accuracy have only minor impact on the quality but less data has to be transferred between memory and execution units, therefore energy consumption as well as execution time is reduced. Image: Playboy Enterprises, Inc.

2 Research

2.5 Groups and Geometry (GRG)

2.5
Groups and Geometry (GRG)

The research group “Groups and Geometry” started its work in June 2015. It cooperates closely with the “Differential Geometry” research group at Heidelberg University. Both groups are headed by Prof. Dr. Anna Wienhard.

Symmetries play a central role in mathematics, as in the natural sciences. Mathematically, symmetries are transformations of an object that leave this object unchanged. These transformations can be composed, i.e. applied one after the other, and then form what is called a group.

In the 19th century, the mathematician Felix Klein proposed a new definition of geometry: Geometry is the study of all properties of a space which are invariant under a given group of transformations. In short: geometry is symmetry. This concept unified classical Euclidean geometry, hyperbolic geometry (then newly discovered), and projective geometry, which has its origins in the study of perspective in art and is not based on the measurement of distances but on incidence relations.

Even more importantly, Felix Klein’s concept fundamentally changed our view of geometry in mathematics and theoretical physics and has remained an important influence until today.

In our research group we investigate various mathematical problems in the fields of geometry and topology that involve the interplay between spaces, such as manifolds or metric spaces, and groups acting as groups of symmetric on them. A special focus lies on the study of (a) deformation spaces of geometric structures and (b) representation varieties.

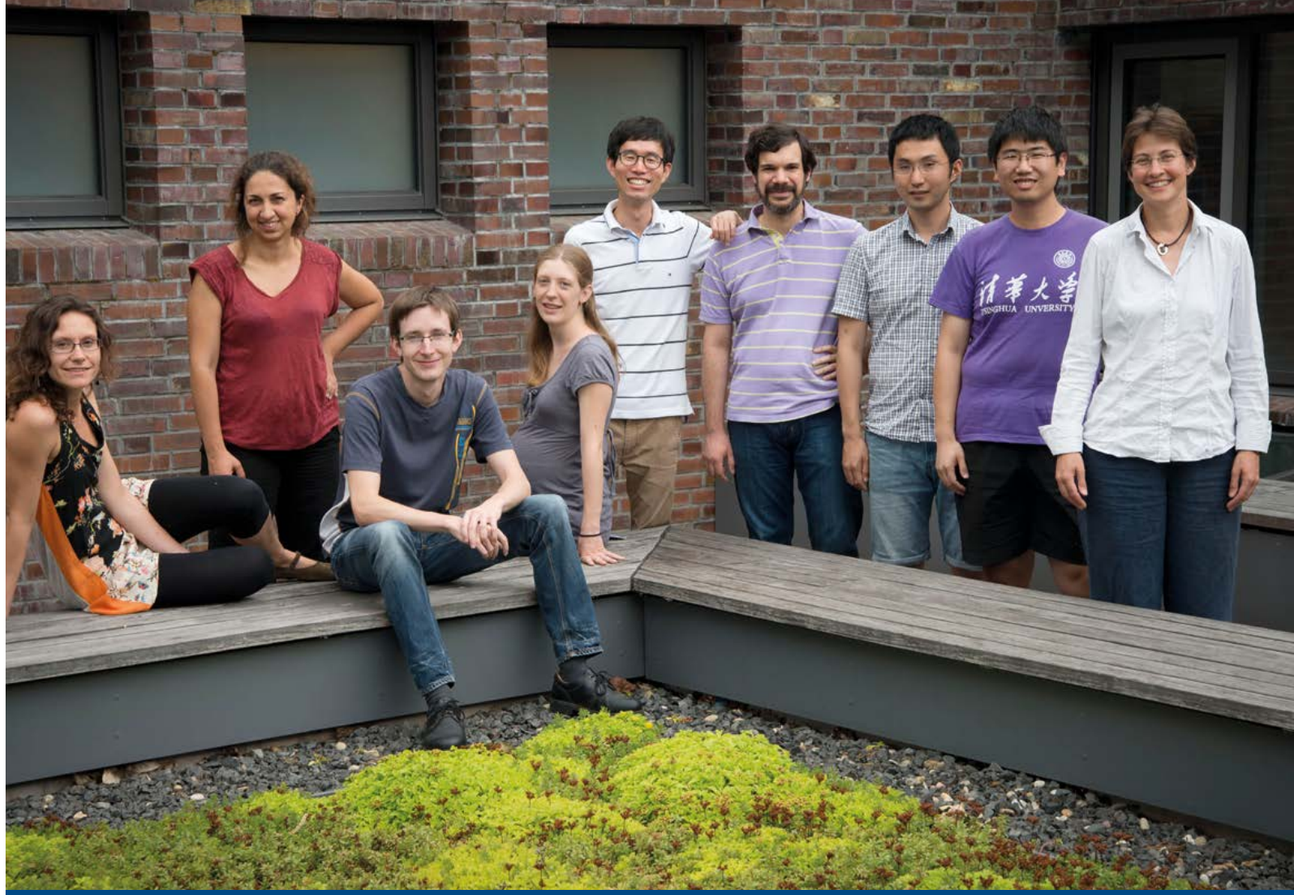
Die Arbeitsgruppe „Gruppen und Geometrie“ startete im Juni 2015. Sie arbeitet eng mit der Arbeitsgruppe „Differentialgeometrie“ an der Uni Heidelberg zusammen. Beide Gruppen werden von Prof. Dr. Anna Wienhard geleitet.

Symmetrien spielen eine zentrale Rolle in der Mathematik als auch in vielen Naturwissenschaften. In der Mathematik verstehen wir unter Symmetrien die Transformationen eines Objektes, die diese invariant lassen. Solche Transformationen lassen sich verknüpfen, d.h. hintereinander ausführen und bilden so eine Gruppe.

Im 19. Jahrhundert entwickelte der Mathematiker Felix Klein einen neuen Begriff der Geometrie: Geometrie ist das Studium der Eigenschaften eines Raumes, die invariant sind unter eine gegebenen Gruppe von Transformationen. Kurz gesagt: Geometrie ist Symmetrie.

Mit diesem Konzept vereinheitlichte Klein die klassische Euklidische Geometrie, die damals gerade neu entdeckte hyperbolische Geometrie, und die projektive Geometrie, die aus dem Studium der perspektivischen Kunst erwuchs und die nicht auf dem Messen von Abständen, sondern auch Inzidenzrelationen beruht. Noch wichtiger ist, dass Felix Kleins Konzept unser Verständnis von Geometrie in der Mathematik und der theoretischen Physik grundlegend verändert hat und bis heute prägt.

Unsere Arbeitsgruppe beschäftigt sich mit verschiedenen mathematischen Forschungsfragen aus dem Gebiet der Geometrie und Topologie, in denen das Zusammenspiel zwischen Räumen und Gruppen, die auf diesen als Symmetriegruppen wirken, zentral ist. Besondere Forschungsschwerpunkte liegen auf dem Studium von Deformationsräumen geometrischer Strukturen und Darstellungsvarietäten.



Group leader

Prof. Dr. Anna Wienhard

Staff members

Florian Stecker *(since October 2015)*
Dr. Ana Peon-Nieto *(since October 2015)*

Visiting scientists

Anna Schilling
Dr. Andreas Ott
Dr. Gye-Seon Lee
Dr. Sourav Ghosh *(since October 2015)*
Nicolaus Treib
Dr. Daniele Alessandrini
Dr. Shinpei Baba

What is geometry?

We all learned at school what geometry is: we learned when two triangles are congruent, we learned the Pythagorean theorem $a^2 + b^2 = c^2$, which tells us the relation between the length s of the sides of a right-angled triangle, we learned that the sum of the inner angles of a triangle is 180 degrees. These are all properties of Euclidean geometry. But we do not have to go far to see that geometric properties in other spaces are very different. We live on the surface of the earth, which we idealize to be a sphere, a round ball. On the sphere the shortest way to pass between two points is to follow a great circle, so pieces of great circles are the "straight lines" in this geometry. Given three points on the sphere, we can form the triangle by joining the points pairwise by "straight lines". If we then measure the sum of the interior angles, we find that they do not add up to 180 degrees, in fact the sum of the interior angles depends on the triangle we choose, but it is always bigger than 180 degrees. Nor does the Pythagorean theorem hold any longer. In spherical geometry – the geometry of our life on the surface of the Earth – many of the properties we consider to be geometric facts are not true. In the 19th century, another species of non-Euclidean geometry was discovered, now referred to as hyperbolic geometry. Here the sum of the interior angles of a triangle is always less than 180 degrees and can be arbitrarily small when the triangle gets big.

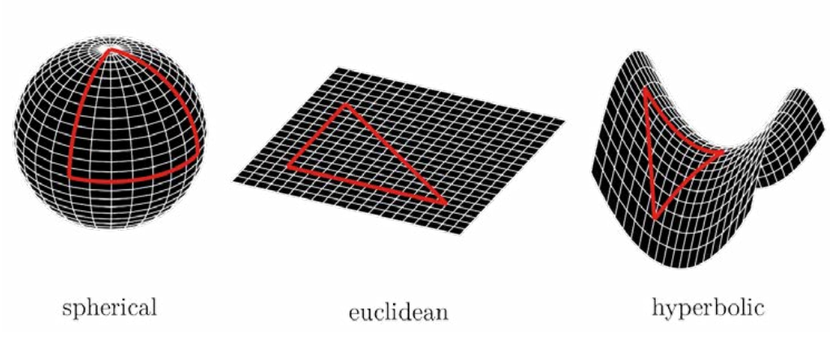
Whereas in Euclidean geometry, the world around every point is flat, in spherical geometry it is round. If you stand at a given point in hyperbolic geometry, the world around you looks like a saddle – or a Pringle chip.

It is actually not so easy to visualize the hyperbolic plane. There are several models on offer, in our illustrations we will be using the hyperboloid model and the Poincaré disc model. Every model correctly represents only some of the geometric properties of the hyperbolic plane (e.g. angles), while distorting others (e.g. distances) at least in our "Euclidean eyes."

Euclidean, spherical, and hyperbolic geometry are all based on measuring distances between points, they are Riemannian geometries. In dimension two, they are the only Riemannian geometries. In dimension three, Euclidean, spherical, and hyperbolic geometry are just three of the eight possible geometries.

The confusion about what geometry actually is increases when we consider geometries that are not based on the notion of a distance between points. One example is affine geometry, which is the geometry of a plane in which we cannot measure distances but only have a notion of parallelism. Another example is projective geometry, the kind of geometry we get when we take a plane and imagine adding points at infinity such that any two parallel lines will intersect at infinity. (Think of a perspective drawing where parallel lines seem to intersect at infinity.) These geometries were developed in the 19th century and queried the classical notion of what geometry is.

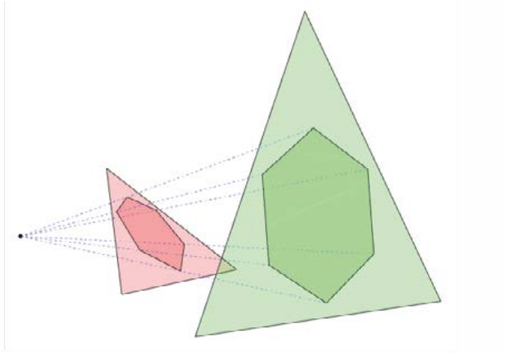
Figure 23: Triangles in the three two-dimensional geometries.



In his "Erlangen program" of 1872, Felix Klein (1849 – 1925) proposed a concept of geometry that unified all known geometries. At the heart of his concept is the idea of focusing not only on a space but on a space and a group of symmetries or transformations acting on that space. Following his proposals, a geometry is given by a space X and a group G of transformations of X, such that for every pair of points x and y in X there exists a transformation in G that sends x to y. Geometric properties of a space or of an object are those properties that do not change under any of the transformations in G, i.e. properties that are invariant under the given symmetries.

This concept is what we understand nowadays by a geometry in mathematics and in theoretical physics. And there are many more interesting and important geometries than the Euclidean, spherical or hyperbolic, affine or projective varieties, such as conformal geometry or Anti-de-Sitter geometry, which play an important role in physics.

Figure 24: Projective Geometry Figure.



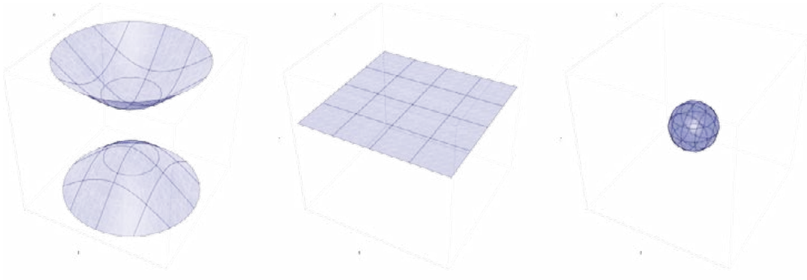
These two shapes look very different to our "euclidean eye", but projectively, as objects in the projective plane, they are congruent.

Projective geometry as a unifying geometry

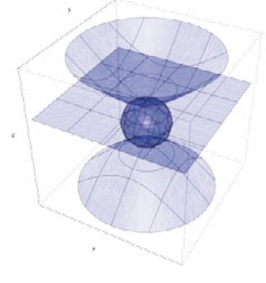
Felix Klein realized the universal features of projective geometry. He described how Euclidean, spherical, and hyperbolic geometry can all be regarded as sub-geometries of projective geometry. If X is the projective space and G the group of projective transformations, then there exist three open subsets E, S, and H of projective space and three subgroups G(E), G(S), G(H) of the group of projective transformations that preserve E, S, and H respectively, such that the geometry given by E and G(E) is Euclidean, the geometry given by S and G(S) spherical, and the geometry given by H and G(H) hyperbolic geometry.

In fact, in projective geometry it is actually possible to "transition" from hyperbolic geometry through Euclidean geometry to spherical geometry. More precisely, using projective transformations, one can rescale hyperbolic geometry and spherical geometry so that they limit on Euclidean geometry.

Figure 25: Subgeometries of projective geometry.



hyperbolic geometry *euclidean geometry* *spherical geometry*



as subgeometries of projective geometry

Similarly, all the eight three-dimensional Riemannian geometries that play a crucial role in Thurston's geometrization program, as completed by Perelmann a few years ago, can also be thought of as sub-geometries of projective geometry. In dimension three, we do not know all the potential transitions between the eight Thurstonian geometries.

In a joint project with Daryl Cooper (UCSB) and Jeff Danciger (UT Austin) [Cooper D et al., 2015] we describe all the potential limit geometries of the so-called affine symmetric sub-geometries of projective geometry. This gives us a precise description of all limit geometries of hyperbolic and spherical geometry in any dimension.

Geometric structures and their deformation spaces

In many situations, we are not interested in the geometry in itself but are faced with a more challenging problem. We are given a topological space, for example the surface of a donut or a pretzel, or the complement of a knot in three-dimensional space, and we would like to endow this space with a given geometry (X,G). This can be thought of as applying a geometric structure to the topological space such that around every point in our space it looks as if we were standing at a point in X.

This problem has a well-known solution for surfaces. Surfaces are classified by their genus, which is the number of "holes" you have in your surface – the sphere has genus 0, the donut, which in mathematics is called a torus, has genus 1, the surface of a standard pretzel you buy at a bakery has genus 3. The uniformization theorem in dimension two states that a surface can be endowed with precisely one of the two-dimensional geometries: with a spherical structure if the genus is 0, a Euclidean structure if the genus is 1. and a hyperbolic structure if the genus is greater than 1.

Figure 26: Surfaces and their structures.

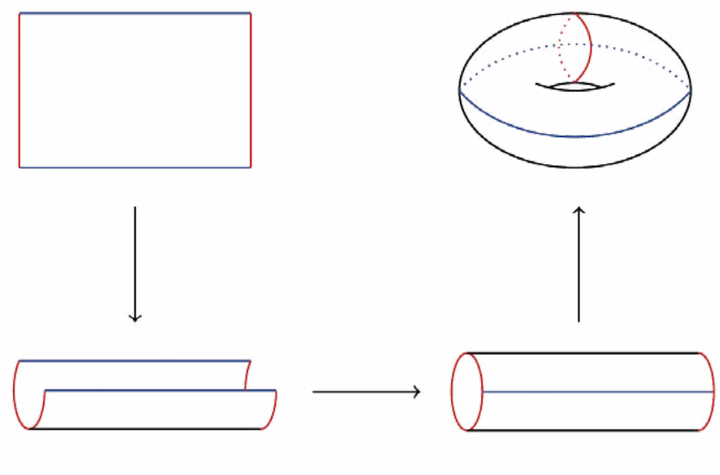
These surfaces carry a spherical, euclidean, hyperbolic structure.

In dimension three and higher, not every manifold can be endowed with a geometric structure. Thurston's geometrization theorem states that in dimension three any manifold can be cut into smaller pieces and these smaller pieces then endowed with exactly one of the eight three-dimensional geometries.

In fact, if a manifold carries a specific geometric structure, it might actually carry this geometric structure in several different ways. Take for example the torus. One way to endow this torus with a Euclidean structure is by constructing it out of a piece of the Euclidean plane. So take a rectangular piece of paper and glue the opposite sides together. This gives you a cylinder. Then you take the ends of the cylinder and glue them together (imagine the sheet of paper to be really stretchy). This gives you a torus, and around every point in this torus you just have a piece of the Euclidean plane, so it is equipped with a Euclidean structure.

You will realize immediately that we could have made the same construction with a parallelogram instead of a rectangle. This will produce a different Euclidean structure on the torus.

Figure 27: Folding a torus.



Folding a torus out of a piece of paper.

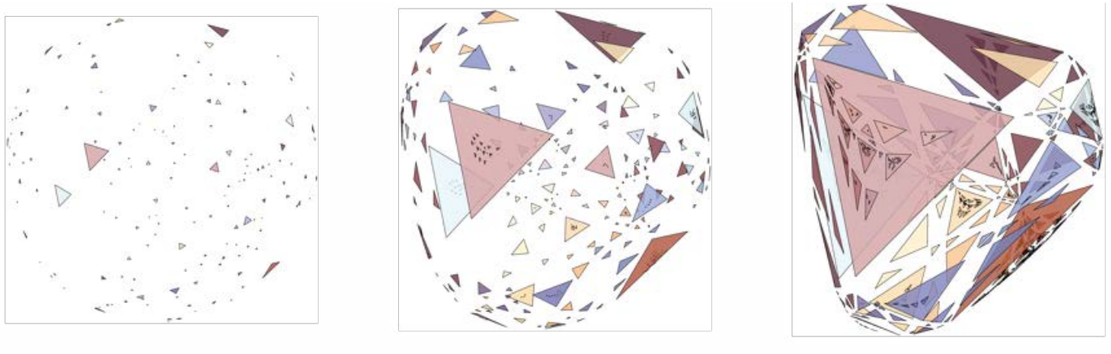
Thurston's geometrization theorem states that in dimension three any manifold can be cut into smaller pieces and these smaller pieces then endowed with exactly one of the eight three-dimensional geometries.

This way, one obtains a two-dimensional space of different Euclidean structures on the torus (if you fix the area of the parallelogram but vary its shape). For the pretzel surface (of genus 3), there is a 12-dimensional space of different hyperbolic structures. In general, for a surface of genus g>1, there is a 6g-6-dimensional space of different hyperbolic structures.

The space of all possible ways of endowing a manifold with a specific geometric structure is called the deformation space. These deformation spaces are very interesting and themselves carry a lot of structure. Sometimes, for example in a sphere, there is no interesting deformation space, just one spherical structure on the sphere. This phenomenon is called rigidity.

For hyperbolic three-dimensional manifolds, a famous rigidity theorem proposed by George Mostow [Strong rigidity of locally symmetric spaces. Annals of Mathematical Studies, Princeton 1973] states that there is a unique hyperbolic structure on this manifold. Since hyperbolic geometry is a sub-geometry of projective geometry, this hyperbolic structure induces a projective structure. In many cases, the hyperbolic structure can be deformed as a projective structure. This is a very interesting phenomenon that has recently been receiving a lot of attention. In joint work with Sam Ballas (UCSB) and Jeff Danciger (UT Austin), Gye-Seon Lee constructed many explicit deformations with totally geodesic boundaries [Samuel A. Ballas, Jeffrey Danciger, Gye-Seon Lee, Convex projective structures on non-hyperbolic three-manifolds, preprint arXiv:1508.04794.].

Figure 28: Deformation spaces: Projective deformations of the hyperbolic structure on a three dimensional hyperbolic manifold of finite volume.



Relation with representation varieties

Faced with a complicated object, mathematicians try to associate to this object invariants, which are easier to understand but contain important information about the object. We have already seen an example of such an invariant: to a surface we associate the genus. This invariant is just an integer. The uniformization theorem tells us that this integer already contains information on the geometric structures our surface may carry.

A slightly more sophisticated invariant is the fundamental group of a topological space. The fundamental group is an algebraic object (a group). It is constructed by considering closed paths on the space. For the torus above, the fundamental group is the group of pairs of integers (n,m), with the composition given by addition (n,m) + (p,q) = (n+p, m+q).

There is a very close relation between the deformation space of geometric structures modeled on (X,G) on a manifold M and maps from the fundamental group of M into the group G, which preserve the composition of group elements. Such maps are called group homomorphisms.

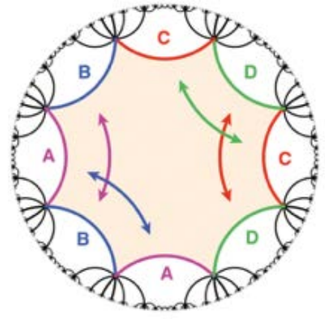
Charles Ehresmann (1905-1972) observed that locally, the space of geometric structures on M looks exactly like the space of all group homomorphisms from the fundamental group of M into G. This provides powerful techniques for studying geometric structures.

The global relation between the space of homomorphisms and the deformation spaces of geometric structures is however quite difficult to determine. In particular it is a very interesting and challenging problem to determine the group homomorphisms arising from geometric structures and to characterize them explicitly.

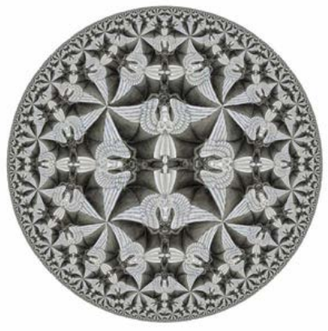
A model example where this problem has a very satisfying answer is again the deformation space of hyperbolic structures on a surface of genus g>1. In this case, the deformation space can actually identified with the space of all homomorphisms, which are injective with a discrete image. In fact, there is a beautiful relation between such homomorphisms and tilings.

A Euclidean structure on a torus, or a hyperbolic structure on a surface of genus g>1, gives rise to periodic tiling of the Euclidean or the hyperbolic plane, respectively. The symmetry group of the tiling is the fundamental group of the surface. And the group homomorphism of the fundamental group of the surface into the group of transformations of the Euclidean or hyperbolic plane is precisely given by realizing the fundamental group as the symmetry group of the tiling.

Figure 29: Tilings



Hyperbolic tiling, giving a surface of genus2. Image: William Goldman/Bill Casselman



An artistic version inspired by Escher's drawings. Image: Jos Leys, www.josleys.com

In recent years, special subsets of the space of homomorphisms of the fundamental group of a surface of genus g>1, more generally, so-called hyperbolic groups, into more general transformation groups, such as transformations of a symplectic vector space, have been discovered. These special subsets are called Hitchin representations, maximal representations, and Anosov representations.

In joint work with Olivier Guichard (IRMA Strasbourg) [Guichard O., Wienhard A., Anosov representations: Domains of discontinuity and applications, Inventiones Mathematicae, Volume 190, Issue 2 (2012)], we have shown that in fact all Anosov representations arise from geometric structures on higher dimensional manifolds. For many of the Hitchin representations these are projective structures on iterated sphere bundles over the surface.

Proper actions

In order to construct a geometric structure, given only a group homomorphism of the fundamental group of a surface into a transformation group G, one basic approach is to find an appropriate space X such that (X,G) is a geometry, and an open subset D of X with a tiling such that the fundamental group is the symmetry group of the tiling.

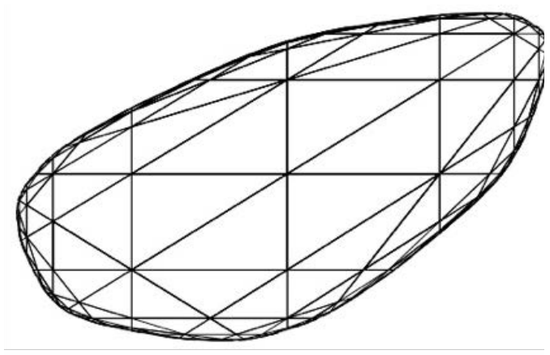


Figure 30: Projective tilings: A tiling of a convex subset of projective space. It associates a projective structure on a surface to a homomorphism of the fundamental group into SL (3,R), which lies in the Hitchin component.
Image: Kelly Delp, Cornell University

The group homomorphism sends every element of the fundamental group to an element of G, hence a transformation of X. Instead of constructing the tiling explicitly, it is sufficient to find a subset D that is preserved by all these transformations coming from elements of the fundamental group, such that the induced action of the fundamental group on D is properly discontinuous. This basically means that complicated elements of the fundamental group will move points in D far away. We call such a subset D a domain of discontinuity.

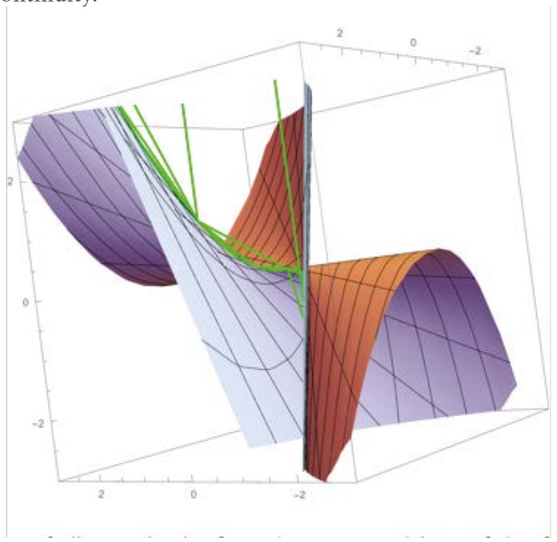


Figure 31: Domains of discontinuity: A domain of discontinuity for a homomorphism of the fundamental group into SL(4,R), which lies in the Hitchin component. It gives a projective structure on a three dimensional manifolds, which is a circle bundle over the surface.

In joint work with Olivier Guichard (IRMA Strasbourg) [Guichard O, Wienhard A, Anosov representations: Domains of discontinuity and applications, Inventiones Mathematicae, Volume 190, Issue 2 (2012)] we have constructed such domains of discontinuity D for all Anosov representations and shown that the tile of the associated tiling of D is actually compact (finite). In joint work with Francois Gueritaud (Université de Lille), Olivier Guichard (IRMA Strasbourg), and Fanny Kassel (Université de Lille) [Gueritaud et al, 2015a] [Gueritaud et al, 2015b], we have established a concrete link between Anosov representations and properly discontinuous actions not only on subsets but on the entire space X. This opens up many interesting questions which we will explore in the future. ■

2 Research

2.6 Molecular Biomechanics (MBM)

2.6
Molecular Biomechanics (MBM)

Why do astronauts suffer from bone loss? The reason is low gravity in space. Bones need mechanical stimulation for regeneration and growth. But how does bone tissue sense mechanical forces? Stem cell differentiation into bone tissue is complex. Several possible candidates have been suggested to work as force sensors in bone stem cells, all of which are proteins. But their interplay on a molecular scale has remained elusive. Our aim is to identify and characterize the sensors for mechanical stress operative in biological systems. To this end, we use and develop molecular simulations on different scales, novel analysis tools, and bioinformatics methods. In 2015, one primary focus has been on two specific mechano-sensors, focal adhesion kinase – involved not least in bone tissue formation – and the von Willebrand factor, a protein in the blood triggering the first stages prior to blood coagulation.

In addition, we have been very active in working together with Heidelberg experimentalists on mechanisms of protein interactions and functions. With Edward Lemke, EMBL, and other colleagues, we were able to formulate a new paradigm of how a disordered protein, nucleoporin, can bind its partners, a process vital to any eukaryotic cell. With Tobias Dick, DKFZ, we were able to decipher a surprisingly simple mechanism through which GAPDH, a well-studied metabolic protein, can be redox-regulated. Apart from molecular dynamics simulations, these studies also employed quantum chemical calculations on the lower length and time scale, and Brownian dynamics simulations as developed in the Molecular and Cellular Modelling group at HITS *(Chapter 2.7)* to reach larger scales.

Warum leiden Astronauten an Knochenschwund? Verantwortlich hierfür ist die geringe Schwerkraft im Weltraum. Knochen benötigen für Regeneration und Wachstum mechanische Stimulation. Aber wie nimmt das Knochengewebe mechanische Kräfte wahr? Stammzelldifferenzierung zu Knochengewebe ist komplex. Als mögliche Kraftsensoren in Knochenstammzellen wurden bereits mehrere potentielle Kandidaten, allesamt Proteine, in Betracht gezogen. Ihre Wechselwirkung auf molekularer Ebene konnte bisher jedoch noch nicht nachgewiesen werden. Unser Ziel ist es, die Sensoren für die mechanische Belastung in biologischen Systemen zu identifizieren und zu charakterisieren. Zu diesem Zweck nutzen und entwickeln wir molekulare Simulationen auf verschiedenen Längenskalen, neuartige Analyse-Tools und Methoden der Bioinformatik. Im Jahr 2015 lag der Schwerpunkt hierbei auf zwei spezifischen Mechanosensoren, der fokalen Adhäsionkinase, die unter anderem an der Knochengewebebildung beteiligt ist, und dem von-Willebrand-Faktor, ein Protein im Blut, das die ersten Schritte auslöst, die der Blutgerinnung vorangehen.

Darüber hinaus bestand eine rege Zusammenarbeit mit Heidelberger Forschern über die Mechanismen von Protein-Interaktionen und Funktionen. Mit Edward Lemke, EMBL, und anderen Kollegen konnten wir bestimmen, wie das ungeordnete Protein Nukleoporin sich an seine Partner bindet, ein Prozess von entscheidender Bedeutung für jede eukaryotische Zelle. Mit Tobias Dick, DKFZ, konnten wir einen überraschend einfachen Mechanismus entschlüsseln, durch den GAPDH, ein gut erforschtes metabolisches Protein, Redox-reguliert wird. Diese Studien umfassten außer Molekulardynamik-Simulationen auch quantenchemische Berechnungen auf der unteren Längen- und Zeitskala und Brownsche Dynamik-Simulationen, wie sie von der MCM Gruppe des HITS entwickelt wurden (Kapitel 2.7), um größere Maßstäbe zu erreichen.

Group leader

Prof. Dr. Frauke Gräter

Staff members

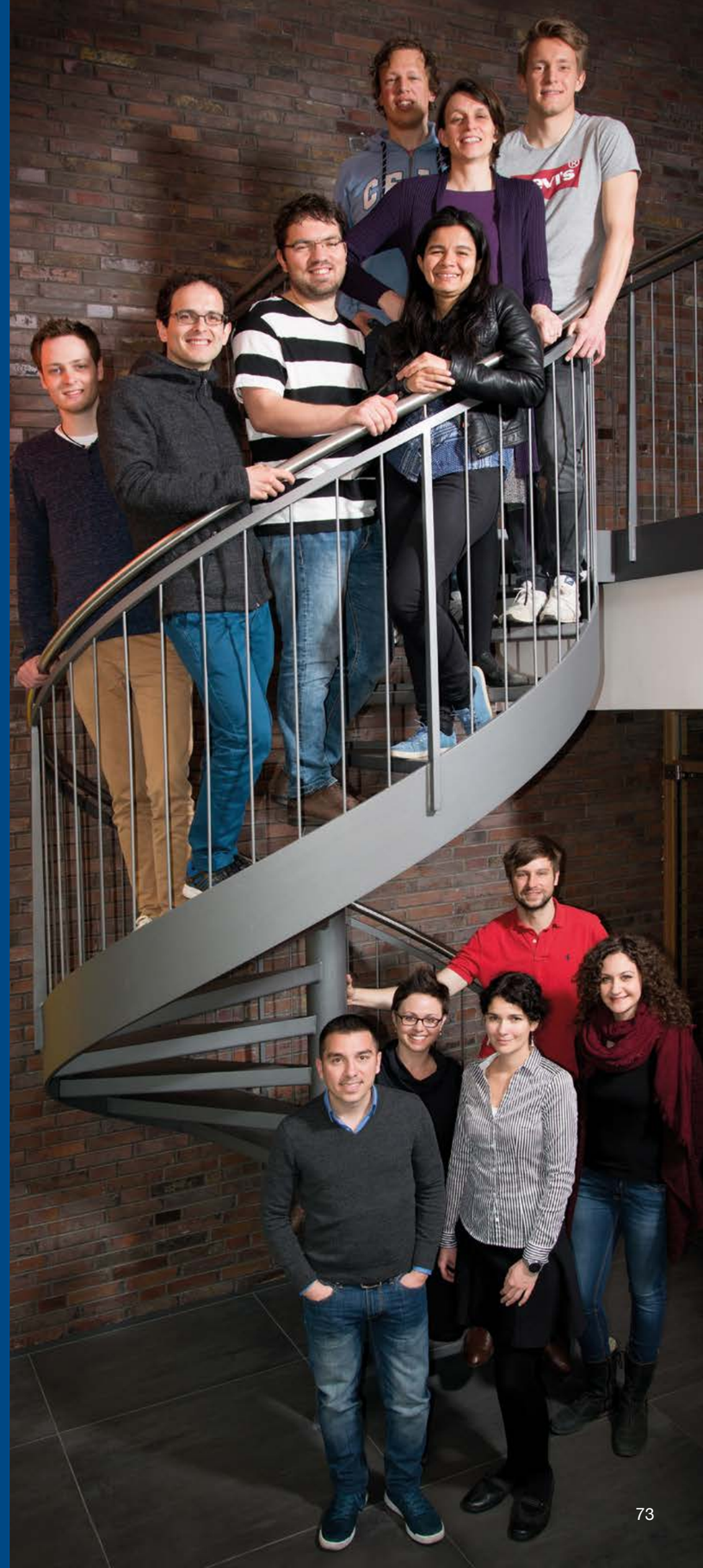
Dr. Camilo Aponte-Santamaria
Dr. Eduardo Cruz-Chu
Dr. Csaba Daday *(since October2015)*
Ana Herrera-Rodriguez
Dr. Katra Kolšek
Dr. Davide Mercadante
Dr. Vedran Miletic
(since November 2015)
Johannes Wagner

Scholarship holder

Jing Zhou *(until August 2015)*

Visiting scientists

Dr. Agnieszka Bronowska
(until April 2015)
Dr. Lipi Thukral
(May – June 2015)



Molecules for sensing force: focal adhesion kinase

Prof. Frauke Gräter, Dr. Jing Zhou

Focal adhesion kinase (FAK), a non-receptor tyrosine kinase, is known as a central regulator of focal adhesions (FAs), which are key cellular locations for mechanosensing via the translation of mechanical signals (outside cells) into biological signals (inside cells). Studies have shown that FAK mediates both force-guided cell migration and strain-induced proliferation. However, the mechanism of force-induced FAK activation has not been clear to date.

The aim was to reveal the mechanism behind FAK mechanical activation and determine how force is transduced through FAK to downstream signaling. FAK activity is regulated by an intramolecular autoinhibitory interaction between two of its domains – central catalytic and N-terminal FERM domain – by blocking the Tyr576/577 phosphorylation site needed for maximal FAK-associated activity.

Recent studies have emphasized the role of phosphoinositide phosphatidylinsositol-4,5-bisphosphate (PIP_2) as a stimulus for FAK activation. The binding of PIP_2 to the FAK basic patch can induce allosteric change in FAK, but not the exposure of Tyr576/577. Thus we hypothesized that the mechanical force as an additional stimulus for FAK activation takes place in 3 steps. First, FAK is tethered between the PIP_2 – enriched membrane and the cytoskeleton. Second, tensile force propagating from the membrane through the PIP2 binding site of the FERM domain and from the cytoskeleton-anchored FAT domain activates FAK by relieving the occlusion of the central phosphorylation site of FAK (Tyr576/577) by the autoinhibitory FERM domain. Third, FAK would be the first mechano-enzyme of FAs allowing the direct transduction of a mechanical signal into an enzymatic reaction that enhances the activation of subsequent signaling events.

To test this hypothesis, we performed extensive force-probe molecular dynamics (FPMD) simulations of the crystal structure containing the FERM-kinase fragment of FAK under various conditions in a mechanical environment. At HITS, Camilo Aponte-Santamaria and Jing Zhou have been involved in these studies, together with Sebastian Sturm from Leipzig University, a guest scientist visiting the MBM group at HITS in 2015. We tested the force-induced FAK activation process for membrane binding at different PIP_2 concentrations in different pulling directions with different pulling velocities and then compared it with the FAK activation mechanism in the system without the membrane. Our simulation results suggest a specific opening of the domain interface due to the lower mechanical stability of the kinase C-lobe with a helix bundle structure. Moreover, our results corroborated the assumption that PIP_2 clustering is needed for FAK activation at cell peripheries *(Figure 32)*.

We then established a mechano-biochemical network model *(Figure 33)* to connect force-dependent FAK kinetics based on extrapolated MD results to the expected downstream RasGTP concentrations. Extrapolation using a theoretical force spectroscopy model provided kinetic parameters for the downstream FAK activation. This hybrid approach combining MD and biochemical network simulations provides direct evidence for a mechano-enzymatic function of FAK in cell signaling.

Our results have initiated single molecule experiments to test our predictions. Should FAK's role as a force sensor be confirmed, this new role will have implications both for our views on force-induced bone stem cell differentiation and for ongoing efforts to design FAK inhibitors against various types of cancer.

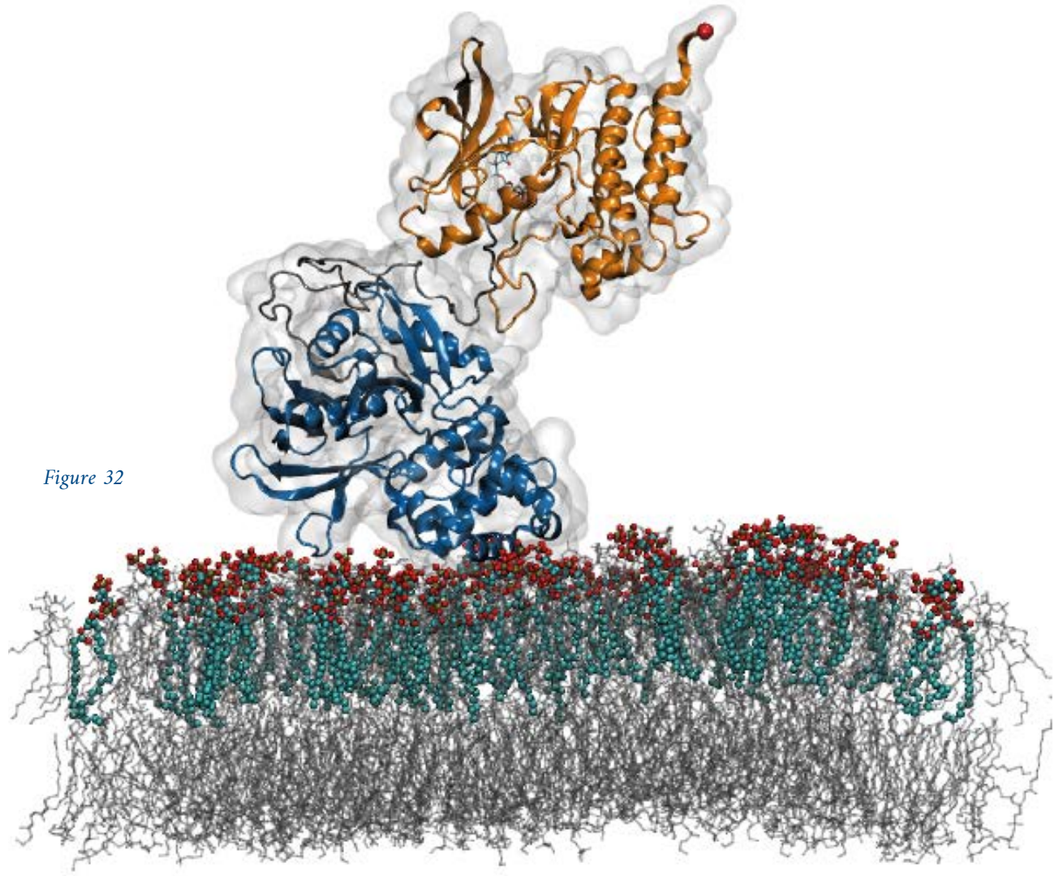


Figure 32

Figure 32: Simulation system of the FAK-membrane complex in a water box. FK-FAK (shown as a cartoon) interacts with the POPE membrane (shown as sticks in gray) via PIP_2 (shown as spheres and colored according to atom types). The force was applied to the Cα atoms of the kinase C-terminal residue (shown as red dots).

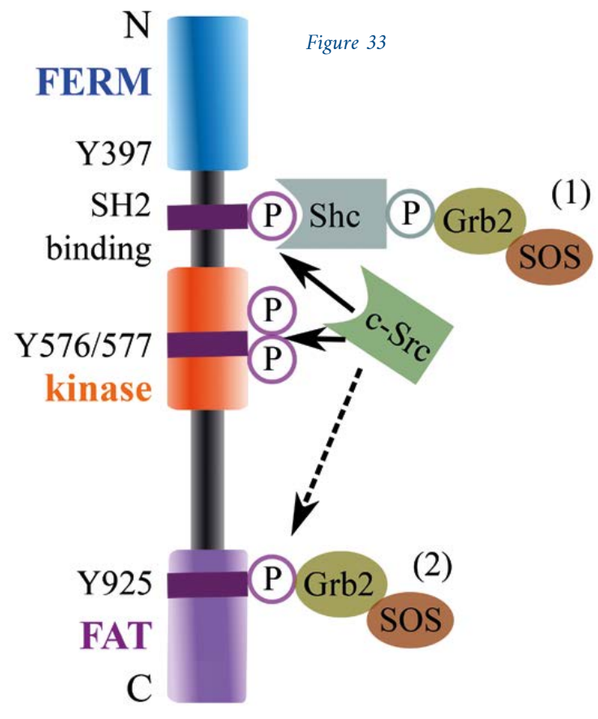


Figure 33

Figure 33: Schematic diagram of the interactions of FAK with Grb2, either directly or via Shc. The SH2 domain of Shc and c-Src can bind to the auto-phosphorylation site Tyr397 in FAK. The SH3 domain of Grb2 can bind to the phosphorylated Tyr317 in Shc and the phosphorylated Tyr925 in FAK.

Old dog, new tricks: GAPDH dehydrogenase as redox sensor at the molecular level

Dr. Agnieszka Bronowska

Many proteins perform more than one function in the cell. This has recently been recognized as a widespread phenomenon with important implications for human health and systems biology. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH), a well-known glycolytic enzyme, has been implicated in several non-metabolic processes, including transcriptional regulation and oxidation sensing in cells. The latter makes GAPDH an attractive anti-cancer target. GAPDH is overexpressed in multiple human cancers, including melanoma, and its expression is positively correlated with the progression of the disease. In this context, the molecular mechanism of GAPDH oxidation by hydrogen peroxide (H_2O_2) was a subject of studies conducted by researchers at HITS and at the German Cancer Research Center (DKFZ).

It has been known for many years that GAPDH is very sensitive to reversible oxidation by hydrogen peroxide, but the mechanism behind this oxidation process has remained elusive. Disruption of the oxidation of the catalytic cysteine C152 residue does not affect the glycolytic activity of GAPDH. Accordingly, some specific mechanism controls the oxidation of C152 by H_2O_2, a mechanism that must be distinct from the reaction between C152 and the substrate, glyceraldehyde 3-phosphate. In a study published in Nature Chemical Biology [Peralta et al., 2015] such a mechanism was described for the first time. It is based on a dedicated proton relay within the GAPDH core controlled by protein dynamics (*Figure 34*). A proposed model involving proton transfer through several key residues to the catalytic C152 was tested by a combination of computational methods, such as quantum mechanical (QM) calculations and all-atom molecular dynamics (MD) simulations. Using these methods, researchers at HITS designed GAPDH mutants, while the DKFZ experimentally tested and validated the proposed model. The generation of mutants in which the glycolytic and peroxidatic activities of GAPDH were decoupled allowed for a direct assessment of the physiological relevance of GAPDH H_2O_2-sensitivity. It was demonstrated that GAPDH's H_2O_2-sensitivity is a key component in the cellular adaptive response to the altered redox environment.

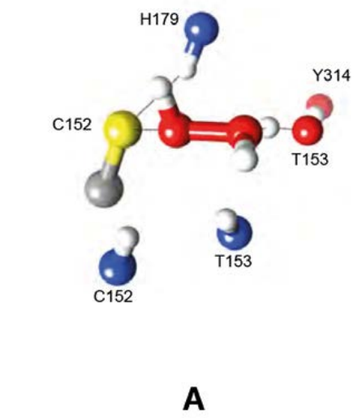


Figure 34: Model of the proton relay within the GAPDH core. Snapshot of the SN2 transition state based on the H_2O_2 reaction path suggested by QM calculations (A). Proposed reaction scheme for the proton relay: nucleophilic substitution is coupled to the protonation of the leaving hydroxyl ion by T153. In turn, T153 is re-protonated by Y314 (B).

Floppy but fast: revealing one of the secrets of nuclear transport by identifying a new paradigm regulating the binding of intrinsically disordered proteins

Dr. Davide Mercadante

Intrinsically disordered proteins (IDPs) belong to a class of molecules that, unlike globular proteins, lack secondary structure elements. Despite this lack of secondary structure, IDPs are involved in key physiological cell processes in which they mediate complex cellular functions.One of the crucial physiological processes in eukaryotic cells – the transport of molecules into and out of the nucleus – is effectively mediated by the interaction between IDPs and globular proteins. The exchange of components between the cytoplasmic and nucleoplasmic compartments is tightly regulated by excluding molecules with a weight above approximately 40 kDa. Any heavier molecule needs to associate to specific nuclear transport receptors (NTRs) in order to be able to cross between those two compartments. Nucleus and cytoplasm are separated by a nuclear membrane in which numerous pores, defined as nuclear pore complexes (NPCs), are embedded, representing the passageway between the two sections. NPCs are ring-shaped supramolecular structures composed of nucleoporins that can either be structured or intrinsically disordered. Structured nucleoporins compose the ring and anchor NPCs to the nuclear membrane, while intrinsically disordered nucleoporins sit inside the central cavity of the pores and form a spaghetti-like mesh able to specifically bind NTRs that ship cargo molecules between compartments, thus providing an example of intrinsically disordered (nucleoporins) and globular (NTRs) proteins.

The interaction between the two partners occurs when FG-repeats of nucleoporins dock onto NTRs surface in specific hydrophobic pockets. Interestingly, FG-nucleoporins show an extremely low level of sequence conservation except for the FG-dipeptides, which are highly conserved and present in multiple copies along the sequence.

A still poorly understood aspect of nucleo-cytoplasmic transport is the mechanism through which nucleoporins bind NTRs, considering the minimalistic nature of their binding moieties (simply FG-dipeptides) and the timescales involved. Indeed, NPC trespassing by NTRs only occurs on the ms timescale but is thought to feature the formation and breakage of many hundreds of interactions between NTRs and nucleoporins.

To answer these questions, we have investigated the interaction of Nucleoporin 153 (Nup153) and the nuclear transport receptor Importinβ by means of computer simulations strictly interfaced with experiments (*Figure 35*).

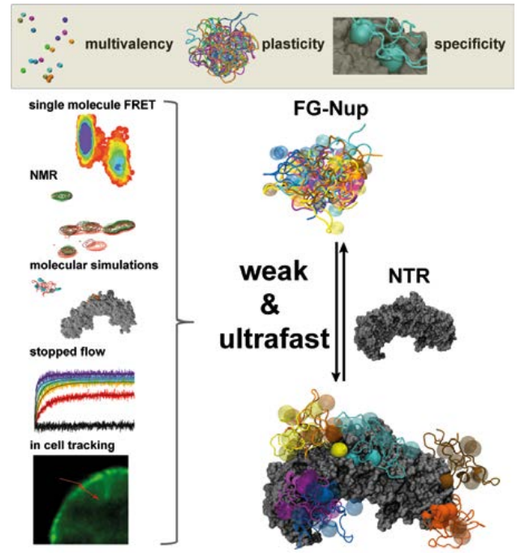


Figure 35: Schematic representation of the approaches adopted to obtain structural, thermodynamic, and kinetic information regarding the binding of Nup153 to Importinβ. Adopted from Milles, Mercadante et al. (2015): "Plasticity of an ultrafast interaction between nucleoporins and nuclear transport receptors". Cell 163:(3) 734-745.

We found that the incubation of a fragment of Nup153 with Importinβ yields numerous and specific binding events in molecular dynamics simulations on the ns timescale. Interestingly, such associations did not follow the induced fit or conformational selection mechanisms described so far. In an induced fit mechanism, the formation of some secondary structure elements is observed upon the formation of the complex, and this was never observed after interaction between Nup153 and Importinβ. On the other hand, in a conformational selection mechanism, some conformations are more prone than others to bind, though the binding between Nup153 and Importinβ was reproduced choosing completely different conformations of the intrinsically disordered nucleoporins. This shows this binding to be different from the behavior of IDPs binding globular proteins as described so far. The binding was indeed found to be solely mediated by the FG-dipeptides without flanking regions taking part in the docking of the partners. It is thus a minimalistic mode of binding involving the shortest-ever recorded binding moieties (a phenylalanine and a glycine residue). It appears to be clear that the minimalistic binding observed is functional in the fast association observable in MD simulations. Brownian Dynamics simulations helped to retrieve the association rates describing the complex formation; the simulations revealed that the association between the partners is among the fastest ever recorded ($k_{on} = 10^8$ M^{-1} s^{-1}) and is energetically driven by the desolvation of the hydrophobic FG-repeats upon binding Importinβ. Subsequently, stopped-flow experiments confirmed the values retrieved in the simulations. To fully resolve the enigma of fast-trespassing NPC, NMR experiments revealed a thermodynamically weak binding in which the FG-repeats bind in the mM concentration range indicating a high k_{off} coupled to a high k_{on}. Taken together, these observations revealed the mechanism by which NTRs could trespass NPCs on so fast a timescale. High association and dissociation are ensured by minimalistic binding moieties that only shortly and yet specifically interact with the NTRs inside the pore during trespassing. The lack of conformational selection in characterizing the binding between the partners is fully in line with the mechanistic nature of the nuclear transport, in which NTRs encounter a series of nucleoporins in a myriad of different conformations. Here, the marked presence of FG-dipeptides on the nucleoporins ensures binding without the need for conformational re-arrangements, which would slow down the binding process. Overall, these findings identify both a new conformationally independent binding mechanism in IDPs and a plausible way for the transport of molecules into and out of the nucleus in eukaryotes.

The von Willebrand Factor (VWF): disentangling blood coagulation with a computer

Dr. Camilo Aponte-Santamaria
Dr. Katra Kolšek

VWF is a giant mechanosensitive protein that plays a key-role in primary hemostasis. It is also called the ‘molecular glue’ of the blood. Too much or too active glue results in thrombosis, while too little or too inactive glue results in bleeding disorders. Triggered by the shear of flowing blood, it cross-links the extra-cellular matrix and blood platelets at sites of vascular injury. Using a multi-disciplinary approach combining simulations and experiments, we have significantly contributed to the functional understanding of a number of VWF domains and to the comprehension of VWF genetic disorders with a view to developing diagnostic tools.

VWF autoinhibition by force-sensitive protein-protein interactions

We describe a new force-sensory mechanism for VWF-platelet binding that addresses the inactivation of VWF by shielding its adhesion sites, combining molecular dynamics (MD) simulations, atomic force microscopy (AFM), and microfluidic experiments. Our simulations demonstrate that the VWF A2 domain targets a specific region in the VWF A1 domain corresponding to the binding site of the platelet glycoprotein Ibα (GPIbα) receptor, thereby causing its blockage. This implies autoinhibition of the VWF for the binding of platelets mediated by the A1-A2 protein-protein interaction *(Figure 36)*. A stretching force dissociates the A1-A2 complex before the unfolding of A2, thus ensuring VWF platelet-binding activation before unfolding-mediated proteolytic cleavage takes place *(Figure 37, inset)*. Microfluidic experiments with an A2-deletion VWF mutant resulted in increased platelet binding, corroborating the key autoinhibitory role of the A2 domain within VWF multimers *(Figure 37)*. Overall, autoinhibition of VWF mediated by force-dependent interdomain interactions provides the molecular basis for the shear-sensitive growth of VWF-platelet aggregates and may be similarly involved in shear-induced VWF self-aggregation and other force-sensing functions in hemostasis [Aponte-Santamaria et al., 2015].

Figure 36

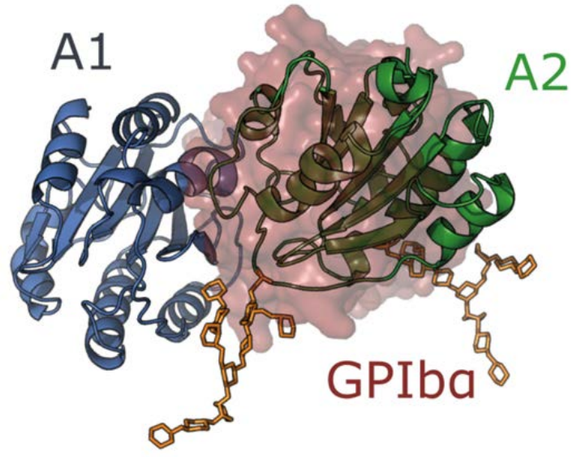


Figure 37

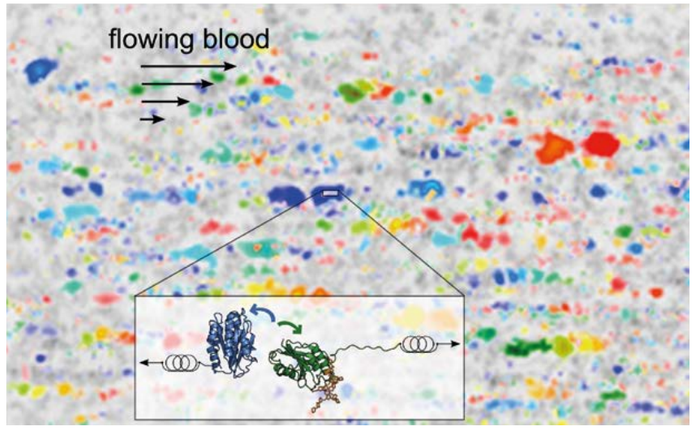
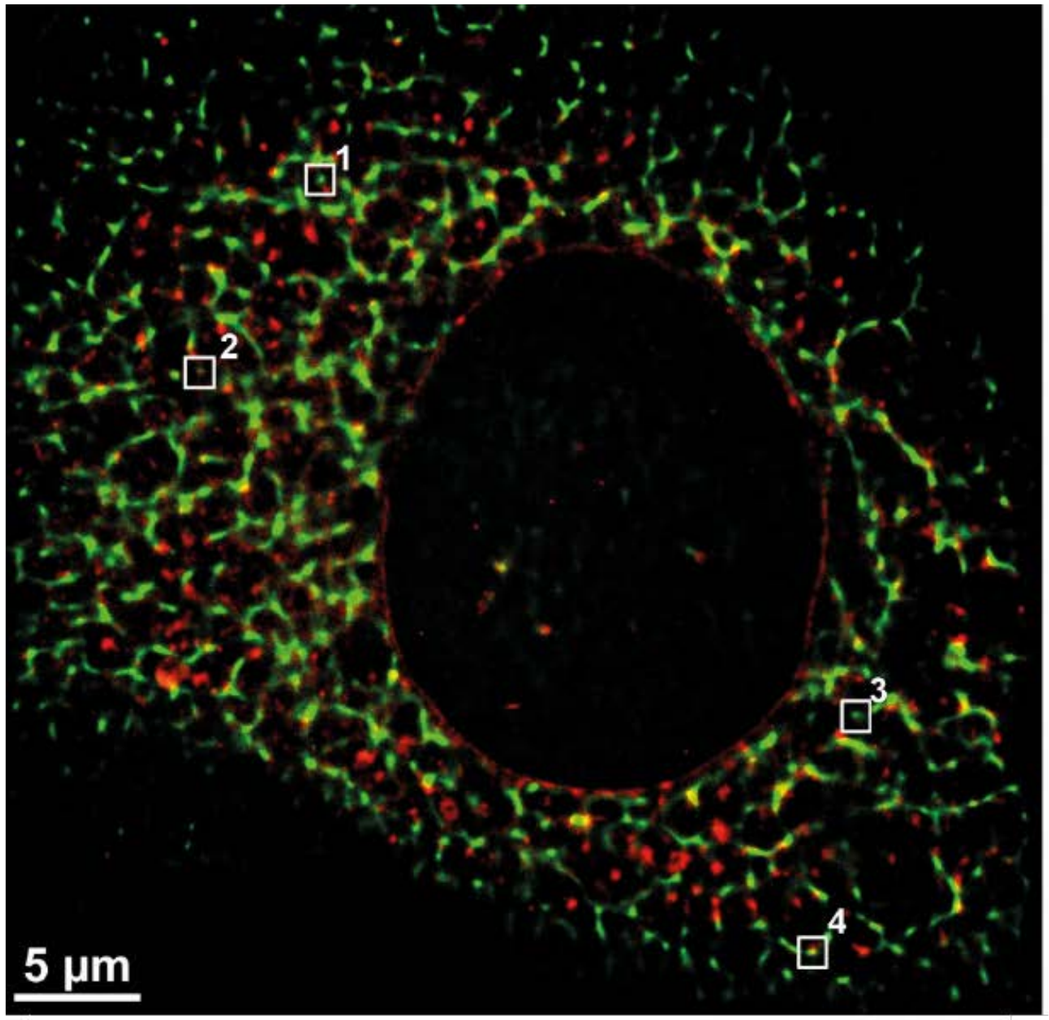


Figure 36/37: Mechanosensitive protein-protein interactions mediate von Willebrand factor (VWF) activation. (36) The VWF A2 domain (green) targets a specific region in the VWF A1 domain (blue) corresponding to the binding site of the platelet glycoprotein Ibα (GPIbα) receptor (red), thereby causing its blockage. (37) Snapshot of blood flowing through microchannels in microfluidic experiments.

Aggregates composed of VWF and platelets rolling through microfluidic channels are highlighted in color. The inset shows the dissociation of the VWF A1-A2 complex induced by an elongational force produced by the shear of blood. Upon dissociation, the A1 domain is ready to bind GPIbα, thereby triggering the formation of rolling aggregates during primary hemostasis.

Molecular mechanism of VWF dimerization

As a crucial activation step, VWF dimerizes in the endoplasmic reticulum by disulphide bond-formation between two C-terminal CK domains. The molecular mechanism of dimerization and the catalytic enzyme involved in disulphide bond-formation have been unknown till now. Our collaborators discovered that PDI, an essential folding protein in the endoplasmic reticulum, binds tightly to VWF with K_D in a nanomolar range and co-localizes with VWF, as made visible by high-resolution microscopy (*Figure 38, top*). To shed more light on this observation, we computationally elucidated the step-by-step mechanism of disulfide bond- formation during the dimerization process by MD and protein-protein docking (*Figure 38, bottom*). CK dimer is formed by three intermolecular disulphide bonds forming a very strong knot impossible to cleave by the harsh environment of flowing blood. PDI catalyzes VWF dimerization by first forming two of the three disulphide bonds. In the last stage, after the conformational rearrangement of terminal 'lids' in the CK domain, the third bond is formed, apparently to protect the first two bonds from reduction. **Read more: S. Lippok, K. Kolsek, A. Loef et al, Blood. DOI: http://dx.doi.org/10.1182/blood-2015-04-641902**



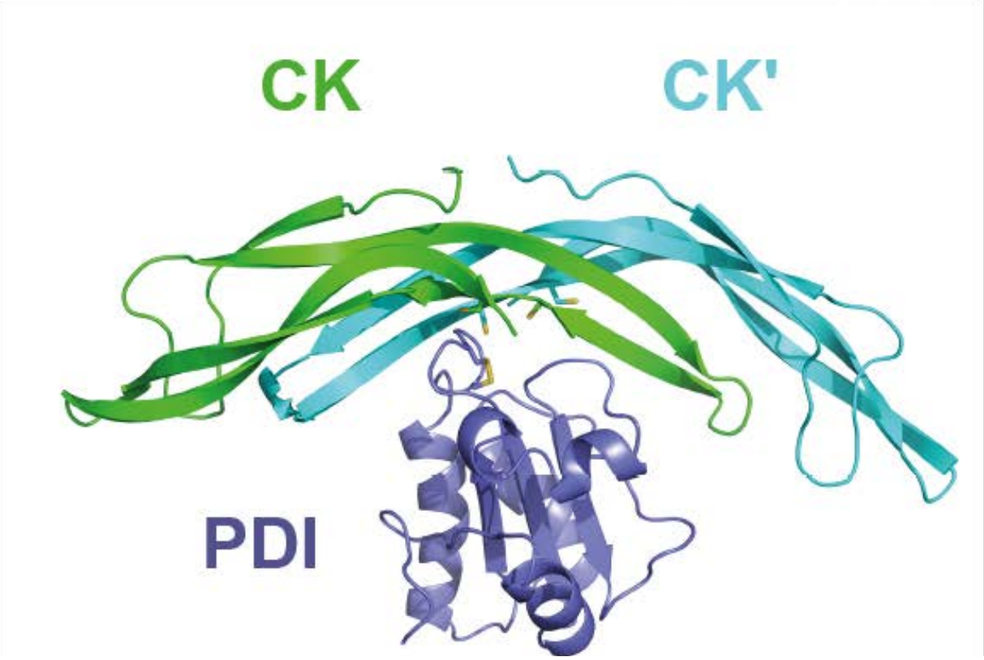


Figure 38: Mechanism of PDI dimerization in CK domain of VWF. (top) PDI and VWF complexes detected by two-color STORM microscopy. VWF (green), PDI (red), and co-localization (yellow) can be observed. (bottom) The enzyme PDI binds to two CK monomers to form a bridge between them. The figure has been adopted from (S. Lippok et al, Blood. DOI: http://dx.doi.org/10.1182/blood-2015-04-641902).

2 Research

2.7 Molecular and Cellular Modeling (MCM)

2.7 Molecular and Celluar Modeling (MCM)

Molecular recognition, binding, and catalysis are fundamental processes in cell function. The ability to understand how macromolecules interact with their binding partners and participate in complex cellular networks is crucial to the prediction of macromolecular function, and to applications such as protein engineering and structure-based drug design.

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 atomic-detail molecular simulations.

This report describes the results achieved this year in three selected projects. They demonstrate the types of methods we develop to study macromolecular interactions and their application to problems in biology, biotechnology, and drug design. The projects are on: (i) prediction of protein-protein interactions and the effects of mutations, (ii) prediction of protein-surface interactions and the effects of fluorescent labels, (iii) multiscale simulation of protein dynamics.

Molekulare Erkennung, Bindung und Katalyse sind grundlegende Prozesse der Zellfunktion. Die Fähigkeit zu verstehen, wie Makromoleküle mit ihren Bindungspartnern interagieren und an komplexen zellulären Netzwerken teilnehmen, ist entscheidend für die Vorhersage von makromolekularen Funktionen und für Anwendungen wie beispielsweise Protein-Engineering und strukturbasiertes Wirkstoffdesign.

In der MCM Gruppe sind wir in erster Linie daran interessiert zu verstehen, wie Moleküle interagieren. 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 sind beispielhaft für die Art von Problemen, die wir in unseren Projekten durch die Entwicklung und Anwendung rechnerischer Methoden zur Untersuchung biomolekularer Strukturen, Dynamik, Wechselwirkungen und Reaktionen behandeln. In enger Zusammenarbeit mit Experimentatoren verwenden wir in interdisziplinären Ansätzen rechnerische Methoden aus den Bereichen der Physik-, Bio- und Chemoinformatik. Das breite Spektrum unserer Methoden reicht dabei von interaktiven 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 zu simulieren, und andererseits ihre Anwendungen in Biologie, Biotechnologie und Medikamentenforschung. Die Projekte beschäftigen sich mit (i) Vorhersage von Proteinwechselwirkungen und die Auswirkungen von Mutationen, (ii) Vorhersage der Protein-Oberflächen-Wechselwirkungen und die Auswirkungen von Fluoreszenzfarbstoffen, (iii) Multiskalensimulation der Dynamik von Proteinen.

Group leader

Prof. Dr. Rebecca Wade

Staff members

Dr. Neil Bruce
Dr. Anna Feldman-Salit *(until Feb. 2015)*
Dr. Jonathan Fuller *(until April 2015)*
Dr. Daria Kokh
Dr. Prajwal Nandekar
Dr. Joanna Panecka
Ina Pöhner
Dr. Stefan Richter
Antonia Stank

Scholarship holders

Ghulam Mustafa *(DAAD Scholar)*
Musa Özboyaci *(HITS Scholarship)*
Mehmet Öztürk *(HITS Scholarship)*
Xiaofeng Yu *(HITS Scholarship, until April 2015)*
Gaurav Ganotra *(HITS Scholarship)*

Visiting scientists

Wiktoria Giedroyc-Piasecka *(June–Sep. 2015)*
E.R. Azhagiya Singam *(Oct.–Dec. 2015)*

Students

Max Horn
Martin Reinhardt *(June–Aug. 2015)*
Sören von Bülow *(from Sept. 2015)*
Max Waldhauer *(April–July 2015)*
Talia Zeppelin *(Erasmus student, until Jan 2015)*

General news

Amongst this year's new group members and visitors we welcomed Dr. Joanna Panecka from Warsaw to work on the NMTrypI (New Medicines for Trypanosomatidic Infections) project. Dr. Prajwal Nandekar returned to HITS for a postdoc, having completed his doctoral studies at the National Institute for Pharmaceutical Education and Research (NIPER), India, and as a DAAD Sandwich Scholar in the MCM group. Two graduate students visited the group for three-month stays: Wiktoria Giedroyc-Piasecka from Wroclaw University of Technology and Singam Azhagiya from the Central Leather Research Institute, Chennai, India. Three Master's students, including Talia Zeppelin, an Erasmus student studying Medicinal Chemistry at Aarhus University in Denmark, did internships in the group. Gaurav Ganotra completed his thesis project on methodology for studying drug-binding kinetics computationally for his Master's degree in Life Science Informatics from Bonn University; he is staying on to do his doctoral studies here. Max Horn and Max Waldhauer successfully carried out their thesis projects for their Bachelor's degrees in Molecular Biotechnology at Heidelberg University. Dr. Xiaofeng Yu and Dr. Michael Berinski both successfully completed their doctoral studies and moved on to new positions in Germany and the U.K., respectively. Also leaving were Dr. Anna Feldman-Salit – returning to Heidelberg University – and Dr. Jonathan Fuller to a position in Switzerland.

The Virtual Liver Network project, supported by the German Federal Ministry for Education and Research (BMBF), finished this year. In this project, we worked on the development of structural bioinformatics and molecular simulation tools for systems biology applications, and we applied these methods to study cross-talk between biochemical pathways, as well as protein-protein and protein-membrane interactions relevant to liver cell interactions and endocytosis. Software tool development included the LigDig webserver [Fuller, 2015], and the introduction of new features in our Brownian dynamics simulation software for macromolecules, SDA [Martinez, 2015] and the webSDA webserver [Yu, 2015b], as well as collaborative work with the SDBV group at HITS on the extraction of kinetic parameters from SABIO-RK.

We also completed an Alexander von Humboldt Foundation Research Group Linkage Programme project on "Plant enzymes from metallopeptidase families M20 and M49" carried out with Dr. Sanja Tomic at the Institute Rudjer Boskovic, Zagreb, Croatia, and Prof. Jutta Ludwig-Müller at the Technical University, Dresden. During this project, two graduate students from Zagreb, Antonija Tomic and Mario Gundic, visited us for research stays in which we collaborated on the coarse-grained simulation of large-scale conformational changes of peptidase enzymes. In part of this work [Tomic, 2015], the application of range of molecular dynamics simulation techniques revealed the interplay between substrate binding and the dynamic equilibrium between open, semi-, and fully closed enzyme states of the human dipeptidyl peptidase III.

Prediction of protein-protein interactions and the effects of mutations

We apply a range of computational approaches, including methods that we develop, to predict interactions between protein molecules and the effects of point mutations on these interactions. Applications this year included:

- Protein modeling to understand how a point mutation discovered by Gudrun Rappold and colleagues (Department of Human Molecular Genetics, Heidelberg University Hospital) to occur in members of a family in Pakistan displaying intellectual disability and epilepsy might exert its effect at the molecular level [Rafiullah, 2015]. The modeling showed that the mutation may impair protein-protein interactions, *(see Figure 39)*.

- Modeling of mutants and binding free energy calculations for a protein complex involved in cell division in yeast in a collaboration with the group of Elmar Schiebel (ZMBH, Heidelberg) [Seybold, 2015]. This study revealed the function of the Kar1 and Cdc31 proteins, crucial components of the yeast spindle pole body (SPB), in SPB duplication. The duplication of the SPB is an essential part of the chromosome segregation process in every cell cycle. At the beginning of the cell division process and SPB duplication, a multiprotein structure called the half-bridge juts out from the SPB and elongates into a bridge structure. The binding free energy calculations for a set of Cdc31 mutants indicated that the Cdc31 proteins stabilize the structure of the SPB bridge to promote its duplication, *(see Figure 40, next page)* This mechanism was supported by subsequent yeast growth measurements for the mutants. The SPB is the functional equivalent of the mammalian centrosome, and thus these results improve our basic understanding of cell division processes, which are of fundamental importance in cancer research.

- Predicting co-chaperone complexes important for efficiently solubilizing stress-induced protein aggregates in humans using our Brownian dynamics-based docking approach [Nillegoda, 2015]. In their native state, proteins are folded. This correctly folded state is at constant risk from external and internal influences. Damaged proteins lose their structure, unfold, and then tend to clump together. If such aggregates form, they can damage the cells and even cause the cells to die. In this study, carried out with the group of Bernd Bukau (ZMBH-DKFZ Alliance, Heidelberg University), the problem addressed was how healthy metazoan (animal) cells eliminate intracellular protein aggregates when they lack the heat-shock protein disaggregase HSP100 of non-metazoan HSP70-dependent protein disaggregation systems. It was discovered that synergetic cooperation be-

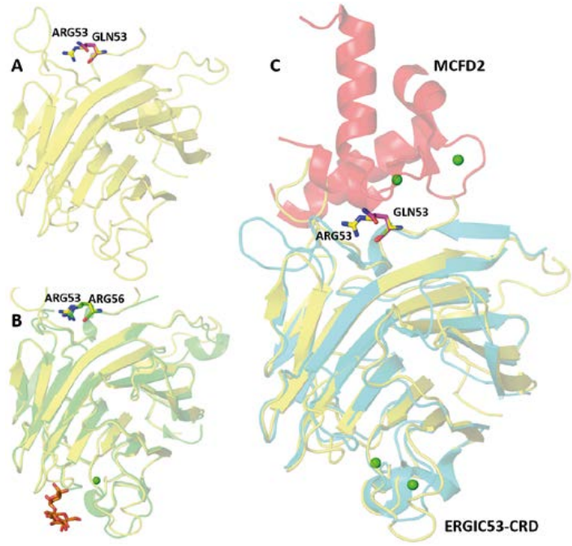


Figure 39: Comparative model of the structure of lectin mannose-binding 2-like protein (LMAN2L) showing (A) the position of the mutation associated with intellectual disability (wild-type: Arg53 (yellow), mutant: Gln53 (pink). (B) Superposition of LMAN2L (yellow) on vesicular integral-membrane protein 36 (Vip36) (green) in complex with a mannose molecule (orange sticks) showing that the mutation site is far from the sugar-binding site. (C) Superposition of LMAN2L (yellow) on the endoplasmic reticulum Golgi intermediate compartment 53-carbohydrate recognition domain (ERGIC53-CRD) (cyan), which is bound to the protein multiple coagulation factor deficiency 2MCFD2 (red) to form the cargo receptor. The mutation site is at the protein-protein interface, suggesting that it may impair the protein interactions of LMAN2L [Rafiullah R, Aslamkhan M, Paramasivam N, Thiel C, Mustafa G, Wiemann S, Schlesner M, Wade RC, Rappold GA, Berkel S. Homozygous missense mutation in the LMAN2L gene segregates with intellectual disability in a large consanguineous Pakistani family. Journal of Medical Genetics (2016) 53:138-144].

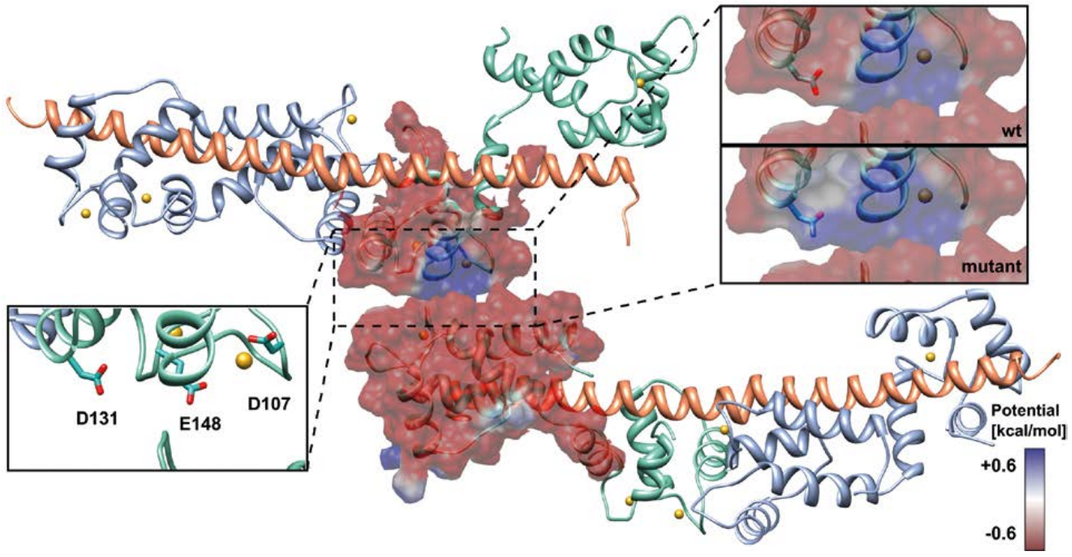


Figure 40: A model of anti-parallel Sfi1-Cdc31 protein interactions in the spindle pole body bridge (Sfi1, orange, Cdc31, green and blue ribbons). Dominant mutants of Cdc31 suppress the essential requirement of Kar1. Modeling of the mutants and binding free energy calculations show how the interfacial mutations stabilize the Cdc31 interactions by increasing electrostatic complementarity [Seybold, 2015].

tween complexed J-proteins (HSP40 proteins) from two different classes unleashed highly efficient protein disaggregation activity in human and nematode HSP70 systems. Biochemical and biophysical experiments together with docking simulations showed that the mixed-class J-protein complexes are transient, involve complementary charged regions conserved in the J-domains and the carboxy-terminal domains of each J-protein class, and are flexible with respect to subunit composition *(see Figure 41)*. Complex formation allows J-proteins to initiate transient higher-order chaperone structures involving HSP70 and interacting nucleotide exchange factors. A network of cooperative class A and B J-protein interactions therefore provides powerful, flexible, and finely regulatable disaggregase activity. This finding may have implications for the processing of harmful protein aggregates in neurodegenerative diseases and ageing processes.

- Predicting how a protein central to anti-tumor defense binds to and inhibits an enzyme critical for glucose-dependent aerobic respiration using our Brownian dynamics-based docking methodology. Georg Gdynia, Wilfried Roth, and colleagues at the German Cancer Research Center (DKFZ, Heidelberg) found that the High Mobility Group Box 1 (HMGB1) cytokine protein eliminates cancer cells by triggering metabolic cell death. It does so by allosterically inhibiting the tetrameric form of the enzyme pyruvate kinase M2 and thereby blocking glucose-driven aerobic respiration. To understand the determinants of this inhibitory activity, we docked the two proteins and investigated the effects of metabolite binding and phosphorylation on the specificity of the interaction, *(see Figure 42)* The docking calculations showed that the HMGB1 B box domain could specifically block the allosteric regulatory binding site on pyruvate kinase M2. This was confirmed experimentally. This work establishes a link between innate tumor defense and tumor metabolism [Gdynia et al., Nature Commun., 2016, in press].

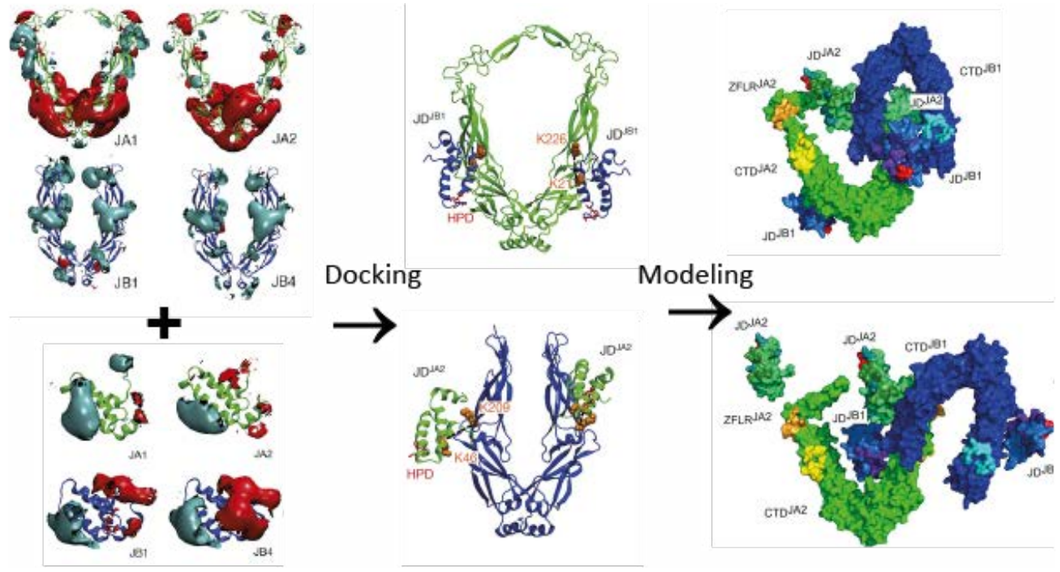
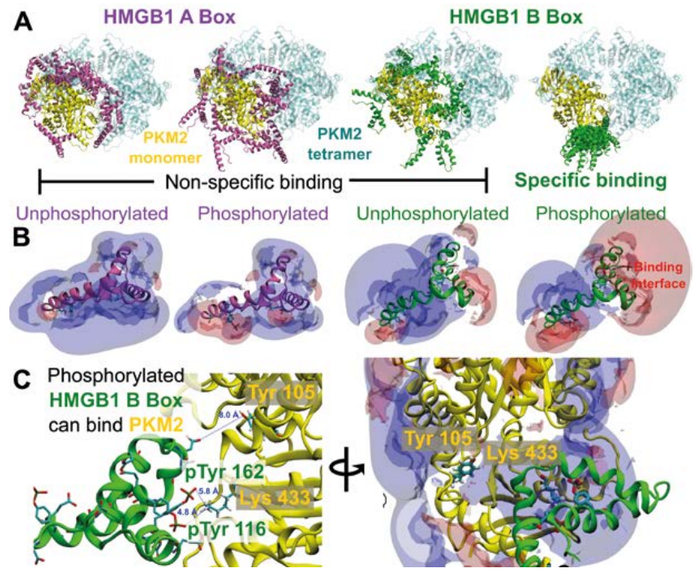


Fig. 41: Docking of class A and class B J-protein co-chaperones for efficient protein disaggregation activity in metazoans. The J-proteins consist of N- and C-terminal domains connected by a flexible linker and the C-terminal domains themselves display high flexibility. We therefore docked the N-terminal J-domains (bottom left) to the C-terminal domains (top left) of the J-proteins for different potential combinations of J-proteins and then derived putative models of J-protein dimers (right). Transient binding occurred with varying degrees of specificity through complementary charged regions. Electrostatic isopotential contours are shown (cyan: positive, red: negative). [Nillegoda, 2015].

Figure 42: HMGB1 is an allosteric inhibitor of tetrameric pyruvate kinase M2. (A) The docking calculations to tetrameric pyruvate kinase M2 (one monomer in yellow, the other three in cyan) were performed for the HMGB1 A box (magenta) and B box (green) with and without two phosphorylated tyrosine residues. (B) Electrostatic isopotential contours (positive: blue; negative: red) of the corresponding HMGB1 A and B boxes. (C) Close-up of the specific complex (right in A) obtained with a phosphorylated HMGB1 B box showing the positive potential of pyruvate kinase M2 (yellow) that is complementary to the negative patch on HMBG1 (green). Further calculations showed that docking was specific only in the absence of the allosteric effector and when pyruvate kinase M2 was unphosphorylated (at Tyr 105). [Gdynia et al., Nature Communications, 2016, in press].



Prediction of protein-surface interactions and the effects of fluorescent labels

Proteins interact with the surfaces of inorganic materials and metals in many biological systems, and these interactions are critical to applications in biotechnology and medicine. Our understanding of the mechanisms and determinants of these interactions at the molecular level is, however, very limited. Computational approaches can complement experiments to provide a detailed picture of protein-surface interactions, (for a review, see [Ozboyaci et al., Q. Rev. Biophys. (2016), 49, e4]). We are developing procedures based on Brownian dynamics and molecular dynamics simulation to simulate protein adsorption to inorganic surfaces. We have applied these to reveal the steps in the recognition and adsorption of several different types of protein to a gold surface. We are also simulating how proteins adsorb to silica and mica surfaces that are negatively charged to predict their binding orientation and how their adsorption behavior depends on environmental conditions such as ionic strength. The adsorption of proteins to charged surfaces can be studied experimentally by attaching a fluorescent label to the protein whose decrease in fluorescence intensity as it approaches the surface can be monitored. We used our SDA software to perform docking computations to evaluate the influence of label attachment on the adsorption process of a protein to a charged surface, *(see Figure 43)* [Romanowska, 2015b]. Although the modified protein had a relatively small fluorescent label attached, it was found that this changes the protein's surface charge distribution, which then disrupts the adsorption onto a charged surface. Therefore, it is important when interpreting the results of experiments with labeled proteins that both protein species, labeled and unlabeled, are explicitly accounted for. We showed that this can be done quite easily with relatively straightforward Brownian Dynamics docking and electrostatic potential calculations when there are significant electrostatic interactions between the protein and the surface. As fluorescent labels are often attached to proteins in order to monitor them experimentally, yet are frequently assumed to have little effect on the protein properties, the implications of this study extend beyond studies of protein-surface interactions.

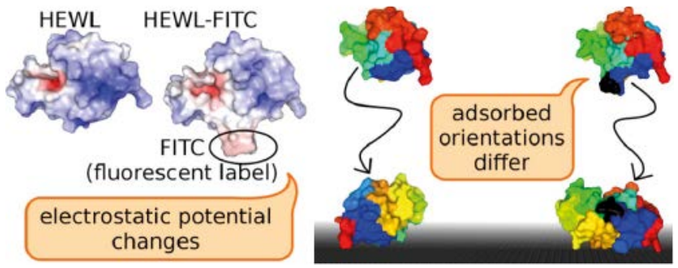


Figure 43: When the label matters: The addition of a fluorescent dye label (fluorescein isothiocyanate (FITC)) to the protein hen egg white lysozyme (HEWL) changes the electrostatic potential of the protein, which is shown mapped on the molecular surface of the protein on the left (positive: blue, negative: red). The addition of the label thus affects the orientation in which the protein adsorbs on an oppositely charged surface in computational docking calculations, as shown on the right where the protein is colored by sequence from blue at the N-terminus to red at the C-terminus, and the label is colored black. These results imply that interpretation of experiments with labelled proteins should explicitly account for the effects of the label, even if it is small compared to the protein itself. Reproduced with permission from [Romanowska et al., Nano Lett. (2015) 15:7508-7513. Copyright (2015) American Chemical Society].

Multiscale simulation approaches for studying protein dynamics

A major challenge in studying protein binding and function is the large range of temporal and spatial scales not only of the protein conformational dynamics but also of the environment, e.g. the lipid membrane in which a membrane protein is embedded. Computationally, this problem can be addressed by using a set of simulation techniques with models at different levels of resolution, where the coarser-grained models are parameterized from computations with the finer-grained models. In collaboration with Lei Liu and Dieter Heermann (Department of Physics, Heidelberg University) [Liu, 2015], such an approach was taken to simulate multi-C_2H_2 zinc-finger domain proteins, transcription factors that bind DNA. Simulations included standard atomic-detail molecular dynamics simulations of single domains and linkers, models with rigid body domains that pivot with respect to each other, and a mesoscale model with each residue represented by 3 – 4 beads, *(see Figure 44)*.

Previously, we developed the first systematic procedure to insert and simulate cytochrome P450 proteins (CYPs) in a phospholipid bilayer, *(see Figure 45, next Page)*. We combined coarse-grained and all-atom molecular dynamics (MD) simulations to achieve efficient sampling of the possible arrangements of the protein in the bilayer. This work brought about a paradigm shift towards studying microsomal CYPs as dynamic structures in their natural, lipid environment rather than solely in artificially solubilized forms. In further studies [Mustafa, 2015], we have improved this procedure and tested it on a number of human CYPs. We tested different variants of the MARTINI coarse-grained model and found that use of the standard non-polar water model with a reaction field treatment of long-range interactions in the GROMACS 5.0.4 implementation provided a suitable protocol for modeling such protein-membrane systems. The CYP-membrane models generated are being used to investigate how ligands access the buried active site of the proteins [Yu, 2015a] and how CYP proteins bind their redox partners.

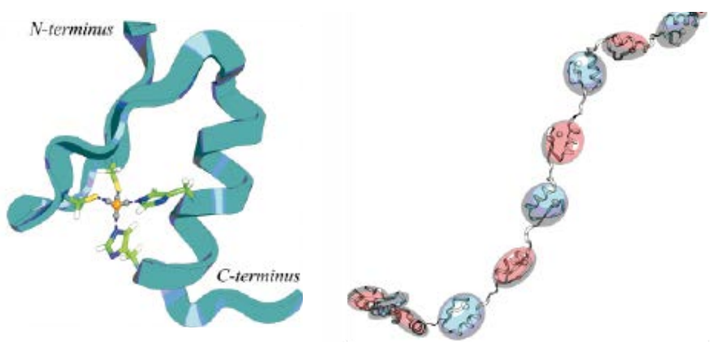
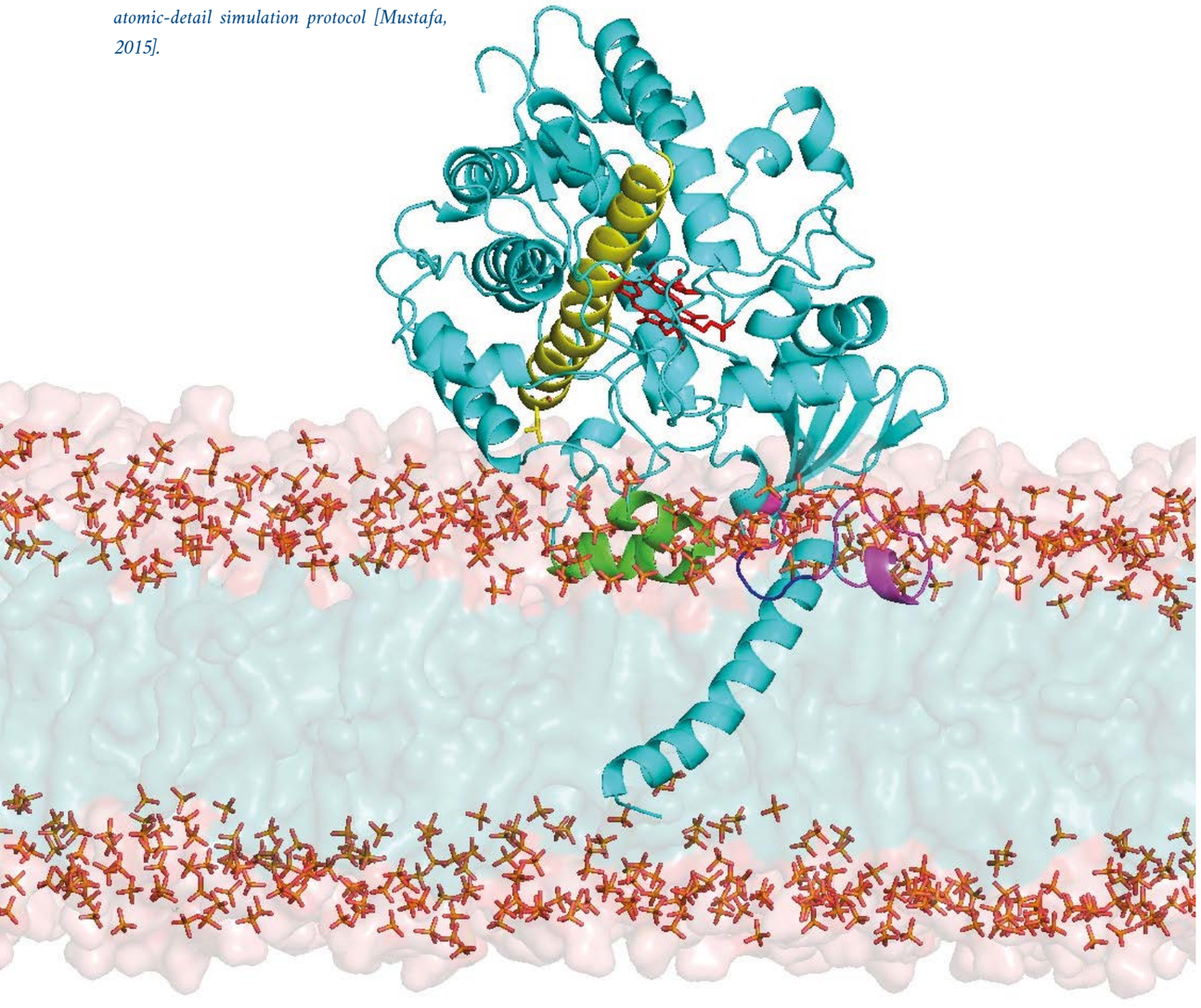


Figure 44: Multiscale modeling of multi-zinc finger domain proteins. Left: A zinc-finger domain (Zinc: orange sphere). Right: Multiple rigid zinc-finger domains connected by flexible linkers [Liu,2015].

Figure 45: Orientation of a cytochrome P450 heme protein (cartoon representation) in a phospholipid bilayer (phosphate groups in red stick representation; lipid tails in cyan) obtained by a combined coarse-grained and atomic-detail simulation protocol [Mustafa, 2015].



Another multiscale approach being pursued is the combination of Brownian and molecular dynamics simulations of protein-surface, protein-protein, and protein-small molecule association, *(see Figure 46).* In Brownian dynamics simulations, the solvent is treated as an implicit continuum solvent, whereas in molecular dynamics simulations it is usually treated in atomic detail [Bruce, 2015]. The combination of methods is being applied for docking and for association rate constant calculations.

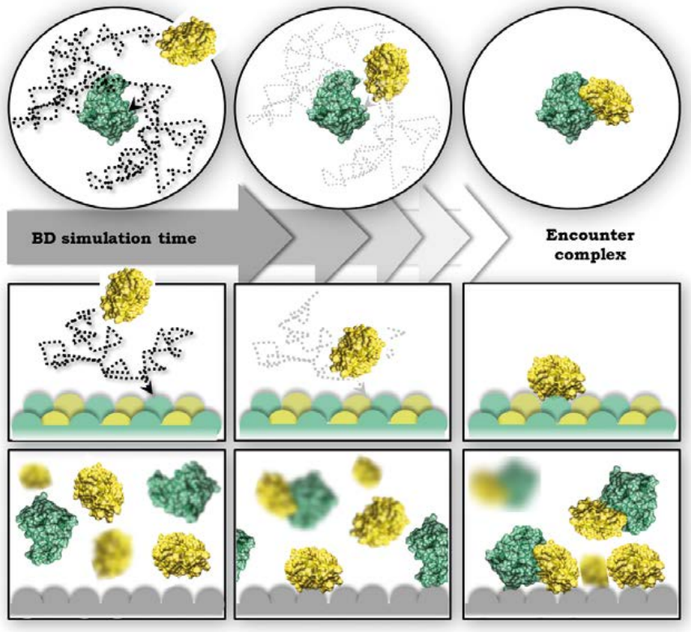
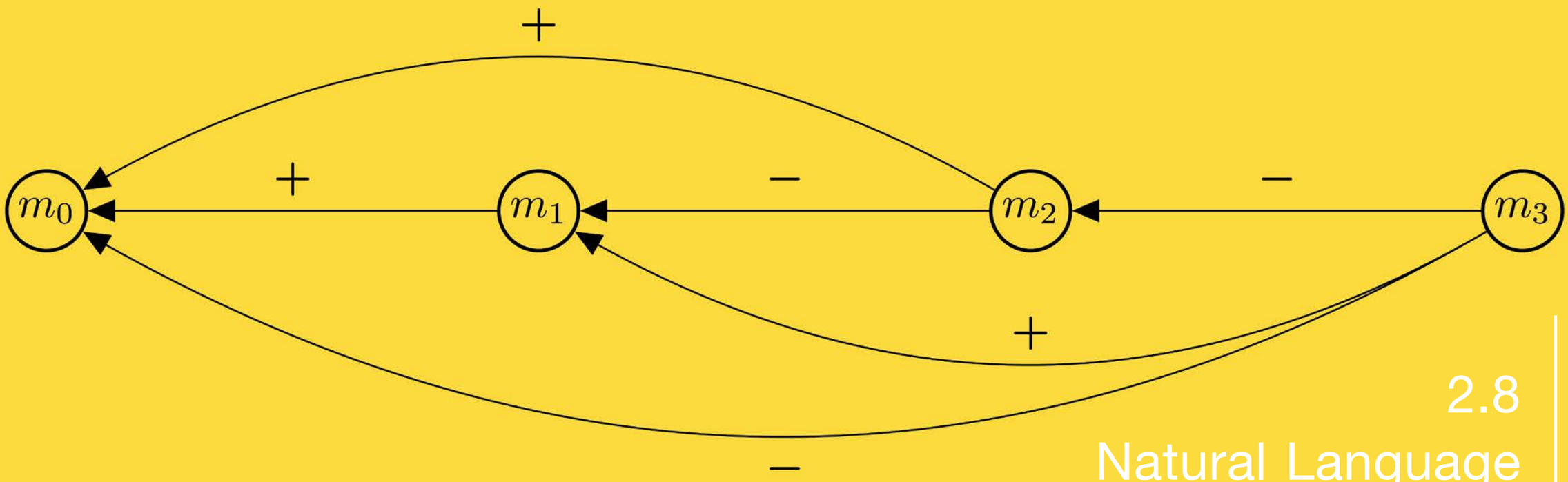


Figure 46: Systems that can be simulated by Brownian dynamics with the SDA software (http://mcm.h-its.org/sda7) to compute rate constants for association or predict diffusional encounter complexes: two macromolecules (top); a protein and a surface (middle), or many proteins diffusing in the presence of a surface (bottom) [Bruce, 2015].

2 Research



2.8 Natural Language Processing (NLP)

2.8 Natural Language Processing (NLP)

The Natural Language Processing (NLP) group develops methods, algorithms, and tools for the automatic analysis of natural language. We focus on discourse processing and related applications such as automatic summarization and readability assessment.

Instead of doing research, group leader Michael Strube spent a good part of the year planning, organizing, and running ACL-IJCNLP 2015, the 53rd Annual Meeting of the Association for Computational Linguistics held in Beijing in late July 2015. The remainder of the year he needed to recover from this experience.

Nevertheless, Michael kept the NLP group members and PhD students busy. In spring 2015 Yufang Hou submitted her PhD thesis entitled "Unrestricted Bridging Resolution," while late in the year Angela Fahrni finally defended hers. Yufang joined IBM Research in Dublin, Ireland, as a research staff member, and she continues to work there on discourse and pragmatics. Angela is a research staff member at IBM Research, Zürich, where she works on machine learning. The current PhD students continued with top-flight publications of their work at NLP and AI venues. With their work on entity linking and event extraction, Benjamin Heinzerling and Alex Judea participated with great success in shared tasks at the Text Analysis Conference.

In 2015, we also started two new projects. As of April we are a member of the DFG-funded Research Training Group AIPHES (Adaptive Preparation of Information from Heterogeneous Sources) together with the Computational Linguistics Dept. at Heidelberg University and the Computer Science Dept. at the TU Darmstadt. In July we started with the project SCAD (Scalable Author Disambiguation for Bibliographic Databases) in conjunction with the online computer science bibliography DBLP Zentralblatt MATH, which provides a similar service for mathematics. Benjamin Heinzerling joined the NLP group as a PhD student in AIPHES, Mark-Christoph Müller is a research associate in SCAD.

Die Natural Language Processing Gruppe (NLP) entwickelt Methoden, Algorithmen und Tools zur automatischen Analyse von Sprache. Wir konzentrieren uns auf die Diskursverarbeitung und verwandte Anwendungen wie automatische Zusammenfassung und Lesbarkeitsbewertung.

Anstatt zu forschen, verbrachte Gruppenleiter Michael Strube einen Großteil des Jahres mit der Planung, Organisation und Durchführung von ACL-IJCNLP 2015, der jährlich stattfindenden Konferenz der Association for Computational Linguistics Ende Juli 2015 in Beijing. Den Rest des Jahres benötigte er anschließend, um sich davon zu erholen.

Dennoch ließ Michael seine Gruppenmitglieder und Doktoranden nicht zur Ruhe kommen. Im Frühling 2015 reichte Yufang Hou ihre Dissertation mit dem Titel „Unrestricted Bridging Revolution" ein. Ende des Jahres verteidigte Angela Fahrni ihre Doktorarbeit. Yufang arbeitet jetzt bei IBM Research in Dublin, wo sie weiterhin zu den Themen Diskurs und Pragmatik forscht. Angela befasst sich mit maschinellem Lernen als wissenschaftliche Mitarbeiterin bei IBM Research in Zürich. Die derzeitigen Doktoranden der NLP Gruppe haben diese Arbeit mit erstklassigen Publikationen und auf zahlreichen Veranstaltungen fortgesetzt. Benjamin Heinzerling und Alex Judea nahmen mit ihrer gemeinsamen Arbeit über Entity Linking und Event Extraction sehr erfolgreich an einer Shared Task auf der Text Analysis Conference teil.

2015 rief die NLP Gruppe außerdem zwei neue Projekte ins Leben: Seit April sind wir Mitglied der DFG-geförderten Graduiertenkollegs AIPHES (Adaptive Preparation of Information from Heterogeneous Sources), gemeinsam mit dem Fachbereich Informatik der TU Darmstadt. Im Juli nahmen wir das Projekt SCAD (Scalable Author Disambiguation for Bibliographic Databases) auf, in Verbindung mit der Online-Fachdatenbank DBLP – Zentralblatt MATH, die einen ähnlichen Service für Mathematik anbietet. Benjamin Heinzerling ist als Doktorand im Rahmen von AIPHES Mitglied der NLP Gruppe, Mark-Christoph Müller ist wissenschaftlicher Mitarbeiter für das SCAD Projekt.

Group leader

Prof. Dr. Michael Strube

Staff members

Sebastian Martschat *(from September 2015)*
Dr. Mark-Christoph Müller *(from November 2015)*

Scholarship holders

Yufang Hou *(until March 2015)*
Alex Judea
Sebastian Martschat *(until August 2015)*
Mohsen Mesgar
Nafise Moosavi
Daraksha Parveen

Visiting scientists

Sebastian Heinzerling
(Research Training Group AIPHES Scholarship, from April 2015)
Caecilia Zirn

Students

Nicolas Bellm *(until July 2015)*
Patrick Claus *(from February to July 2015)*
Benjamin Heinzerling *(from February to March 2015)*

Coreference Resolution

This year we devised a machine-learning framework that allows for unified representation of approaches to coreference resolution [Martschat and Strube, 2015]. As a foundational natural language processing task, coreference resolution has received considerable research attention. Machine-learning models range from simple binary classification models to sophisticated structured prediction approaches.

In our analysis of popular approaches to coreference resolution, we have observed that they can be understood as structured prediction models predicting structures that are not contained in the data. To understand this better, let us consider two examples. The mention-pair approach models coreference resolution as labeling pairs of mentions as either coreferent or non-coreferent. Hence, the approach can be understood as predicting a *labeled list* of mention pairs. The antecedent-tree approach casts coreference resolution as predicting a *tree* that represents all anaphor-antecedent decisions in the document. What both outputs – labeled lists and trees – have in common is that they are not annotated in the data. Instead, the data only contains assignments of mentions to entity identifiers.

Hence these approaches tackle coreference resolution by predicting structured objects that are not annotated in the data. These arecalled *latent structures*. This perspective allows for a unified view of machinelearning approaches to coreference resolution as latent structured prediction.
In this unified view, the approaches take as an input a document and as an output a pair consisting of the predicted latent structure and the coreference relations identified in the document. Typically, the coreference relations are obtained by extracting the corresponding information from the predicted latent structure.

To represent the latent structures we settled on a graph-based approach. In particular, we employed directed labeled graphs. They constitute an expressive formalism which is intuitive and is also sufficient to express most of the popular approaches to coreference resolution.

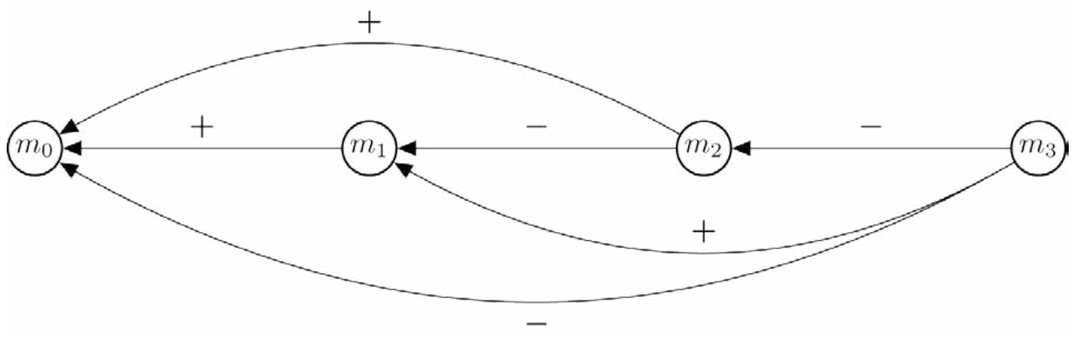
Figure 47 shows how the mention-pair approach is modeled in our framework. Each edge corresponds to a decision for a pair of mentions that is either "+" (coreferent) or "-" (non-coreferent). We learn parameter vectors with a variant of the structured perceptron algorithm.

We implemented three of the most popular models in our framework: the mention-pair model, two variants of the mention-ranking model, and antecedent trees. By representing the approaches in our framework, we can make the structural differences and similarities between the models explicit. Several of the implemented models achieve state-of-the-art performance.

We then performed a detailed analysis of the models. To do so, we employed the coreference resolution error analysis method, that we proposed last year. This method extracts recall and precision errors and represents these as pairs of mentions. By comparing the errors made by the models, we saw that the mention-ranking model improves over the mention-pair model mainly due to improved anaphoricity detection. In contrast to the mention-pair model, the mention-ranking model is able to model anaphoricity detection structurally. Moreover, we showed that the antecedent-tree model relies on the same structure as the mention-ranking model, but takes a per-document perspective, while the ranking model considers each anaphor in isolation. This broader perspective results in more cautious decisions, leading to higher precision but lower recall.

Our current work concentrates on how to extend the framework to entity-based approaches.

Figure 47: Modeling of the mention-pair approach.



Detecting Non-Coreferent Mentions

Coreference resolution means clustering referring expressions in a text so that each resulting cluster represents an entity. It is a very challenging task in natural language processing and a great deal remains to be done.

Text spans referring to an entity are called mentions. Mentions are the primary objects in a coreference resolution system. Like most of the work on coreference resolution, we only consider mentions that are noun phrases. However, not all noun phrases are mentions. With respect to the coreference resolution problem, different noun phrases have different functions. A noun phrase may not refer to any entity at all and only be used in a sentence to fill a syntactic position. The pronoun *it* in the sentence *it is raining* is an example of a non-referential noun phrase. Noun phrases which do refer to an entity (mentions) can be further divided into two categories: mentions referring to entities that only appear once in the discourse (i.e. singletons) and mentions realizing entities that have been referred to more than once in the text (i.e. coreferent mentions). We refer to both singletons and non-referential phrases as non-coreferent mentions.

A large number of mentions that appear in a text are non-coreferent. For instance, in the OntoNotes English development set, more than 80% of mentions are singletons. Pruning this huge number of non-coreferent mentions will result in a smaller and more balanced search space for coreference resolution. Therefore search space pruning is a promising way of improving coreference resolution.

The latent-ranking model is the best-performing model for coreference resolution to date [Martschat and Strube, 2015]. If we use gold mentions, the latent-ranking model by [Martschat and Strube 2015] achieves an overall score of 80% on the CoNLL 2012 English test set. This result shows that once we have the ideal pruned search space, the ranking model with the current set of features is reasonably capable of finding corresponding entities of mentions. The substantial results gap (17%) between gold mentions and system mentions implies that search space pruning is a promising direction for further improvements in coreference resolution.

With all this in mind, we are now designing and implementing a singleton detection model. We have incorporated our model into various coreference resolution systems, including the Stanford rule-based coreference resolver and our in-house coreference resolver, i.e. Cort [Martschat and Strube, 2015]. The singleton detection model improves the Stanford results by 2 percent. Similarly, it improves the results produced by the pairwise and ranking coreference models of Cort by 1 and 0.5 percent respectively.

An important aspect of our model is that it only uses shallow features for detecting non-coreferent and coreferent mentions. However, this shallow approach significantly outperforms a state-of-the-art singleton detection model heavily based on carefully designed linguistic features.

Entity Linking

When given a text, computers do not have any conception of the things mentioned in it. This leads to confusion when different things are referred to in the same way. For example, *Obama* can, among other things, refer to the current U.S. president or a town in Japan. Furthermore, a given entity can be referred to in various ways. For example, *the current U.S. president*, *Barack Hussein Obama*, and *the 44th U.S. president* all refer to the same person. Entity linking tackles this problem by automatically identifying mentions of *entities*, such as persons, organizations, or locations, and *linking* those mentions to their corresponding entries in a knowledge base. The knowledge base provides rich, computer-readable information about these entities, which can then be used in downstream NLP applications such as search, question answering, or automatic summarization.

Visual Error Analysis for Entity Linking

The NLP group has created an error analysis tool for entity linking. Evaluation of entity-linking systems is usually done in terms of metrics that measure the overall error. They are well suited for comparing systems, but provide little feedback to developers of entity-linking systems looking to understand their systems better. Our error analysis tool visualizes various types of errors in a concise fashion and enables developers to quickly identify the system components that contribute most to the overall error and should accordingly be improved.

HITS at TAC KBP 2015

The NLP group participated in the English entity discovery and linking (EDL) track of the Knowledge Base Population (KBP) shared task at the Text Analysis Conference (TAC) 2015. In the EDL track, systems are required to discover and link mentions of five entity types, namely persons, locations, organizations, geo-political entities, and facilities. In addition to linking, systems also need to recognize whether an entity mentioned in a text is not contained in the knowledge base (NIL, for Not In Lexicon) and cluster all mentions of that entity across documents in the test corpus. The test corpus consists of texts from two genres: one formal (newswire) and one informal (discussion forum).

In previous years, the NLP group participated with a sophisticated system that tackles many of the required tasks both globally and jointly. This is achieved by formulating dependencies and interactions between the tasks in a probabilistic modeling language and then searching a possibly very large space to find a mathematically optimal solution. This approach was motivated by the shortcomings of prior pipeline approaches performing the tasks in sequence. Though simpler and faster, pipelines suffer from error propagation and the fact that information learned at a later stage cannot be used in earlier stages.

Aiming to get the best of both worlds, we created a modular, hybrid system that first takes easy decisions in a pipeline-like fashion but also allows for interaction between earlier and later stages by repeatedly running some of the components.

The tougher problems still remaining are then solved by our last year's system, which can now benefit from additional information and a much smaller search space, since the easier problems have already been solved by the pipeline-like components.

The eight participants, including teams from IBM, Carnegie Mellon University, and Rensselaer Polytechnic Institute, were compared in three categories. The HITS team came 7th in entity typing (we did not optimize for this category), first in linking and NIL classification,and second in NIL clustering.

Event Extraction

Event Extraction is an information extraction task. Given a collection of texts, find mentions of interesting events and the times, places, and entities playing a role in those events. Consider the following example.

In *Baghdad*, a *cameraman* **died** when an *American tank* **fired** on the *Palestine Hotel.*

Two events can be found in this example. A DIE event is indicated by the word **died**. *Cameraman* is the victim in this DIE event, *Baghdad* the place, and *American tank* the instrument. An ATTACK event is indicated by the word **fired**. Here, *cameraman* and *Palestine hotel* are the targets, and *Baghdad* and *American tank* are again the place and the instrument.

Tackling the full task involves finding event mentions, entity mentions, and temporal expressions. It also involves identifying the roles played by the mentioned entities in the events.

The task is even more complicated. Roles don't always have to be filled, e.g. there is no mention of the time at which the DIE and ATTACK events took place. Entities may play roles in different events, e.g. *cameraman* as the victim of DIE and the target of ATTACK. Also, multiple entities may play the same role in one event, e.g. *cameraman* and *Palestine hotel* as targets in the ATTACK event.

The Event Extraction setting we use was developed for the Automatic Content Extraction conference in 2005 (ACE). 33 event types were defined and organized in 8 broad categories, e.g., CONFLICT and JUSTICE. ACE also defined the semantic types of entity that can play roles in events, e.g., persons, locations, and facilities. Over 500 documents were annotated for entity and event mentions.

We started from the observation that Event Extraction is similar to another well-known task, Frame-semantic Parsing, where the goal is to find semantic frames in texts. Frame semantics is an influential theory of meaning. Although the two tasks aim to find different things, the structure of Event Extraction and Frame-semantic Parsing is identical. A sequence of words indicates frames/events, which in turn have roles that may be filled by various mentions. From a computational point of view, the two tasks differ only in the features used by the statistical models. Based on this observation, we decided to retrain a state-of-the-art frame-semantic parser, SEMAFOR, to extract events.

SEMAFOR combines two statistical models to perform the actual parsing. The first model predicts whether a sequence of words will trigger an event or not. We had to change this model slightly because frame semantics has no notion of potential triggers. A verb like 'die,' for example, is associated with different frames but always triggers one of them. In event extraction, the occurrence of 'die' does not guarantee that it will also trigger an event of interest. We had to introduce a 'null event' to enable the trigger model to return a negative result. The second model predicts the roles of the recognized events that are filled, i.e., given an event like ATTACK and a role like target, it predicts whether there is a non-empty sequence of words playing that role in that event. In terms of Event Extraction, this model has two major weaknesses. First, there cannot be multiple role fillers. In our example, the target role of ATTACK would either be filled by *cameraman* or by *Palestine hotel*. Second, the way role-filler candidates are computed is not suited to Event Extraction. With various modifications its coverage is no higher than just under 80%. We were forced to use gold entity mentions in our first experiments.

By retraining SEMAFOR, we have come close to the state of the art in Event Extraction. The next step will be to change the way role-filler candidates are computed and to enable the argument model to predict multiple role-fillers.

Assessing readability

Readability refers to the ease with which readers understand texts. The readability assessment task quantifies text understanding difficulty. In previous research, many features have been used to assess readability. These range from simple features such as sentence- and word-length to complex semantic features like coherence. In a coherent text, each sentence is connected to other sentences. These connections between sentences make the text easier to understand.

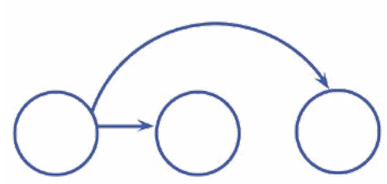
We model the connections between sentences in a text and measure the coherence of the text in terms of these connections. We use the readability assessment task to test our coherence model. We use a dataset containing articles from the Wall Street Journal corpus. A human readability score from 1 to 5 is assigned to every article in the dataset. These human scores enable us to measure the power of the coherence features.

We use the entity-graph coherence model to represent texts. The entity graph captures the distribution of entities over sentences by means of a bipartite graph consisting of two sets of nodes: sentence nodes and entity nodes.

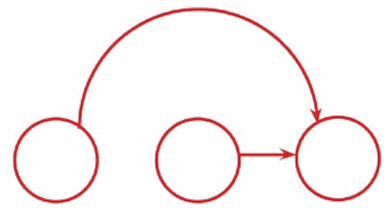
To put it simply, we can use the one-mode projection graph of the entity graph to obtain a graph whose nodes only represent sentences and edges to connect sentence nodes sharing at least one entity node in the entity graph. We represent every text by its one-mode projection graph and then use this graph to induce coherence features.

We propose using the frequency of all occurring subgraphs in a one-mode projection graph as connectivity features of that graph. In the texts, these features would capture the frequency of different coherence patterns modeled by the subgraphs. All 3-node subgraphs are defined as the smallest subgraph that can capture coherence patterns.
Our interpretation of these 3-node subgraphs is as follows, *(see Figure 48)*:

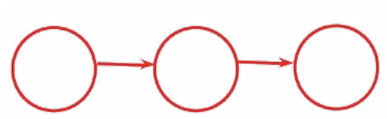
Figure 48 : Interpretation of 3-node subgraphs.



sg1



sg2



sg3

sg1 demonstrates that some entities are mentioned in one sentence and the subsequent sentences are about these entities.

sg2 shows that two unconnected sentences are related to each other by the following sentence.

sg3 reminds us of the linguistic phenomenon "shift in the focus of attention". The reader's focus of attention is on the entity that connects the current sentence to the preceding sentences. The absence of a connection between the first and the last sentence nodes indicates that the connecting entity changes between sentences, i.e. the topic of the text is shifting.

We compute the Pearson correlation coefficient between the frequency of these 3-node subgraphs and the human readability scores. *sg1* is positively and *sg2* negatively correlated with human readability scores. The main difference between these two subgraphs is the order of sentences. This confirms that the right order of sentences leads to a more coherent and readable text.

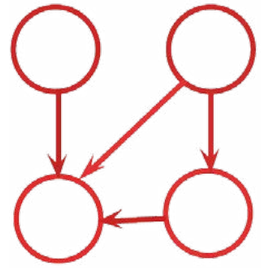
sg3 is negatively correlated with the human readability scores. Too many shifts in the topic make the text hard to read.

We also compute the frequency of the 4-node subgraphs and then check the correlation between the frequency of these subgraphs and the human readability scores, *(see Figure 49)*.
First, we find a stronger correlation between the frequency of 4-node subgraphs and readability than between the frequency of 3-node subgraphs and readability. Second, among the subgraphs with 4 edges, subgraph (a) is the most negatively correlated subgraph, and (b) is the most positively correlated one. This shows that coherence is not only a question of the number of connections but also of the connectivity style of sentence nodes in the graphs.

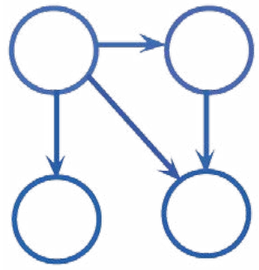
We also use these coherence features to rank pairs of texts based on the readability level of texts. Accuracy comparison shows that when the frequency of 4-node subgraphs is used as a coherence feature, the performance of the system is higher than that of the system using the frequency of 3-node subgraphs. The system with 4-node subgraphs outperforms the state-of-the-art system on this dataset by by 5%.

These coherence features can be used in different NLP applications. Text summarizers can use the patterns to extract not only important sentences but also well-connected sentences from the text to put in the final summary. In this way, the final summary would be more coherent and more readable.

Figure 49: Correlation of 4-node subgraphs.



(a) Negatively correlated



(b) Positively correlated

Automatic Summarization

In Natural Language Processing, research is a cumulative activity. The work of downstream researchers depends on access to upstream discoveries. The main goal in summarizing scientific articles is to generate a summary of a long article that satisfies three requirements:

- *Importance:*
  The summary should contain important information from the article.

- *Non-redundancy:*
  The sentences in the summary should contain non-redundant information. Information should be diverse among the sentences of a summary.

- *Coherence:*
  Although the sentences of a summary should carry important and diverse information from the given article, they should be connected to one another in such a way that the summary is coherent and easy to read.

In our work we focus on incorporating coherent patterns into the final summary. We use a graph-based representation for the input documents. The graph is the bipartite *entity graph* introduced by [Guinaudeau and Strube: Graph-based local coherence modeling. In Proc. of the 51st Ann. Meeting of the ACL (Vol 1: Long Papers), Sofia, Bulgaria, 4 – 9 August 2013, pp. 93 – 103.] for evaluating local coherence. Then we maximize importance, non-redundancy, and pattern-based coherence.

We calculate the importance of sentences using the Hubs and Authorities algorithm [Kleinberg, 1999]. We believe that if the final summary gives unique information in every sentence, then the summary is non-redundant. We incorporate entity information for non-redundancy. The last criterion for establishing coherence is to assume that the final summary will be coherent if the connection structure across the sentences of the summary is similar to coherence patterns frequently encountered in the corpus.

We applied our approach to the *PLOS Medicine* dataset introduced by [Parveen and Strube, 2015]. This dataset contains 50 scientific articles in *XHTML* format. It is easy to deal with the *XHTML* format, i.e. extracting information from *XHTML* is much more convenient than from *PDF* formatted documents. In this dataset every scientific article is accompanied by a summary written by an editor who is not one of the authors of the article. The editor's summary is written by an editor of the month. This editor's summary has a broader perspective than the authors' abstract and is usually well structured. Hence we can use the editor's summary as a *gold standard* summary for evaluation.

We also applied our method on the *DUC 2002* dataset, which is a standard dataset of news articles originally created and annotated for the Document Understanding Conference. This dataset is designed for generic text summarization of new articles. It contains 567 news articles for summarization and every article is accompanied by at least two human summaries. However, the articles in this dataset are considerably shorter than the articles in *PLOS Medicine.*

We evaluate our method in two ways. First, we use the *ROUGE* score to compare it with other models. *ROUGE* is a standard evaluation metric in automatic summarization. Second, we evaluate the readability of the output summaries by way of human assessment. For this purpose, we give summaries from different systems to human subjects and ask them to rank the summaries on the basis of readability.

In future, we shall be working on the generation of sectional summaries. Here the goal is to summarize specific sections of scientific papers instead of the whole paper, which may contain information irrelevant to the user.

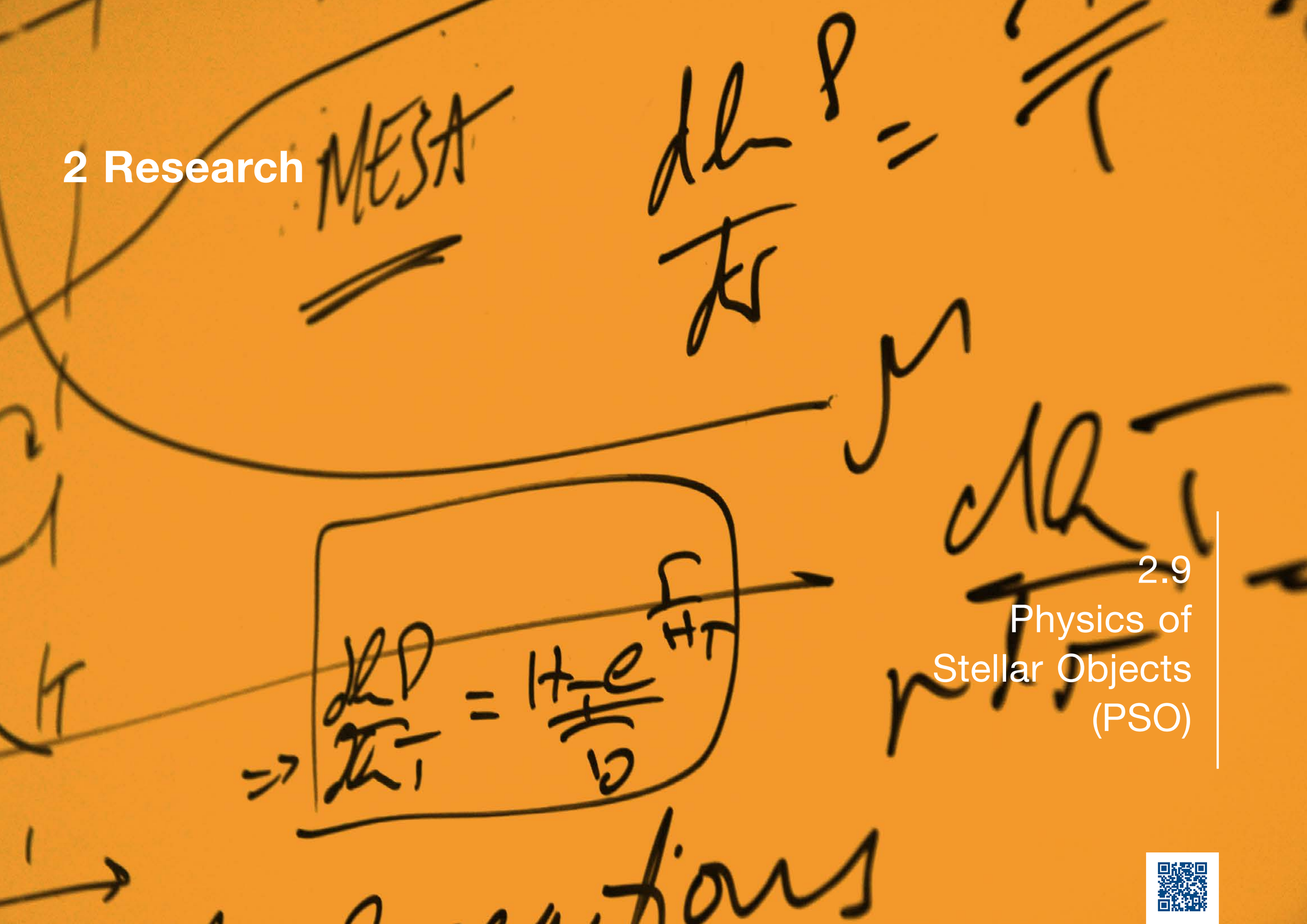
Author Disambiguation

The year 2015 also saw the kick-off of the Leibniz-funded project "Scalable Author Disambiguation for Bibliographic Databases" *(SCAD)*. In this project, we collaborate with two world-leading scholarly online bibliographic databases: DBLP (http://dblp.uni-trier.de), which focuses on computer science, and zbMATH (http://www.zbmath.org), which focuses on mathematics. Together, we aim to provide significant improvements in *author identification*. Given a set of publications with their respective author names, author identification is the task of deciding which of these author names actually refer to the same individual. Put another way, author identification attempts to cluster all publications authored by the same individual.

Getting author identification right is extremely important. User surveys and usage statistics of bibliographic databases show that targeted author queries are predominant in the navigation patterns of users searching for scholarly material.
Scientific organizations and policy-makers often rely on author-based statistics as a basis for critical action. Universities and research agencies often use publication and citation statistics in their hiring and funding decisions.

While the author disambiguation task may initially sound trivial, the practical problems it poses are immense. With several thousand new publications per month, and often only sparse author information (*John Smith*, or sometimes only *J. Smith*), maintaining high data quality is a challenging and time-consuming task for the human editors at DBLP and zbMATH. The problem is further aggravated by the fact that some authors publish under different versions of their names.

Author identification is already an established research area in computer science. From an NLP perspective, it is very similar to *entity disambiguation*. Given the NLP group's expertise in this field, we plan to apply our graph-based and network approaches to this task and to supplement them with methods for computing semantic similarity, which is another area of research the NLP group focuses on. Another thing that makes this project unique is the fact that the collaboration of DBLP and zbMath provides unprecedentedly direct access to large amounts of manually curated data.


MESA

2.9
Physics of Stellar Objects (PSO)

"We are stardust" – the matter we are made of is largely the result of processing the primordial material formed in the Big Bang. All heavier elements originate from nucleosynthesis in stars and gigantic stellar explosions. How this material formed and how it is distributed over the Universe is a fundamental concern for astrophysicists. At the same time, stellar objects make the Universe accessible to us by way of astronomical observations. Stars shine in optical and other parts of the electromagnetic spectrum. They are fundamental building blocks of galaxies and larger cosmological structures.

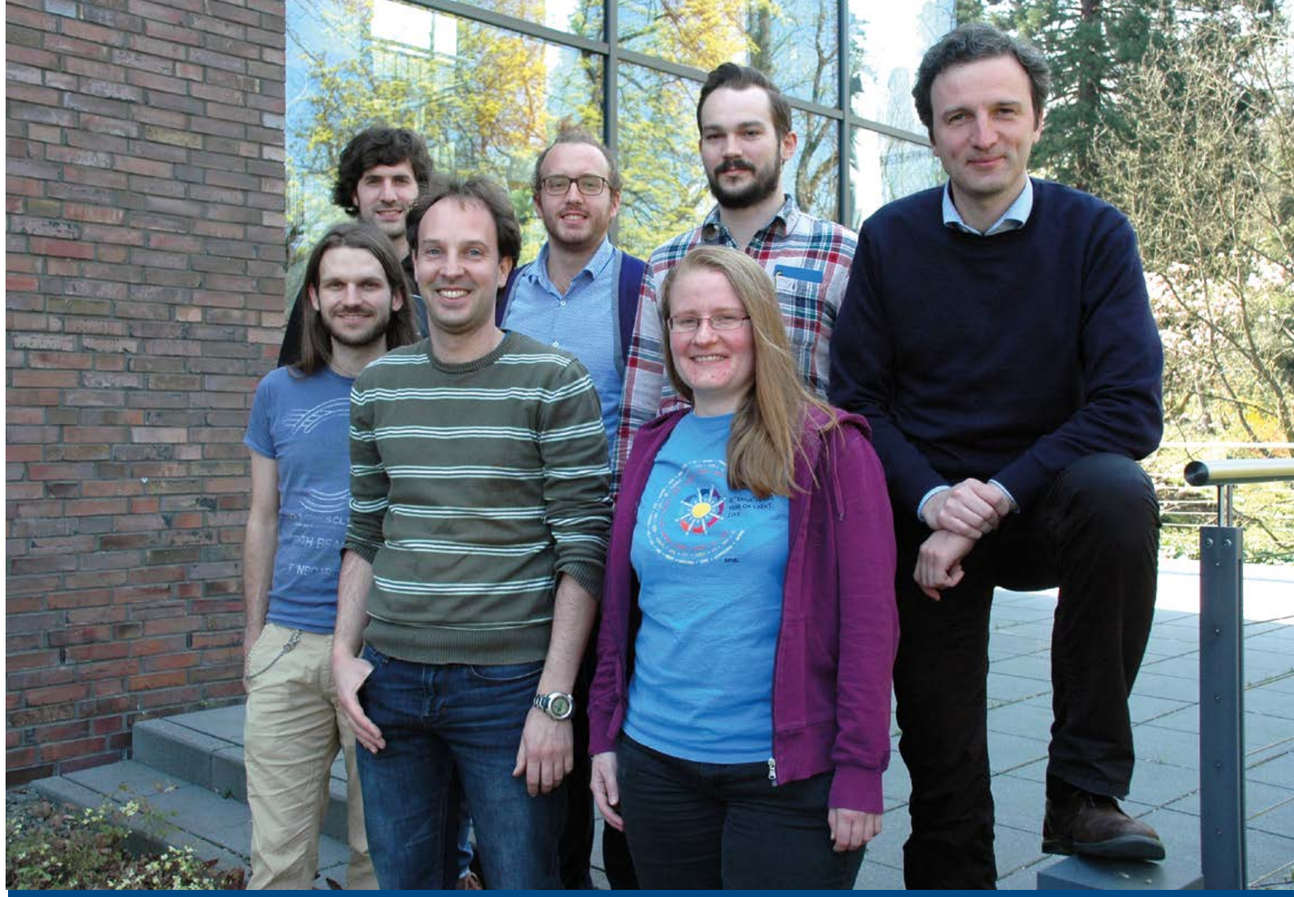
With numerical simulations, we seek to understand the processes going on in stars and stellar explosions. Newly developed numerical techniques and the ever growing power of supercomputers facilitate the modeling of stellar objects in unprecedented detail and with unexampled precision. A primary goal of our group is to model the thermonuclear explosions of white dwarf stars leading to the astronomical phenomenon known as Type Ia supernovae. These are the main sources of iron in the Universe and have been instrumental as distance indicators in cosmology, leading to the spectacular discovery of the accelerating expansion of the Universe. Multi-dimensional fluid-dynamical simulations in combination with nucleosynthesis calculations and radiative transfer modeling provide a detailed picture of the physical processes in Type Ia supernovae but are also applied to other kinds of cosmic explosions.

Classical astrophysical theory describes stars as one-dimensional objects in hydrostatic equilibrium. However, simplifying assumptions limit the predictive power of such models. With newly developed numerical tools, we explore dynamical phases in stellar evolution in three-dimensional simulations. Our aim is to construct a new generation of stellar models based on an improved description of the physical processes taking place in stars.

„Wir sind Sternenstaub" – die Materie, aus der wir geformt sind, ist zum großen Teil das Ergebnis von Prozessierung des primordialen Materials aus dem Urknall. Alle schwereren Elemente stammen aus der Nukleosynthese in Sternen und gigantischen stellaren Explosionen. Eine fundamentale Frage ist, wie dieses Material geformt wurde und wie es sich im Universum verteilt. Gleichzeitig machen stellare Objekte das Universum für uns in astronomischen Beobachtungen überhaupt erst sichtbar. Sterne scheinen im optischen und anderen Teilen des elektromagnetischen Spektrums. Sie sind fundamentale Bausteine von Galaxien und aller größeren kosmologischen Strukturen.

Wir streben mit numerischen Simulationen ein Verständnis der Prozesse in Sternen und stellaren Explosionen an. Neu entwickelte numerische Techniken und die stetig wachsende Leistungsfähigkeit von Supercomputern ermöglichen eine Modellierung stellarer Objekte in bisher nicht erreichtem Detailreichtum und mit großer Genauigkeit. Ein Hauptziel unserer Gruppe ist die Modellierung von thermonuklearen Explosionen weißer Zwergsterne, die zum astronomischen Phänomen der Supernovae vom Typ Ia führen. Diese sind die Hauptquelle des Eisens im Universum und wurden als Abstandsindikatoren in der Kosmologie eingesetzt, was zur spektakulären Entdeckung der beschleunigten Expansion des Universums führte. Mehrdimensionale strömungsmechanische Simulationen kombiniert mit Nukleosyntheserechnungen und Modellierung des Strahlungstransports ergeben ein detailliertes Bild der physikalischen Prozesse in Typ Ia Supernovae, werden aber auch auf andere Arten von kosmischen Explosionen angewendet.

Die klassische astrophysikalische Theorie beschreibt Sterne als eindimensionale Objekte im hydrostatischen Gleichgewicht. Die hierbei verwendeten vereinfachenden Annahmen schränken jedoch die Vorhersagekraft solcher Modelle stark ein. Mit neu entwickelten numerischen Hilfsmitteln untersuchen wir dynamische Phasen der Sternentwicklung in dreidimensionalen Simulationen. Unser Ziel ist es, eine neue Generation von Sternmodellen zu entwickeln, die auf einer verbesserten Beschreibung der in Sternen ablaufenden physikalischen Prozesse basiert.



Group leader

Prof. Dr. Friedrich Röpke

Staff members

Dr. Andreas Bauswein *(since August 2015)*
Dr. Philipp Edelmann

Scholarship holder

Dr. Samuel Jones *(Alexander von Humboldt Scholarship, since May 2015)*

Visiting scientists

Aron Michel
Kai Marquardt
Sebastian Ohlmann

Student

Sabrina Gronow *(since October 2015)*

Stellar explosions: Which stars explode as Type Ia supernovae?

Type Ia supernovae are among the most spectacular astrophysical events. Their extremely luminous appearance and – by astronomical standards – the remarkable uniformity in their brightness have led to their use as cosmic distance indicators. With Type Ia supernovae, it is possible to measure distances throughout large parts of the Universe with high precision. Such measurements, performed in the late 1990s, showed that the Universe is expanding faster and faster. This discovery was awarded the Nobel Prize in Physics 2011, but it has puzzled physicists to the present day. The cause of the acceleration was attributed to an unknown "dark" energy form that makes up about 70 per cent of all energy in today's Universe.

But the objects that led to this ground-breaking discovery, Type Ia supernovae, themselves remain enigmatic. Despite all the efforts undertaken, extensive observational surveys have not yet been able to identify the origin of these bright cosmic events. We see them as astronomical objects that in the course of a few days brighten until they outshine their entire host galaxies composed of hundreds of billions of stars. Then, after another few days, they fade away.

Supernovae are generally attributed to the explosion of stars. While all other supernova classes are associated with the release of gravitational binding energy as the cores of massive stars collapse, Type Ia supernovae are believed to form as a result of thermonuclear explosions of fairly light, but compact, stars – so-called white dwarfs. These mark the endpoint in the evolution of light and intermediate-mass stars (such as our Sun). Having burned all their hydrogen and helium fuel, their cores are composed of a mixture of carbon and oxygen, or, in somewhat heavier stars, a mixture of oxygen and neon. In principle, energy can be produced in nuclear fusion up to the formation of iron group elements, but the low mass of the stars prevents the triggering of more advanced burning stages. Accordingly, no thermal pressure is generated that supports these stars against their own gravity. They contract until they find a new equilibrium state where gravity is balanced by a quantum-mechanical effect: the degeneracy of the electrons in the stellar plasma. The resulting "Fermi pressure" stabilizes the stars when they have reached a radius typical of planets, while their total mass is still approximately that of the Sun. Extremely high densities in their material are thus unavoidable. This stabilization is effective up to the so-called Chandrasekhar mass limit, about 1.4 solar masses.

The challenge to Type Ia supernova research is to find out why the thermonuclear explosion occurs. The quantum-mechanical stabilization is eternal, and a white dwarf is a rather unspectacular object – unless it has a close binary companion that interacts with it and triggers a thermonuclear explosion. Several scenarios have been proposed for this, differing in the assumed nature of the companion star, the physics of the interaction, the transport of material between the two stars, and the mass of the white dwarf at the onset of the explosion. These are explored in hydrodynamical explosion simulations. A large variety of such scenarios and parameters have been simulated by members of the PSO group over the past years. To this end, specialized codes have been developed that follow the explosion dynamics in three-dimensional simulations, making use of some of the world's fastest supercomputers. The laws of fluid dynamics allow for two distinct modes of propagation of a thin combustion wave: a subsonic flame called "deflagration" and a supersonic "detonation." These modes lead to different patterns in the structure of the explosion debris. Subsequent to such explosion simulations, radiative transfer calculations are performed by collaborators, who are able to predict synthetic observables from the supernova models. This allows direct comparison with astronomical data and hence a validation of the modeling assumptions. The rich variety of different explosion scenarios is motivated by recent observational results. Although the majority of Type Ia supernovae are rather uniform in their characteristics, peculiar objects have been increasingly observed and distinct sub-classes of objects have been identified with properties diverging widely from the normal events.

One parameter that has rarely been considered in past research is the chemical composition of the exploding white dwarf star. Most models have assumed it to consist of an equal-by-mass mixture of carbon and oxygen. The composition, however, is important for the dynamics of the explosion and hence for the characteristics of the predicted supernova event. It determines how much nuclear energy can be released in "burning" this stellar fuel material into iron group elements. Starting from carbon releases more energy than starting from a mixture of oxygen and neon.

A standard scenario for producing normal Type Ia supernovae is a thermonuclear deflagration triggering near the center of a carbon-oxygen white dwarf star close to the Chandrasekhar mass. It later turns into a detonation giving rise to a powerful explosion. The carbon-to-oxygen ratio in the material of the star is known to affect the explosion energetics. It is expected that realistic progenitor stars have carbon-depleted cores. The observational imprints of the effect, however, are weak [Ohlmann, S. T.; Kromer, M.; Fink, M.; Pakmor, R.; Seitenzahl, I. R.; Sim, S. A. & Röpke, F. K.: The white dwarf's carbon fraction as a secondary parameter of Type Ia supernovae, Astronomy & Astrophysics 572, A57 (2014)]. Recent simulations by the PSO group [Marquardt, 2015] have focused on the question of what the outcome of thermonuclear explosions of oxygen-neon white dwarfs would look like. Although not as numerous as carbon-oxygen white dwarfs, these stars should also contribute to observed explosive events. Indeed, if one considers systems in which mergers of two white dwarfs lead to thermonuclear explosions, the contribution of systems involving oxygen-neon white dwarfs could be in the vicinity of 10%. But are the predicted observables consistent with observed Type Ia supernovae? To answer this question, the PSO group performed simulations that followed detonations in sub-Chandrasekhar mass oxygen-neon white dwarfs. As these are more massive in form than their carbon-oxygen counterparts, detonations in them lead to extremely powerful and bright events. The observables predicted from the simulations of the PSO group resemble those of the brightest events in the observed sample. The spectra agree somewhat better with observations than those predicted from carbon-oxygen white dwarfs.

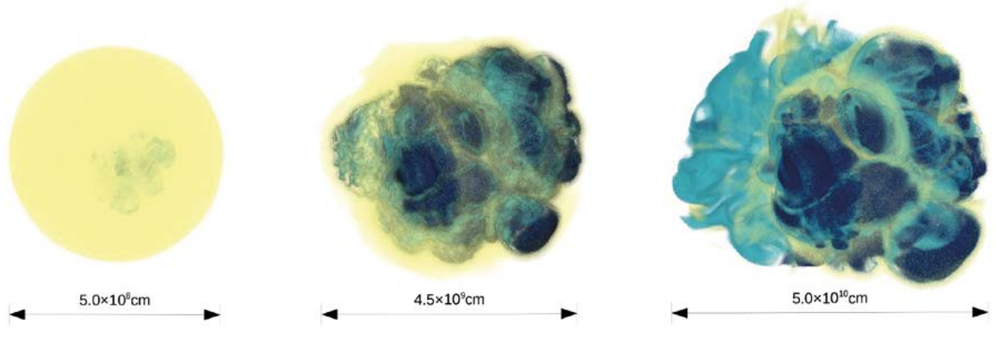


Figure 50: Simulation of a thermonuclear explosion in a hybrid carbon-oxygen-neon white dwarf. Snapshots are taken 0.85, 2.5, and 10 seconds after ignition. The colors indicate the chemical composition of the ejecta material. Figure taken from [Kromer, 2015].

A very interesting possibility arises from an intermediate case where the exploding white dwarf has a carbon-rich core but its outer shell consists of oxygen and neon. Thermonuclear explosions in such stars were recently investigated by the PSO group [Kromer, 2015]. When approaching the Chandrasekhar mass limit and triggering a deflagration flame in its center, the resultant explosions in such stars are predicted to lead to rather peculiar events. The flame burns through the carbon-rich core but quenches on reaching the outer shell because burning the oxygen-neon material does not release enough energy to sustain the deflagration flame, *see Figure 50, previous page*. The result is a very weak explosion, and faint events are expected from this explosion scenario. The synthetic observables predicted from the hydrodynamical explosion simulations of the PSO group are in good agreement with the observed extremely faint Type Iax supernovae, a subclass of Type Ia events. They thus provide a promising explanation for these enigmatic objects.

Since every chemical element has its own "fingerprint", astronomers can reconstruct their abundances in the stars of our Galaxy.

Stellar alchemy: How to make gold in the Universe

A star eight times more massive than our Sun ends its life in a spectacular supernova explosion. While the outer layers of the star are disrupted in the course of the explosion, the inner core collapses under its own weight and forms an ultra-compact neutron star or a black hole. Many stars exist in double systems with the components orbiting each other. Consequently, two such supernova explosions may lead to a neutron star orbiting a black hole or a double neutron-star system.

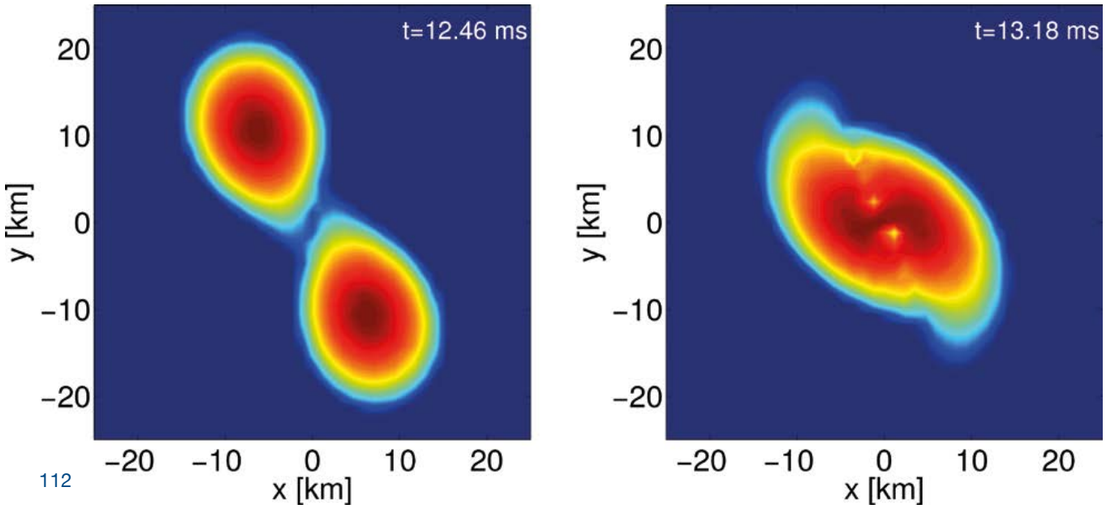


Figure 51: Simulation of two neutron stars orbiting and merging. The figure shows the density in the equatorial plane (rotation counter-clockwise). A small fraction of stellar material becomes unbound during merging. Nucleosynthetic processes forge heavy elements like gold in the ejecta. (Simulation by A. Bauswein, HITS)

These binary systems are of great interest because the two components slowly approach each other. After many millions of years, this process leads to the coalescence of the binary *Figure 51.* The violence of an event like a merger between two neutron stars or a neutron star and a black hole provides extreme conditions under which heavy nuclei can be created. In fact, it is suspected that such mergers may be the origin of gold, platinum, uranium, and other heavy elements. It is also to be expected that the light produced by nearby mergers be observable with telescopes, which may help us to understand the merging process and to finally clarify whether the gold in a wedding ring was forged billions of years ago by a neutron star merger.

The largest telescopes are searching for signals from such mergers, and special instruments are being built to detect the vibrations of space-time caused by them. Those vibrations were already predicted by Albert Einstein as a consequence of his theory of General Relativity. These so-called gravitational waves have recently been directly observed. Two highly sensitive gravitational-wave detectors, combined in the Advanced LIGO in the USA, were able to ultimately confirm this cornerstone of Einstein's theory. The detected signal is attributed to the merger of two black holes. But do scientists actually have any prospects of catching a glimpse of a signal from a neutron star merger? The answer depends on the rate of such mergers, which unfortunately is not very well known. An approach to estimating the probability of these events has been worked out by an international collaborative venture led by HITS's researcher Andreas Bauswein [Bauswein,A.; Ardevol Pulpillo,R.; Janka,H.-T.; Goriely,S., Nucleosynthesis constraints on the neutron star-black hole merger rate, Astrophysical Journal Letters 795, L9 (2014)].

We know how abundant some heavy elements are in the Universe. Even a tiny admixture of a heavy element in a star leads to a particular pattern in the spectrum of light it emits. Since every chemical element has its own "fingerprint", astronomers can reconstruct their abundances in the stars of our Galaxy. This gives an estimate of the total mass of heavy elements in the Galaxy.

Given this knowledge, the international research team followed up a simple idea. If one knows the amount of heavy elements produced by one merger event, one can easily deduce how many events are needed to obtain the total amount observed. Dividing this number by the age of the Galaxy gives an estimate of the merger rate, and here an order-of-magnitude estimate is already very welcome. However, there is a snag. How can we know how many heavy elements are produced by one merger event if no such event has ever been directly observed? The answer, of course, lies in theoretical models for the mergers. The team led by Andreas Bauswein conducted high-performance computer simulations of the merging process to determine their nucleosynthesis yields. They employed the technique of so-called smooth-particle hydrodynamics, which is well-suited to resolving the relatively small amounts of matter relevant for this study. The analysis is based on over 100 different simulations for neutron star-black hole mergers and mergers of double neutron stars. It is important to consider all possible initial configurations of such binary systems, i.e. all possible combinations of the masses of their components.

The calculations reveal that a single merger typically produces heavy elements about 1/100 the mass of the Sun. The joint project also confirmed that in the outflows from mergers the different heavy elements are produced with the same relative abundance as observed in nature. Every merger event produces the impressive amount of about one Jupiter-mass of gold in addition to roughly similar amounts of other heavy elements. This material, however, is spread by the violence of the coalescence, is strongly diluted, and mixes with the surround-

ing matter, which forms stars. By these complex processes the Galaxy has been continuously enriched by heavy elements throughout its history. A portion of this matter forms planets like the Earth, and the heavy elements forged in the course of mergers can be found in mines and rivers.

With their analysis, the scientists from HITS, the Aristotle University Thessaloniki, the Max Planck Institute for Astrophysics in Garching, and the Université Libre de Bruxelles predict one merger every 10,000 to 100,000 years in our Galaxy, *see Figure 52*. Since the upcoming super-sensitive instruments can observe many more Galaxies, they have a good chance of detecting several merger events per year. So excitement is growing as the new detectors become operational, and the scientific community has justified hopes for settling the perennial question about the origin of heavy elements.

As a spin-off from the set of performed calculations, the research team also analyzed the detailed gravitational-wave signals of merger events [Bauswein, 2015a] [Bauswein, 2015b]. Interestingly, they were able to identify certain patterns in the expected signals that reveal details about the merging process. For instance, it is possible to deduce whether the merging was relatively smooth or whether the binary components collided violently. The dynamics of the merging process depend crucially on the properties of the stellar matter, which at the relevant densities of neutron stars are governed by nuclear physics. Thus, the observation and the understanding of gravitational-wave signals also provide important insights for nuclear physics, especially in regimes that are not accessible in terrestrial experiments.

This important link to high-density matter physics has been strengthened by a detailed study of the capabilities of analyzing the data from future gravitational-wave detectors [Clark, 2015]. In this type of measurement, it is a formidable challenge to find a signal buried away in the very noisy raw data from the instrument. Collaborative research involving HITS, the Aristotle University Thessaloniki, and the Georgia Institute of Technology has now confirmed that the expected hidden signal patterns can be extracted from a future measurement. The evidence was adduced by employing templates of signals produced by simulations at HITS. These templates were hidden in available noisy detector data not actually containing any real signals. Existing and newly developed analysis tools were able to recover what was hidden and to extract the relevant parameters. Accordingly, future gravitational-wave measurements promise more insights than the mere confirmation of the existence of these waves. Notably, they will shed light on fundamental interactions of nuclear matter.

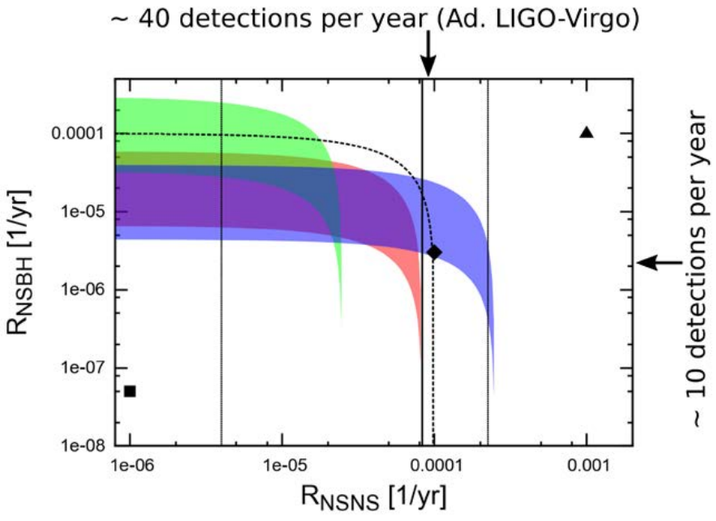


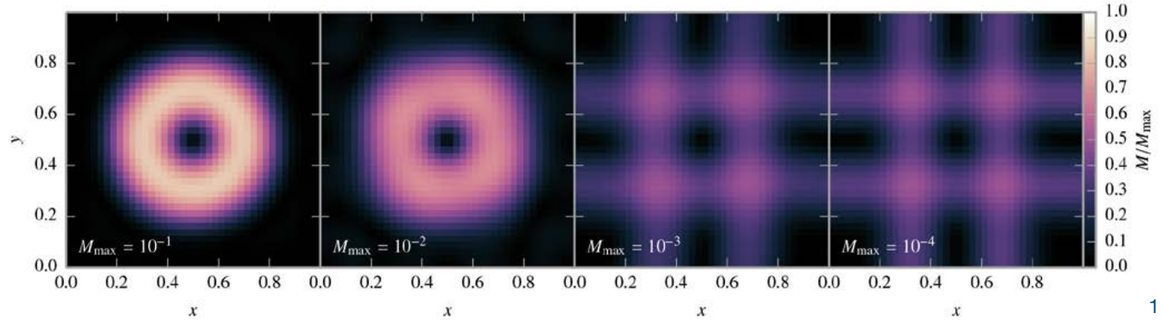
Figure 52: Estimated Galactic rates of neutron star-neutron star collisions vs. neutron star-black hole mergers. The colored regions bracket the estimated rate (measured in events per year per Galaxy) and illustrate that existing and upcoming gravitational-wave instruments like Advanced LIGO have a good chance of detecting gravitational radiation from compact object mergers. For instance, a Galactic rate of 10^{-4} neutron star-neutron star mergers per year corresponds to roughly 40 detections per year with the Advanced LIGO-Virgo network at design sensitivity. (Figure taken from Bauswein,A.; Ardevol Pulpillo,R.; Janka,H.-T.; Goriely,S., Nucleosynthesis constraints on the neutron star-black hole merger rate, Astrophysical Journal Letters 795, L9 (2014)).

Stellar evolution: How to model slow fluid dynamics

Stars spend the majority of their lives in rather quiescent states. At their cores, they process hydrogen by nuclear fusion on time scales of millions to billions of years. The energy thus produced is then transported to the surface by radiation and the convective flow of the plasma making up the star. These flows are quite well described by the laws of fluid dynamics, the *Navier-Stokes equations*. Despite this fact, stars have been traditionally described using simple, static, spherically symmetric models. As observations of stars become more abundant and detailed, it is increasingly obvious that we need to take a step forward towards truly dynamic, multidimensional stellar models. The hydrodynamic modeling of stars and the development of new numerical methods for astrophysical flows are major research topics in the PSO group.

Hydrodynamic flows show a degree of self-similarity. By the mathematical process of non-dimensionalization it can be seen that the characteristic behavior of a flow depends mainly on a few dimensionless numbers. As long as viscosity only acts on scales much smaller than those of interest, the sole important number characterizing the flow regardless of scale is the *Mach number M*. It is the ratio of local fluid velocity and sound speed. The solutions behave fundamentally differently depending on whether they are in the supersonic ($M \gg 1$), transonic ($M \approx 1$), or subsonic ($M \ll 1$) regimes. This is also reflected in the fact that quite different numerical methods are used in the different regimes. Shocks develop in supersonic flows, which means that these schemes must be able to handle discontinuities. On the other side of the spectrum, the low Mach-number regime, solutions are typically smooth. In the limit of $M \to 0$ the equations display peculiar behavior. Sound waves decouple completely from the rest of the flow, referred to as the *incompressible solution*. This poses a major problem for numerical codes as they often make the incompressible part of the solution create artificial sound waves, which causes an unphysically high dissipation of energy and makes the solution unusable. This problem is easily demonstrated in a simple test, the *Gresho vortex*. This is a rotating vortex stabilized by a radial pressure gradient. The analytic solution in this case is stationary, i.e. it remains in its initial condition. *Figure 53* shows what happens when the Roe solver, a standard method for compressible hydrodynamics, is used to simulate the vortex. We see that it completely dissolves after just one turnover at Mach numbers below 10^{-3}.

Figure 53: Gresho vortex after one full rotation computed using the original Roe solver. The panels show different initial Mach numbers. The color scale has been adjusted to the respective initial Mach number. Figure taken from Mizcek et al, 2015, A&A, 576, A50, reproduced with permission © ESO.



There are many numerical schemes designed specifically for the low Mach-number limit that avoid this problem. The major downside of this approach is that these schemes produce a physically incorrect result once the Mach number approaches unity. For problems involving flows with Mach numbers from about 1 to 10^{-6}, some other scheme needs to be used. Such situations are quite common in the astrophysics of stellar interiors. A typical example is the boundary between a region of a star where energy is transported predominantly by turbulent convection ($M \sim 10^{-2}$) and a more quiescent region in which energy is mainly transported by radiation ($M \sim 10^{-6}$).

A careful analysis of the Roe scheme reveals the cause of the excessive dissipation. The artificial viscosity term, which must be added to stabilize the discretization, does not scale with the Mach number in the same way as the physical terms. This means that for the Mach number tending to zero, the artificial viscosity terms start to dominate the solution completely, as we saw in *Figure 53*. While it is still stable, the result is subject to much higher dissipation resulting in less turbulent mixing and energy transport.

In our work [Misczek, 2015], we propose a small, but important change to the original Roe scheme, which makes it behave correctly right down to very low Mach numbers of even 10^{-10}. This technique, called flux preconditioning, has been used in the past, but mainly for steady-state problems. The novelty in our approach is the form of the preconditioning matrix, which avoids some of the problems involved in earlier proposals. Specifically, it is compatible with gravity source terms, which are ubiquitous in astrophysics. It also shows improved convergence behavior when used with implicit time discretization. *Figure 54* shows the improvement gained by this change of the Roe solver. The solution is now almost identical to the initial condition and practically independent of the Mach number, which is as it should be.

Tests of this new scheme have also been performed for non-stationary flows. *Figure 55* shows a seeded Kelvin-Helmholtz instability at Mach number 10 – 2 before it enters the chaotic phase. Here the upper and lower region of the fluid move to the right, while the central region moves to the left, causing shear at the interface. To have a reference solution, we compute the solution not only at the test resolution of 128 × 128 but also at 1024 × 1024. We see that even at low resolution our new scheme yields results much closer to the high-resolution reference. The original Roe flux smears out most of the fine structure of the instability.

This new scheme has been integrated into our main simulation code for stellar hydrodynamics, SLH (Seven-League Hydro). Apart from a choice of several numerical schemes suited for simulations of flows involving all Mach numbers, SLH can also use very long time steps due to implicit time discretization. Thus it is an excellent tool for studying flows in the deep interiors of stars with high accuracy over long time-scales. ■

Figure 54: Gresho vortex after one rotation computed using the original Roe solver. The panels show different initial Mach numbers. The color scale has been adjusted to the respective initial Mach number. Figure taken from Mizcek et al, 2015, A&A, 576, A50, reproduced with permission © ESO.

Figure 55: A Kelvin-Helmholtz instability before it enters the chaotic phase. The color signifies a passive scalar, i.e. a field advected with the flow that does affect dynamics in any way. The top panels have been computed in low resolution, the bottom ones are reference solutions at much higher resolution. The left panels show the original Roe scheme, the right panels the flux preconditioned Roe scheme. Figure taken from Mizcek et al, 2015, A&A, 576, A50, reproduced with permission © ESO.

Figure 54

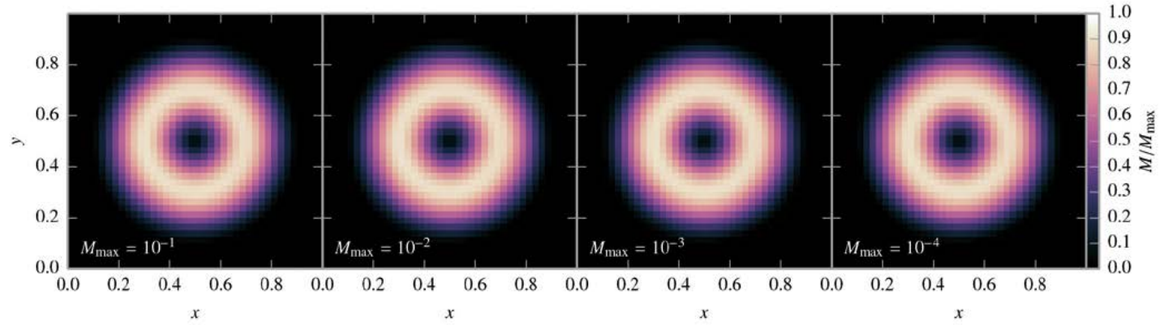
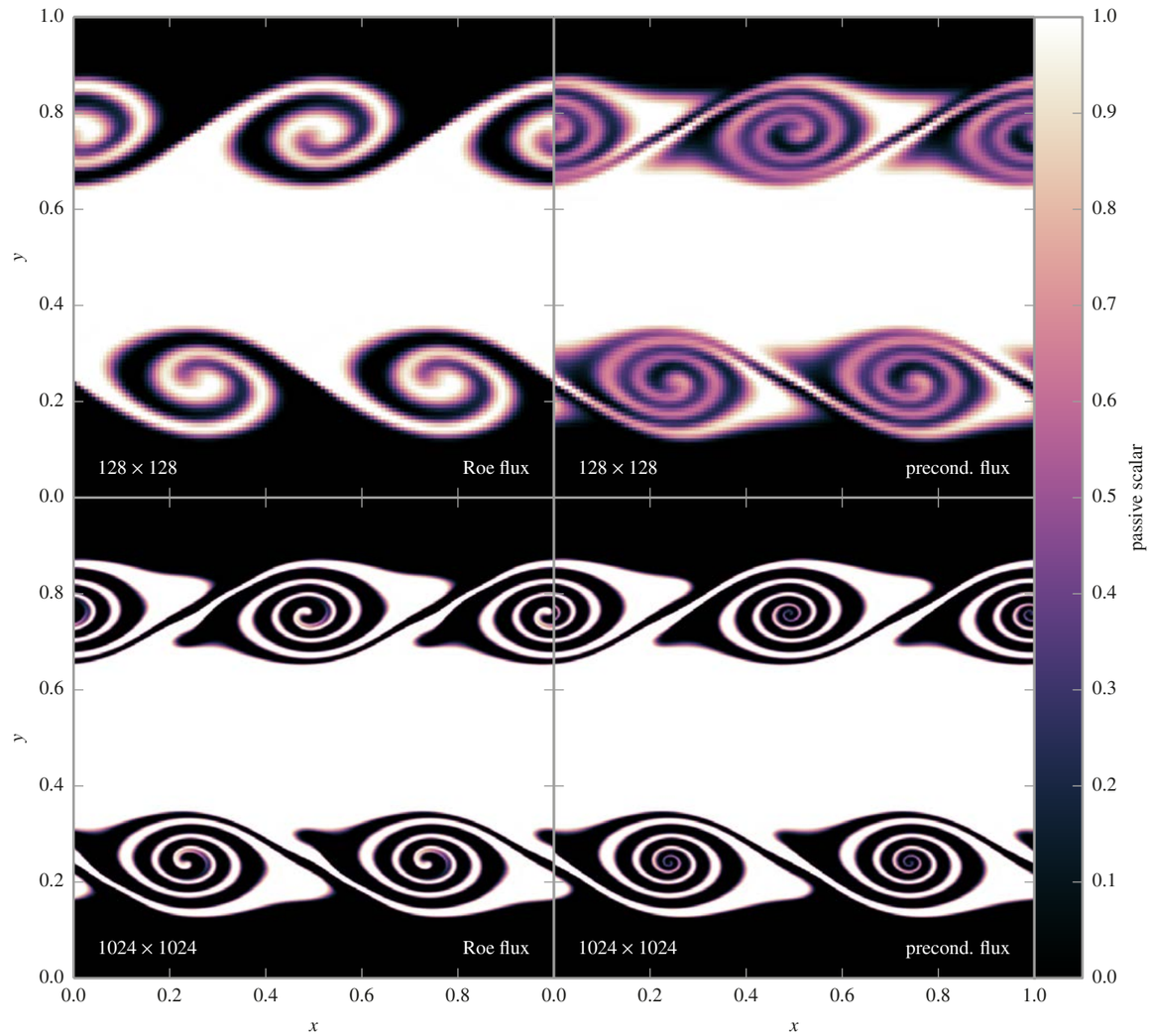


Figure 55



2 Research

2.10 Scientific Computing (SCO)

2.10
Scientific Computing (SCO)

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 analyses
- scientific software quality & verification

Secondary research interests include, but are not limited to:

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

In the following section, we outline our current research activities. Our research is situated at the interface(s) between computer science, biology, and bioinformatics. The overall goal is to devise new methods, algorithms, computer architectures, and freely available/accessible tools for molecular data analysis and to 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*
- *Analysen von „Next-Generation" Sequenzdaten*
- *Qualität & Verifikation wissenschaftlicher Software*

Sekundäre Forschungsgebiete sind unter anderem:

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

Im Folgenden beschreiben wir unsere Forschungsaktivitäten. Unsere Forschung setzt an der Schnittstelle zwischen Informatik, 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."

Group leader

Prof. Dr. Alexandros Stamatakis

Staff members

Andre Aberer
Dr. Tomáš Flouri
Diego Darriba
Dr. Paschalia Kapli

Scholarship holders

Kassian Kobert *(HITS Scholarship)*
Alexey Kozlov *(HITS Scholarship)*
Jiajie Zhang *(HITS Scholarship)*
Lucas Czech *(HITS Scholarship)*

Visiting scientists

Prof. Dr. Mark Holder *(until July 2015)*
Dr. Emily Jane McTavish *(until August 2015)*
Rebecca Harris *(from October until December 2015)*

Student

Sarah Lutteropp

What happened at the Lab in 2015?

The following is an account of the important events at the lab in 2015. In winter 2014/2015, Alexis, Tomas, Alexey, Kassian, Mark, and Andre taught a class entitled "Introduction to Bioinformatics for Computer Scientists" at the Karlsruhe Institute of Technology (KIT). As in previous years, we again received a very positive teaching evaluation from the students. Based on this and our previous teaching evaluations, we received in summer 2015 a certificate from the Dean of the Computer Science department at KIT for excellence in teaching.

In summer 2015, Tomas and Alexis again taught our recently devised practical programming course at KIT and worked with two teams of students to design useful, open-source software for biologists. Together with the students, we wrote a paper about the code we had developed and published it via the biorxiv preprint server [Baudis et al., 2015]. For this practical course we also received a learning quality index of 100 out of 100 based on the student evaluations. (see http://sco.h-its.org/exelixis/web/teaching/courseEvaluations/Practical_2015_Evaluation.pdf). Finally, in summer 2015 we also taught our main seminar "Hot Topics in Bioinformatics" again.

The year was also very important for our former PhD student Jiajie Zhang, who successfully defended his PhD thesis at the University of Lübeck in March 2015. By that time, Jiajie had already started working at a Start-Up company in the UK that is developing novel DNA sequencing technology. The same company had already hired our former SCO PhD student Fernando Izquierdo-Carrasco the year before.

The lab member who has spent the longest time working with Alexis, Andre J. Aberer, also defended his PhD in late 2015 at the Computer Science department of the Karlsruhe Institute of Technology. Andre, still a first year Master's student at that time, had attended the first full university course Alexis ever taught back in winter 2008/2009 in Munich. Since then he has worked with Alexis on a student research project, his Master's thesis, and his PhD thesis. He now works in the financial sector in Frankfurt.

Toward the end of 2015, Constantin Scholl, a student of the Information Engineering and Management program at KIT submitted an outstanding Bachelor thesis on a theoretical computer science topic.

We were also happy to welcome our new student programmer, Sarah Lutteropp, from KIT, a previous winner of the "Bundeswettbewerb Informatik". She will also be doing her Master's thesis with us and will join the lab as a PhD student some time in the fall 2016.

In 2015, we again hosted a visiting PhD student via our visiting PhD student program. Rebecca Harris, a biologist from the University of Washington at Seattle, stayed with us from mid-October to mid-December.

We had to say good-bye to Prof. Mark Holder and Dr. Emily Jane McTavish, who left us in July and August respectively. Mark spent a whole sabbatical year with us, while Emily, Mark's Post-Doc, stayed with us for 9 months on a Humboldt fellowship. We are still working with them on joint research projects and papers.

Another highlight in 2015 was the course on computational molecular evolution that took place for the 7th time on the campus of the European Bioinformatics Institute at Hinxton, UK. Again, Alexis co-organized and taught at this event. Paschalia also took part as a teaching assistant.

Finally, we sold several commercial licenses for the PEAR software [Zhang et al., 2014] to biotech companies around the world (mainly in the US). Note that the software is free for academic use. The license fees will be used to fund internships at labs abroad for KIT computer science students doing their Master theses with us.

Introduction

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, i.e., for disentangling evolutionary events 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, manual reconstruction of trees is no longer feasible. This is why evolutionary biologists have to rely on computers for phylogenetic and population-genetic analyses.

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 key problem that needs 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. This gap has further widened in 2015 (see http://www.genome.gov/images/content/cost_per_genome_oct2015.jpg) due to the arrival of a new generation of sequencers on the market, not least from the company our two former PhD students Fernando and Jiajie are now working for.

Accordingly, as computer scientists, we are facing a scalability challenge, i.e., we are constantly trying to catch up with the data avalanche and to 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 technology is changing rapidly. Accordingly, the output of these machines in terms of the length and quality of the sequences they can generate is also changing constantly. This requires the continuous development of new algorithms and tools for filtering, puzzling together, and analyzing these molecular data. 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, e.g., 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 identify endangered species that need to be protected, determined on the basis of 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 (e.g., a human population) can be used to identify mutations leading to specific types of cancer or other serious diseases.

Based on the prolegomena, one key challenge for computer science is scaling existing analytic methods to the huge new datasets produced by next-generation sequencing methods.

Also, these codes need to be capable of leveraging the computational resources provided by supercomputers. In 2014 we achieved major breakthroughs in the development of scalable phylogenetic inference tools with the release of the ExaML (Exascale Maximum Likelihood) and ExaBayes (Exascale Bayesian Inference) open-source codes for large-scale phylogenomic inference on supercomputers. In 2015, we further improved and maintained these tools and also incorporated new models of evolution [Aberer et al., 2015, Kozlov et al., 2015].

Another area of major interest, concern, and a new direction for research, is scientific software quality and data analysis pipeline complexity. In this new line of research, we focus not only on implementation quality and errors, but also on conceptual or mathematical errors occurring before the actual coding stage.

One main motivation for investigating this further is that many truths we nowadays believe in are based on badly maintained, low-quality scientific software.

Also, with the advent of Next Generation Sequence (NGS) data, bioinformatics analysis pipelines have become increasingly complex. In the good old 'Sanger sequencing days', the analysis pipeline was rather straightforward, once the sequences were available. For a phylogenetic study, it consisted of the following steps: align -> infer tree -> visualize tree.

For NGS data and huge phylogenomic datasets, such as the insect transcriptome or bird genome analyses we published in late 2014, pipelines have become a great deal longer and more complex. They also require user expertise in an increasing number of bioinformatics areas (e.g., orthology assignment, read assembly, dataset assembly, partitioning of datasets, divergence times inference, etc.). In addition, these pipelines typically require a plethora of helper scripts, usually written in languages such as perl (a language highly susceptible to coding errors due to the lack of typing) or python to transform formats, partially automate the workflow, and connect the components.

Our main concern is that if each code or script component used in such a pipeline has a probability of being 'buggy' P (bug), the probability that there is a bug in the pipeline increases dramatically with the number of components.

If detected too late, errors in the early stages of pipelines (e.g., NGS assembly) for large-scale data analysis projects can have a dramatic impact on all downstream analyses (e.g., phylogenetic inferences, dating). In short, they will all have to be repeated!

In fact, this has happened in every large-scale data analysis project we have been involved in so far.

To indulge in a little heresy, we will now outline two new research projects that focus on partially erroneous bioinformatics operations on computers. Potentially, the errors and inconsistencies we detected may have dramatic effects on the biological results reported.

We conclude, on a more optimistic note, with the description of a project that aims at identifying and correcting errors in sequence databases.

A critical review of Gotoh's algorithm

Pairwise sequence alignment is perhaps the most fundamental bioinformatics operation. The algorithm is used to assess the degree of similarity between two DNA sequences / fragments. Such an assessment of sequence similarity is frequently the very first analytic step in quantitative biological data analyses, for instance, to determine which gene a newly obtained DNA sequence belongs to, or which known species a novel bacterial strain is most similar to.

The task of comparing two sequences is not as straightforward as one might think. An optimal algorithm for the so-called global alignment problem was described in 1970 by Needleman and Wunsch. Note that, in this context, optimal refers to finding the best possible alignment between two sequences (the one with the highest similarity score). In 1982 Gotoh presented an improved algorithm with a lower so-called time complexity. In theoretical computer science, lower time complexity means that the relative amount of operations as a function of the input data size (in our case the lengths of the two sequences we want to align) required to calculate the result on a processor is provably lower than that of a previous algorithm for that task. Gotoh's algorithm for pair-wise sequence alignment is frequently cited (1510 citations, Google Scholar, January 2016), taught and, most importantly, both used and implemented. While implementing the algorithm, we discovered in Gotoh's paper two mathematical mistakes or imprecise formulations (the debate on which of these is the case is still going on) that may induce sub-optimal sequence alignments.

First, there are some minor indexing mistakes in the so-called dynamic programming algorithm which immediately become apparent when implementing it. In computer science, we use dynamic programming algorithms to fill a matrix that stores intermediate results.

In our case, this matrix stores alignments of all possible sub-sequences (reading the sequences from left to right) of the two sequences we want to align in order to progressively build an optimal alignment. This minor mistake we found occurs in the matrix indices, but it can easily be spotted and corrected by any experienced programmer intending to implement the algorithm.

However, Gotoh's original paper contains a second, more profound problem connected with the dynamic programming matrix initialization. Note that, to start filling a dynamic programming matrix, the topmost row and the leftmost column need to be initialized by appropriate values that will guarantee the correctness of the algorithm. Note further that the term correctness here means that the algorithm will always, and under all possible input parameters, yield the optimal pairwise sequence alignment.

The issue with the initialization is not so straightforward to detect. It can easily be missed and will then find its way into actual implementations. To this end, we have tried to quantify the extent to which this imprecise formulation or mere error has been propagated. For instance, this initialization error is present in standard bioinformatics text books, including the widely used books by Gusfield and Waterman.

To also assess the extent to which this error has found its way into university courses, we scrutinized freely available undergraduate lecture slides. We found that 8 out of 31 lecture slides contained the mistake (one set of slides has already been corrected after the initial publication of the preprint describing this work [Flouri et al., 2015c]), while 16 out of 31 simply omit parts of the initialization, thus giving an incomplete description of the algorithm.

Finally, we also inspected ten source codes and conducted tests on open-source implementations of Gotoh's algorithm.

We found that five implementations were incorrect, i.e., yielded sub-optimal pairwise sequence alignments under certain input parameters.

Note that not all the bugs we identified are due to the errors/imprecise statements in Gotoh's paper. Three implementations rely on additional constraints on the input data that limit their generality. Accordingly, only two of the ten implementations we scrutinized yielded formally correct results.

Fortunately, the error traceable to Gotoh's lack of precision is straightforward to resolve, and we provide a correct open-source reference implementation. We believe, though, that enhancing the awareness of these errors is crucial, since the impact of incorrect pairwise sequence alignments that, as we

have said, typically represent one of the very first stages in any bioinformatics data analysis pipeline can have a detrimental impact on downstream analyses such as multiple sequence alignment, orthology assignment, phylogenetic analyses, divergence time estimates, etc.

However, due to the complexity of modern data analysis pipelines and the code size of the tools that rely on this algorithm, expecting to be able to quantify the impact of this error on downstream analyses and published biological results would be utterly naïve. We are currently collaborating with some authors of potentially affected software to determine whether the error is present. On a more positive note, some slides and implementations have already been corrected following the preprint publication and its rapid distribution via social media.

A critical review of tree-viewing software

Along the same lines we investigated another topic that we had long suspected might pose problems. As already mentioned, one of the main research topics of our lab is to develop tools for the reconstruction of phylogenetic trees. Once those trees have been inferred, they are typically visualized by so-called tree viewers so that biologists can interpret them, draw evolutionary conclusions, and write their papers. Figures of such tree visualizations form part of almost every evolutionary biology paper.

Evidently, such tree viewers are highly cited, and erroneous representations of the trees might induce incorrect biological conclusions. The problem we detected is not associated with the trees as such, but concerns the meta-data associated with them. In evolutionary analyses we do not only infer a plain tree but also so-called support values. Such support values are associated with each inner branch (branches connecting two inner nodes/ancestors) of the tree and essentially represent the degree to which we believe that the represented branch is correct. Note that, if we cut a tree at an inner branch, this will generate two disjoint subtrees and hence two disjoint sets of species. Thus, the support value for the specific inner branch tells us how confident we are that the inferred split is correct. Those support values are essential for the interpretation of trees and for drawing biological conclusions about the evolution of the species we are analyzing. Note that it is nowadays almost impossible to publish phylogenetic trees without such support values.

As a consequence, the incorrect display of support values on trees, that is, incorrect mapping of support values to inner branches may have an impact on the interpretation of phylogenetic analyses.

The problem we have unraveled (see [Czech and Stamatakis, 2015]) is associated with rather unfortunate data format definitions and implicit meta-data semantics. Presumably for historical reasons, branch support values are typically stored as node labels in the widely-used Newick format for storing phylogenetic trees.

However, as already pointed out, support values are attributes of branches (splits) in the unrooted phylogenetic trees generated by common tree inference programs. Therefore, storing support values as node labels can potentially lead to incorrect support-value-to-bipartition mappings when manually rooting or re-rooting those initially unrooted trees in tree viewers. Note that, in most publications, the mathematically unrooted tree objects are displayed as rooted trees based on some biological background knowledge.

The correctness of the support-value-to-bipartition mapping after re-rooting depends on the largely – and unfortunately – implicit semantics of tree viewers for interpreting node labels. To assess the potential impact of these ambiguous and predominantly implicit semantics of support values, we analyzed 10 different popular tree viewers.

We found that most of them exhibit incorrect or unexpected behavior of some kind when re-rooting trees with support values.

We also found that the popular Dendroscope tool (over 1000 citations, Google Scholar January 2016) interprets Newick node labels as simply that, i.e., node labels in Newick trees. However, if they are meant to represent branch support values, the support- value-to-branch mapping is incorrect when trees are re-rooted with Dendroscope.

Unfortunately, we were not able to investigate more than two empirical studies using Dendroscope for visualization.

We initially contacted the authors of 14 papers citing Dendroscope, published in journals such as Nature, PLOS, BMC, and JBC. Of the authors contacted, five replied, but only two were ultimately able to provide us with the trees used to generate the visualizations in their publications.

This was a frustrating experience with respect to the reproducibility of science, since all we were doing was to ask the authors of published papers to provide us with plain text files of the trees they had published.

Of the two trees we did receive, one was correctly displayed, but the other was incorrectly displayed regarding the support values. We illustrate the incorrect support value mapping produced by Dendroscope for this phylogenetic study in *Figure 56.* In sub-figure (a) we display the original Newick tree used to generate the figure in the respective paper.

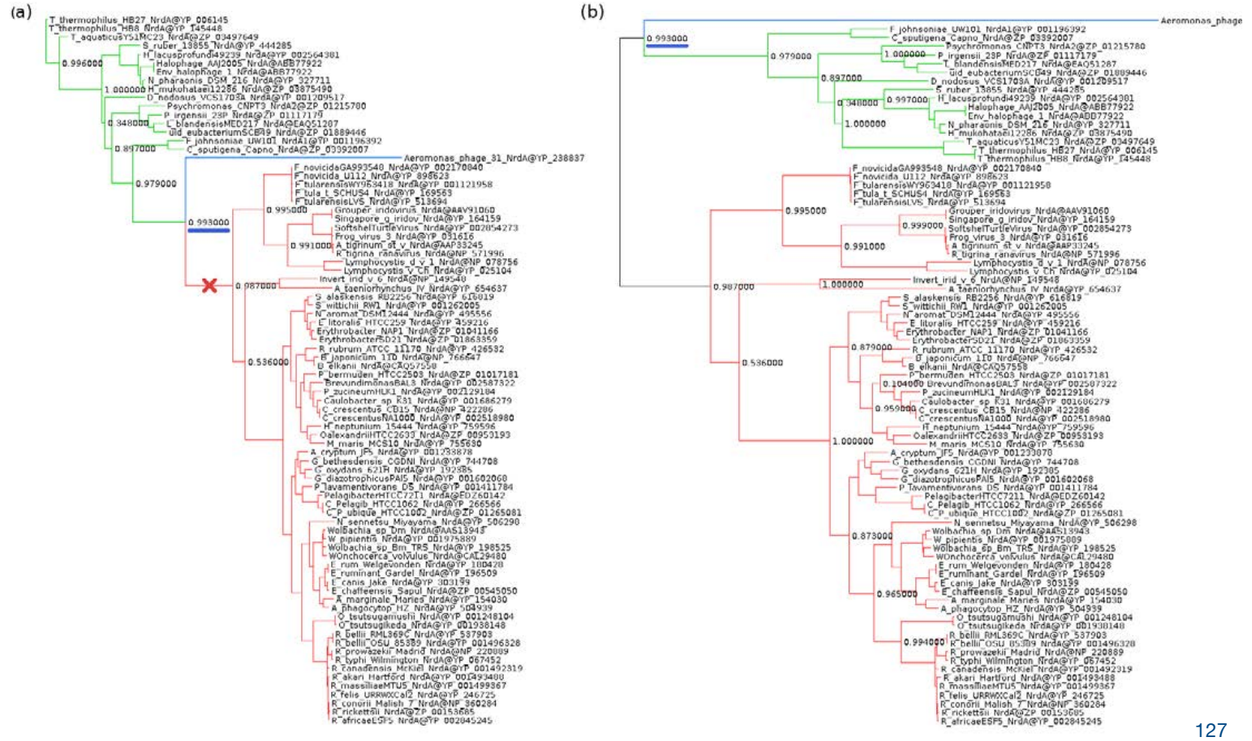


Figure 56: Visualizing the original phylogenetic tree from an empirical phylogenetic study. Sub-figure (a) shows the original tree with the branch used for re-rooting marked by a red cross. Sub-figure (b) shows the re-rooted tree with incorrectly placed branch support values (e.g., the underlined one). We have colored appropriate subtrees in red, blue, and green to highlight their respective positions after re-rooting.

We have marked the branch used for re-rooting the tree with a red cross. We have also colored the subtrees so that their corresponding positions in the re-rooted tree are easy to see. Sub-figure (b) shows the re-rooted tree, which is topologically identical to the one presented in the empirical paper. However, the branch support values between the old and the new root node in our *Figure 56* are not mapped to the same inner branch in sub-figures (a) and (b). For example, in sub-figure (a) the underlined support value refers to the split between the green species and the subtree comprising the blue and red species, whereas in sub-figure (b) it refers to the split between the red species and the subtree containing the green and blue species.

Fortunately, this incorrect mapping did not have any impact on the conclusions of the paper (pers. comm. by the author), and the author of Dendroscope has already promised to work on a fix. Also, part of the user community has been alerted to this tree viewer issue via social media.

As a general solution, we suggest that (i) branch support values should be stored exclusively as meta-data associated to branches (and not nodes), and (ii) if this is not feasible, tree viewers should include a user dialogue that explicitly forces users to define whether node labels are to be interpreted as node or branch labels, prior to displaying the tree.

Detecting and correcting mislabeled sequences

DNA sequences that are stored and freely accessible in public databases are mostly annotated by the submitting authors without any further validation. This setup can generate so-called erroneous taxonomic sequence labels. To provide an extreme example, a sequence that was actually obtained from a mouse might be annotated as a sequence belonging to a virus, or vice versa. Such erroneous taxonomic sequence labels are also commonly called mislabeled sequences.

They constitute a real data curation problem that exponential DNA data accumulation makes extremely hard to resolve (see, for instance, the following popular science article: http://the-scientist.com/?articles.view/articleNo/41821/title/Mistaken-Identities/).

Mislabeled sequences difficult to identify and can induce downstream errors because new sequences are typically annotated or identified by using existing ones (and often using the pairwise alignment methods mentioned above).

Furthermore, the taxonomic mislabelings in reference sequence databases can bias so-called meta-genetic studies that rely on the correctness of the reference taxonomy. In meta-genetic studies, one may, for instance, sample the DNA of all bacterial sequences in the human gut. Thus, one ends up with a huge number of anonymous bacterial sequences (typically more than 100,000) that need to be taxonomically classified to disentangle the bacterial diversity of the gut.

Despite significant efforts to improve the quality of taxonomic annotations, the curation rate is low because so far the manual curation process has been so labor-intensive.

In this project, we developed the SATIVA open-source software (available at http://sco.h-its.org/exelixis/web/software/sativa/index.html), which is a method that uses phylogenetic approaches to automatically identify taxonomically mislabeled sequences (also called "mislabels" for short) using statistical models of evolution.

We adapted our well-established Evolutionary Placement Algorithm (EPA) to detect sequences whose taxonomic annotation is not consistent with the underlying phylogenetic signal. Also, SATIVA can automatically propose a corrected taxonomic classification for the putative mislabels it has identified.

By using simulated data with deliberately mislabeled sequences, we showed that our method achieves a high degree of accuracy, both in identifying and correcting mislabels. Furthermore, an analysis of four de-facto standard microbial reference databases (Greengenes, LTP, RDP and SILVA) revealed that they currently only contain between 0.2% and 2% of mislabeled sequences. This rather low error rate indicates that microbial databases are well-curated, which is a relief. However, for larger and less well-curated databases, such asGenBank, we believe that our tool will help to efficiently identify and correct a larger number of putative mislabels. The main contributions of SATIVA are: (i) we can economize on curation man-hours, and (ii) we provide an evolutionary framework for detecting mislabels that takes into account the phylogenetic signal. For ease of use, we have also integrated our tool into the popular ARB software for microbial data analysis. *Figure 57* displays a phylogenetic tree with two mislabels identified by SATIVA and displayed in ARB.

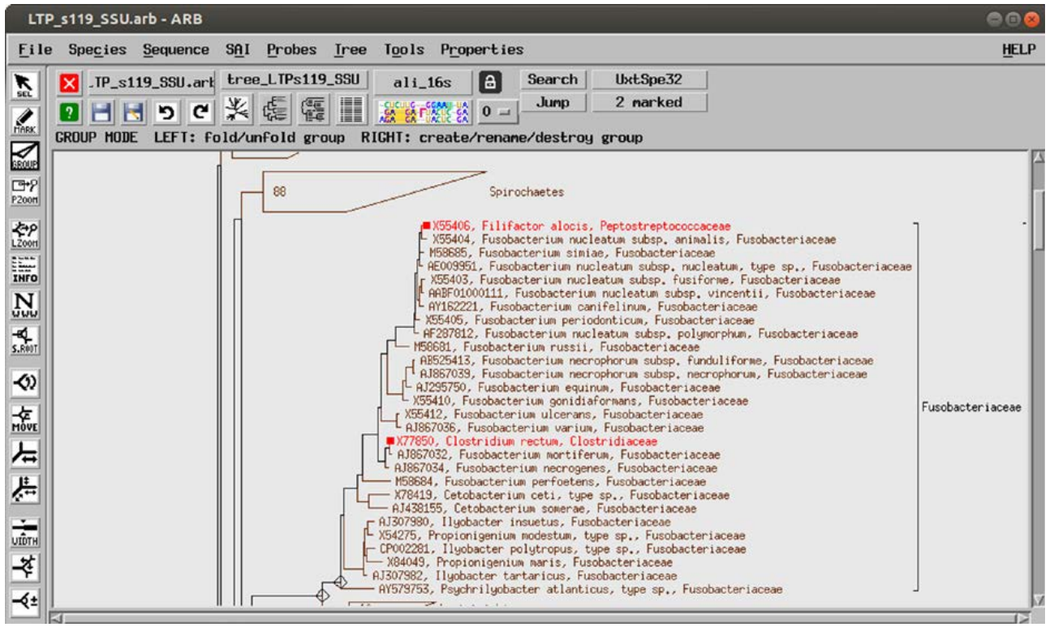


Figure 57: SATIVA as integrated into ARB. The ARB tree viewer has highlighted two putatively mislabeled sequences identified by our algorithm.

2 Research

2.11 Scientific Databases and Visualization (SDBV)

2.11
Scientific Databases and Visualization (SDBV)

Our main interest is the structuring and curation of biological data, i.e. making data useful. Curated data can be reused and contextualized. We provide data-management services for science based on the systems SEEK and Excemplify that we are in the course of developing. These enable users to curate their own data and exchange it with colleagues. We also provide SABIO-RK, a database containing curated biochemical reaction kinetics data. These different viewpoints on data curation enable us to refine our services and do basic research on them.

The sister of curation is standardization. Standardization seeks to add value to data by defining reference structures and thus facilitating curation. This connection motivates our strong presence in standardization initiatives and projects like NormSys.

We provide these services in a variety of projects that confront us with different aspects of data-management problems and enable us to adapt and diversify our software accordingly. In terms of projects, the main changes in 2015 were the end of the Virtual Liver Network and the start of de.NBI, the German Network for Bioinformatics Infrastructure.

In the past two years we have started working towards wider scientific dissemination, notably with the Discover the Liver portal providing basic information on the liver and discoveries from the Virtual Liver for laypersons and the Operations Explorer that provides hospital data for journalists. Both pose questions that are interesting for the core concern behind our work: How can we make data reusable and easy to understand?

Der Schwerpunkt unserer Arbeit liegt in der Strukturierung und Kuratierung von Daten und somit in deren Nutzbarmachung. Kuratierte Daten können wiederverwertet und in Beziehung gesetzt werden. Wir bieten Datenmanagementdienste für Wissenschaftler an, auf der Grundlage der in unserer Gruppe entwickelten Systeme SEEK und Excemplify. Diese ermöglichen den Nutzern, ihre Daten zu kuratieren und mit Kollegen auszutauschen. Außerdem bieten wir mit SABIO-RK eine kuratierte Datenbank für biochemische Reaktionskinetik-Daten. Diese unterschiedlichen Blickwinkel auf Datenkuratierung ermöglichen uns, die Services zu verbessern und die Grundlagen für diese Services zu erforschen.

Die Schwester der Kuratierung ist die Standardisierung. Standardisierung ermöglicht Kuratierung durch die Definition von Referenz-Strukturen für Daten. Diese Verbindung motiviert unsere Präsenz in Standardisierungsinitiativen und Normungsorganisationen, beispielsweise im Rahmen des NormSys-Projekts.

Wir bringen unsere Services in verschiedene Projekte ein. Dadurch lernen wir neue Aspekte der Datenmanagementprobleme kennen und können die Funktionalität unserer Software anpassen und erweitern. Die wichtigsten Änderungen bezüglich der Projekte im Jahr 2015 waren das Ende des Virtuellen Leber Netzwerks, sowie der Beginn von de.NBI, dem Deutschen Netzwerk für Bioinformatik-Infrastruktur.

In den letzten zwei Jahren haben wir uns auch verstärkt der Verbreitung wissenschaftlicher Inhalte für die Öffentlichkeit gewidmet und Plattformen dafür entwickelt: Das Discover the Liver-Portal mit Wissenswertem zur Leber und neuen Erkenntnissen aus der Virtuellen Leber für Laien, und der Operations Explorer mit Krankenhaus-Daten für Journalisten. Beide Systeme lassen uns Fragen stellen, die für unser Kerninteresse wichtig sind: Wie machen wir Daten wiederverwendbar und einfach verständlich?

Group leader

Priv.-Doz. Dr. Wolfgang Müller

Staff members

Lihua An *(until July 2015)*
Martin Golebiewski
Ron Henkel *(since July 2015)*
Dr. Iryna Ilkavets *(until December 2015)*
Renate Kania
Dr. Olga Krebs
Quyen Nguyen
Dr. Maja Rey
Ivan Savora *(until March 2015)*
Dr. Andreas Weidemann
Dr. Ulrike Wittig

Students

Alexander Nikolaew *(since May 2015)*
Jill Zander
Zhen Dong

Infrastructures for systems biology

In the interdisciplinary field of systems biology, researchers from academia, hospitals, and industry work together and compile data originating both from different experimental technologies (genomics, transcriptomics, proteomics, metabolomics, functional assays, etc.) and from the treatment of patients. The aim behind this is to set up computer-simulatable models of organisms, their parts, and their functions that can help to better unravel the biological processes going on in cells, tissues, organs, and the entire organism. Examples for such consortia, where experimentalists, theoreticians, and clinicians work together to construct and simulate complex computer models are large-scale research initiatives like the German Virtual Liver Network (VLN: http://www.virtual-liver.de) and European networks like ERASysAPP (ERA-Net for Systems Biology Applications) or NMTrypI (New Medicines for Trypanosomatidic Infections).

The core concern of systems biology is to obtain, integrate, and analyze complex data sets from multiple sources. All these data and models and the corresponding metadata (data describing the data) have to be consistently structured, stored, compared, interrelated, and finally integrated to assemble the pieces of the jigsaw puzzle and obtain the complete picture. This calls for web-accessible data management infrastructures enabling collaborators to share and exchange information and data as and when it is produced, throughout the entire iterative cycle of experimentation, modeling, and exploitation of the results obtained. In conjunction with its collaboration partners, SDBV develops and maintains data-management platforms that help to structure, exchange, integrate, and publish experimental data, models, workflows, and additional information pertaining to them. The group is involved in major national and European consortia that either aim to provide the infrastructural backbone for systems biology research (like ISBE, FAIRDOM or de.NBI) or are doing research of their own based on the infrastructure provided by our group (like VLN, SBEpo or NMTrypI). The group not only develops the technical infrastructure for this collaboration, it also provides curation and data stewardship as a service.

ERASysAPP & FAIRDOM

With partners in Germany, Switzerland, and the UK, the joint project FAIRDOM (www.fair-dom.org) was set up to establish a pan-European infrastructure amalgamating data- and model-management expertise for systems biology and offering this as a service to the scientific community. In this project, which is jointly funded by BMBF (D), BBSRC (UK), and SystemsX (CH), we build on the experience and expertise of two systems biology management platforms: SEEK [Wolstencroft et al., 2015], mainly developed in collaboration between SDBV and partners at the University of Manchester, and openBIS, mainly developed by our partners in Switzerland.

FAIRDOM aims to provide the infrastructure for "Findable, Accessible, Interoperable, Reusable" (FAIR) data shared within and between systems biology research networks and operates two parallel activities: (a) the Technical Development Track to create and support the joint openSEEK platform, incorporated tools, and modeling services and (b) the Community Track to establish a pan-European community, to deliver support services for the ERASysAPP research network (www.erasysapp.eu), to build a knowledge network, and to provide training activities.

As a central instance of the openSEEK software, the FAIRDOMHub (www.fairdomhub.org) has been adopted and implemented by several research networks for their data- and model-management. All current ERASysAPP projects and former SysMO (Systems Biology of Microorganisms) research groups have access to their shared assets (data, models, SOPs, etc.) through FAIRDOMHub. However, the system has also been well adopted beyond the research initiatives that we are part of: 36 projects, 136 institutes, and 535 people have been registered so far. During 2015 we visited all EraSysAPP projects and presented FAIRDOM offerings, refined data-management plans, and steps towards realization. In addition, we organized two meetings with our data-management contacts within the projects (PALs). At each meeting, new requirements or suggestions for improvement were collected and funneled back into the development process.

The SDBV group is responsible for hosting, maintaining, and administrating FAIRDOMHub and various project-specific uses of SEEK instances. This year, we performed several improvements to SEEK, including enhancements of modularity and system architecture plus a complete overhaul of look and feel to modernize and update the system and the underlying software framework. SEEK has been extended to support the export and coherent packaging of structured content compatible with the Research Object specifications (www.researchobject.org). Export from SEEK as Research Objects enables the exchange and migration of content between SEEK installations as well as the snapshotting of content for publication supplements or archiving in general public repositories. Additional SEEK developments for enhanced functionality include the creation of Digital Object Identifiers (DOI) directly from SEEK for published assets (e.g. to refer to this content in a publication) using the DataCite service, as well as the implementation of a workflow for publishing content. Together with the improvements mentioned, the new SEEK 1.0 release from December 2015 is also capable of supporting multiple autonomous research projects and independent self-management.

The core concern of systems biology is to obtain, integrate, and analyze complex data sets from multiple sources. All these data and models and the corresponding metadata (data describing the data) have to be consistently structured, stored, compared, interrelated, and finally integrated to assemble the pieces of the jigsaw puzzle and obtain the complete picture.

German Network for Bioinformatics Infrastructure (de.NBI)

The German Network for Bioinformatics Infrastructure initiative (http://www.denbi.de) funded by the Federal Ministry of Education and Research (BMBF) started in March 2015. The kick-off meeting in Bielefeld, where the Coordination and Administration Unit (CAU) of the network is located, assembled the 23 project partners organized in eight Service Centers. Three service centers deal with medical, microbial, and plant genomics, two are involved in proteomics and RNA bioinformatics. One service center specializes in integrative bioinformatics, two others concentrate on data management and data collection. The decision-making body of de.NBI is the Central Coordination Unit (CCU) made up of the de.NBI coordinator and the unit coordinator of each of the eight Service Centers *(Figure 58)*. They are supported by recommendations from five Special Interest Groups (SIGs). These SIGs are small discussion groups of representatives from all de.NBI units working on topics related to web presence, service and service monitoring, training and education, infrastructure and data management and de.NBI development.

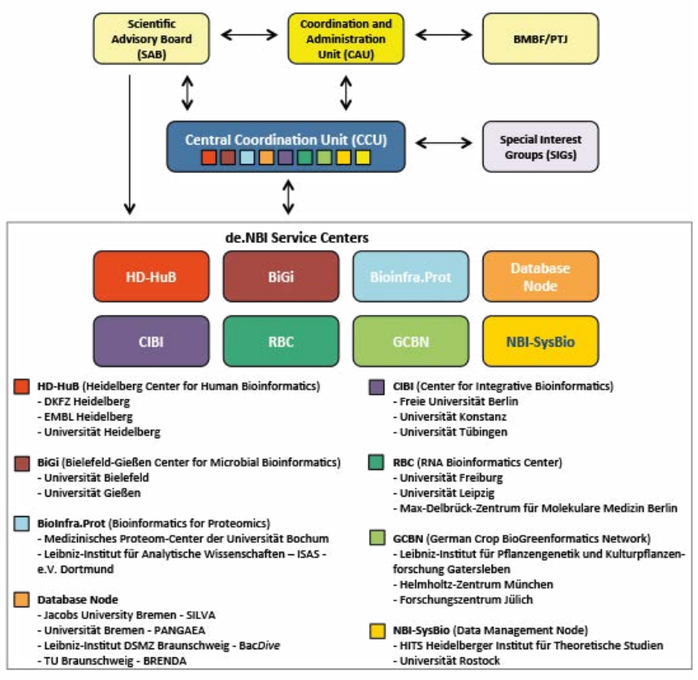


Figure 58: Organigram of the de.NBI project.

In conjunction with the SEMS group at the University of Rostock, SDBV represents the Data Management Node (NBI-SysBio) of the de.NBI network. The central mission of our Service Center is to support the systems biology cycle with standardized data-management solutions *(Figure 59)*. Accordingly, a de.NBI-program of our own as part of FAIRDOM Hub (https://fairdomhub.org/programmes/6) was set up to provide a platform (SEEK) for the German systems biology community and others on which to manage, enrich, and share data, models, and SOPs (Standard Operating Procedures) confidentially. In addition, NBI-SysBio administers the SEEK software and the services assisting all kinds of systems biology projects making full use of the platform.

To improve the support for storing, version control, and the search for models in SEEK, Ron Henkel in Rostock is currently working on the development of MaSyMoS (Management System for Models and Simulations) and its integration into SEEK. MaSyMoS is a tool for seeking out computational models and simulation descriptions and is currently being extended to work with visual representations of models encoded in SBGN (Systems Biology Graphical Notation). A strategy has been developed for the integration of SED-ML (Simulation Experiment Description Markup Language) into SEEK via an extension of the simulation tool JWS Online Simulator and for storing SED-ML files via the MaSyMoS graph representation.

Furthermore, NBI-SysBio supports the systems biology cycle in finding curated and structured kinetic data needed for modeling by offering access to our biochemical reactions database SABIO-RK.

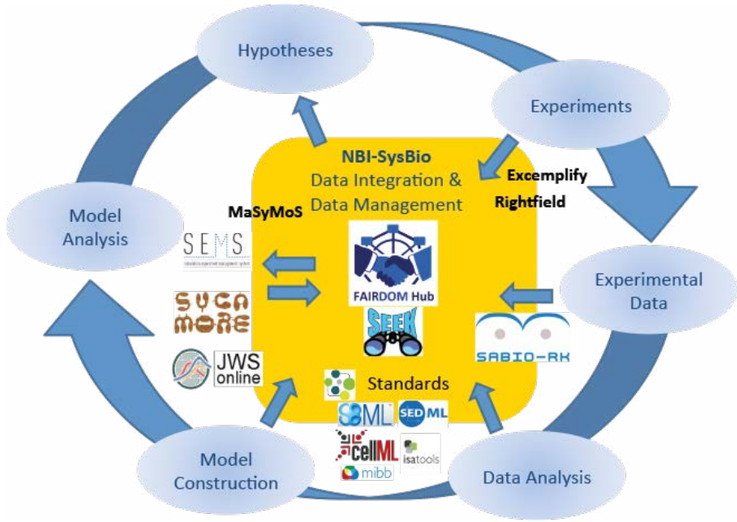


Figure 59: Role of the Data Management node NBI-SysBio in the systems biology cycle.

SABIO-RK

SABIO-RK (http://sabio.h-its.org) is a web-accessible, manually curated database that has been established as a resource for accessing quantitative data describing kinetic properties of biochemical reactions. Its special focus is on supporting modelers in creating simulatable computer models of biochemical reaction networks. SABIO-RK's data are mainly based on information reported in the scientific literature, manually extracted by experts, and stored in a structured format. In addition, SABIO-RK also offers direct automatic data upload from lab experiments and from computer models. The database contains annotations to controlled vocabularies and ontologies. It is interlinked with other biological databases. A flexible way of exporting database search results in a table-like format is provided. Users can tailor their own custom-made export format by selecting properties of the entries in the result set the user wants to export. Both the export of data (different datasets can be exported together) and the import of data (e.g. from computer models) are possible via the SBML (Systems Biology Markup Language) format.

One task of de.NBI project has been to extend the existing ways in which SABIO-RK collects feedback from users with a view to improving the SABIO-RK database content and complying more efficiently with user requirements. In the case of an empty query result, the user is presented with an opportunity to directly request addition of the corresponding data. Similarly, the user can also proactively ask via SABIOs Services for curation of individual papers or, say, pathway- and organism-associated data. These user requests are visible to the public as a curation priority list *(Figure 60, next page)*. Additionally, database queries are now logged for measuring purposes.

The database now (as of December 2015) contains more than 53,000 different entries related to 5,232 publications from about 270 different journals. The list of kinetic parameters comprises more than 41,600 velocity constants (e.g. Vmax, kcat, rate constants), about 43,000 Km or S_half values, and about 12,000 inhibition constants (Ki and IC50).

Figure 60: SABIO-RK Curation Service with link to a request form for users.

Infrastructure for Systems Biology Europe (ISBE)

The SDBV group has been participating in the preparatory phase of ISBE (http://project.isbe.eu), a pan-European project aimed at providing and maintaining a sustainable, integrated, and distributed infrastructure for systems biologists all over Europe. Key areas of support within ISBE will consist, broadly speaking, of high-end expertise in modeling and data-generation technologies, and the storage, access, and integration of data and models produced from systems approaches.

The primary output of the preparatory phase (2012 to 2015) was a comprehensive business plan describing infrastructure operation and funding, supplemented by detailed strategies for the upcoming implementation and running phase of ISBE. Data and model management turned out to be the vital core in such an infrastructure, as data structuring, standardization, and integration is regarded as one of the main tasks. Together with UK partners at the University of Manchester, the European Bioinformatics Institute (EMBL-EBI), and Imperial College London, we took part in the activities of the model- and data-management work package. This was responsible for surveying the state of the art and best practices for model and data management in systems biology, collaborating in standardization activities, and finally developing a strategic implementation plan for sustainable and integrated data and model management.

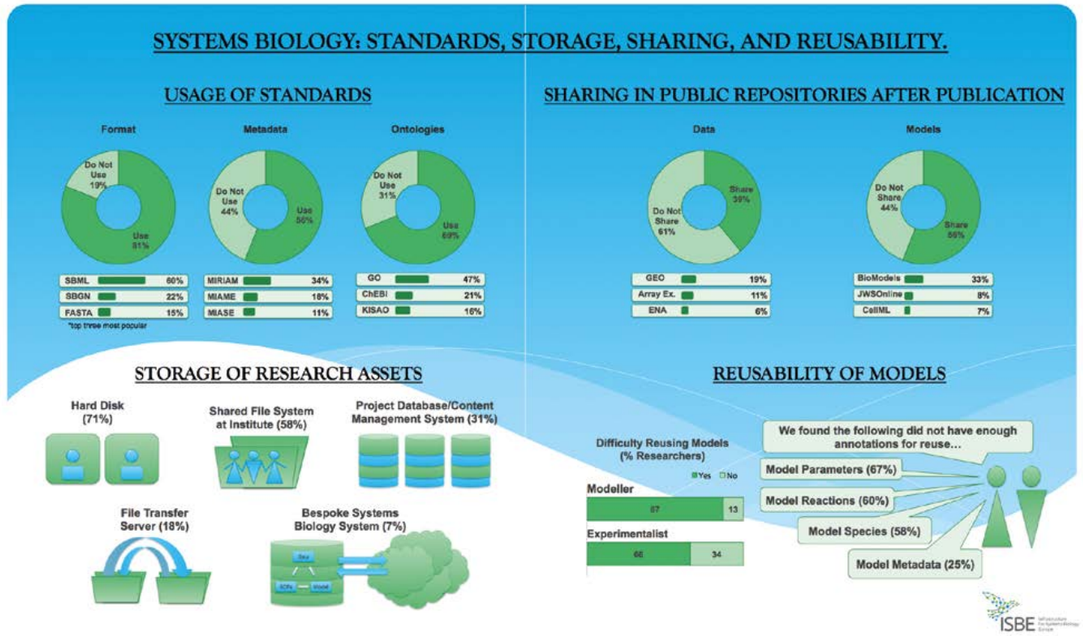


Figure 61: Summary of the ISBE survey results.

In collaboration with our work-package partners, we surveyed the community to evaluate the uptake of available standards and current practices of researchers in data and model management. For this purpose, we designed a questionnaire that was widely distributed in the systems biology community and addressed the following four key areas:

- standards usage
- data and model storage before publication
- sharing in public repositories after publication
- reusability of data, models, and results.

The outcome of the survey based on our analysis of the replies is valuable for the implementation of ISBE. End 2015, the results were successfully published in Molecular Systems Biology [Stanford, Wolstencroft, Golebiewski, Kania, et al., 2015]. *Figure 61* gives a summary of the survey results.

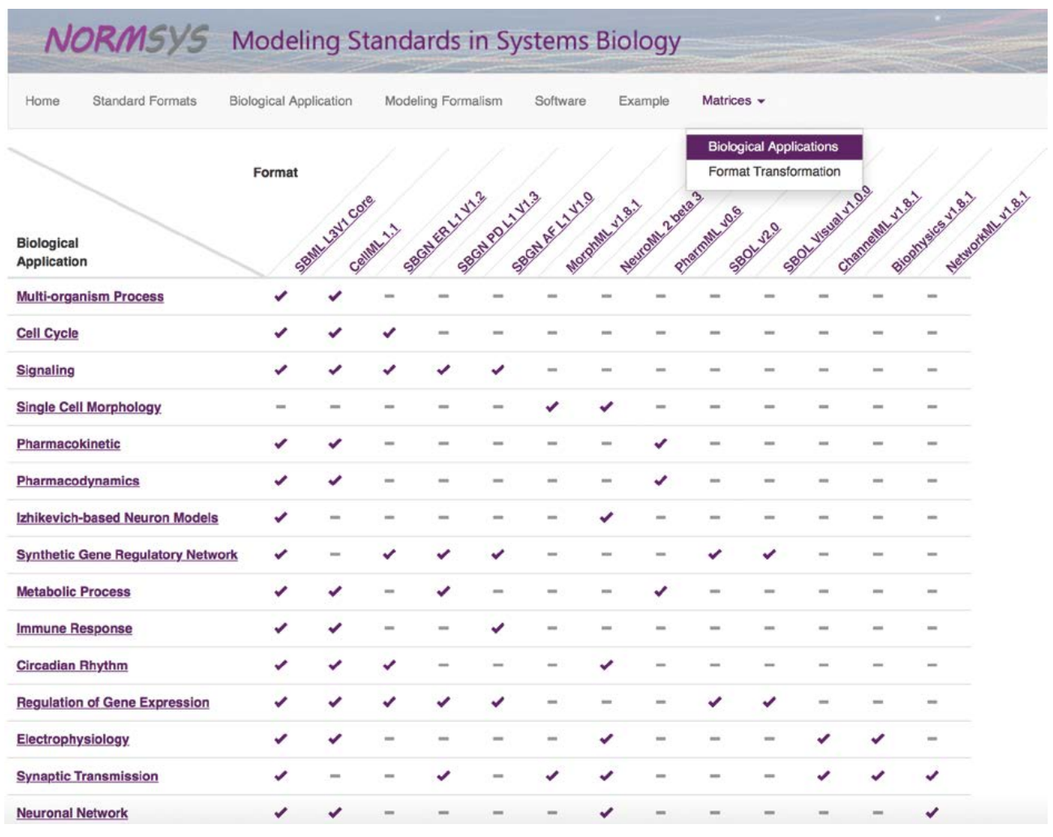
Modeling standards in systems biology (NormSys)

In systems biology, consistent structuring and description of data and computer models with their underlying experimental or modeling processes is only possible by applying interoperable standards for formatting and describing data, workflows, and models, as well as the metadata describing the interconnection between all these. The NormSys consortium (http://www.normsys.de), which we coordinate, brings different stakeholders together to further harmonize and unify standardization in systems biology: researchers from academia and industry with their grass-roots standardization initiatives like the Computational Modeling in Biology Network (COMBINE: http://www.co.mbine.org), their scientific journals and research funding agencies, and standardization institutions like DIN or ISO (International Organization for Standardization). To factor in requirements from different stakeholders, we collaborate with partners from academia (University of Potsdam) and private enterprises (LifeGlimmer GmbH, Berlin). The project is funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) and aims to enhance and promote the formal standardization of existing community standards for modeling by bridging existing gaps between stakeholder groups.

To achieve this goal, we actively contribute both to initiatives like COMBINE [Hucka et al., 2015], where we are on the board of coordinators and to the committee work of standardization bodies like the Technical Committee for Biotechnology Standards at ISO (ISO/TC 276) and its national German counterpart at DIN. In April 2015, Martin Golebiewski was appointed to head the international work-group for "data processing and integration" (WG5) of ISO/TC 276. He has already presided over its German counterpart at DIN since 2014. Our NormSys collaboration partners are also contributing to the work of these groups. In this context, we are currently working at ISO to create an international framework standard referring to existing standards for data processing and modeling in the life sciences, in order to establish an ISO standard as a hub for such community standards and provide a guideline for their application.

A prerequisite for this ISO standard is a comprehensive inventory of community standards for modeling in systems biology and related fields [Schreiber, Bader, Golebiewski et al., 2015]. To survey these, we are developing the NormSys registry for modeling standards (http://normsys.h-its.org), which was first released in October 2015 (Figure 62). It provides a single access point for consistent information about model exchange formats such as SBML, CellML, SBGN, SED-ML, NeuroML for neuroscience models, SBOL (Synthetic Biology Open Language), the Pharmacometrics Markup Language (PharmML), and others. The publicly available platform not only lists the standards, but also compares their major features, their potential fields of biological application and potential use cases (including model examples), as well as their similarities, relationships, commonalities, and differences. This NormSys registry is a convenient information source, especially for experimentalists, modelers, and software developers planning to apply the standard formats for their different purposes. It provides them with detailed information, plus links to the webpages, specifications, and web services associated with the formats.

Figure 62: The NormSys registry provides details about modeling standards in systems biology and indicates potential fields of biological application.



New Medicines for Trypanosomatidic Infections (NMTrypI)

NMTrypI (http://www.nmtrypi.eu/) is an EU-funded project assembling experts from 14 partners in Europe and disease-endemic countries. Its aim is to develop new innovative drugs against Trypanosomatidic infections i.e. sleeping sickness, leishmaniasis, and Chagas disease. Within the NMTrypI project, SDBV is responsible for the central data management. SEEK is used for this purpose, and additional features are being developed to meet the special requirements for handling chemical compound information within this project.

New features include the automatic detection of UniprotKB accession numbers based on regular expressions. These UniprotKB identifiers are highlighted in Excel tables and a popup menu is available offering links to the Uniprot database (http://www.uniprot.org/) and StringDB (http://string-db.org/). For chemical compounds detected in Excel tables based on project internal nomenclature, a popup menu is shown that contains a search option or a link with a view to creating a compound summary report.

The compound summary report collects all data for one compound from all accessible Excel tables and displays this information in one single document. Each user gets a compound report that reflects the data they are allowed to see. A small symbol next to the compound identifier links up with the visualization of the chemical compound by drawing the chemical structure based on an available SMILES string.

To visualize the data of an Excel table, an interactive heat map can be created by selecting different columns in the table. For better representation, the intervals between the data values can be modified by a slider *(see Figure 63, next page)*.

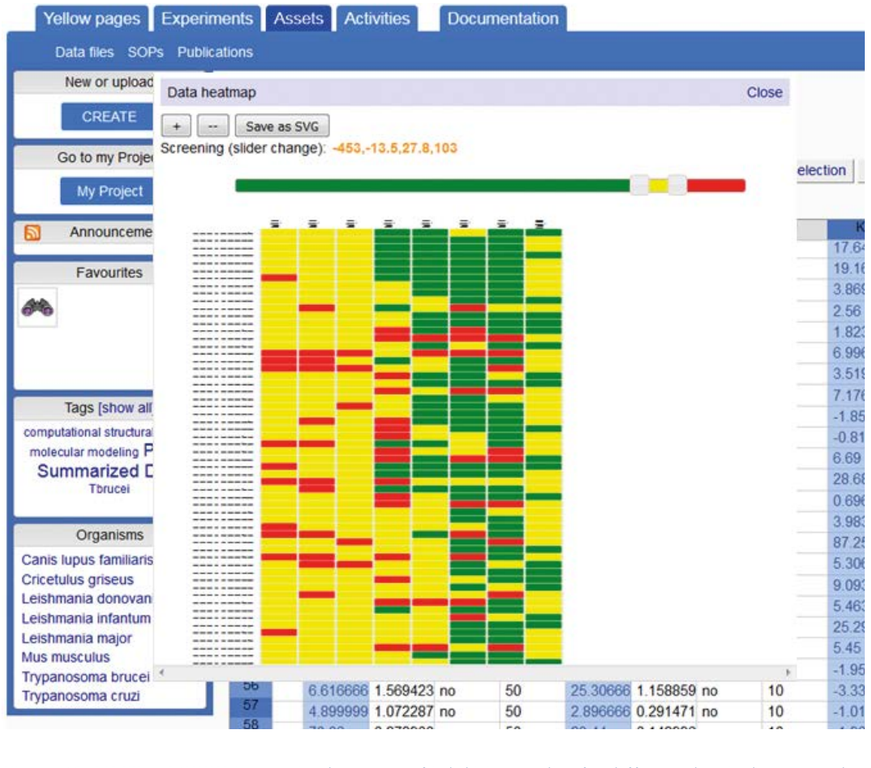


Figure 63: NMTrypI SEEK interactive heat map of inhibition studies for different chemical compounds.

Operations Explorer

The Operations Explorer has been developed to support data journalists in their investigative work on the German health system. The project was initiated in collaboration with Volker Stollorz (Science Media Center, Cologne) and was initially funded by the Robert Bosch Foundation. The project is maintained in conjunction with the WDR (Westdeutscher Rundfunk, West German Broadcasting) and the TV- and movie-production company Längengrad in Cologne. The Operations Explorer enables the browsing and interlinking of statistical data from different origins, including data from ICD-10 (International Classification of Diseases) and OPS ("Operationen- und Prozedurenschlüssel"), taxonomic data, district population data, and socio-economic data.

In 2015 the search and visualization capabilities of the Operations Explorer were further extended by the development of two additional modules: processing of data from German hospitals (IQM data (Quality Medicine Initiative) and processing of data sets from the AOK, one of the biggest German compulsory health insurance funds ("Allgemeine Ortskrankenkasse").

Inspired by the Operations Explorer, we began to think seriously about how to best visualize so-called bivariate maps. Bivariate maps are used to accurately and graphically illustrate the relationship between two variables. The most common version of bivariate maps is choropleth maps (i.e. maps colored/textured to visualize numbers such as density of diagnoses), which visualize two-dimensional data. Experimenting with such maps, the subjective impression that remained was that our first go at bivariate maps was rather hard to read.

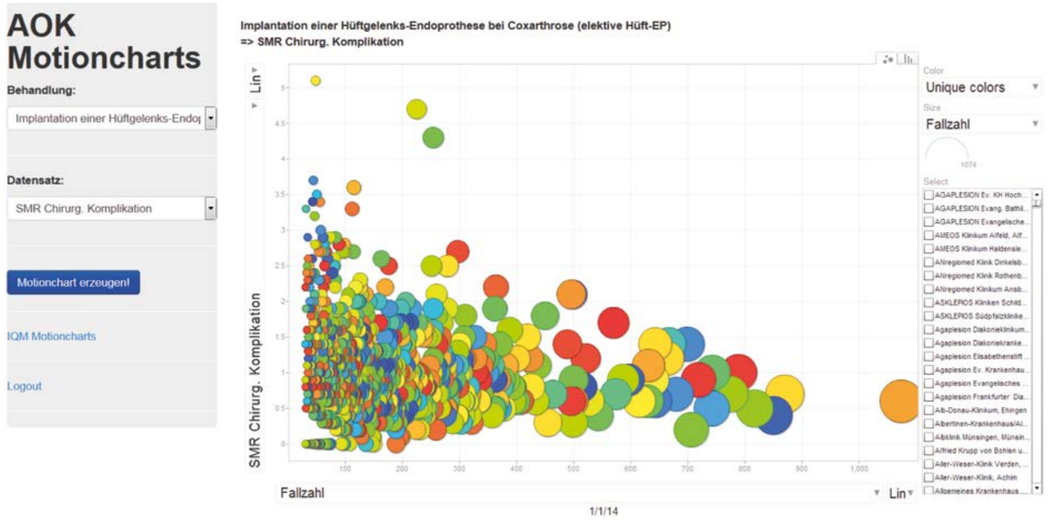


Figure 64: Visualization of statistical data on the German health system as bubble motion charts. On the left navigation panel, patient treatment can be selected in combination with the criteria under investigation (in this example, implantation of a hip-joint endoprosthesis versus standard mortality rate, SMR). Based on the underlying data set, an animated image is generated displaying the values for each hospital. The x-axis displays the number of treatments, the y-axis the SMR values. The size of the bubbles also corresponds to the number of cases. As a third dimension, change over time can be viewed ('motion' of the bubbles). Note the three bubbles in the upper left part (yellow, pale green, bright green). At these hospitals some 50 to 230 hip-joint endoprostheses were implanted in 2014. However, the standard mortality rate at these hospitals was much higher (4.3 – 5.1) compared to the average standard mortality rate (about 0.8 – 1.0). Moving the mouse pointer over the bubbles reveals the exact SMR values and case numbers as well as the name of the hospitals.

In a project conducted in conjunction with DHBW Mosbach, Jill Zander is investigating the type of bivariate map that would best fit our purposes. An initial set of experiments has been performed based on maps from the extensive literature. These yielded a preferred bivariate map type. However, during the experiments it became clear that the preferred bivariate map type probably also depends on the use case and the specific user group. This will be part of a new set of experiments followed by user tests involving the journalists using the Operations Explorer.

Prospects

2015 was an exciting year, with big new projects demonstrating the impact of the group as an important European partner for systems biology data management. 2016 will see some more mature projects (NormSys, FAIRDOM, de.NBI), some phasing out (NMTrypI, SBEpo), and the start of the German LiSyM network (Liver Systems Medicine). This will build on the success of the Virtual Liver Network, which ended in 2015 after more than five years of funding.

Our long-term commitment to ensuring the accessibility of data and metadata over and above the lifetime of project-funding periods will be an important prerequisite for sustainable research in systems biology based on the results achieved so far. Naturally, our ongoing work serving the scientific community with high quality biochemical data in SABIO-RK and the curation efforts underlying it will continue as before.

We shall be facing interesting challenges, from software sustainability to structuring interfaces between clinical and non-clinical data, from self-curation to professional curation, and from server-side data manipulation to immediately understandable in-browser data visualization.

2 Research

2.12 Theoretical Astrophysics (TAP)

2.12
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 shortly after the Big Bang until today. We are especially interested in how galaxies form, ultimately producing magnificent systems like our own Milky Way, a busy megalopolis with more than a hundred billion stars. We also aim to constrain the properties of dark matter and dark energy, the two enigmatic matter and energy components that dominate today's universe and pose some of the most fundamental problems in modern physics.

A prominent role in our work is played by numerical simulations on a variety of scales, both of the collisionless and the hydrodynamic type. To this end, we develop novel numerical schemes that can be used efficiently on very large supercomputers, with the goal of exploiting them to their full capacity in our attempt 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 its comparison with theoretical models. With our simulations, we can focus our research on how diverse physical processes relevant in structure formation interact in a complex and highly non-linear fashion. A current priority in our group is to incorporate aspects of physics into our models that, though known to be important, have frequently been neglected in the past. These include supermassive black hole formation, cosmic rays, or radiative transfer. In this report we highlight some of the results we have come up with in the past year.

Die Theoretische Astrophysik Gruppe am HITS versucht die Physik der kosmischen Strukturentstehung während der letzten 13.5 Milliarden Jahre, vom Urknall bis heute, zu verstehen. Unser besonderes Interesse gilt der Entstehung von Galaxien, welche schließlich zur Bildung von großartigen Systemen wie unserer Milchstraße führt, einer geschäftigen Metropole mit mehr als einhundert Milliarden Sternen. Wir arbeiten auch an einer Bestimmung der Eigenschaften der Dunklen Materie und der Dunklen Energie, jenen rätselhaften Komponenten, die den heutigen Kosmos dominieren und die zu den fundamentalsten Problemen der modernen Physik gehören.

Eine besonders wichtige Rolle in unserer Arbeit spielen numerische Simulationen auf verschiedenen Skalen. Zu diesem Zweck entwickeln wir 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.

Mit der Hilfe von Simulationen sind wir insbesondere in der Lage, das komplexe und nichtlineare Zusammenspiel verschiedener physikalischer Prozesse zu studieren. Eine aktuelle Priorität in unsere Gruppe besteht darin, Physik in unsere Modelle einzubauen, die zwar als wichtig erachtet wird, die aber bisher vernachlässigt wurde, etwa superschwere Schwarze Löcher, kosmische Strahlen oder Strahlungstransport. In diesem Bericht stellen wir beispielhaft einige Ergebnisse unserer Arbeit im vergangenen Jahr vor.

Group leader

Prof. Dr. Volker Springel

Postdocs

Dr. Robert Grand
Dr. Rüdiger Pakmor
PD Dr. Christoph Pfrommer
Dr. Christine Simpson
Dr. Dandan Xu
Dr. Martin Sparre *(since March 2015)*

Students

Andreas Bauer
Kevin Schaal
Rainer Weinberger
Christian Arnold
Jolanta Zjupa
Svenja Jacob *(since Nov. 2015)*
Sebastian Bustamante *(since Oct. 2015)*
Denis Yurin *(until Feb. 2015)*

Visiting scientist

Lee Rosenthal *(since Sept. 2015)*

Formation and evolution of disk galaxies

Among the most important and interesting objects populating the Universe are so-called disk galaxies. They are made up of stars, dust, and interstellar gas, and this material is organized in an approximately disk-like shape that spins around its symmetry axis and is embedded in a dark matter halo. The most prominent example of this type of object is our own galaxy, the Milky Way. For decades, the formation of disk galaxies has been a puzzle for cosmologists – numerical simulations invariably yielded far too massive galaxies with an overly large central bulge and only a very small disk component. In recent years, however, substantial progress has been made in improving the physical realism of the simulation predictions.

The difficulties in modeling galaxy formation ab initio in numerical simulations primarily stem from the multi-scale and multi-physics nature of the problem, making it intrinsically very complex. To obtain realistic results, simulations need to address physical processes acting on very small scales – much smaller than the typical size of a galaxy – that nevertheless have a large impact on the global properties of the galaxy. Additionally, there are a large variety of physical processes that are still poorly understood and often neglected in theoretical studies, even though they may play a key role in galaxy evolution. One example is the dynamics of magnetic fields, which may play a role in regulating star formation. Compared to plain gas dynamics, which has typically been employed in computing galaxies so far, the simulation of magnetic fields poses additional mathematical and numerical challenges.

Using our AREPO code on the SuperMUC computer at the LRZ in Garching, we have carried out the first successful attempt to include a realistic modeling of magnetic fields in cosmological simulations of disk galaxy formation. What distinguishes AREPO from other astrophysical codes is its moving and fully dynamic mesh. The code does not partition the simulated universe via a fixed grid but rather uses a movable and deformable mesh, which allows a very accurate processing of the vastly different scales of length and mass occurring in cosmological simulations. The moving mesh approach combines the accuracy of traditional grid codes in modeling gas dynamics with the high adaptivity of particle-based codes, enabling it to track the clustering of matter particularly well.

We have succeeded in computing a set of fully cosmological simulations of 30 halos hosting realistic disk galaxies [Grand et al., 2015]. *Figure 65* shows an example of the star and gas distribution in a representative simulation of the set. In these calculations, collectively named the Auriga Project, all the simulated objects include magnetic fields evolving self-consistently from the initial conditions up to the final simulation time. In addition to magnetic fields, another distinctive trait of the project is the large number of simulated objects, enabling us to probe how the formation of disk galaxies is linked to the assembly history and the properties of the hosting halos. Also, a subsample of the full simulation set has been simulated at different resolutions without changing the galaxy formation model parameters. This has enabled us to investigate whether the galaxy properties are robust with respect to large variations in numerical resolution. At the finest resolution level, our runs improve by a factor of eight in mass over previous achievements and are among the highest-resolved cosmological simulations of galaxy formation carried out to date.

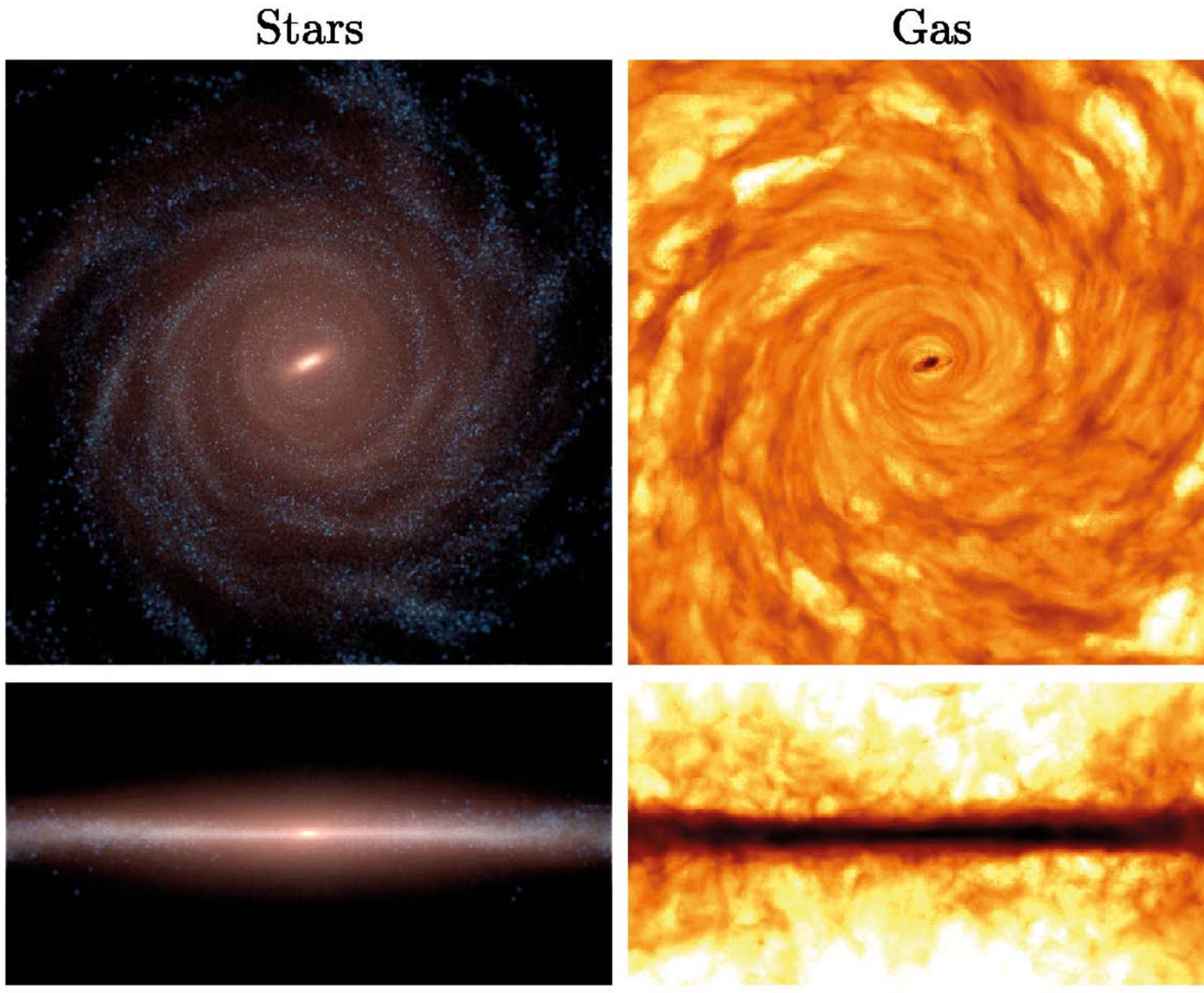


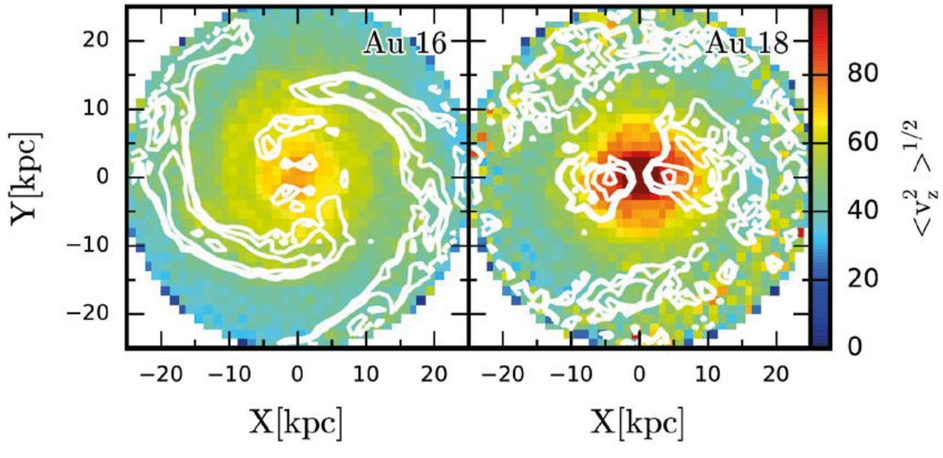
Figure 65: Example of one of our simulated Auriga galaxies in our set of magneto-hydrodynamic cosmological simulations. In face-on and edge-on projections, the figure shows stellar (left column) and gas (right column) projections at the end of the simulation after ~ 14 billion years of evolution.

The high resolution enabled us to resolve galactic substructure and the associated dynamic influence on the evolution of the stellar disks in systems representative of a globally successful galaxy formation model in a fully cosmological setting. In particular, the galaxies develop well-formed bars and spiral arms, the associated dynamic influence of which is thought to play an important role in shaping the structural properties of the disk.

In one of our studies [Grand et al., 2015], we have analyzed the vertical structure of the stellar disks in an attempt to identify the dynamic mechanisms responsible for scattering stars into vertically thickened distributions, currently a hotly debated topic in the galactic astronomy community. From the Auriga simulations we were able to establish that the stellar bar is the most ubiquitous vertical star scattering source, increasing the vertical velocity dispersion of star particles gradually over long timescales *(see Figure 66)*. Significant satellite interactions increase the dispersion dramatically on short timescales but are less prevalent. Interestingly, we found it to be unlikely that thick disks in the outer galaxy originate in high velocity dispersion stars from the inner galaxy moved there through a process termed "radial migration." This was previously believed to be a likely contributor to thick disk formation. In addition, we found that in nearly all cases the vertical structure of the disk is dominated by the formation of new stars, which are born progressively closer to the mid plane of the disk over time, thereby creating a vertically thin disk. In many cases, this process is dominant over the heating of preexisting stars.

One important aspect of the disks, the formation of spiral arms, is directly related to stellar dynamics within the disk, which can now be sampled with an appropriate number of star particles in a fully cosmological setting. Simulations of disk galaxies have long been known to produce transient, winding spiral arms that rotate at the same speed as the stars and have a lifetime roughly comparable to the dynamic time of the galaxy. Such properties, though in agreement with some observational studies of external galaxies, are at odds with the most widely accepted theory of spiral arms. This is Spiral Density Wave Theory, which predicts that spiral arms should be long-lived features that rotate rigidly around the disk and therefore never wind up. With their unprecedented resolution, our simulations are an ideal testbed for studying the formation of spiral arms in a variety of cosmological environments, ranging from quiescent to more active evolutionary histories.

Figure 66: Face-on maps of the vertical velocity dispersion of stars for a galaxy with no bar (left) and a strong bar (right). Over-density contours are overlaid in white. Note that the velocity dispersion is much larger in the strong bar case than in the galaxy with no bar.



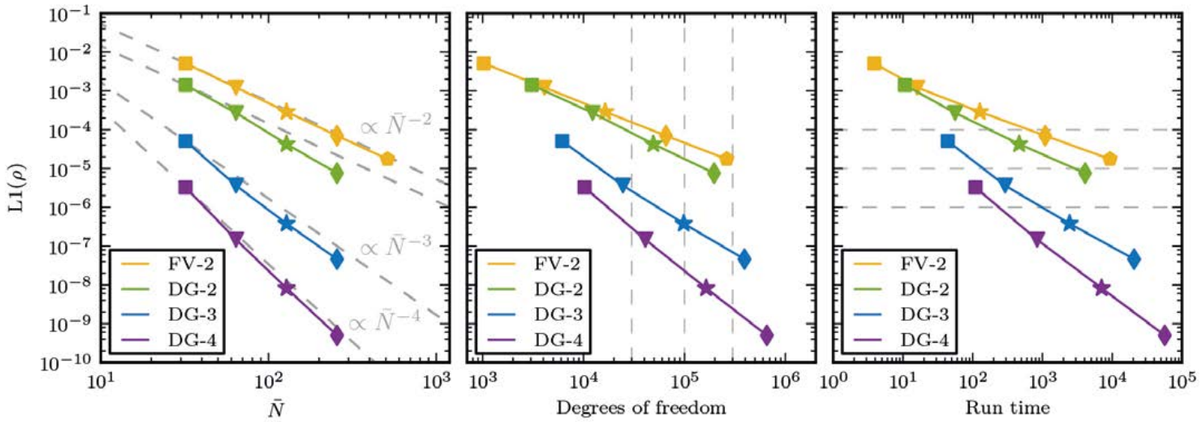
High-order discontinuous Galerkin hydrodynamics in astrophysics

In astrophysics, most numerical work on solving the Euler equations of ideal hydrodynamics has so far been carried out with two basic types of code. On the one hand, there is the broad class of Eulerian methods that utilize classical hydrodynamic solvers operating on fixed Cartesian grids or on meshes that can adjust their resolution in space with the adaptive mesh refinement (AMR) technique. On the other, there are pseudo-Lagrangian discretizations in the form of smoothed particle hydrodynamics (SPH), which are further flexible and popular tools in studying many astrophysical problems.

Some of the main advantages and drawbacks of these methods become apparent if we recall the fundamental difference in their numerical approach, which is that grid codes discretize space, whereas SPH decomposes a fluid in terms of mass elements. The traditional discretization of space used by Eulerian methods yields good convergence properties and boasts high accuracy and efficiency for many problems. Furthermore, calculations can be straightforwardly distributed onto parallel computing systems, often allowing high scalability provided there is only little communication between the cells. By contrast, the discretization of mass used by SPH results in natural resolution adjustment in converging flows so that most of the computing time and available resolution are dedicated to dense, astrophysically interesting regions. Moreover, Lagrangian methods can handle gas with high advection velocities without suffering from large errors. However, both methods also have substantial weaknesses, ranging from problems with the Galilean invariance of solutions in the case of grid codes to a suppression of fluid instabilities and noise in the case of SPH.

This has motivated us to explore another class of numerical methods, so-called discontinuous Galerkin (DG) schemes, which can be used for a broad range of partial differential equations. DG is a finite element approach incorporating several aspects of classic finite volume (FV) methods. The partial differential equation is solved in a weak formulation by means of local basis functions, yielding a global solution that is in general discontinuous across cell interfaces. The approach requires communication only between directly neighboring cells and allows for the exact conservation of physical quantities, including angular momentum. Importantly, the method can be straightforwardly implemented with arbitrary spatial order since it also directly solves for higher-order moments of the solution. Unlike standard FV schemes, this higher-order accuracy is achieved without requiring large spatial stencils, making DG particularly suitable for the utilization of massive parallel systems with distributed memory because of its favorable compute-to-communicate ratio and enhanced opportunities for hiding communication behind local computations.

In order to thoroughly explore the utility of DG in real astrophysical applications, we have developed TENET, a new MPI-parallel DG code that solves the Euler equations on an AMR grid to arbitrary spatial order [Schaal et al, 2015]. In our method, the solution within every cell is given by a linear combination of Legendre polynomials, and the propagation in time is accomplished with an explicit Runge-Kutta (RK) time integrator. A volume integral and a surface integral have to be computed numerically for every cell in every timestep. The surface integral involves a numerical flux computation that we carry out with a Riemann solver, much as in standard Godunov methods. In order to cope with physical discontinuities and spurious oscillations, we use a simple minmod limiting scheme.



In Figure 67 we show the results of a convergence study for an isentropic vortex, which provides for a powerful test of the accuracy of hydrodynamic schemes. This flow has a stationary analytic solution enabling us to assess the accuracy through a suitably defined L1 error norm. The measured error norm decreases with resolution, meeting the expected convergence rates given by the dashed lines. At a given resolution, the second-order DG-2 code achieves higher precision than the FV-2 code, reflecting the larger number of flux calculations involved in DG. The convergence rates are equal, however, as expected at second order. The higher-order DG formulations show much smaller errors, which also drop more rapidly as resolution increases. In the middle panel of the figure we show the same data but replace the linear resolution on the horizontal axis with the number of degrees of freedom (DOF). This number is closely related to the used memory of the simulation. At the same number of DOF, the FV-2 and DG-2 codes achieve similar accuracy. Moreover, using higher-order DG methods significantly increases the precision obtained per DOF in this problem.

Finally, in the right-hand panel, we compare the efficiencies of the different schemes by plotting the precision obtained as a function of the measured run time, which directly indicates the computational cost. In this way, we shed light on the question of the relative speed of the methods (or the computational cost required) for a given degree of precision. In terms of efficiency, the higher-order schemes (DG-3, DG-4) show a significant improvement over the second-order methods.

Figure 67: Left panel: L1 error norm as a function of linear resolution for the two-dimensional isentropic vortex test. Each data point corresponds to a simulation, and different colors indicate the different methods. Middle panel: the same simulation errors as a function of degrees of freedom (DOF), which is an indicator for the memory requirements. Right panel: L1 error norm versus the measured run time of the simulations.

Kelvin-Helmholtz instability is one of the most important fluid instabilities in astrophysics. For example, it plays an important role in the stripping of gas from galaxies falling into a galaxy cluster. The instability is triggered by shear flows, often also involving fluids with different densities, and grows exponentially until the primary billows break, subsequently leading to a turbulent mixing of the two phases. We illustrate the power of our new adaptive DG code by simulating Kelvin-Helmholtz instability at high resolution, as shown in *Figure. 68*. The panel top left shows the density for the whole two-dimensional simulation box, and the following panels zoom in repeatedly. Thanks to the higher order and the avoidance of reconstruction steps, the DG scheme shows only little diffusion and mixing, allowing the formation of very fine structures. Smaller KH instabilities arise on top of the large-scale waves demonstrating the fractal nature of this test problem. Self-similar instabilities are present across different scales, and an ending of this pattern is only set by the limited resolution. The adaptive mesh used by the calculation is overlaid in the bottom-right panel, revealing three different AMR levels in this sub-box.

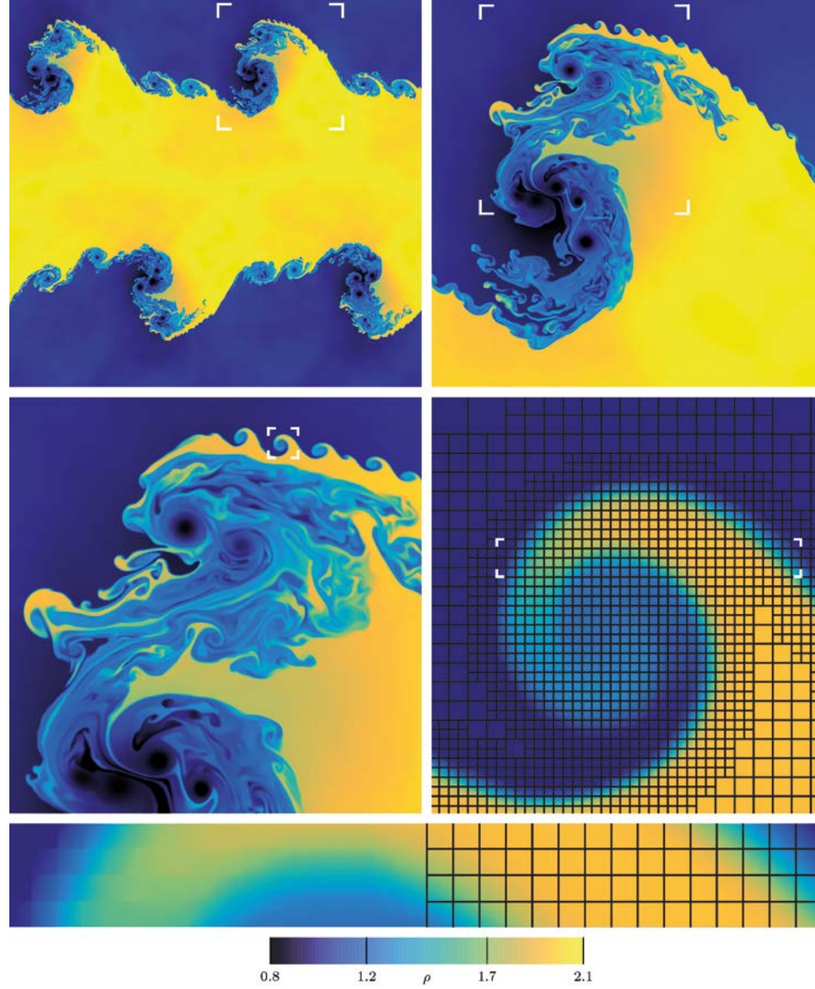


Figure 68: High-resolution Kelvin-Helmholtz simulation with 4-th order DG and AMR at time t = 0.8. The simulation starts with 64^2 cells (level 6) and refines down to level 12, corresponding to an effective resolution of 4096^2. The mesh refinement approach makes it possible to resolve fractal structures created by secondary billows on top of the large-scale waves.

Hydrogen reionization in the Illustris universe

When the Universe was almost 400,000 years old, it recombined, and most of the gas became neutral, leaving behind only a tiny residual free electron fraction. Yet in the present Universe it is well established that the intergalactic medium (IGM) is highly ionized, as has been inferred from the absorption spectra of nearby quasars. Hence there must have been an "epoch of reionization" (EoR) sometime in between, when photons emitted by stars and possibly quasars ionized the intergalactic hydrogen again. This is believed to first happen to hydrogen at redshifts $z \sim 7-10$, with helium being ionized considerably later at $z \sim 3$ by the harder radiation of quasars. The duration of the transition process and the nature of the source population ultimately responsible for reionization are subjects of much observational and theoretical research. A particularly exciting prospect is that an observational breakthrough in this field may be imminent, notably through direct mapping of the EoR with 21-cm observations.

Recently, cosmological hydrodynamic simulations of galaxy formation such as the Illustris project have advanced to a state where they produce realistic galaxy populations

simultaneously at z = 0 and at high redshifts throughout a representative cosmological volume. This is achieved with coarse sub-resolution treatments of energetic feedback processes that regulate star formation, preventing it from exceeding the required-low overall efficiency. One important manifestation of feedback is the presence of galactic winds and outflows that substantially modify the distribution of the diffuse gas in the circumgalactic medium and the IGM. This in turn also influences gas clumping and recombination rates in models of cosmic reionization. It is thus particularly interesting to test whether in principle detailed models of galaxy growth such as Illustris are also capable of delivering a successful description of cosmic reionization, and if so, what assumptions are required to achieve this.

To address this problem, we used a sequence of high time-resolution snapshots of the high-resolution Illustris simulation and combined them with a GPU-accelerated radiative transfer scheme capable of accurately evolving ionizing radiation for an arbitrary number of sources [Bauer et al., 2015]. We were particularly interested in the question of whether the star formation history predicted by Illustris can reionize the universe early enough to be consistent with observational constraints and how the reionization transition proceeds in detail in this scenario. By implementing two different radiative transfer methods, we could also evaluate how well they inter-compare, thus providing an estimate for the systematic uncertainties associated with these radiative transfer methods. The main sources of ionizing photons considered in our model are ordinary stellar populations in young galaxies, which (arguably) appear to be the most likely sources responsible for reionization. In our work we have mainly been interested in testing this hypothesis based on direct adoption of the stellar populations forming in the Illustris simulation. *Figure 69* gives a visual impression of the binned gas density field in Illustris at z = 7 and of the clustered galaxy population representing our source population.

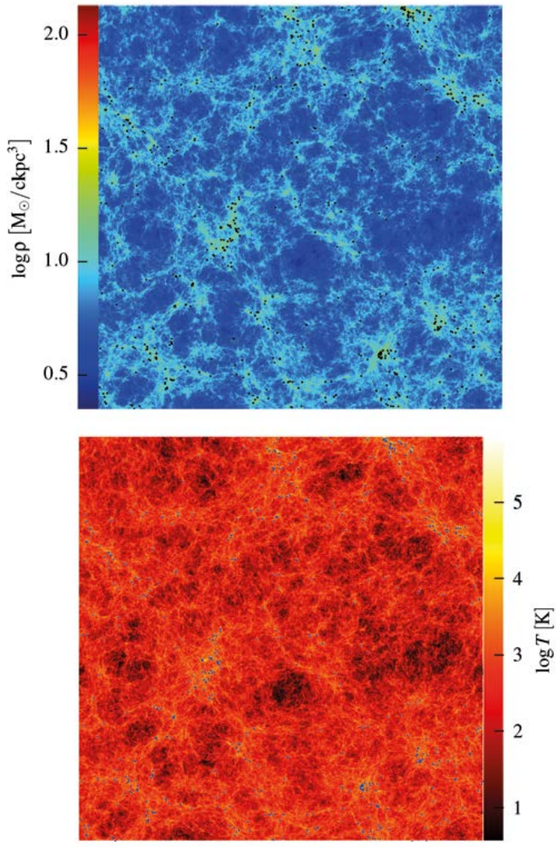


Figure 69: Projected slices through the binned gas and temperature fields of the Illustris simulation at redshift z = 7, just before reionization. The top panel shows the gas density field, with overlaid circles giving the locations of galaxies identified at this time. The bottom panel gives the corresponding mass-weighted temperature field.

The progress of reionization in different environments is visualized in *Figure 70*, where we compare two regions around very massive halos with a more average environment around a typical medium-sized halo and an underdense region. The different projections show the four regions at six different output times. Reionization starts inside the most massive halos first, quickly ionizing the surrounding regions. Compared to such a high-density environment, the onset of reionization is considerably delayed around a medium mass halo. More drastically, reionization of the underdense region only begins when the denser regions have almost completed their reionization transition. The visual impression is thus qualitatively consistent with an inside-out reionization scenario in which halos in high-density regions are affected first, and lower density voids are reionized rather late, for the most part after the reionization of overdense gas.

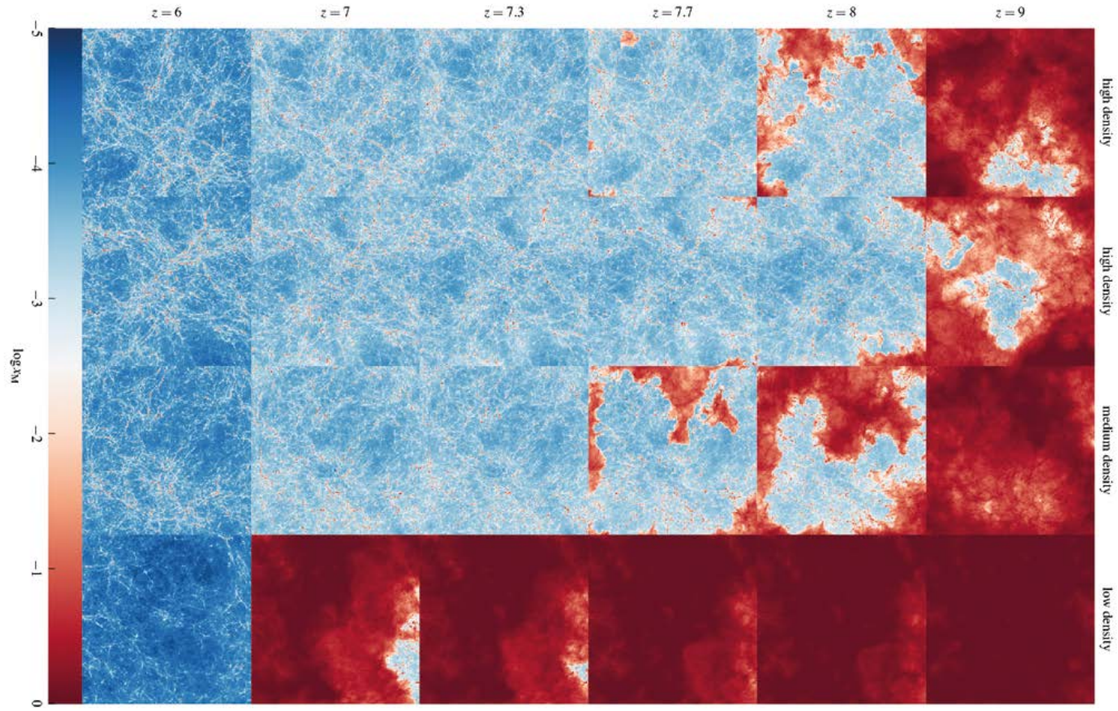


Figure 70: Progression of reionization as seen in the neutral hydrogen fraction in slices through selected sub-volumes in Illustris, each with a side-length of 21.3 comoving Mpc. Each row shows the time evolution of a different, randomly selected environment in our best model; the two rows on top correspond to an average density higher than the mean, the other columns have medium and low mean density, as labeled.

In *Figure 71* we show the resulting optical depth of our reionization simulations as a function of the integration redshift for three of our models. The most recent 2015 constraint from the PLANCK satellite is shown as a horizontal line, together with the ±1σ uncertainty region (shaded). Within the error bars, our models with a variable escape fraction are comfortably compatible with the 2015 Planck Collaboration results. In particular, our best fiducial model predicts an optical depth of $\tau = 0.065$, which is in very good agreement with the most recent Planck 2015 data, whereas all of our other models prefer the low side of the range determined by Planck. It is very promising that the notorious tension between galaxy formation simulations and optical depth inferred from cosmic microwave background measurements appears to be almost completely resolved with the 2015 Planck data and the Illustris simulation.

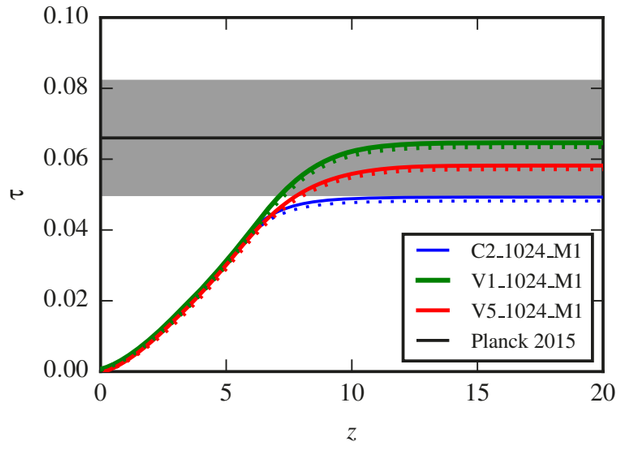


Figure 71: Cumulative optical depth for Thompson scattering on free electrons, integrated out to the redshift specified on the horizontal axis. Solid lines include electrons from doubly ionized helium (assuming that they contribute for z < 3), while dotted lines assume hydrogen and one helium electron only. The horizontal line with ±1σ uncertainty region (shaded) marks the newest 2015 constraints by the Planck Collaboration. Out best fiducial model is in very good agreement with optical depth inferred from these precision measurements.

3 Centralized Services

Group Leader

Prof. Dr. Andreas Reuter *(acting)*

Staff members

Christina Blach *(office)*
Christina Bölk-Krosta *(controlling)*
Benedicta Frech *(office)*
Ingrid Kräling *(controlling)*
Kerstin Nicolai *(controlling)*
Rebekka Riehl *(human resources and assistant to managing director)*
Stefanie Szymorek *(human resources)*

We tend to think of an administrative body as that part of an organization that year in, year out does the same routine jobs in much the same way in accordance with a firmly established set of rules or guidelines. But right from its beginnings the HITS administration has never fitted into this category, and 2015 was no exception in this respect.

Of course all sectors of the Institute (including ten research groups at the beginning of the year and twelve at the end) have to be given fully professional administrative support in dealing with personnel, project controlling, contracts, administrative organization, travel fees et cetera. But this year again, there were also a number of special projects that were crucially dependent on assistance from the Institute's administration.

The most challenging of these was the relaunch of the Institute website. The thoroughgoing overhaul of the site and its transfer to a new platform was coordinated jointly by the administration and the press department, and implementation was undertaken in close cooperation with the IT service unit and the research groups.

The year 2015 saw the arrival of two new research groups, the Physics of Stellar Objects group (PSO) in January and the (associate) Groups and Geometry outfit (GRG) in June. As anticipated, the biggest problem was to give these groups the space they needed while at the same time ensuring that the foreseeable requirements of more new groups joining us next year can be fulfilled.

In January, HITS organized a curtain-raiser symposium ushering in the 20th anniversary of the Klaus Tschira Stiftung. The event was attended by Baden-Württemberg's science minister Theresia Bauer, the rectors of three neighboring universities (Heidelberg, KIT Karlsruhe, Mannheim), and many other high-ranking guests. Providing an overview of the broad activity spectrum of the Institute, it was also of interest for the HITS staff members themselves *(see Chapter 5.3)*.

A second symposium in late November was a farewell event occasioned by Andreas Reuter's imminent departure as managing director. At the heart of the event were the talks given by some of his closest associates from the fields of academia and politics, including last year's Turing Award winner Michael Stonebraker of MIT *(see Chapter 5.9)*.

With problems of space becoming more acute, the decision was made to rent additional areas in the new Mathematikon building on the university campus. Preparations for the effective incorporation of this "outpost" into the ongoing operations of the Institute will keep the administration on their toes in the coming year as well.

With all this on the cards, there is little danger of the HITS administration turning into the kind of body adequately described by the first sentence of this report.

Group Leader

Dr. Ion Bogdan Costescu

Staff members

Dr. Bernd Doser *(Software Developer)*
Dr. Christian Goll *(System Administrator)*
Cristian Huza *(System Administrator, from January 2015)*
Norbert Rabes *(System Administrator)*
Andreas Ulrich *(System Administrator)*
Dr. Nils Wötzel *(Software Developer)*

Student

Philipp Grüner

We started the year with a bang: In mid-January we relaunched the HITS website with a new Content Management System and visual theme. This makes the site responsive to the type of device used for browsing. It also provides flexibility for adding new information quickly and comfortably and offers each researcher the opportunity to have a personal web page. The new website is connected to the major social networks and includes a live Twitter feed.

Another type of media offering insights into our activities is live video streaming, which we use, for example, to broadcast HITS colloquia and the Heidelberg Laureate Forum presentations. To improve video quality and increase the number of simultaneous viewers, we have installed a new MediaSite-based streaming solution. This runs locally and uses HITS' connection to the Internet, thus lowering the long-term maintenance costs and giving us more control over the streaming conditions

In their research, most HITS scientists rely on access to powerful computers and to large and fast storage systems. To increase computational power, we have extended our HPC cluster by further 1024 Intel Ivy Bridge cores. The nodes in which these cores reside are connected to the rest of the cluster with Infiniband and form a full-bi-section-bandwidth island together with the other nodes equipped with the same processor generation. This setup reduces the fragmentation of HPC resources and caters for computational jobs with a higher degree of parallelization.

The large number of processor cores available and the mixture of processor architectures, number of cores per node, and memory sizes provide the HITS scientists with the opportunity to run a variety of applications. However, this also requires careful planning of the scheduling policies to ensure efficient use of the available resources. To assist in these decisions, we have started using XDMoD, a cluster activity monitor and flexible statistics generator, which provides significant insights into the usage patterns of the different research groups, both for the IT support staff and for the users themselves.

The storage systems connected to the HPC cluster have undergone significant changes. The Panasas parallel storage was starting to show its age and was underperforming in metadata-rich operations, notably when accessing small files (for example,

during the compilation of a large software package) or when storing many files in a single directory (as is frequently required in bioinformatics, computational biochemistry/biophysics, or computational linguistic research projects) – precisely the types of operations that have been figuring more and more prominently in the day-to-day work of the institute. To maintain the high level of performance expected by our users, most of the scientific data (around 150 TB stored in several tens of millions of files) was migrated from the Panasas storage to one of the existing BeeGFS-based parallel storage systems – a complex operation, but one that we have made almost transparent to our users and achieved with minimal downtime. The remaining files, i.e. the users home directories, will be migrated to a new storage system in early 2016. To cope both with the migrated data and with the new data produced by this year's projects, the storage capacity of one of the BeeGFS file-systems was increased to more than 1.1 PB. As usual with BeeGFS, the addition of a storage server has also enhanced overall performance in accessing the data. However, a new bioinformatics project has also shown the need for an increase in the number of files that can exist simultaneously on the file system. While the software in question has now been modified to generate fewer data files, the growing amount of files is a continuing trend. We have therefore planned a replacement for the respective BeeGFS metadata server, which will be installed early next year.

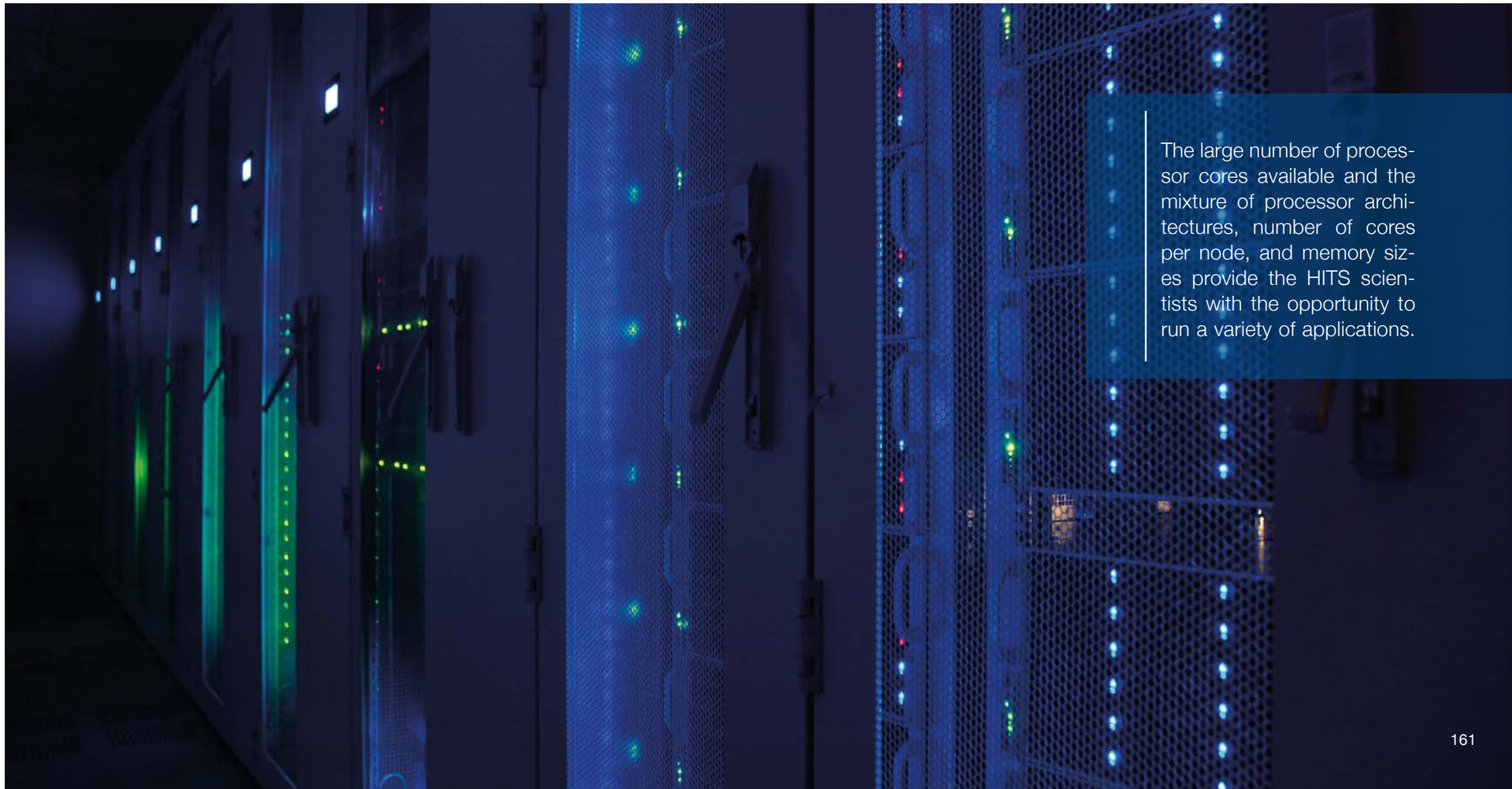
The scientific software development services offered by the ITS group started in 2014 and have successfully continued during this year. Seven projects, including one with funding from a third-party agency, have been brought to completion and follow-up projects are already planned for two of them. The seven projects cover areas ranging from linear algebra and linear/nonlinear solvers to computational biophysics and from standardization in systems biology to managing reaction-kinetic data. In short, the decision in favor of in-house professional software developers has proved to be the right one.

To round up the activities with a direct scientific impact, the educational offers of the ITS group have also continued in 2015 with the organization of an internal workshop titled "Modern Software Development with C++."Attendance was high, and we hope that the knowledge thus acquired will be successfully applied in the current and future scientific software development projects stemming from HITS.

Apart from the user-visible changes, a lot of effort has been put into maintaining and extending the institute's IT infrastructure. Using the central configuration management system introduced last year, we have converted a significant proportion of our servers to an automated setup. This has decreased the time needed to set up and maintain the servers and enables us to track all configuration changes. However, in connection with the increase in the number of research groups at HITS, it has also led to a growing number of virtual servers that we need to manage. Partly due to this and partly due to a change in the licensing conditions, we and our colleagues at Villa Bosch have had to reorganize and extend our server virtualization environment. This has involved not only adding further physical hosts but also increasing the capacity of the DataCore virtual storage and extending the network backbone, which is used both for internal connections within the institute and for the redundant connections to the nearby Villa Bosch.

Given the higher number of servers, the backup has also been extended and reorganized. For the virtualized servers referred to above, we have started using Veeam, which offers native virtualization support and, with it, a significantly increased backup speed. The classical backup system has not been forgotten – a combined software and hardware upgrade of one of its components has led to faster compression and block-level deduplication, also reducing the backup time. Luckily, we have only rarely had to use the restore capabilities of these systems.

As indicated above, quite a number of changes are already planned for early 2016. In addition, we hope to improve some of the IT-centered internal processes and interaction with the end-users by introducing new documentation and ticket systems.



The large number of processor cores available and the mixture of processor architectures, number of cores per node, and memory sizes provide the HITS scientists with the opportunity to run a variety of applications.

4 Communication and Outreach

4 Communication and Outreach

HITS Communications team in 2015 (f.l.t.r.): Anna Giulia Deutsch, Peter Saueressig, Isabel Hartmann.

The Alumni meeting on July 4: Right: Martin Pippel (CBI) during his talk. Left: The participants after the boat trip.

Head of Communications

Dr. Peter Saueressig

Members

Isabel Hartmann
(public relations)

Elisa Behr
(student until July 2015)

Anna Giulia Deutsch
(student from October 2015)

The HITS communications team is pursuing two different goals at the same time. On the one hand, we are engaged in communicating "HITS" as a brand name for a small but excellent research institute with an international, interdisciplinary, and inspiring atmosphere. On the other, we are doing our best to get the media to take due note of the excellence of our scientists in their respective areas.

In 2015, HITS welcomed two new research groups that will certainly help to achieve this twofold goal. The "Physics of Stellar Objects" (PSO) group, headed by Friedrich Röpke, is a perfect complement for the work of Volker Springel's TAP group on large-scale processes like galaxy formation, while the "Groups and Geometry" (GRG) group, headed by ERC-winning mathematician Anna Wienhard, brings additional expertise to supplement the work already being done in the two existing mathematics groups.

As in the previous year, we are happy to announce that two HITS scientists are still up there among the group of highly cited researchers worldwide. According to the "Highly Cited Researchers" report by the Thomson Reuters Group, Volker Springel (TAP) and Tilmann Gneiting (CST) again rank among the scientists most cited in their subject field and year of publication, which is an important indicator for the scientific impact of a publication. Of the 3,126 most frequently cited scientists worldwide, 184 have a primary or secondary affiliation with a German institution.

Among other excellent research publications, scientists from the MBM group published a paper in "Cell" about new findings on so-called disordered proteins, thus helping to resolve a long-standing paradox in cell biology *(see Chapter 2.6)*. Moreover, researchers from the DMQ group were awarded one of the most prestigious awards in the field of applied mathematics for their paper on computer-supported simulation of flow processes *(see Chapter 2.4)*.

Journalist in Residence program

Although the star of science journalism is waning, we still believe that an important prerequisite for successful communication is the development of reliable and sustainable journalistic contacts. 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".

In June 2015, HITS announced the second international call for applications. At the World Conference of Science Journalists in Seoul, South Korea, head of communications Peter Saueressig presented the program at the conference dinner. By the deadline in September, 40 candidates from 23 countries had applied. A jury consisting of science journalists and scientists from universities, Max Planck Institutes, and HITS, selected radio journalist Michael Stang from Cologne/Germany to join us in 2016 and the Indian science journalist TV Padma from New Delhi in 2017.

Outreach activities

Another important part of our work consists of outreach events angled at different target groups. On January 23, HITS organized a scientific symposium on the occasion of the 20th anniversary of the Klaus Tschira Stiftung, with distinguished speakers from the fields of science and politics and various presentations from all research groups *(see Chapter 5.3)*. In March, we played host to software legend Stephen Wolfram, who gave a splendid public talk in the crowded Studio Villa Bosch *(see Chapter 5.4)*. In April, HITS participated in the national "Girls' Day", offering a half-day, hands-on workshop for schoolgirls *(see Chapter 5.5)*. In July, scientists from four HITS groups presented hands-on stations at "Explore Science" *(see Chapter 5.6)*. Shortly after, the institute participated at the International Summer Science School Heidelberg and welcomed two students from Rehovot / Israel *(see Chapter 5.7)*. In August, HITSters again hosted a group of young researchers participating in the Heidelberg Laureate Forum *(see Chapter 5.8)*. On November 27, HITS organized a scientific symposium in honor of Andreas Reuter, with distinguished computer science experts and a talk by Michael Stonebraker, winner of the ACM A.M. Turing Award 2014 *(see chapter 5.9)*.

Finally, HITS is also eager to keep in touch with its alumni via various online and social media channels and by providing former employees with print products like "The Charts" newsletter and the Annual Report. We also organized an alumni meeting in July, complete with talks on current HITS research and a boat trip on the Neckar river.

Peter Saueressig presenting the Journalist in Residence program at the World Conference of Science Journalists in Seoul, June 9, 2015.

5 Events

5.1.1 Joint Wellcome Trust-EMBL-EBI Advanced Course on Computational Molecular Evolution

April 13–24, 2015, Hinxton (UK)

The need for effective and well-informed analysis of biological sequence data is increasing with the explosive growth of biological sequence databases. A molecular evolutionary framework is central to many bioinformatics approaches used in these analyses, for example de novo gene finding from genomic sequences. Additionally, explicit use of molecular evolutionary and phylogenetic analyses provide important insights in their own right, such as analysis of adaptive evolution in viruses that provides clues about their interaction with host immune systems.

For the 7th time already, our advanced course took place at the European Bioinformatics Institute in Hinxton, UK, and equipped graduate and postgraduate researchers with the theoretical knowledge and practical skills required to carry out molecular evolutionary analyses on sequence data, including data retrieval and assembly, alignment techniques, phylogeny reconstruction, hypothesis testing, and population-genetic approaches. The course covered the analysis of both protein and nucleotide sequences, including NGS data.

Participants were given an opportunity for direct interaction with some of the world's leading scientists and the authors of famous analysis tools in evolutionary bioinformatics, like Olivier Gascuel, Nick Goldman, Bruce Rannala, Alexandros Stamatakis, and Ziheng Yang.

Co-organizer Alexandros Stamatakis (SCO group leader) taught at the event. The committee received approximately 80 applications for the 40 places available. SCO team-member Paschalia Kapli participated as a teaching assistant.

5.1.2 ISO Technical Committee ISO/TC 276 Biotechnology

Meeting in Shenzhen (China),
April 13–18, 2015

The ISO Technical Committee ISO/TC 276 Biotechnolgy during the meeting in Shenzhen, China. (photo: BGI)

The International Organization for Standardization (ISO) provides definitions of standards that are recognized worldwide, not least in the life sciences, where they are of major value for industrial, agricultural, and medical applications. The ISO Technical Committee for Biotechnology Standards (ISO/TC 276), in which more than 100 experts from 23 countries work together to establish new standards in biotechnology and related fields, met for its plenary and work-group meetings in Shenzhen (China) from April 13–18, 2015. During this event, hosted by BGI (Beijing Genomics Institute), a new work-group called "Data Processing and Integration" (WG5) was founded by the committee. The new work-group aims to define standards for the formatting, transfer, and integration of life-science data and corresponding computer models. One main objective is to design an ISO standard as a framework connecting existing "de facto" standards defined by scientific grass-roots initiatives and currently used by researchers. This will standardize the concerted interplay between such community standards and define interfaces between different data formats so as to facilitate the exchange and combination of data and computer models.

At the meeting in Shenzhen, Martin Golebiewski from the SDBV group was appointed the convenor (i.e. leader) of this new ISO work-group. He is already part of the board of coordinators of the international grass-roots standardization initiative COMBINE (Computational Modeling in Biology Network) and the coordinator of the German NORMSYS project, which focuses on standardizing modeling formats in systems biology. This ensures a close liaison of the work-group with scientific standardization initiatives. The secretariat of the new ISO committee is sited at the German Institute for Standardization (DIN), and around 50 experts from different European countries, as well as from Japan, the US, South Korea, Iran, and China are already collaborating in this setting. In 2015 the ISO/TC 276 work-group "Data Processing and Integration" met twice already, for the inaugural meeting in Berlin (Germany) on July 15–16 and for a second meeting co-located with meetings of all other work-groups for biotechnology standards in Tokyo (Japan) on October 26–30. Besides the work referred to earlier on an ISO framework for community standards, the new committee has already identified exciting fields of activity, for example a budding collaboration with the Moving Picture Experts Group (MPEG) on standards for genome compression.

5.1.3 8th International Biocuration Conference

At the 8th International Biocuration Conference, April 23–26, 2015 in Beijing, Renate Kania and Dr. Ulrike Wittig held a workshop entitled "Money for Biocuration – Strategies, Ideas & Funding." The purpose of organizing this workshop was to exchange experience and new ideas with regard to the funding problems that some databases are faced with. Even though funding for research infrastructures is becoming more widely available, maintaining and feeding existing databases can still be a problem, even if they are well-established. Recognition of the importance of curation still lags behind awareness of the necessity for data generation, and this is also reflected by funding policies.

In their preparations for the workshop, the HITS researchers set up a survey to obtain more detailed information about the current funding situation that different databases and curators find themselves in. The results of the survey were presented in the first talk in the workshop by Ulrike Wittig, together with an overview of alternative funding/business models. In the second talk, Donghui Li (TAIR, Phoenix Bioinformatics, USA) described how TAIR had switched to a non-profit subscription model after their NSF funding ended in 2013. Finally, Alex Bateman (Protein & Protein Family Services, EBI, UK and head of the ISB) talked about activities at EBI, notably the collaboration between different protein databases, and shared his thoughts with the audience on the importance of funding infrastructures and services. Workshop participants discussed the advantages and disadvantages of alternative business models such as subscription funding. The researchers also talked about what the International Society for Biocuration (ISB) could do to help improve the funding situation for biocuration and how individual curators could help.

In future there will be a forum for discussion and knowledge exchange on the website of the ISB. The talks given in Beijing will be made available there, too.

Renate Kania (left) and Ulrike Wittig presenting their findings at the workshop.

5.1.4 ACL-IJCNLP 2015

Michael Strube (NLP group leader) served as program co-chair for the flagship conference of the Association for Computational Linguistics, the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Joint Conference on Natural Language Processing. The conference was held in Beijing, China, July 26–31, 2015, at the China National Convention Center in the Olympic Park.

Michael Strube addressing the audience in Beijing.
(Photo: Chinese Information Processing Society of China)

Organizing such a big and important event is always a group effort that takes more than a year. Michael and his co-chair Chengqing Zong (Chinese Academy of Sciences) were supported by 37 area chairs and about 900 reviewers selecting from 1340 submissions the 318 papers to be presented at the conference. Half of the papers were presented orally in five parallel tracks. The others were presented in two 3-hour poster sessions along with system demonstrations and a student poster session. The conference also featured talks by two keynote speakers, Marti Hearst (UC Berkeley) and Jiawei Han (University of Illinois at Urbana-Champaign), eight half-day tutorials, 15 workshops, and the associated Conference on Natural Language Learning.

Michael was particularly proud of his success in organizing the poster session. While the academic community now knows how to organize oral sessions, poster sessions have mostly been hit-or-miss affairs. Based on his previous experience with badly organized poster sessions, Michael developed the theory that for a poster session with 100 posters and about 1200 attendees each poster needs its own virtual box measuring four by four meters. This worked beautifully and created a good atmosphere, because everyone had access to the posters without bumping into each other. ■

5.1.5 SHENC symposium on the “von Willebrand factor”

The 1st International SHENC symposium entitled “Function of von Willebrand factor in primary and secondary hemostasis” took place from September 13–15, 2015 at the Hotel Hafen, Hamburg. The meeting assembled international experts from various fields of science, ranging from theoretical physics and biophysics to biomedicine, to discuss the newest developments in our understanding of the von Willebrand factor. The von Willebrand factor is not only a critical factor guiding hemostasis and involved in thrombosis and bleeding disorders, it is also a fascinating biomolecule from a physics point of view, as it switches functions in the shear flow present in blood vessels.

The meeting showcased some of the more recently solved von Willebrand factor puzzles, such as its role in inflammation and its catch-bond-like function. The symposium was organized by the DFG-funded research group SHENC, for which Frauke Gräter, group leader of the MBM group at HITS, acts as one of the coordinators. ■

5.1.6 Workshop on Reproducible and Citable Data and Models

ERASysAPP (ERA-Net for Applied Systems Biology, a cooperative project of 16 funding agencies/partners) gave two overlapping teams the opportunity to hold a workshop on data citation (Goble from Uni Manchester, Müller/SDBV) and model reproducibility (Waltemath from Uni Rostock, Müller/SDBV). Together, we convinced ERASysAPP that collating these workshops would be a good idea because the topics are so well matched. Citing data and models that are not reproducible is much less interesting. So our topic was: How do we make models reproducible and link them to their underlying data, and how does this fit in with the bigger picture of modern science?

A web of data managers on the beach at Warnemünde.
(Photo: TMZ, University of Rostock)

The workshop took place September 14 – 16 in Warnemünde near Rostock. We combined talks from distinguished researchers in the field of data and model management with hands-on sessions at which tools of the FAIRDOM project were used to demonstrate the concepts.

The workshop was directed both towards data-management researchers and prospective users of complex data-management solutions. The talks gave the big picture, starting from the publisher’s view of data citation and moving on from there to licensing and the importance of context, the OpenAIRE infrastructure, and the Research Data Alliance. The model reproducibility talks were anything but abstract, detailing why it is hard to reproduce models and what progress is being made in simplifying model reproducibility.

In hands-on sessions, users were given the opportunity to create and publish a package consisting of data and reproducible models for the Zenodo repository by using Research Objects and creating a Digital Object Identifier to make these data citable.

The workshop was attended by audiences from many walks of life, and speakers and organizers learned a lot from each other. The open-minded atmosphere encouraged people to ask questions, and the venue was congenial enough to prompt participants to stay together and go on discussing what they had heard until well into the evening. We are looking forward to organizing similar workshops in the near future. ■

5.1.7 ISESO Symposium

On November 9 and 10, 2015, the International Symposium on Energy System Optimization (ISESO 2015) took place at the Studio Villa Bosch conference center in Heidelberg.

More than 50 participants from Europe, North and South America discussed how methods from different disciplines, such as mathematics, electrical engineering, economics, and operations research can help us tackle the many challenges facing our energy systems.

Keynote speakers were Prof. Claudio Cañizares, University of Waterloo/Canada, talking about energy management systems for local and small grids; Dr. Tanja Clees, Fraunhofer Institute SCAI, St. Augustine, who presented a system for simulation, analysis, and optimization of energy networks; and Prof. Shmuel S. Oren, UCLA Berkeley / USA, who focused on smart markets for smart electricity grids.

The Symposium was initiated by a consortium of partners from HITS, Heidelberg University, and

The ISESO organizers.

Karlsruhe Institute of Technology (KIT) working together on the research project "New Approaches to Integrated Energy Systems and Grid Modeling" funded by the DFG (German Research Foundation).

The symposium was organized by Prof. Wolf Fichtner (KIT), Prof. Vincent Heuveline (Heidelberg University / HITS), Prof. Thomas Leibfried (KIT), Dr. Valentin Bertsch (KIT), Dr. Michael Schick (HITS), Philipp Gerstner (Heidelberg University / HITS), and Dr. Michael Suriyah (KIT). ■

5.1.8 International Conference on Systems Biology (ICSB) and COMBINE tutorial workshop

Singapore, November 23 – 26, 2015

Systems biologists from all around the world gathered at the 16th International Conference on Systems Biology (ICSB) that took place in Singapore November 23 – 26, 2015 (http://icsb15.apbionet.org). In this context Martin Golebiewski (SDBV) chaired the session "Big Data for Systems Biomedicine – Data and modeling standards" of the main conference and also organized the one-day COMBINE tutorial workshop "Modeling and Simulation Tools in Systems Biology." The data and modeling standards session, which around 60 scientists attended, started with the talk "NormSys – Harmonizing Standardization Processes for Model and Data Exchange in Systems Biology" presented by the session chair. This was followed by Akira Funahashi (Keio University, Japan) talking about the CellDesigner modeling tool for biochemical networks. The other two talks in this session were given by Riza Theresa Batista-Navarro (University of Manchester, UK) on "Bridging text-mined biomolecular events to pathway models using approximate subgraph matching techniques" and Andrei Zinovyev (Institut Curie, Paris, France) on the "Atlas of cancer signaling networks: Google maps of cancer biology."

The COMBINE tutorial workshop (http://co.mbine.org/events/tutorial2015), which was organized at the National University of Singapore (NUS) as a satellite of the ICSB conference, showed the mainly young attending scientists how to set up computer models of biological networks and simulate these models in different systems-biology platforms. Lectures, software demonstrations, and hands-on sessions provided the attendees with the necessary skills for creating, simulating, and handling such models. An international team of tutors instructed the scientists on how to use modeling and simulation tools and databases, and demonstrated the use of standardized formats in describing and exchanging the models. The SDBV group presented their SABIO-RK database for reaction kinetics data and the SEEK system (FAIRDOM-Hub) for integrated data and model management. HITSter Antonia Stank (MCM) demonstrated the web-accessible tools LigDig, SYCAMORE, and webPIPSA and their uses for molecular modeling. Other software tools introduced at this tutorial workshop included COPASI for simulation and analysis of biochemical networks and their dynamics and the BioModels database for computational models, as well as the CellDesigner platform and Pathway Commons, an online resource for searching and visualizing biological pathways. As an important prerequisite for the information and data exchange amongst scientists, domain-specific community standard formats for data and models were demonstrated to be crucial for exchange between the different platforms, tools, and databases. Such community standards for modeling in the life sciences are defined by the COMBINE network (http://co.mbine.org) that was co-hosting this tutorial workshop. ■

Martin Golebiewski

Keynote address "Systems Medicine and Proactive P4 Medicine: Transforming Healthcare" presented by Leroy Hood (Seattle, WA, USA) at ICSB 2015 in Singapore.

Lars Lindberg Christensen (MSc)
ESO, Garching February 9, 2015:
ESO now and in the Future

Dr. Nick Goldman
EMBL European Bioinformatics Institute, UK
March 16, 2015: Information Storage in DNA

Dr. Stephen Wolfram
Wolfram Research, Oxfordshire, UK March 27, 2015: The Future of Computation and Knowledge

Graham T. Johnson (PhD)
University of California, San Francisco, USA
April 23, 2015: Towards Whole Cells Modeled in 3D Molecular Detail and Community Curated with cellPACK

Dr. Larry Krumenaker
Freelance Science Writer May 18, 2015:
Attitudes about Science in Asia

Dipl.- Phys. Raoul Heyne
Energie-Forschungszentrum Niedersachsen (EFZN), June 15, 2015: Elektromobilttät und Energiewende

Prof. Dr. Ralf G. Herrtwich
Director Driver Assistance and Chassis Systems, Daimler AG July 20, 2015 Making Bertha Drive:
A T(o)uring Test

Michael J. Koch
Linguist, Hamburg University „Vortrag zur Sommerpause“ on August 10, 2015: Karl von den Steinen, ein Polynesischer Mythos

Prof. Petros Koumoutsakos
Chair for Computational Science, ETH Zürich, Switzerland September 21, 2015: High Performance Computing in the Time of Uncertainty

Pat Helland
Salesforce.com Inc., San Francisco, USA
October 19, 2015: Subjective Consistency

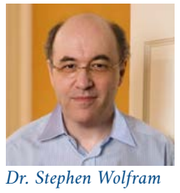
Prof. Dr. Henrik Kaessmann
ZMBH Heidelberg, November 16, 2015:
Functional Evolution of Mammalian Genomes

Prof. Zan Luthey-Schulten
University of Illinois at Urbana-Champaign, USA
December 15, 2015: Simulations of Cellular Processes: from Single Cells to Colonies

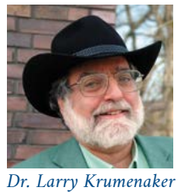
Prof. Klaus Schulten
University of Illinois at Urbana-Champaign, USA
December 15, 2015: Towards an Atomic Level Description of a Living Cell – The Photosynthetic Chromatophore of Purple Bacteria, a Milestone

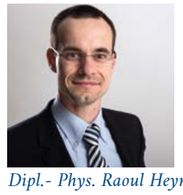

Lars Lindberg Christensen (MSc)


Dr. Nick Goldman


Dr. Stephen Wolfram


Graham T. Johnson (PhD)


Dr. Larry Krumenaker

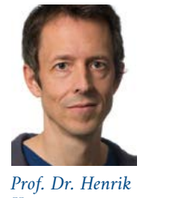

Dipl.- Phys. Raoul Heyne


Prof. Dr. Ralf G. Herrtwich


Michael J. Koch


Prof. Petros Koumoutsakos


Pat Helland


Prof. Dr. Henrik Kaessmann


Prof. Zan Luthey-Schulten


Prof. Klaus Schulten

The Klaus Tschira Stiftung (KTS) was founded by Klaus Tschira in 1995. Over the years, it has become one of the biggest private nonprofit foundations in Europe. For their anniversary year, the KTS institutions drew up a special program. The first event was a scientific symposium on January 23, hosted by HITS at the Studio Villa Bosch.

“The KTS has promoted research in natural sciences from the very beginning,” said Klaus Tschira in his welcoming speech. “Unlike most research centers, HITS does not focus on just one discipline – it has always been our intention to make it an interdisciplinary institute. I am firmly convinced that enormous potential for new discoveries lies on the borderlines of traditional fields of research.” This was one of the last public speeches Klaus Tschira made before his untimely death in March 2015.

Theresia Bauer, Minister for Science, Research and the Arts in Baden-Württemberg, gave a welcome speech as well. “HITS is an outstanding example of how private funding can support excellent research,” she said. “We are extremely grateful for the work of the Klaus Tschira Stiftung. The successful cooperation between the Heidelberg Institute for Theoretical Studies, the Karlsruhe Institute of Technology, and Heidelberg University proves that HITS has evolved into one of the best research centers in Baden-Württemberg.”

Other speeches of welcome were held by Prof. Bernhard Eitel and Prof. Holger Hanselka, rector of Heidelberg University and president of the Karlsruhe Institute of Technology (KIT). Astrophysicist Prof. Mark Vogelsberger (MIT Massachusetts Institute of Technology) and biophysicist Prof. Jeremy Smith (Oak Ridge National Laboratory, Tennessee) gave talks on scientific subjects. Both scientists cooperate closely with researchers from HITS. Their presentations showed how so-called “data-driven” research provides new opportunities for science by using mathematical modeling and computer simulations.

During the event, HITS research groups presented their fields of research – from astrophysics to molecular biology – with hands-on demonstrations, posters, and videos. The piano trio Aisthesis/KlangForum Heidelberg provided an exquisite musical setting for the symposium.

f.l.t.r.: Prof. Rebecca Wade (HITS scientific director), Prof. Andreas Reuter (HITS managing director), Prof. Holger Hanselka (president of KIT Karlsruhe), Theresia Bauer (MWK Baden-Württemberg), Dr. h.c. Dr.-Ing. E.h. Klaus Tschira (founder and managing director of HITS), Prof. Bernhard Eitel (rector of Heidelberg University). photo: Klaus Tschira Stiftung

HITS scientists of the TAP group present their research to Minister Theresia Bauer. Photo: Klaus Tschira Stiftung

5.4
Stephen Wolfram at HITS

On March 27, HITS had the honor of welcoming the distinguished scientist, technologist and entrepreneur Dr. Stephen Wolfram.

Wolfram has devoted his career to the development and application of computational thinking. He is the original developer of "Mathematica", the author of "A New Kind of Science", and the man behind the "Wolfram Alpha" search engine. He visited the institute to get an idea of the scientific goals pursued by the various HITS research groups. For this purpose, HITSters from all groups met Wolfram for a round-table discussion. Later in the afternoon, he gave a special colloquium talk on "The Future of Computation & Knowledge" in the crowded Carl Bosch Auditorium of Studio Villa Bosch, with many guests from academia and industry. "Simulations are not only convenient but essential for research," he stated in his talk, which was moderated by Prof. Volker Springel (TAP). ■

The video of the talk is available on the HITS YouTube Channel: https://www.youtube.com /user/TheHITSters

Stephen Wolfram meeting the HITSters.

Visitors at the special colloquium.

Stephen Wolfram explaining his new software.

5.5
Girls' Day
Premiere at HITS

On April 23, HITS participated for the first time in the national "Girls' Day". During this event, which takes place once a year, various institutions offer workshops for girls to introduce them to jobs in which women are still underrepresented.

The goal of the event is to broaden the mind of young girls and to get them interested in, say, a MINT subject, such as research at HITS. Three different research groups offered small-scale, hands-on workshops to show the girls what the daily work and life of a researcher looks like.

Sarah Lutteropp from the Scientific Computing (SCO) group brought along an Egyptian criminal case that the girls were asked to solve with the help of computers. In the workshop, the girls had to save the life of the young pharaoh in the story by using a computer program to find the right antidote for snake venom. After some warm-up exercises and programming practice, the girls were able to save the pharaoh from his awful fate and learned a lot about how computers can facilitate research work.

In the workshop by the Molecular and Cellular Modeling (MCM) group, the girls found out how molecules are structured and how this helps researchers to devise new medicines. Using 3D models and computer programs, they learned how to model a certain drug so that it can "dock" on to a specific molecule in the body.

The Computational Statistics (CST) group showed the girls how to write a small computer program calculating probabilistic weather forecasts. The girls quickly learned how much science is needed to make weather forecasts reliable.

After the workshops, the girls had lunch with scientists and were taken on a tour of the institute premises. The highlight: a guided tour of the HITS server room. On this occasion, sixteen girls between ten and fifteen years visited HITS to participate in the workshops and find out what it's like to be a HITS researcher. The event was a complete success both for the girls and the scientists, so HITS will definitely be taking part again in 2016. ■

Daria Kokh (MCM) showing some 3D molecules.

Group picture after a successful day.

This summer, as in previous years, HITS participated in the Explore Science event, which once again took place in Mannheim's Luisenpark. The event is angled at children, secondary-school students, and their families. Organized by the Klaus Tschira Stiftung, it assembles any number of hands-on stations, exhibitions, and presentations designed to get youngsters interested in the natural sciences. Explore Science went on from July 8–12. Its motto was "Physics: Pure Motion!", and it attracted over 50,000 visitors from all over the region.

The HITS researchers took up the topic and illustrated it in three interactive hands-on stations showing that motion can be found everywhere – in cells just as much as in water or wind. The Molecular and Cellular Modeling (MCM) group built wobbly molecules made of paper to show the children that the molecules in our body are flexible and that this quality is extremely important for the organism as a whole. The scientists used simple paper strips and little paper balls to create colorful, wiggly molecules. They then encouraged the children to construct molecules of their own in various shapes and colors, which they could take home as souvenirs.

On the subject of motion in the air, the Astroinformatics (AIN) and Data Mining and Uncertainty Quantification (DMQ) groups built a flow channel consisting of small shelves and over 5,000 drinking straws. The children could then release artificial fog into the channel to visualize the airstream. A small toy airplane was used to demonstrate how readily a stream of air adapts even to the smallest objects. The children saw and learned at first-hand how objects divide the airstream and how even the slightest motion of the airplane can stir the airflow in the twinkling of an eye.

The third station revolved around the element water. The Computational Statistics (CST) group had dreamed up a small but exciting game to show the visitors how the motion of water can be predicted. To this end, the statisticians constructed a dam with building blocks. The children were then given toy money with which they could buy different sized dam blocks. After that, they were asked to cast dice deciding randomly how much water would be routed into the reservoir behind the dam. The aim of the game was to get through three rounds without the dam being flooded. Though the game is quite simple, it requires a lot of statistical information, and after a few rounds even the youngest visitors could figure out the best strategy for winning the game.

Left: Fabian Krüger (CST) playing the flood game with some children. Below: Children using strips of paper to construct their own molecules with the help of Antonia Stank (MCM).

Dennis Kügler (AIN) explaining the airflow channel.

In summer 2015, HITS once again welcomed two students in the framework of the International Summer Science School Heidelberg (ISH).

The project enables young people from Heidelberg's twinned cities and partner organizations who are interested in the natural sciences to get practical inside impression of the laboratories of renowned research institutes here in Heidelberg. This year's students visiting the HITS research outfit were Helly Hayoun and Yarden Hadar from Rehovot / Israel. Both girls were supervised by HITS scientists Dr. Davide Mercadante from the Molecular Biomechanics (MBM) group and Antonia Stank from the Molecular and Cellular Modeling (MCM) group. During their three weeks stay at the institute, the girls learned basic skills in computational biology and experienced what it is like to live and work as a scientist at an international research institute.

F.l.t.r.: Antonia Stank, Yarden Hadar, Helly Hayoun, Dr. Davide Mercadante.

The 3rd Heidelberg Laureate Forum took place on August 23–28, 2015. Laureates of the most prestigious awards in mathematics and computer science (Alan Turing Awards, Fields Medal, Nevanlinna Prize) came together in Heidelberg to meet young researchers from all over the world.

As in the year before, there were several talks given by the laureates, young researchers and laureates interacted in various framework activities, and the young researchers had an opportunity to get to know the Heidelberg research landscape.

A highlight of HLF 2015 was the Hot Topic Session. The format was introduced in 2014 and deals every year with a different big issue related to science and society that experts and laureates discuss at the Forum. This year's Hot Topic session was coordinated by Michele Catanzaro, author of "Networks: A Very Short Introduction." He is a highly accomplished freelance science journalist and also was HITS "Journalist in Residence" 2014. The topic up for discussion was "Brave New Data World." The panelists looked into fundamental questions like how secure our data is, how intellectual property is evolving, and whether we should regulate this "brave new data world". The session gave a broad overview of how scientists see these very pressing issues that our society is facing. To watch an online video of the session, go to the website of the Heidelberg Laurate Forum.

HITS is a co-initiator of this networking event and supports the HLF with its scientific expertise. HITS and its scientists also hosted an event in the framework of the "young researchers' day". This special program enables young researchers to visit and link up with various scientific institutes in Heidelberg. On August 26, HITS welcomed 25 young scientists from all over the world to the institute. The HITSters presented their work and were keen to hear the views of their international guests.

The next Heidelberg Laurate Forum will take place on September 18 – 23, 2016.

©Heidelberg Laureate Forum Foundation / C. Flemming

5.9
Symposium in honor of Andreas Reuter

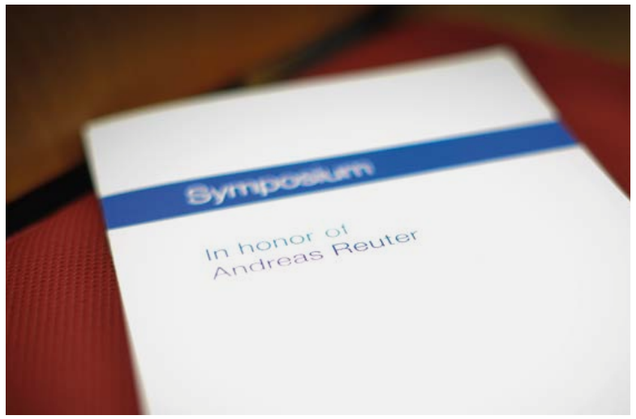
On November 27, 2015, HITS organized a symposium in honor of Andreas Reuter at the Studio Villa Bosch. His scientific career spans more than four decades of computer science, a period of fundamental progress in IT research and development. Hence the event was also an excellent opportunity to invite several distinguished experts on database research and high performance computing, some of whom have been involved with Andreas in these fields for many years.

In two sessions they traced his contributions to the scientific community, notably the textbook "Transaction Processing – Concepts and Techniques" that he wrote with Turing Award-winner James "Jim" Gray in 1992 – "the Bible of transactions processing," as Dr. C. Mohan put it. Dr. Mohan is an IBM Fellow at the Almaden Research Center in San Jose, USA, where Andreas Reuter spent some time in the eighties. He gave a speech on "Hype and Reality in Data Management".

Via videoconference, Prof. Michael Stonebraker, 2014 Turing Award-winner, sent his greetings to Andreas Reuter and delivered a speech on problems with large data volumes called "The 800 Pound Gorilla of Big Data". Prof. Tony Hey, chief data scientist of the Science and Technology Facilities Council, UK and a member of the HITS Scientific Advisory Board (SAB), gave a brilliant talk titled "Stretch Limos, Transaction Processing and e-Science". (Both talks can be watched on the HITS YouTube channel).

Two other members of the HITS SAB addressed the audience: SAB chair Prof. Dieter Kranzlmüller, director of the Leibniz Supercomputing Center, Munich University, and Dr. Heribert Knorr, former director-general, Ministry of Science, Research and the Arts Baden-Württemberg, Stuttgart. Other speakers were Beate Spiegel, managing director of the Klaus Tschira Stiftung, Renate Ries, head of communications for the Klaus Tschira Stiftung, and HITS group leaders Priv.-Doz. Dr. Wolfgang Müller (SDBV) and Prof. Vincent Heuveline (DMQ). The music for the event was provided by pianist J. Marc Reichow (Klangforum Heidelberg), who played works by Händel, Mozart, and Busoni.

In his closing remarks, Andreas Reuter expressed his gratitude to all present and reminisced about the two brightest minds he had met in his life, Jim Gray and Klaus Tschira. After his retirement from the HITS board of directors in April 2016, he will still be active in science and science management as scientific and managing director of EML European Media Laboratory, scientific chair of the Heidelberg Laureate Forum, and senior professor of Heidelberg University.



The Supernova project group at HITS (f.l.t.r.): Dr. Volker Gaibler, Dr. Kai Polsterer, Dr. Dorotea Dudaš.

ESO Supernova – Planetarium and Visitor Center

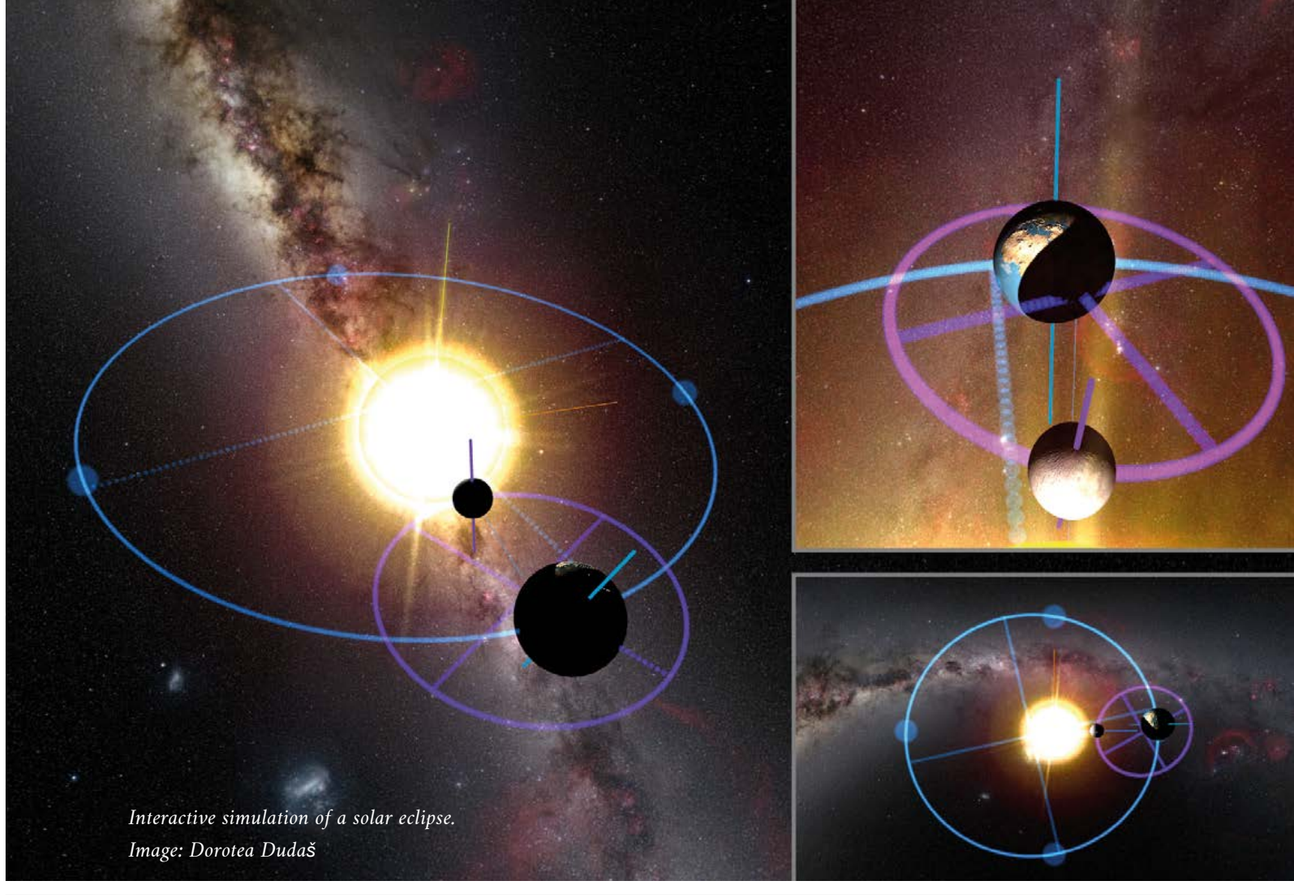
Space, stars, and the universe exert an almost magical attraction on many people. While astronomy is an exciting and rapidly evolving science, acquainting the public with the latest knowledge from astronomical research is often difficult, and the pressure of daily work makes it a rare activity for most scientists working in the field.

The ESO Supernova visitor center in Garching near Munich is designed to bridge the gap between astronomical research and the interested public, enabling visitors to catch up on the latest state of play in astronomy, reporting on recent discoveries, and inspiring visitors with the fascination of the universe we live in.

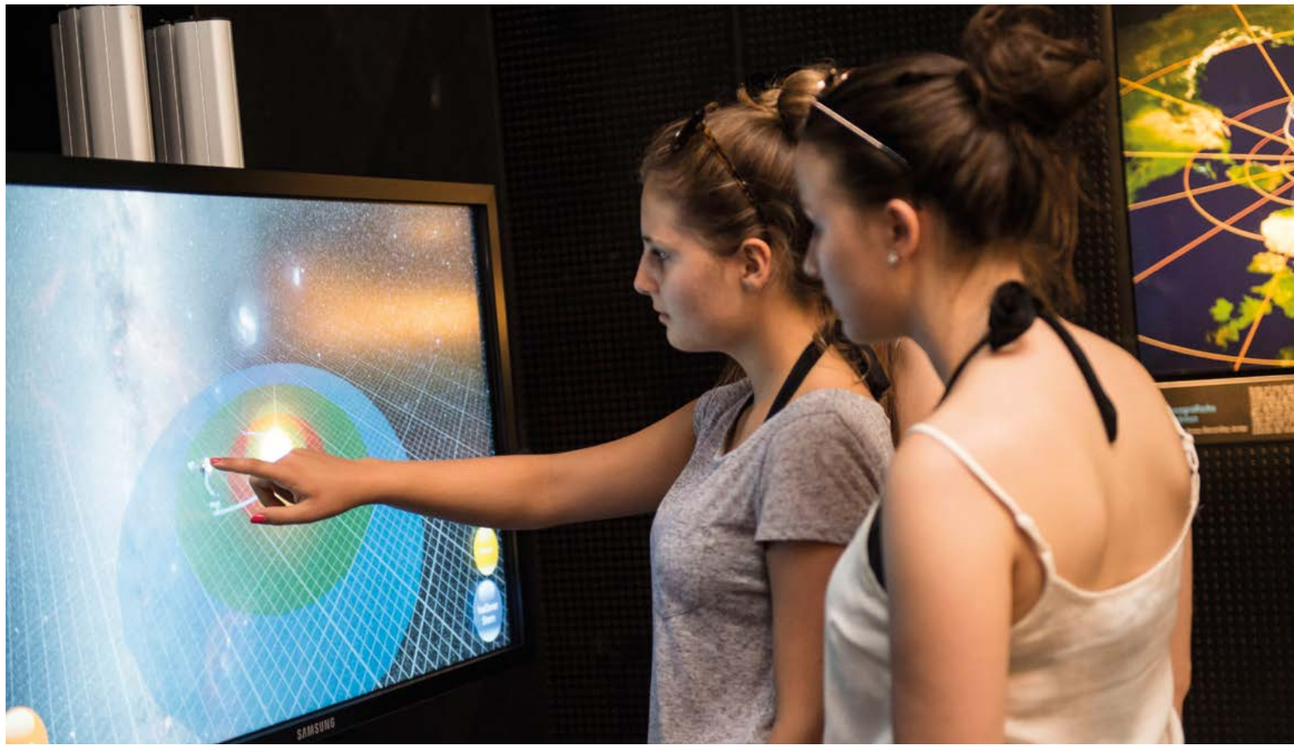
The Supernova project group at HITS brings its scientific expertise to bear on the decision process about what goes into the Garching exhibitions. And that's not all, by any means. It also devises and develops interactive exhibits specifically for the ESO Supernova that will enable visitors to immerse themselves in a broad range of topics by means of interactive computer simulations, virtual reality, and state-of-the-art computer graphics. This way, they can discover astronomy for themselves, deepen their understanding of the science, and share their fascination with others.

The two developers, Dr. Dorotea Dudaš and Dr. Volker Gaibler, have a wealth of professional expertise at their finger-tips, ranging from computer graphics and numerical mathematics to astronomy and theoretical astrophysics. Project manager is Dr. Kai Polsterer, leader of the Astroinformatics junior group.

The center is the fruit of a collaboration agreement between the Heidelberg Institute for Theoretical Studies (HITS) and ESO. The Klaus Tschira Stiftung (KTS) is funding the construction of the premises, and ESO will be running the facility. The ESO Supernova is due for completion in early 2017 and will open its doors to the public in the course of that same year.



Interactive simulation of a solar eclipse.
Image: Dorotea Dudaš



Visitors trying out the prototype of a hands-on station for the new ESO Supernova.
Photo: Klaus Tschira Stiftung

7
Publications

Aberer AJ, Stamatakis A, Ronquist F. An efficient independence sampler for updating branches in bayesian markov chain monte carlo sampling of phylogenetic trees. Systematic Biology online (2016) 65:161 – 176, online: Jul 30.2015.

Aponte-Santamaría C, Huck V, Posch S, Bronowska AK, Grässle S, Brehm MA, Obser T, Schneppenheim R, Hinterdorfer P, Schneider SW, et al. Force-sensitive autoinhibition of the von Willebrand factor is mediated by interdomain interactions. Biophysical journal (2015) 108:2312 – 2321.

Arnold C, Puchwein E, Springel V. The Lyman α forest in f(R) modified gravity. Monthly Notices of the Royal Astronomical Society (2015) 448:2275 – 2283.

Banfield JK, Wong OI, Willett KW, Norris RP, Rudnick L, Shabala SS, Simmons BD, Snyder C, Garon A, Seymour N, Middelberg E, Andernach H, Lintott CJ, Jacob K, Kapińska AD, Mao MY, Masters KL, Jarvis MJ, Schawinski K, Paget E, Simpson R, Klöckner HR, Bamford S, Burchell T, Chow KE, Cotter G, Fortson L, Heywood I, Jones TW, Kaviraj S, López-Sánchez ÁR, Maksym WP, Polsterer KL, Borden K, Hollow RP, Whyte L. Radio Galaxy Zoo: host galaxies and radio morphologies derived from visual inspection. Monthly Notices of the Royal Astronomical Society (2015) 453:2326 – 2340.

Baran S, Lerch S. Log-normal distribution based EMOS models for probabilistic wind speed forecasting. Quarterly Journal of the Royal Meteorological Society (2015) 141:2289 – 2299.

Barton C, Flouri T, Iliopoulos CS, Pissis SP. Global and local sequence alignment with a bounded number of gaps. Theoretical Computer Science (2015) 582:1–16.

Battaglia N, Bond JR, Pfrommer C, Sievers JL. On the Cluster Physics of Sunyaev-Zel'dovich and X-Ray Surveys. IV. Characterizing Density and Pressure Clumping due to Infalling Substructures. Astrophysical Journal (2015) 806:43.

Baudis N, Barbera P, Graf S, Lutteropp S, Opitz D, Flouri T, Stamatakis A. Two independent and highly efficient open source tkf91 implementations. bioRxiv (2015).

Bauer A, Springel V, Vogelsberger M, Genel S, Torrey P, Sijacki D, Nelson D, Hernquist L. Hydrogen reionization in the Illustris universe. Monthly Notices of the Royal Astronomical Society (2015) 453:3593 – 3610.

Becerra F, Greif TH, Springel V, Hernquist LE. Formation of massive protostars in atomic cooling haloes. Monthly Notices of the Royal Astronomical Society (2015) 446:2380 – 2393.

Beckmann A, Xiao S, Müller JP, Mercadante D, Nüchter T, Kröger N, Langhojer F, Petrich W, Holstein TW, Benoit M, et al. A fast recoiling silk-like elastomer facilitates nanosecond nematocyst discharge. BMC biology (2015) 13:3.

Bruce NJ, Kokh DB, Ozboyaci M, Wade RC. Modelling of Solvation Effects for Brownian Dynamics Simulation of Biomolecular Recognition. In "Computational Trends in Solvation and Transport in Liquids-Lecture Notes, IAS Series, Schriften des Forschungszentrums Jülich", Eds. Sutmann G, Grotendorst J, Gommpper G, Marx D. Forschungszentrum Jülich GmbH, Jülich, Germany (2015), vol. 28, pp. 259 – 280.

Cooper D, Danciger J, Wienhard A. Limits of geometries (2015). Preprint, arXiv:1408.4109.

Cooper AP, Gao L, Guo Q, Frenk CS, Jenkins A, Springel V, White SDM. Surface photometry of brightest cluster galaxies and intracluster stars in ΛCDM. Monthly Notices of the Royal Astronomical Society (2015) 451:2703 – 2722.

Czech L, Stamatakis A. Do phylogenetic tree viewers correctly display support values? bioRxiv (2015).

Darriba D, Flouri T, Stamatakis A. The state of software in evolutionary biology. bioRxiv (2015).

Diehl R, Siegert T, Hillebrandt W, Krause M, Greiner J, Maeda K, Röpke FK, Sim SA, Wang W, Zhang X. SN2014J gamma rays from the 56Ni decay chain. Astronomy & Astrophysics (2015) 574:A72.

Ehm W, Gneiting T, Jordan A, Krüger F. Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings (with discussion and reply). Journal of the Royal Statistical Society Series B (2016).

Elliott G, Ghanem D, Krüger F. Forecasting conditional probabilities of binary outcomes under misspecification. Review of Economics and Statistics (2015).

Feldmann K, Scheuerer M, Thorarinsdottir TL. Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. Monthly Weather Review (2015) 143:955 – 971.

Fissler T, Ziegel JF, Gneiting T. Expected shortfall is jointly elicitable with value-at-risk: Implications for backtesting. Risk.net, January 2016, pp.58 – 61; online: December 2015.

Flouri T, Giaquinta E, Kobert K, Ukkonen E. Longest common substrings with k mismatches. Information Processing Letters (2015a) 115:643 – 647.

Flouri T, Izquierdo-Carrasco F, Darriba D, Aberer A, Nguyen LT, Minh B, Von Haeseler A, Stamatakis A. The phylogenetic likelihood library. Systematic Biology (2015b) 64:356 – 362.

Flouri T, Kobert K, Rognes T, Stamatakis A. Are all global alignment algorithms and implementations correct? bioRxiv (2015c).

Fontanot F, Baldi M, Springel V, Bianchi D. Semi-analytic galaxy formation in coupled dark energy cosmologies. Monthly Notices of the Royal Astronomical Society (2015) 452:978 – 985.

Fuller JC, Martinez M, Henrich S, Stank A, Richter S, Wade RC. LigDig: a web server for querying ligand-protein interactions. Bioinformatics (2015) 31:1147 – 1149.

Gao L, Theuns T, Springel V. Star-forming filaments in warm dark matter models. Monthly Notices of the Royal Astronomical Society (2015) 450:45 – 52.

Garcia-Prada O, Peon-Nieto A, Ramanan S. Higgs bundles and the Hitchin-Kostant-Rallis section (2015). Preprint, arXiv: 1511.02611

Garg D, Skouloubris S, Briffotaux J, Myllykallio H, Wade RC. Conservation and Role of Electrostatics in Thymidylate Synthase. Sci. Rep. (2015) 5:17356.

Genel S, Fall SM, Hernquist L, Vogelsberger M, Snyder GF, Rodriguez-Gomez V, Sijacki D, Springel V. Galactic Angular Momentum in the Illustris Simulation: Feedback and the Hubble Sequence. Astrophysical Journal (2015) 804:L40.

Gerstner P, Schick M, Heuveline V. A multilevel domain decomposition approach for solving time constrained optimal power flow problems. Preprint Series of the Engineering Mathematics and Computing Lab (2015) 4:1 – 32.

Gianniotis N, Kügler SD, Tino P, Polsterer KL, Misra R. Autoencoding Time Series for Visualisation. In "European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning" (2015) pp. 495 – 500.

Grand RJJ, Bovy J, Kawata D, Hunt JAS, Famaey B, Siebert A, Monari G, Cropper M. Spiral- and bar-driven peculiar velocities in Milky Way-sized galaxy simulations. Monthly Notices of the Royal Astronomical Society (2015a) 453:1867–1878.

Grand RJJ, Kawata D, Cropper M. Impact of radial migration on stellar and gas radial metallicity distribution. Monthly Notices of the Royal Astronomical Society (2015b) 447:4018 – 4027.

Grand JJ, Springel V, Gomez FA, Pakmor R, Campbell DJR, Jenkins A. Vertical disc heating in Milky Way-sized galaxies in a cosmological context. ;onthly notices of the Royal Astronomical Society, preprint (2015c), arXiv:1512.02219.

Guéritaud F, Guichard O, Kassel F, Wienhard A. Anosov representations and proper actions (2015a). ArXiv:1502.03811, to appear in Geometry and Topology.

Guéritaud F, Guichard O, Kassel F, Wienhard A. Compactification of certain Clifford-Klein forms of reductive homogeneous spaces (2015b). Preprint, arXiv:1506.03742.

Guichard O, Kassel F, Wienhard A. Tameness of Riemannian locally symmetric spaces arising from anosov representations (2015). Preprint, arXiv:1508.04759.

Hansen LV, Thorarinsdottir TL, Ovcharov E, Gneiting T, Richards D. Gaussian random particles with flexible Hausdorff dimension. Advances in Applied Probability (2015) 47:307 – 327.

Heinzerling B, Strube M. Visual error analysis for entity linking. In "Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics: System Demonstrations", Beijing, China, 26-31 July 2015 (2015) pp. 37 – 42.

Hemri S. Multi-model combination and seamless prediction. In "Handbook of Hydrometeorological Ensemble Forecasting", Eds. Duan Q, Pappenberger F, Thielen J, Wood A, Cloke HL, Schaake JC. Springer-Verlag, Berlin (2015).

Hemri S, Lisniak D, Klein B. Multivariate postprocessing techniques for probabilistic hydrological forecasting. Water Resources Research (2015) 51:7436 – 7451.

Henriques BMB, White SDM, Thomas PA, Angulo R, Guo Q, Lemson G, Springel V, Overzier R. Galaxy formation in the Planck cosmology - I. Matching the observed evolution of star formation rates, colours and stellar masses. Monthly Notices of the Royal Astronomical Society (2015) 451:2663 – 2680.

Hoecker M, Polsterer KL, Kugler SD, Heuveline V. Clustering of complex data-sets using fractal similarity measures and uncertainties. In "Computational Science and Engineering (CSE), 2015 IEEE 18th International Conference on". IEEE (2015) pp. 82 – 91.

Hucka M, Nickerson DP, Bader G, Bergmann FT, Cooper J, Demir E, Garny A, Golebiewski M, Myers CJ, Schreiber F, Waltemath D, Le Novère N. Promoting coordinated development of community-based information standards for modeling in biology: the combine initiative. Frontiers in Bioengineering and Biotechnology (2015) 3.

Hunt JAS, Kawata D, Grand RJJ, Minchev I, Pasetto S, Cropper M. The stellar kinematics of corotating spiral arms in Gaia mock observations. Monthly Notices of the Royal Astronomical Society (2015) 450:2132 – 2142.

Inserra C, Sim SA, Wyrzykowski L, Smartt SJ, Fraser M, Nicholl M, Shen KJ, Jerkstrand A, Gal-Yam A, Howell DA, Maguire K, Mazzali P, Valenti S, Taubenberger S, Benitez-Herrera S, Bersier D, Blagorodnova N, Campbell H, Chen TW, Elias-Rosa N, Hillebrandt W, Kostrzewa-Rutkowska Z, Kozłowski S, Kromer M, Lyman JD, Polshaw J, Röpke FK, Ruiter AJ, Smith K, Spiro S, Sullivan M, Yaron O, Young D, Yuan F. OGLE-2013-SN-079: A Lonely Supernova Consistent with a Helium Shell Detonation. Astrophys. J. Lett. (2015) 799:L2.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. Phylogenomic analyses data of the avian phylogenomics project. GigaScience (2015) 4:1 – 9.

Judea A, Strube M. Event extraction as frame-semantic parsing. In "Proceedings of STARSEM 2015: The Fourth Joint Conference on Lexical and Computational Semantics", Denver, Col., 4 – 5 June 2015 (2015) pp. 159 – 164.

Kobert K, Salichos L, Rokas A, Stamatakis A. Computing the internode certainty and related measures from partial gene trees. bioRxiv (2015).

Kosenko D, Hillebrandt W, Kromer M, Blinnikov SI, Pakmor R, Kaastra JS. Oxygen emission in remnants of thermonuclear supernovae as a probe for their progenitor system. Monthly Notices of the Royal Astronomical Society (2015) 449:1441 – 1448.

Kozlov AM, Aberer AJ, Stamatakis A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics (2015) 31:2577 – 2579.

Kratzke J, Schick M, Heuveline V. Fluid-structure interaction simulation of an aortic phantom with uncertain young's modulus using the polynomial chaos expansion. Applied Mechanics and Materials (2015) 807:34 – 44.

Kratzke J, Schoch N, Weis C, Mueller-Eschner M, Speidel S, Farag M, Beller C, Heuveline V. Enhancing 4d pc-mri in an aortic phantom considering numerical simulations. In "Proceedings of SPIE Medical Imaging 2015: Physics of Medical Imaging" (2015).

Kromer M, Ohlmann ST, Pakmor R, Ruiter AJ, Hillebrandt W, Marquardt KS, Röpke FK, Seitenzahl IR, Sim SA, Taubenberger S. Deflagrations in hybrid CONe white dwarfs: a route to explain the faint Type Iax supernova 2008ha. Monthly Notices of the Royal Astronomical Society (2015) 450:3045 – 3053.

Krüger F, Clark TE, Ravazzolo F. Using entropic tilting to combine BVAR forecasts with external nowcasts. Journal of Business & Economic Statistics (2015a).

Krüger F, Nolte I. Disagreement versus uncertainty: Evidence from distribution forecasts. Journal of Banking & Finance (2015b).

Kügler SD, Gianniotis N, Polsterer KL. Featureless classification of light curves. Monthly Notices of the Royal Astronomical Society (2015a) 451:3385 – 3392.

Kügler SD, Polsterer K, Hoecker M. Determining spectroscopic redshifts by using k nearest neighbor regression. I. Description of method and analysis. Astronomy & Astrophysics (2015b) 576:A132.

Kügler, SD, Gianniotis, N, Polsterer, KL. An explorative approach for inspecting Kepler data, MNRAS (2016), 455:4399-4405, doi:10.1093/mnras/stv2604; online: December 9, 2015.

Kulkarni G, Hennawi JF, Oñorbe J, Rorai A, Springel V. Characterizing the Pressure Smoothing Scale of the Intergalactic Medium. Astrophysical Journal (2015) 812:30.

Lamberts A, Chang P, Pfrommer C, Puchwein E, Broderick AE, Shalaby M. Patchy Blazar Heating: Diversifying the Thermal History of the Intergalactic Medium. Astrophysical Journal (2015) 811:19.

Leaché AD, Banbury BL, Felsenstein J, de Oca AnM, Stamatakis A. Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for inferring snp phylogenies. Systematic Biology (2015) 64:1032 – 1047.

Li W, Baldus IB, Gräter F. Redox potentials of protein disulfide bonds from free-energy calculations. The Journal of Physical Chemistry B (2015) 119:5386 – 5391.

Liu L, Wade RC, Heermann DW. A multiscale approach to simulating the conformational properties of unbound multi-C2H2 zinc finger proteins. Proteins(2015) 83:1604 – 1615.

Liu ZW, Tauris TM, Röpke FK, Moriya TJ, Kruckow M, Stancliffe RJ, Izzard RG. The interaction of core-collapse supernova ejecta with a companion star. Astronomy & Astrophysics (2015) 584:A11.

Louet M, Seifert C, Hensen U, Gräter F. Dynamic allostery of the catabolite activator protein revealed by interatomic forces. PLoSComputational Biology (2015) 11:e1004358.

Marinacci F, Vogelsberger M, Mocz P, Pakmor R. The large-scale properties of simulated cosmological magnetic fields. Monthly Notices of the Royal Astronomical Society (2015) 453:3999-4019.

Marquardt KS, Sim SA, Ruiter AJ, Seitenzahl IR, Ohlmann ST, Kromer M, Pakmor R, Röpke FK. Type Ia supernovae from exploding oxygen-neon white dwarfs. Astronomy & Astrophysics (2015) 580:A118.

Martinez M, Bruce NJ, Romanowska J, Kokh DB, Ozboyaci M, Yu X, Öztürk MA, Richter S, Wade RC. SDA 7: A modular and parallel implementation of the simulation of diffusional association software. J. Comput. Chem. (2015) 36:1631 – 1645.

Martschat S, Claus P, Strube M. Plug latent structures and play coreference resolution. In "Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics: System Demonstrations", Beijing, China, 26 – 31 July 2015 (2015) pp. 61 – 66.

Martschat S, Göckel T, Strube M. Analyzing and visualizing coreference resolution errors. In "Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstration Session", Denver, Col., 31 May – 5 June 2015 (2015) pp. 6 – 10.

Martschat S, Strube M. Latent structures for coreference resolution. Transactions of the Association for Computational Linguistics (2015) 3:405 – 418.

McTavish EJ, Hillis DM. How do snp ascertainment schemes and population demographics affect inferences about population history? BMC Genomics (2015) 16:1 – 13.

McTavish EJ, Hinchliff CE, Allman JF, Brown JW, Cranston KA, Holder MT, Rees JA, Smith SA. Phylesystem: a git-based data store for community-curated phylogenetic estimates. Bioinformatics (2015) 31:2794 – 2800.

McTavish EJ, Steel M, Holder MT. Twisted trees and inconsistency of tree estimation when gaps are treated as missing data – the impact of model mis-specification in distance corrections. Molecular Phylogenetics and Evolution (2015) 93:289 – 295.

Mendoza-Temis JdJ, Wu MR, Langanke K, Martınez-Pinedo G, Bauswein A, Janka HT. Nuclear robustness of the r process in neutron-star mergers. Phys. Rev. C (2015) 92:055805.

Mercadante D, Milles S, Fuertes G, Svergun DI, Lemke EA, Gräter F. Kirkwood-buff approach rescues overcollapse of a disordered protein in canonical protein force fields. The journal of physical chemistry B (2015) 119:7975 – 7984.

Mesgar M, Strube M. Graph-based coherence modeling for assessing readability. In "Proceedings of STARSEM 2015: The Fourth Joint Conference on Lexical and Computational Semantics", Denver, Col., 4-5 June 2015 (2015) 309 – 318.

Meyer-Hübner N, Suriyah M, Leibfried T, Slednev V, Bertsch V, Fichtner W, Gerstner P, Schick M, Heuveline V. Time constrained optimal power flow calculations on the german power grid. In "Proceedings of International ETG Congress, Bonn" (2015).

Miczek F, Röpke FK, Edelmann PVF. New numerical solver for flows at various Mach numbers. Astronomy & Astrophysics (2015) 576:A50.

Milles S, Mercadante D, Aramburu IV, Jensen MR, Banterle N, Koehler C, Tyagi S, Clarke J, Shammas SL, Blackledge M, et al. Plasticity of an ultrafast interaction between nucleoporins and nuclear transport receptors. Cell (2015) 163:734 – 745.

Mocz P, Vogelsberger M, Pakmor R, Genel S, Springel V, Hernquist L. Reducing noise in moving-grid codes with strongly-centroidal Lloyd mesh regularization. Monthly Notices of the Royal Astronomical Society (2015) 452:3853 – 3862.

Muñoz DJ, Kratter K, Vogelsberger M, Hernquist L, Springel V. Stellar orbit evolution in close circumstellar disc encounters. Monthly Notices of the Royal Astronomical Society (2015) 446:2010 – 2029.

Mustafa G, Nandekar PP, Yu X, Wade RC. On the application of the MARTINI coarse-grained model to immersion of a protein in a phospholipid bilayer. J. Chem. Phys. (2015) 143:243139.

Nelson D, Genel S, Vogelsberger M, Springel V, Sijacki D, Torrey P, Hernquist L. The impact of feedback on cosmological gas accretion. Monthly Notices of the Royal Astronomical Society (2015) 448:59 – 74.

Nelson D, Pillepich A, Genel S, Vogelsberger M, Springel V, Torrey P, Rodriguez-Gomez V, Sijacki D, Snyder GF, Griffen B, Marinacci F, Blecha L, Sales L, Xu D, Hernquist L. The illustris simulation: Public data release. Astronomy and Computing (2015) 13:12 – 37.

Nillegoda NB, Kirstein J, Szlachcic A, Berynskyy M, Stank A, Stengel F, Arnsburg K, Gao X, Scior A, Aebersold R, Guilbride DL, Wade RC, Morimoto RI, Mayer MP, Bukau B. Crucial HSP70 co-chaperone complex unlocks metazoan protein disaggregation. Nature (2015) 524:247 – 251.

Ovcharov E. Existence and uniqueness of proper scoring rules. Journal of Machine Learning Research (2015) 16:2207 – 2230.

Pan YC, Foley RJ, Kromer M, Fox OD, Zheng W, Challis P, Clubb KI, Filippenko AV, Folatelli G, Graham ML, Hillebrandt W, Kirshner RP, Lee WH, Pakmor R, Patat F, Phillips MM, Pignata G, Röpke F, Seitenzahl I, Silverman JM, Simon JD, Sternberg A, Stritzinger MD, Taubenberger S, Vinko J, Wheeler JC. 500 days of SN 2013dy: spectra and photometry from the ultraviolet to the infrared. Monthly Notices of the Royal Astronomical Society (2015) 452:4307 – 4325.

Parveen D, Ramsl HM, Strube M. Topical coherence for graph-based extractive summarization. In "Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing", Lisbon, Portugal, 17 – 21 September 2015 (2015) pp. 1949 – 1954.

Parveen D, Strube M. Integrating importance, non-redundancy and coherence in graph-based extractive summarization. In "Proceedings of the 24th International Joint Conference on Artificial Intelligence", Buenos Aires, Argentina, 25 – 31 July 2015 (2015) pp. 1298 – 1304.

Peon-Nieto A. Cameral data for SU(p+1,p) Higgs bundles. Preprint, arxiv:1506.01318.

Peralta D, Bronowska AK, Morgan B, Dóka É, Van Laer K, Nagy P, Gräter F, Dick TP. A proton relay enhances h2o2 sensitivity of gapdh to facilitate metabolic adaptation. Nature chemical biology (2015) 11:156 – 163.

Polsterer KL, Gieseke F, Gianniotis N, Kügler SD. Analyzing Complex and Structured Data via Unsupervised Learning Techniques. IAU General Assembly (2015) 22:2258115.

Polsterer KL, Gieseke F, Igel C. Automatic Galaxy Classification via Machine Learning Techniques: Parallelized Rotation/Flipping INvariant Kohonen Maps (PINK). In "Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV)", Eds. Taylor AR, Rosolowsky E (2015) vol. 495 of Astronomical Society of the Pacific Conference Series, p. 81.

Richardson D, Hemri S, Bogner K, Gneiting T, Haiden T, Pappenberger F, Scheuerer M. Calibration of ECMWF forecasts. ECMWF Newsletter (2014/15) 142:12 – 16.

Rodriguez-Gomez V, Genel S, Vogelsberger M, Sijacki D, Pillepich A, Sales LV, Torrey P, Snyder G, Nelson D, Springel V, Ma CP, Hernquist L. The merger rate of galaxies in the Illustris simulation: a comparison with observations and semi-empirical models. Monthly Notices of the Royal Astronomical Society (2015) 449:49 – 64.

Romanowska J, Kokh DB, Fuller JC, Wade RC. Computational Approaches for Studying Drug Binding Kinetics. In "Kinetics and Thermodynamics of Drug Binding Kinetics and Thermodynamics of Drug Binding", Eds. Keserü GM, Swinney DC. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, (2015a) vol. 65, pp. 7508 – 7513.

Romanowska J, Kokh DB, Wade RC. When the Label Matters: Adsorption of Labeled and Unlabeled Proteins on Charged Surfaces. Nano Lett. (2015b) 15:7508 – 7513.

Rosa M, Kokh DB, Corni S, Wade RC, Di Felice R. Docking of DNA Duplexes on a Gold Surface. J. Self-Assembly and Molecular Electronics (SAME) (2015) 3(7):1 – 18.

Sales LV, Vogelsberger M, Genel S, Torrey P, Nelson D, Rodriguez-Gomez V, Wang W, Pillepich A, Sijacki D, Springel V, Hernquist L. The colours of satellite galaxies in the Illustris simulation. Monthly Notices of the Royal Astronomical Society (2015) 447:L6 – L10.

Salo-Ahen OMH, Tochowicz A, Pozzi C, Cardinale D, Ferrari S, Boum Y, Mangani S, Stroud RM, Saxena P, Myllykallio H, Costi MP, Ponterini G, Wade RC. Hotspots in an Obligate Homodimeric Anticancer Target. Structural and Functional Effects of Interfacial Mutations in Human Thymidylate Synthase. J. Med. Chem. (2015) 58:3572 – 3581.

Sanderson MJ, McMahon MM, Stamatakis A, Zwickl DJ, Steel M. Impacts of terraces on phylogenetic inference. Systematic Biology (2015) 64:709 – 726.

Schaal K, Bauer A, Chandrashekar P, Pakmor R, Klingenberg C, Springel V. Astrophysical hydrodynamics with a high-order discontinuous Galerkin scheme and adaptive mesh refinement. Monthly Notices of the Royal Astronomical Society (2015) 453:4278 – 4300.

Schaal K, Springel V. Shock finding on a moving mesh – I. Shock statistics in non-radiative cosmological simulations. Monthly Notices of the Royal Astronomical Society (2015) 446:3992 – 4007.

Schefzik R. Multivariate discrete copulas, with applications in probabilistic weather forecasting. Publications de'l Institut de Statistique de'l Université de Paris (2015) 59:87 – 116.

Schoch N, Engelhardt S, Zimmermann N, Speidel S, Simone RD, Wolf I, Heuveline V. Integration of a biomechanical simulation for mitral valve reconstruction into a knowledge-based surgery assistance system. In "Proceedings of SPIE Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling" (2015).

Schreiber F, Bader GD, Golebiewski M, Hucka M, Kornmeier B, Le Novère N, Myers CJ, Nickerson DP, Sommer B, Waltemath D, Weise S. Specifications of standards in systems and synthetic biology. Journal of Integrative Bioinformatics (2015) 12:258.

Seitenzahl IR, Herzog M, Ruiter AJ, Marquardt K, Ohlmann ST, Röpke FK. Neutrino and gravitational wave signal of a delayed-detonation model of type Ia supernovae. Phys. Rev. D (2015) 92:124013.

Seitenzahl IR, Summa A, Krauß F, Sim SA, Diehl R, Elsässer D, Fink M, Hillebrandt W, Kromer M, Maeda K, Mannheim K, Pakmor R, Röpke FK, Ruiter AJ, Wilms J. 5.9-keV Mn K-shell X-ray luminosity from the decay of 55Fe in Type Ia supernova models. Monthly Notices of the Royal Astronomical Society (2015) 447:1484 – 1490.

Seybold C, Elserafy M, Ruthnick D, Ozboyaci M, Neuner A, Flottmann B, Heilemann M, Wade RC, Schiebel E. Kar1 binding to Sfi1 C-terminal regions anchors the SPB bridge to the nuclear envelope. J. Cell Biol. (2015) 209:843 – 861.

Sijacki D, Vogelsberger M, Genel S, Springel V, Torrey P, Snyder GF, Nelson D, Hernquist L. The Illustris simulation: the evolving population of black holes across cosmic time. Monthly Notices of the Royal Astronomical Society (2015) 452:575 – 596.

Simpson CM, Bryan GL, Hummels C, Ostriker JP. Kinetic Energy from Supernova Feedback in High-resolution Galaxy Simulations. Astrophysical Journal (2015) 809:69.

Snyder GF, Torrey P, Lotz JM, Genel S, McBride CK, Vogelsberger M, Pillepich A, Nelson D, Sales LV, Sijacki D, Hernquist L, Springel V. Galaxy morphology and star formation in the Illustris Simulation at z = 0. Monthly Notices of the Royal Astronomical Society (2015) 454:1886 – 1908.

Song C, Schick M, Heuveline V. A polynomial chaos method for uncertainty quantification in blood pump simulation. In "Proceedings of the 1st International Conference on Uncertainty Quantification in Computational Sciences and Engineering (UNCECOMP)" (2015) .

Sparre M, Hayward CC, Springel V, Vogelsberger M, Genel S, Torrey P, Nelson D, Sijacki D, Hernquist L. The star formation main sequence and stellar mass assembly of galaxies in the Illustris simulation. Monthly Notices of the Royal Astronomical Society (2015) 447:3548 – 3563.

Stamatakis A. Using raxml to infer phylogenies. Current Protocols in Bioinformatics (2015) pp. 6 – 14.

Stanford NJ, Wolstencroft K, Golebiewski M, Kania R, Juty N, Tomlinson C, Owen S, Butcher S, Hermjakob H, Le Novère N, Mueller W, Snoep J, Goble C. The evolution of standards and data management practices in systems biology. Molecular Systems Biology (2015) 11.

Suresh J, Bird S, Vogelsberger M, Genel S, Torrey P, Sijacki D, Springel V, Hernquist L. The impact of galactic feedback on the circumgalactic medium. Monthly Notices of the Royal Astronomical Society (2015) 448:895 – 909.

Tomic A, Berynskyy M, Wade RC, Tomic S. Molecular simulations reveal that the long range fluctuations of human DPP III change upon ligand binding. Mol. BioSyst. (2015) 11:3068 – 3080.

Torrey P, Snyder GF, Vogelsberger M, Hayward CC, Genel S, Sijacki D, Springel V, Hernquist L, Nelson D, Kriek M, Pillepich A, Sales LV, McBride CK. Synthetic galaxy images and spectra from the Illustris simulation. Monthly Notices of the Royal Astronomical Society (2015) 447:2753 – 2771.

Torrey P, Wellons S, Machado F, Griffen B, Nelson D, Rodriguez-Gomez V, McKinnon R, Pillepich A, Ma CP, Vogelsberger M, Springel V, Hernquist L. An analysis of the evolving comoving number density of galaxies in hydrodynamical simulations. Monthly Notices of the Royal Astronomical Society (2015) 454:2770 – 2786.

Travaglio C, Gallino R, Rauscher T, Röpke FK, Hillebrandt W. Testing the Role of SNe Ia for Galactic Chemical Evolution of p-nuclei with Two-dimensional Models and with s-process Seeds at Different Metallicities. Astrophys. J. (2015) 799:54.

Wagner JA, Mercadante D, Nikić I, Lemke EA, Gräter F. Origin of orthogonality of strain-promoted click reactions. Chemistry: A European Journal (2015) 21:12431 – 12435.

Wellons S, Torrey P, Ma CP, Rodriguez-Gomez V, Vogelsberger M, Kriek M, van Dokkum P, Nelson E, Genel S, Pillepich A, Springel V, Sijacki D, Snyder G, Nelson D, Sales L, Hernquist L. The formation of massive, compact galaxies at z = 2 in the Illustris simulation. Monthly Notices of the Royal Astronomical Society (2015) 449:361 – 372.

Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, Weidemann A, Bittkowski M, An L, Shockley D, Snoep JL, Mueller W, Goble C. Seek: a systems biology data and model management platform. BMC Systems Biology (2015) 9:1 – 12.

Xu D, Sluse D, Gao L, Wang J, Frenk C, Mao S, Schneider P, Springel V. How well can cold dark matter substructures account for the observed radio flux-ratio anomalies. Monthly Notices of the Royal Astronomical Society (2015) 447:3189 – 3206.

Yu X, Cojocaru V, Mustafa G, Salo-Ahen OMH, Lepesheva GI, Wade RC. Dynamics of CYP51: implications for function and inhibitor design. J. Molec. Recogn. (2015a) 28:59 – 73.

Yu X, Martinez M, Gable AL, Fuller JC, Bruce NJ, Richter S, Wade RC. webSDA: a web server to simulate macromolecular diffusional association. Nucl. Acids Res. (2015b) 43:W220 – W224.

Yurin D, Springel V. The stability of stellar discs in Milky Way-sized dark matter haloes. Monthly Notices of the Royal Astronomical Society (2015) 452:2367 – 2387.

Zhou J, Aponte-Santamaría C, Sturm S, Bullerjahn JT, Bronowska A, Gräter F. Mechanism of focal adhesion kinase mechanosensing. PLoS Computational Biology (2015) 11:e1004593.

Zhou J, Bronowska A, Le Coq J, Lietha D, Gräter F. Allosteric regulation of focal adhesion kinase by pip 2 and atp. Biophysical journal (2015) 108:698 – 705.

Zhu C, Pakmor R, van Kerkwijk MH, Chang P. Magnetized Moving Mesh Merger of a Carbon-Oxygen White Dwarf Binary. Astrophysical Journal (2015) 806:L1.

Degrees

Andre J. Aberer:
"Algorithmic Advancements and Massive Parallelism for Large-Scale Datasets in Phylogenetic Bayesian Markov Chain Monte Carlo" Ph.D. Thesis, Karlsruhe Institute of Technology and HITS: Alexandros Stamatakis, 2015.

Andreas Bauer:
"Reionization in the Illustris Universe and Novel Numerical Methods" Ph.D. Thesis, Physics, Heidelberg University and HITS: Volker Springel, 2015.

Nicolas Bellm:
"Analyse von Twitter-Nachrichten zum US-Präsidentschaftswahlkampf 2012" Master's Thesis, Department of Modern Languages, Heidelberg University and HITS: Michael Strube, 2015.

Michael Berinski:
"Macromolecular Interactions: in silico prediction of the structures of complexes between proteins" Ph.D. Thesis, Faculty of Informatics and Mathematics, Johann Wolfgang Goethe Universität, Frankfurt am Main, and HITS: Rebecca Wade, 2015.

Angela Fahrni:
"Joint Discourse-aware Concept Disambiguation and Clustering" Ph.D. Thesis, Neuphilologische Fakultät, Heidelberg University and HITS: Michael Strube, 2015.

Gaurav Kumar Ganotra:
"Methods to Compute and Investigate Drug Binding Kinetics" Master's Thesis, Life Science Informatics, University of Bonn and HITS: Daria Kokh & Rebecca Wade, 2015.

Eric Hildebrand:
"Constructing Hierarchical Timelines of Evolving News Topics from Twitter" Master's Thesis, Department of Modern Languages, Heidelberg University and HITS: Michael Strube, 2015.

Max Horn:
"Clustering and Scoring the Druggability of Transient Protein Pockets" Bachelor's Thesis, Molecular Biotechnology, Heidelberg University and HITS: Antonia Stank & Rebecca Wade, 2015.

Svenja Jacob:
"Streaming cosmic rays in cool core clusters" Master's Thesis, Physics, Heidelberg University and HITS: Christoph Pfrommer, 2015.

Dennis Kügler:
"On the application of machine learning approaches in astronomy: Exploring novel representations of high-dimensional and complex astronomical data" Ph.D. Thesis, Astronomy, Heidelberg University and HITS: Kai Polsterer, 2015.

Stefan Lambert:
"Combining Forecasts for the United Kingdom Economy" Diploma Thesis, Mathematics, Karlsruhe Institute of Technology and HITS: Tilmann Gneiting, 2015.

Sandeep Patil:
"Multi-scale simulations of silk mechanics" PhD Thesis, RWTH Aachen and HITS: Frauke Gräter, 2015.

Roman Schefzik:
"Physically Coherent Probabilistic Weather Forecasts Using Multivariate Discrete Copula-Based Ensemble Postprocessing Methods" Ph.D. Thesis, Mathematics, Heidelberg University and HITS: Tilmann Gneiting, 2015.

Constantin Scholl:
"The divisible load balance problem with shared cost and its application to phylogenetic inference" Bachelor Thesis, Karlsruhe Institute of Technology and HITS: Alexandros Stamatakis, 2015.

Sebastian Schulz:
"Spectral Networks – A story of Wall-Crossing in Geometry and Physics" Master Thesis, Heidelberg University and HITS: Anna Wienhard, 2015.

Anja Summa:
"Emotion Recognition from Microblogs in the Urban Context with Spatio-Temporal Information" Master's Thesis, Department of Modern Languages, Heidelberg University and HITS: Michael Strube, 2015.

Max Waldhauer:
"Brownian Dynamics Simulations of Chymotrypsin Inhibitor 2 in Concentrated Protein Solutions" Bachelor's Thesis, Molecular Biotechnology, Heidelberg University and HITS: Neil Bruce & Rebecca Wade, 2015.

Xiaofeng Yu:
"Multiscale Simulations of Cytochrome P450 Systems" Ph. D. Thesis, Faculty of Biosciences, Heidelberg University and HITS: Rebecca Wade, 2015.

Jiajie Zhang:
"Models and Algorithms for Phylogenetic Marker Analysis" Ph.D. Thesis, University of Lübeck and HITS: Alexandros Stamatakis, 2015.

Jing Zhou:
"Molecular force-sensing mechanism of focal adhesion kinase" PhD Thesis, Heidelberg University and HITS: Frauke Gräter, 2015.

Mengfei Zhou:
"Cross-lingual Semi-Supervised Modality Tagging" Bachelor's Thesis, Department of Modern Languages, Heidelberg University and HITS: Michael Strube, 2015.

Lectures, Courses and Seminars

Camilo Aponte-Santamaría:
Bioinformatics course, Heidelberg University, January 26 – 30, 2015.

Michael Bromberger:
Wolfgang Karl Praktikum "Hardware-system design" Karlsruhe Institute of Technology. Summer term 2015, winter term 2015/16.

Tilmann Gneiting:
Lecture course on "Forecasting: Theory and practice I" Karlsruhe Institute of Technology, October 2014 – February 2015.

Tilmann Gneiting, Alexander Jordan:
Lecture course and Exercises on "Forecasting: Theory and practice II" Karlsruhe Institute of Technology, April – July 2015, Summer term 2015.

Tilmann Gneiting, Sebastian Lerch:
Seminar on "Statistical forecasting" Karlsruhe Institute of Technology, October 2015 – February 2016.

Martin Golebiewski, Antonia Stank:
COMBINE Tutorial "Modelling and Simulation Tools in Systems Biology" National University of Singapore (NUS), Singapore, November 25, 2015

Frauke Gräter:
Bioinformatics course 2015 (with Prof. Rebecca Wade) Heidelberg, Germany, January 2015.

"Deciphering protein function by molecular simulations" HBIGS course Heidelberg, Germany, April 2015.

"Computational molecular biophysics" Lecture with practicals, Heidelberg University, from Oct 2015 to Feb 2016, for physics and biology master students (with Rebecca Wade).

"Biomolecular simulations" (with Prof. Rebecca Wade), Contribution to lecture at the Graduate Physics Days, Heidelberg University, October 2015.

Vincent Heuveline:
Vorlesung "Uncertainty Quantification", Heidelberg University. Winter term 2015 / 16.

Vorlesung „Numerische Mathematik 3 – Numerische Methoden der Kontinuumsmechanik", Heidelberg University. Summer term 2015.

Seminar "IT-Cybersecurity" Heidelberg University. Winter term 2015 / 16.

Seminar "Uncertainty Quantification" Heidelberg University. Summer term 2015.

PhD Colloquium, Engineering Mathematics and Computing Lab (EMCL) Trier, Germany, September 21 – 25, 2015.

Olga Krebs, Wolfgang Müller:
Data management practicals at 7th International Practical Course in Systems Biology. University of Gothenburg, Sweden, June 1 – 12, 2015.

Wolfgang Müller:
Bachelor-Seminar zur Medieninformatik "Games, Interaktion und Lernen" Otto-Friedrich-Universität Bamberg, Lehrstuhl für Medieninformatik, summer term 2015.

Rüdiger Pakmor, Volker Springel:
"Cosmic Explosions" Department of Physics and Astronomy, Heidelberg University (October 2014 – February 2015).

Christoph Pfrommer:
"The Physics of Galaxy Clusters" Department of Physics and Astronomy, Heidelberg University, (October 2015 – February 2016).

"Cosmology" Department of Physics and Astronomy, Heidelberg University (October 2014 – February 2015).

"The Hertzsprung-Russel Diagramm" Lecture at University Erlangen-Nürnberg, May 4, 2015.

"How to determine the mass of extrasolar planets" Lecture at University Oslo, November 11, 2015.

Andreas Reuter:
Seminar „Datenbanken für Wissenschaftliches Arbeiten (in German)" Heidelberg University, summer term 2015.

Friedrich Röpke (lecturer), S. Ohlmann (tutor), P. Edelmann (tutor):
"Computational Astrophysics" Heidelberg University, summer term 2015.

F. Röpke (lecturer), A. Bauswein (lecturer), A. Michel (tutor):
"White Dwarfs, Neutron Stars and Black Holes – compact objects in astrophysics" Heidelberg University, winter term 2015 / 16.

Michael Schick:
Vorlesung „Mathematik / Informatik B" Heidelberg University. Summer term 2015.

Seminar „Uncertainty Quantification" Heidelberg University. Summer term 2015.

Volker Springel, Rüdiger Pakmor:
"Fundamentals of Simulation Methods" Department of Physics and Astronomy, Heidelberg University (October 2015 – February 2016).

Volker Springel, Karl-Heinz Brenner, Joachim Wambsganß:
Seminar on "Computational Methods in Optics and Astrophysics", Department of Physics and Astronomy, Heidelberg University (April 2015 – July 2015).

Alexandros Stamatakis, Alexey Kozlov, Tomas Flouri, Kassian Kobert, Andre Aberer, Mark Holder:
Regular class: Introduction to Bioinformatics for Computer Scientists, Karlsruhe Institute of Technology (KIT), winter term 2014 / 15.

Alexandros Stamatakis:
Main seminar: Hot Topics in Bioinformatics, KIT, summer term 2015.

Alexandros Stamatakis, Tomas Flouri:
Programming Practical: Hands-on Bioinformatics Practical, KIT, summer term 2015.

Alexandros Stamatakis, Alexey Kozlov, Tomas Flouri, Diego Darriba, Lucas Czech:
Regular class: Introduction to Bioinformatics for Computer Scientists, KIT, winter term 2015 / 16

Alexandros Stamatakis, Paschalia Kapli:
Joint Wellcome Trust-EMBL-EBI Advanced Course on Computational Molecular Evolution, Sanger Centre, Hinxton, UK, 13 – 24 April 2015.

Michael Strube:
PhD Colloquium, Department of Computational Linguistics, Heidelberg University (October 2014 – February 2015).

Seminar: "Entity Linking", Department of Computational Linguistics, Heidelberg University (October 2014 – February 2015).

PhD Colloquium, Department of Computational Linguistics, Heidelberg University (April 2015 – July 2015).

Rebecca Wade:
Module 4," Protein Dynamics and Biomolecular Recognition: Insights from Simulations" M.Sc. Molecular Cell Biology, Heidelberg University, Feb 27, 2015.

Module 3,"Protein Modeling" M.Sc. Molecular Cell Biology, Heidelberg University, Apr 30, May 5, 2015.

Ringvorlesung "Structure and Dynamics of Biological Macromolecules", "Electrostatics, solvation and protein interactions", B.Sc. Biosciences, Heidelberg University, June 18, 2015.

Ringvorlesung „Biophysik", "Receptor-Ligand Interactions: Structure and Dynamics", B.Sc. Molecular Biotechnology, Heidelberg University, Dec 12, 2015.

Rebecca Wade and Frauke Gräter:
"Biomolecular Simulation" lecture course, Physics Graduate Days, Heidelberg University, Oct 5 – 9, 2015.

Rebecca Wade, Neil Bruce, Anna Feldman-Salit, Ghulam Mustafa, Musa Ozboyaci, Ina Pöhner, Stefan Richter, Antonia Stank (MCM), Frauke Gräter, Camilo Aponte-Santamaria, Eduardo Cruz-Chú, Jing Zhou (MBM):
"B.Sc. Biosciences practical course "Grundkurs Bioinformatik" Heidelberg University, 26 – 30 January 2015.

Anna Wienhard:
Hauptseminar Geometrie, Heidelberg University, Spring and Fall, 2015.

Anna Wienhard, Gye-Seon Lee:
Introduction to Riemannian Geometry, Seminar, Heidelberg University, Spring 2015.

Anna Wienhard, Andreas Ott:
Juniorseminar Geometry, Heidelberg University, Spring and Fall 2015.

Anna Wienhard, Anna Marciniak-Czochra, Andreas Ott:
Structures and Mathematics, Seminar, Heidelberg University, Fall 2015.

Anna Wienhard, Daniele Alessandrini:
Themen der Geometrie, Lecture Course, Heidelberg University, Spring 2015.

Peter Zaspel:
Vorlesung „Mathematik / Informatik A", Heidelberg University, winter term 2015 / 16.

9.1 Guest Speaker Activities

Camilo Aponte-Santamaria:

"Elucidating the mechanism of function of biomolecules through molecular dynamics simulations", XIV Latin American Workshop on Nonlinear Phenomena (LAWNP), Cartagena de Indias, Colombia, September 2015; Seminar, Simulation of physical systems group, National University of Colombia, Bogotá, Colombia, October 2015.

"Mechanosensitive Von Willebrand Factor Protein-Protein Interactions Regulate Hemostasis", Seminar, Biological physics group, Boston University, Boston, USA, February 2015; Seminar, Theoretical soft materials group, MIT, Cambridge, USA, June 2015; Seminar, Biomolecular dynamics workgroup, TUM, Munich, Germany. November 2015.

"Molecular driving forces defining lipid positions around Aquaporin-0", Seminar, Molecular Simulation Center, Calgary University, Calgary, Canada, February 2015; Seminar, MD group, Groningen University Groningen, The Netherlands, March 2015.

Andreas Bauswein:

"Gravitational waves and heavy elements from neutron-star mergers", Invited colloquium talk at the University of Ferrara, Italy, September 2015.

Agnieszka Bronowska:

"Adventures with Structure-Based Design", INVISTA Performance Technologies, Redcar, UK, February 2015.

"The devil is in the detail: halogen bonding, protein dynamics, and the structure-based design", School of Chemistry, Newcastle University, Newcastle UK, May 2015.

"Drug development in academia", Proteros GmbH, Munich, May 2015.

"Enthalpy, entropy, and structure-based design" School of Chemistry, Newcastle University, Newcastle, UK, November 2015.

Neil Bruce:

"Molecular Modelling and Simulation", Joint CIBF-HBP Science Workshop, Prato, Italy, May 27 – 29, 2015.

"Bioinformatics and Molecular Simulation Approaches for Constraining Systems Biology Models", CECAM Workshop, Computational Approaches to Chemical Senses, Forschungszentrum Jülich, Germany, September 9 – 11, 2015.

Tilmann Gneiting:

"Statistical post-processing of ensemble forecasts: Current developments and future directions", European Centre for Medium-Range Weather Forecasts (ECMWF), Reading (UK), February 11, 2015.

"Evaluating forecasts: Why proper scoring rules and consistent scoring functions matter", International Journal of Forecasting Editor's Invited Lecture, International Symposium on Forecasting, Riverside, California (United States), June 23, 2015. (remote lecture via video).

"Of quantiles and expectiles: Consistent scoring functions, Choquet representations, and forecast rankings", Ordinary Meeting, Royal Statistical Society, London (UK), December 9, 2015.

Martin Golebiewski:

"Standardising activities in systems biology and beyond: COMBINE and ISO", ERASysAPP workshop "Networking academia industry", Belval, Luxembourg, October 7 – 9, 2015.

Frauke Gräter:

"Structural determinants of silk mechanics from multi-scale simulations", Linz Winter Workshop on Biophysics, Linz, Austria, January 2015.

"Visualization of the NAWIK", Panel discussion at National Center for Science Communication at KIT, Karlsruhe, Germany, April 2015.

"Phase segragation in silk fibers upon stretch from multi-scale modelling" Mini-symposium on Silk Mechanics at the European Solid Mechanics Conference, Madrid, Spain, July 2015.

"Protein dynamics and function from simulations" Lab retreat – Melchior lab, Bingen, Germany, July 28 – 29, 2015.

"Function of von Willebrand factor in primary and secondary hemostasis", International Conference, Hamburg, Germany, September 2015.

Stephan Hemri:

"Copula based approaches to model dependence structures of hydrologic forecasts", Swiss Federal Institute for Forest, Snow and Landscape Research, Birmensdorf , Switzerland, October 27, 2015.

Katra Kolšek:

"In silico identification of endocrine disrupting chemicals – endocrine disruptome", Endocrine disrupting chemicals – from molecule to man, Ljubljana, Slovenia, April 2015.

Fabian Krüger:

"Probabilistic forecasting and predictive model assessment based on MCMC output", University of Oslo, Oslo (Norway), September 29, 2015.

"Combining density forecasts under various scoring rules: An analysis of UK inflation", Conference on Computational and Financial Econometrics (CFE), London (UK), December 14, 2015.

Sebastian Lerch:

"Forecaster's dilemma: Extreme events and forecast evaluation", University of Exeter, Exeter (UK), March 26, 2015; Met Office, Exeter (UK), March 27, 2015.

Davide Mercadante:

"Traversing the way to the nucleus: structural and kinetic dissection of Nup153-Importin-β interaction", Department of Chemistry, The University of Auckland, Auckland, New Zealand, January 2015.

Wolfgang Müller:

"FAIRDOM: Standards-compliant data management" ERASysAPP workshop "Networking academia industry". Belval, Luxembourg, October 7-9, 2015.

Rüdiger Pakmor:

"Magnetohydrodynamics on a moving mesh and its application to galaxies and WD mergers", Niels Bohr Institute, Copenhagen, Denmark, July 13, 2015.

Ana Peon-Nieto:

"Higgs bundles for real groups and the Hitching system", AIM, San Jose, California, September 2015; Conference Fifty Years of the Narasimhan-Seshadri Theorem, Chennai Mathematical Institute, India, October 2015.

"Une construction camérale pour des fibres de Higgs pour des groupes quasi-déployés", Algebraic geometry seminar, University of Rennes, France, September 2015.

"Fibrés de Higgs, groupes réels et le système de Hitchin", Geometry, topology and dynamics seminar, University of Marseille, France, March 2015; Algebra, geometry and topology seminar, Nice University, France, March 2015; Algebraic geometry seminar, Lille University, France, April 2015.

Christoph Pfrommer:

"Cosmic ray feedback in galaxies and cool core clusters", Astronomy Colloquium, University of Wisconsin, Madison, USA, February 26, 2015; Cosmology Colloquium, Perimeter Institute, Waterloo, Canada, April 20 – 21, 2015; Fluids Seminar, Canadian Institute for Theoretical Astrophysics, Toronto, Canada, Apr 22 – 23, 2015; Astrophysics Seminar, Ben Gurion University, Israel, May 20, 2015; Colloquium at Max-Planck Institute for Radioastronomy, Bonn, Germany, October 2, 2015.

"Interfacing High-Energy Astrophysics and Cosmological Structure Formation", Seminar at Niels Bohr Institute, Copenhagen, Denmark, April 9, 2015; Colloquium at University of Erlangen/ Nürnberg, Erlangen, May 4, 2015; Seminar at Oslo University, November 11, 2015.

"The impact of cosmic rays on galaxy formation: passive spectators or active drivers?", Accelerating Cosmic Ray Comprehension at Princeton University, USA, April 13 – 17, 2015.

"Self-interacting dark matter", Cosmology Discussion, Perimeter Institute, Waterloo, Canada, April 20 – 21, 2015; Cosmology Seminar, Canadian Institute for Theoretical Astrophysics, Toronto, Canada, April 22 – 23, 2015.

"The physics and cosmology of TeV blazars", Astrophysics Seminar, Weizmann Institute, Israel, May 17, 2015; Nonthermal Processes in Astrophysical Phenomena, University of Minneapolis, Minnesota, USA, June 10 – 12, 2015.

"Radio galaxies in clusters: cosmic weather stations or novel probes of cluster physics?", Nonthermal Processes in Astrophysical Phenomena, University of Minneapolis, Minnesota, USA, June 10 – 12, 2015.

"AGN feedback: mechanical versus cosmic-ray heating", Intra-cluster medium physics and modelling at Max-Planck Institute for Astrophysics, Garching, June 15 – 17, 2015.

"On the cluster physics of Sunyaev-Zel'dovich and X-ray surveys", Intra-cluster medium physics and modelling at Max-Planck Institute for Astrophysics, Garching, June 15 – 17, 2015.

"Cosmic rays and magnetic fields in galaxies", Cosmic magnetic fields at Ringberg Castle, Germany, July 1 – 4, 2015.

"Radio halos and relics in galaxy clusters", Cosmic magnetic fields at Ringberg Castle, Germany, July 1 – 4, 2015.

"Large-scale shocks and extragalactic cosmic rays", International Meeting on Bayesian Inference of the Galactic magnetic field at ISSI Bern, Switzerland, October 28, 2015.

Friedrich Röpke:

"Stars on fire – Simulating astrophysical burning in stellar evolution and thermonuclear supernovae", Astrophysical Colloquium, Universität Göttingen, Göttingen, Germany, February 12, 2015.

"Typ Ia Supernovae – wie explodierende Sterne unser kosmologisches Weltbild erschütterten", public talk at the Astronomie Stiftung Trebur, Trebur, Germany, 17 April 2015.

"Modeling low Mach number flows in astrophysical systems with preconditioned compressible schemes", (together with Wasilij Barsukow), invited talks at the workshop "Asymptotic Preserving and Multiscale Methods for Kinetic and Hyperbolic Problems"; University of Wisconsin-Madison, Madison, WI, USA, May 8, 2015.

Invited talk at "Higher Order Numerical Methods for Evolutionary PDEs: Applied Mathematics Meets Astrophysical Applications", Banff International Research Station, Canada, May 13, 2015; Physics Colloquium, Technische Universität Darmstadt, Darmstadt, Germany, June 19, 2015.

"Modeling Type Ia supernova explosions", Invited talk at "Nuclear Physics in Astrophysics VII", York, UK, May 19, 2015; Invited talk at the 14th Marcel Grossmann Meeting, Session BN4, Rome, Italy, July 16, 2015; Talk at the AAAstro Seminar, Macquarie University, Sydney, November 6, 2015.

"Modeling Type Ia supernovae", invited talk at Irsee Symposium "Symmetries and Phases in the Universe", Irsee, Germany, June 22, 2015.

"Turbulent flames in astrophysical combustion", invited talk at the 25th ICDERS meeting, Leeds, UK, August 4, 2015.

"Modeling Type Ia supernova explosions and their nucleosynthesis", invited talk at the annual meeting of the German Astronomical Society, Splinter F: "Chemical Oddballs in the Galaxy", Kiel, September 18, 2015.

"Type Ia supernova simulations", talk at workshop "Nucleosynthesis away from stability", Goethe-Universität Frankfurt, Frankfurt, Germany, October 8, 2015.

Volker Springel:

"Forming Galaxies on a Supercomputer", Astrophysical Colloquium, Research Institute in Astrophysics and Planetology, Toulouse, France, January 2015.

"Hydrodynamical simulations of galaxy formation", 3rd Annual DAGAL Meeting, Max-Planck Institute for Astronomy, Heidelberg, March 2015.

"The feedback conundrum: What physics regulates galaxy and star formation?", 3rd Annual DAGAL Meeting, Max-Planck Institute for Astronomy, Heidelberg, March 2015.

"Exploring the physics of galaxy formation with supercomputers", Physics Colloquium, Heidelberg University, June 2015.

"Das Universum im Supercomputer", Carl Friedrich v. Siemens Stiftung, Schloss Nymphenburg, Munich, June 2015.

"Cosmic magnetism in simulations of structure formation", Magnetic Field Workshop, Schloss Ringberg, Tegernsee, July 2015; Astronomical Colloquium, Astronomisches Recheninstitut, Heidelberg University, Heidelberg, July 2015.

"Simulating cosmic structure formation in a dark universe", Astrophysics Colloquium, Albert-Einstein Institute, Potsdam-Golm, September 2015.

"Cosmic Structure Formation on a Moving Mesh", COMPUTE Colloquium, Lund University, Sweden, October 2015.

"Supercomputer-Simulationen der kosmischen Strukturbildung", Bundesweite Lehrerfortbildung, Haus der Astronomie, Heidelberg, November 2015.

"Simulating Cosmic Structure Formation", Keynote Lecture, Platform for Advanced Scientific Conference (PASC2015), ETH Zurich, Switzerland, May 2015.

Alexandros Stamatakis:

"1KITE, a worldwide HPC Initiative: Challenges and Problems", HPC Forum, LRZ, Munich, Germany, October 2015.

"Evolutionsbiologie auf dem SuperMUC" Inauguration of the SuperMUC-2, LRZ, Munich, Germany, June 2015.

"Computational Biology as Computational Science: Challenges & Problems", Hennig XXXIV meeting, New York, USA, June 2015; COS Symposium 2015 - Darwin 2.0: New Tools to Go Through Time, University of Heidelberg, Germany, June 2015; CADMOS day, EPFL, Lausanne, Switzerland, June 2015.

Michael Strube:

"The Dark Side of NLP: When Linguistics is Used to Monitor and Profile You", Digital Humaties Colloquium, Trier University, January 2015.

"The (Non-)Utility of Semantics for Coreference Resolution", Keynote Speech at the Workshop on "Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem 2015) at EMNLP 2015, Lisbon, Portugal, September 2015.

Rebecca Wade:

"Exploring protein dynamics for ligand design", Computationally Driven Drug Discovery 4th Meeting, Angelini Research Center, Pomezia, Italy, February 24 – 26, 2015.

"Modelling of Solvation Effects for Brownian Dynamics Simulation of Biomolecular Recognition", Jülich CECAM School 2015 on "Computational Trends in Solvation and Transport in Liquids", Forschungszentrum Jülich, Germany, March 23 – 27, 2015.

"Modern simulation approaches to study biomolecular interactions" EMBO Course on "Modern biophysical methods for protein-ligand interactions", University of Oulu, Oulu, Finland, June 2 – 4, 2015.

"Exploring protein dynamics for ligand design" Center for Bioinformatics, Hamburg University, Hamburg, Germany, September 8, 2015.

Anna Wienhard:

Mini-course "Anosov representations" Introductory Workshop: Dynamics on Moduli Spaces of Geometric Structures Mathematical Sciences Research Institut (MSRI), Berkeley, January 20 – 23, 2015.

"Anosov representations", Workshop Zariski-dense Subgroups, IPAM, Los Angeles, Feb 9, 2015; Proper actions and compactifications of locally homogeneous manifolds, Conference on Groups, geometry, and three-manifolds, Berkeley, May 22, 2015.

"Geometric compactifications of locally symmetric spaces of infinite co-volume", Workshop New Perspectives on the Interplay between Discrete Groups in Low-Dimensional Topology and Arithmetic Lattices, MFO Oberwolfach, June 27, 2015.

"Projective structures and higher Teichmüller spaces, Conference Classical and quantum hyperbolic geometry and topology", Orsay, July 7, 2015.

"Geometrie durch Symmetrie", Jahresversammlung der Leopoldina, Halle, September 19, 2015.

"Dozentenvortrag Geometrie", Vorkurs Mathematik, Universität Heidelberg, October 9, 2015; Discrete subgroups of Lie groups, Mathematisches Kolloquium FU Berlin, October 22, 2015.

"Special Lecture Series, "Higher Teichmüller Theory", Center for Mathematical Sciences, Technion, Haifa, November16 – 19, 2015.

Jiajie Zhang, Alexey Kozlov:

"Models and Algorithms for Phylogenetic Marker Analysis", DNA Workshop of the German Ornithologists' Society, Heidelberg, Germany, February 2015.

9.2 Presentations

Demos

Martin Golebiewski, Antonia Stank:
Workshop "COMBINE Tutorial on Modelling and Simulation Tools in Systems Biology", International Conference on Systems Biology (ICSB), Singapore, November 25, 2015.

Martin Golebiewski, Wolfgang Müller:
ERASysAPP workshop "Reproducible and Citable Data and Models", Rostock-Warnemünde, Germany, September 14 – 16, 2015.

Martin Golebiewski:
"SABIO-RK – Reaction Kinetics Database", COMBINE Tutorial "Modelling and Simulation Tools in Systems Biology", National University of Singapore (NUS), Singapore, November 25, 2015.

Kassian Kobert:
"Computing the Internode Certainty using partial gene trees", 25th Workshop on Mathematical and Statistical Aspects of Molecular Biology (MASAMB), Helsinki, Finland, April 2015.

Stefan Richter:
"Ligand Egress from Cytochrome P450 – Stereovision Demonstration", Think beyond the limits – 20 Jahre KTS, Villa-Bosch Studio, January 23, 2015.

"Dancing Molecules / Tanzende Moleküle", Explore Science (www.explore-science.info), Luisenpark Mannheim, July 8 – 12, 2015.

Posters

Stephan Hemri:
"Trends in the predictive performance of raw ensemble weather forecasts", Poster, European Geosciences Union General Assembly, Vienna (Austria), April 12 – 17, 2015.

Camilo Aponte-Santamaría:
"Force-sensitive autoinhibition of the von Willebrand factor mediated by inter-domain interactions", 1st SHENC Symposium. Hamburg, Germany, September 2015.

Neil J. Bruce, Michael Martinez, Xiaofeng Yu, Julia Romanowska, Daria B. Kokh, Musa Ozboyaci, Mehmet Ali Öztürk, Stefan Richter, Rebecca C. Wade:
"Simulation of Diffusional Association (SDA 7): Brownian Dynamics Simulations of Macromolecular Association", Jülich CECAM School: "Computational Trends in Solvation and Transport in Liquids," Forschungszentrum Jülich, Germany, March 23 – 17, 2015.

Jonathan C. Fuller, Michael Martinez, Stefan Henrich, Stefan Richter, Antonia Stank, Rebecca C. Wade:
"LigDig: a web server for querying ligand-protein interactions" International Conference on Systems Biology (ICSB), Singapore, November 22 – 23, 2015.

Wiktoria Giedroyc-Piasecka, Rebecca C. Wade, Edyta Dyguda-Kazimierowicz, W. Andrzej Sokalski:
"Protein-ligand interaction analysis as an enhancement tool for inhibitor design", Modeling Interactions in Biomolecules VII, Prague, Czech Republic, September 14 - 18, 2015.

Martin Golebiewski, Lihua An, Iryna Ilkavets, Olga Krebs, Stuart Owen, Quyen Nguyen, Natalie Stanford, Andreas Weidemann, Ulrike Wittig, Katy Wolstencroft, Jacky L. Snoep, Wolfgang Mueller and Carole Goble:
"Data Needs Structure: Data and Model Management for Distributed Systems Biology Projects", 16th International Conference on Systems Biology (ICSB), Singapore, November 23 – 26, 2015.

Ron Henkel, Dagmar Waltemath, Wolfgang Müller:
"NBI-SysBio: Tailor-made model management solutions for systems biology", de.NBI 1st SAB Meeting including Workshop and Plenary Meeting, Berlin, Germany, November 26 – 27, 2015.

Ana Herrera-Rodriguez:
"Stretching single peptides under a uniform flow", Workshop on computer simulation and theory of macromolecules. Hünfeld, Germany, April 2015.

"Self-assembly of silk proteins under a uniform flow", Conference on Polymers and Self-Assembly: From Biology to Nanomaterials. Rio de Janeiro, Brazil, October 2015.

Vincent Heuveline, Petridis Kosmas, Philipp Glaser:
"Uncertainty Quantification in Industry", 2nd GAMM AGUQ Workshop on Uncertainty Quantification, Chemnitz, Germany, September 10 – 11, 2015.

Iryna Ilkavets, Martin Golebiewski, Ivan Savora, Andreas Kebelmann, Sylvain Belahniche, Lihua An, Andreas Weidemann, Quyen Nguyen, Jill Zander, Stuart Owen, Alexander Klug, Carole Goble, Adriano Henney and Wolfgang Müller:
„Seamless data management and public outreach for a large-scale systems biology network: The Virtual Liver experience", International Conference on Systems Biology of Human Disease (SBHD 2015), Heidelberg, Germany, July 6 – 8, 2015.

Alexander Jordan:
"Test for equal predictive accuracy using proper scoring rules", Poster, 2nd Heidelberg-Mannheim Stochastics Colloquium, Heidelberg (Germany), November 26, 2015.

Philipp Kämpfer:
"Improved C. briggsae genome assembly", Worm 2015 – 20th international C.elegans Meeting, Los Angeles, June 24 – 28, 2015.

"A hybrid graph approach for short-read assembly, Sequencing, Finishing, and Analysis in the Future (SFAF)", Santa Fe, May 27 – 29, 2015.

Daria Kokh, Daria Kokh, Marta Amaral, Joerg Bomke, Matthias Dreyer, Matthias Frech, Maryse Lowinski, Alexey Rak, Rebecca C. Wade:
"Study of protein-small molecule binding kinetics using enhanced sampling molecular dynamics simulations", 2nd NovAlix Conference "Biophysics in Drug Discovery", Strasbourg, France, June 9 – 12, 2015.

Katra Kolšek:
"Molecular mechanism of C-terminal VWF-CK domain dimerization by protein disulfide isomerase", 1st International SHENC Symposium. Hamburg, Germany, September 2015.

"Modelling Cysteine Disulfide Exchange Reactions by Molecular Dynamics Simulations", Molecular and Chemical Kinetics Workshop. Berlin, Germany. September 2015.

"Towards Dynamic Disulfides in Molecular Dynamic Simulations", Workshop on computer simulation and theory of macromolecules. Hünfeld, Germany, April 2015.

"Molecular determinants governing von Willebrand factor dimerization during primary hemostasis", XVII. Linz Winter Workshop. Linz, Austria, January 2015.

Olga Krebs, Katy Wolstencroft, Natalie Stanford, Norman Morrison , Martin Golebiewski, Stuart Owen, Quyen Nguyen, Jacky Snoep, Wolfgang Mueller and Carole Goble:
"Semantic interoperability of systems biology data and models", The International Conference on Biomedical Ontology 2015, Lisbon, Portugal, July 27 – 30, 2015.

Olga Krebs, Rostyk Kuzyakiv, Katy Wolstencroft, Natalie Stanford, Martin Golebiewski, Stuart Owen, Quyen Nguyen, Finn Bacall, Norman Morrison, Jakub Straszewski, Caterina Barillari, Lars Malmstroem, Bernd Rinn, Jacky Snoep, Wolfgang Müller and Carole Goble:
"FAIRDOM approach for systems biology data and models", ERASysAPP workshop" Networking academia industry", Belval, Luxembourg, October 7 – 9, 2015.

Davide Mercadante:
"Conformational plasticity and ultra-short linear motifs as a strategy to mediate extremely fast protein association", Conference on Molecular and Cellular Kinetics, Berlin, Germany, September 2015.

Ghulam Mustafa, Prajwal P. Nandekar, Xiaofeng Yu and Rebecca C. Wade:
"Multiscale Molecular Dynamics Simulations of Different Isoforms of CYP450 and Ligand Exit and Entrance Mechanism", Free Energy Workshop (FEW2015), Münster, March 9 – 11, 2015.

Ghulam Mustafa, Prajwal P. Nandekar, Xiaofeng Yu and Rebecca C. Wade:
"Multiscale Simulations of Enzymes in a Bilayer: A Comparative Analysis of MARTINI Coarse Grained Models", Martini Coarse Graining Workshop in University of Groningen, Groningen, Netherlands, August 23 – 28, 2015.

Prajwal P. Nandekar, Abhay T. Sangamwar and Rebecca C. Wade:
"Identification of New Anticancer Compounds Through Computer-Aided Drug Designing Approaches", Bayer IT for Life Science Workshop, Leverkusen, Germany, May 12 – 13, 2015.

Mehmet Ali Öztürk, Rebecca C. Wade:
"Brownian Dynamics Simulations of Linker Histone – Nucleosome Binding", Molecular Modeling Workshop 2015, Erlangen, Germany, 9 – 11 March 2015; Computer Simulation and Theory of Macromolecules, Hünfeld, Germany, April 17 – 18, 2015.

Sebastian Ohlmann:
European Week of Astronomy and Space Science, in Tenerife, Spain, June 22 – 26, 2015.

Joanna Panecka, Ina Pöhner, Talia Zeppelin, Francesca Spyrakis, Maria Paola Costi, Rebecca C. Wade:
"Comparative analysis of targets and off-targets for the discovery of anti-trypanosomatid folate pathway inhibitors", 10th European Workshop in Drug Design (X EWDD), Certosa di Pontignano, Siena, Italy, May 17 – 22, 2015.

Olga Pivovarova, Iryna Ilkavets, Carsten Sticht, Sergei Zhuk, Veronica Murahovschi, Sonja Lukowski, Stephanie Döcke, Jennifer Kriebel, Tonia de las Heras Gala, Anna Malashicheva, Anna Kostareva, Martin Stockmann, Harald Grallert, Christian von Loeffelholz, Norbert Gretz, Steven Dooley, Andreas F.H. Pfeiffer, Natalia Rudovich:
"Modulation of Insulin Degrading Enzyme activity and liver cell proliferation", 51st EASD (European Association for the Study of Diabetes) Annual Meeting, Stockholm, Sweden, September 14 – 18, 2015.

Ina Pöhner, Joanna Panecka, Rebecca Wade:
"What determines docking performance in drug discovery? A case study of PTR1, an anti-parasitic target?", 11th German Conference on Chemoinformatics, Fulda, Germany, November 8 – 10, 2015.

Kai Polsterer, Mark Taylor:
"Virtual Observatory Virtual Reality", ADASS 2015, Sydney, Australia, October 25 – 29, 2015.

Maja Rey, Renate Kania, Dagmar Waltemath, Ulrike Wittig, Andreas Weidemann, Wolfgang Müller:
"de.NBI-SysBio: Standards-based Systems Biology Data Management", 10th CeBiTec Symposium: Bioinformatics for Biotechnology and Biomedicine, Bielefeld, Germany, March 23 – 25, 2015.

Maja Rey, Ron Henkel, Andreas Weidemann, Renate Kania, Dagmar Waltemath, Wolfgang Müller:
"NBI-SysBio: Standards-based Management of Systems Biology data, models, SOPs", de.NBI 1st SAB Meeting including Workshop and Plenary Meeting, Berlin, Germany, November 26 – 27, 2015.

Roman Schefzik:
Poster, Nonparametric Copula Day, Garching near Munich (Germany), June 12, 2015.

Poster, Workshop on Dependence and Risk Measures, Milan (Italy), November 12 – 13, 2015.

Poster, 2nd Heidelberg-Mannheim Stochastics Colloquium, Heidelberg (Germany), November 26, 2015.

Rainer Weinberger:
Poster on "Modeling AGN Radio Mode Feedback in the Moving Mesh Code Arepo", Unveiling the AGN-Galaxy Evolution Connection conference, Puerto Varas, Chile, March 9 – 13, 2015.

Ulrike Wittig, Lei Shi, Meik Bittkowski, Maja Rey, Renate Kania, Martin Golebiewski, Wolfgang Müller:
"Data Upload into SABIO-RK via SBML", Beilstein ESCEC Symposium, Rüdesheim, Germany, September 14 – 18, 2015.

Xiaofeng Yu, Ghulam Mustafa, Vlad Cojocaru, Galina Lepesheva, and Rebecca C. Wade:
"Multiscale Simulations of Cytochrome P450 Enzymes", 4th Annual CCP-BioSim Conference, Frontiers of Biomolecular Simulation, Leeds, UK, January 7 – 9, 2015.

Jolanta Zjupa:
Poster on "Angular Momentum Properties of Haloes in the Illustris Simulation", Theoretical and Observational Progress on Large-scale Structure of the Universe, ESO, Garching, July 20 – 24, 2015.

Talks

Camilo Aponte-Santamaría:
"Mechanosensitive Von Willebrand Factor Protein-Protein Interactions Regulate Hemostasis", Oral contribution, 59th Annual Meeting of the Biophysical Society, Baltimore, USA, February 2015.

Andreas Bauswein:
"Neutron star mergers", NewCompStar School: Dense Matter in Compact Stars: Experimental and Observational Signatures, Bucharest, Romania, September 2015.

Michael Bromberger:
"Exploiting approximate computing methods in FPGAs to accelerate stereo correspondence algorithms", First Workshop on Approximate Computing (WAPCO), HiPEAC Conference, Amsterdam, Netherlands, January 2015.

"Accelerating Applications using Approximate Computing", 2. Arbeitstreffen der Fachgruppe Architektur hochintegrierter Schaltungen, Darmstadt, Germany, June 8, 2015.

"Integration of Approximate Computing in today's Systems: From Hardware to Algorithmic Level", Workshop on Approximate Computing, Paderborn, Germany, October 16, 2015.

Philipp Edelmann:
"Multidimensional hydrodynamics simulations to improve our understanding of stellar evolution", Talk at conference "Nuclear Physics in Astrophysics VII", York, UK, May 18, 2015.

Kira Feldmann:
"Spatial EMOS for TIGGE temperature forecasts", Talk, Mini Symposium on Statistical Postprocessing of Ensemble Forecasts, Heidelberg (Germany), July 15, 2015.

"Spatial EMOS for ECMWF temperature forecasts" Talk, ECMWF Workshop on Calibration and Verification of Ensemble Forecasts, Reading (UK), August 19, 2015.

Philipp Gerstner:
"A Domain Decomposition Approach for Solving Optimal Economic Power Flow Problems in Parallel", EURO 2015, Glasgow, United Kingdom, July 12 – 15, 2015.

Philipp Gerstner, Vincent Heuveline:
"Towards Parallel Solvers for Optimal Power Flow Problems", 7th KoMSO Challenge Workshop, Heidelberg, Germany, October 8 – 9, 2015.

Nikolaos Gianniotis:
"Probabilistic models for structured data", Seminar at the Max Planck Institute for Astronomy, Heidelberg, June 2015.

"Autoencoding Time-Series for Visualisation", Astroinformatics 2016, Dubrovnik, Croatia, October 2015.

"Autoencoding Astronomical Time Series for Visualisation", Big Data in Astronomy Workshop at Tel Aviv Univesity, Israel, December 2015.

Martin Golebiewski:
"Data Management and Standardisation in Distributed Systems Biology Research", BioMedBridges Knowledge Exchange Workshop: Data strategies for research infrastructures, ESO Headquarters, Garching, Germany, February 19, 2015.

"Data processing and integration", ISO/TC 276 Biotechnology Plenary Meeting, Shenzhen, China, April 13 – 18, 2015.

"Standardization in Systems Biology: Building a Bridge Between Grassroots Standardization Initiatives and Standardization Bodies", HARMONY 2015: The Hackathon on Resources for Modeling in Biology, Wittenberg, Germany, April 20 – 24, 2015.

"The NormSys registry for modeling standards in systems and synthetic biology", COMBINE 2015: 6th Computational Modeling in Biology Network Meeting, Salt Lake City, Utah (USA), October 12 – 16, 2015.

"NormSys – Harmonizing standardization processes for model and data exchange in systems biology", 16th International Conference on Systems Biology (ICSB), Singapore, November 23 – 26, 2015.

"COMBINE and its Standards", COMBINE Tutorial "Modelling and Simulation Tools in Systems Biology", National University of Singapore (NUS), Singapore, November 25, 2015.

Robert Grand:
"Star formation Harvard-Heidelberg Conference on Star Formation, Harvard University", Cambridge, USA, May 2015.

"Galactic archeology conference", Potsdam, Bad-Honnef, June 2015.

"SFB Seminar on Galaxy Formation", Astronomisches Recheninstitut, Heidelberg University, Heidelberg, July 2015.

"Radial Migration in Disk Galaxies", Virgo Consortium Meeting, Durham University, Durham, UK, July 2015.

"Disk heating processes in the Auriga Simulations", Virgo Consortium Meeting, Leiden University, Netherlands, December 2015.

Stephan Hemri:
"EMOS for total cloud cover", Talk, Mini Symposium on Statistical Postprocessing of Ensemble Forecasts, Heidelberg (Germany), July 15, 2015; Talk, ECMWF Workshop on Calibration and Verification of Ensemble Forecasts, Reading (UK), August 19, 2015.

Ron Henkel:
"Creating a habitat for computational biological models", Klausurtagung Wirtschaftsinformatik Universität Rostock, Hohen Schönberg, November 18 – 20, 2015.

"NBI-SysBio: Standards-based management of Systems Biology data, models SOPs", de.NBI 1st SAB Meeting including Workshop and Plenary Meeting, Berlin, Germany, November 26 – 27, 2015.

Vincent Heuveline:
"Energy-aware mixed precision iterative refinement for linear systems on GPU-accelerated multi-node HPC clusters", 26th PARS-Workshop, Potsdam University, May 7 – 8, 2015.

"MSO in den Anwendungsfeldern: Energieforschung, BMBF Förderprogramm Mathematik für Innovationen in Industrie und Dienstleistungen", Bonn, Germany, June 23, 2015.

"Algorithms, System and Data Centre Optimisation for Energy Efficient HPC, ICT-Energy Training Day", Bristol, United Kingdom, September 14, 2015.

"Optimal Storage Operation with Model Predictive Control in the German Transmission Grid", International Symposium on Energy System Optimization (ISESO), Heidelberg, November 9 – 10, 2015.

Maximilian Hoecker:
"Clustering of Complex Data-Sets Using Fractal Similarity Measures and Uncertainties", IEEE 18th International Conference on Computational Science and Engineering (CSE), Porto, Portugal, October 21 – 23, 2015.

Svenja Jacob:
"Streaming cosmic rays in cool core clusters" Transregio Winter School, Passo del Tonale, Italy, December 6 – 11, 2015.

Samuel Jones:
Invited seminar, Konkoly Observatory, Hungarian Academy of Sciences, Budapest, Hungary, December 11, 2015.

Invited review talk, Fifty-one Ergs (FOE), NCSU, Raleigh, North Carolina, USA, June 1, 2015.

Talk 2 at JINA-CEE workshop "Galactic evolution, Nuclear Astrophysics and Stellar Hydrodynamics (GNASH)" University of Victoria, Victoria, BC, Canada, May 27, 2015.

Talk 1 at JINA-CEE workshop "Galactic evolution, Nuclear Astrophysics and Stellar Hydrodynamics (GNASH)" University of Victoria, Victoria, BC, Canada, May 25, 2015.

Alexander Jordan:
"Of quantiles and expectiles: Consistent scoring functions, Choquet representations, and forecast rankings". Talk, Karlsruhe Institute of Technology, Karlsruhe (Germany), 14 April 2015; Talk, Workshop for Young Scientists on Elicitability, Propriety and Related Topics, Bern (Switzerland), May 29, 2015.

Olga Krebs:
"Standards-compliant data management for systems biology" 7th International Practical Course in Systems Biology, Gothenburg, Sweden, June 3, 2015.

Fabian Krüger:
"Using entropic tilting to combine BVAR forecasts with external nowcasts" Talk, Jahrestagung des Vereins für Socialpolitik, Münster (Germany), 8 September 2015.

"Disagreement and forecast combination" Talk, Heidelberg University, Heidelberg (Germany), November 18, 2015.

Dennis Kügler:
"Featureless Classification of Light Curves", ITA Blackboard Colloquium, Heidelberg, June 2015.

"Featureless Classification of Light Curves", Splinter on E-Science & Virtual Observatory at the Annual Meeting of the Astronomische Gesellschaft, Kiel, September 2015.

"Featureless Classification of Light Curves", AstroInformatics 2016, Dubrovnik, Croatia, October 2015.

Sebastian Lerch:
"Forecaster's dilemma: Extreme events and forecast evaluation", PICO presentation (interactive screen presentation), European Geosciences Union General Assembly, Vienna (Austria), April 12 – 17, 2015; Talk, 35th International Symposium of Forecasters, Riverside (United States), June 21 – 24, 2015; Talk, Karlsruhe Institute of Technology, Karlsruhe (Germany), October 27, 2015.

"Probabilistic forecasting and predictive model assessment based on MCMC output", Talk, Heidelberg University, Heidelberg (Germany), April 29, 2015.

"Similarity-based semi-local estimation of EMOS models", Talk, Minisymposium on Statistical Postprocessing of Ensemble Forecasts, Heidelberg (Germany), July 15, 2015; Talk, University of Debrecen, Debrecen (Hungary), September 17, 2015.

"Probabilistic verification of extreme weather events", Talk, ECMWF Workshop on Calibration and Verification of Ensemble Forecasts, Reading (UK), August 19, 2015.

Sebastian Lerch, Fabian Krüger:
"Probabilistic forecasting and predictive model assessment based on MCMC output", Workshop for Young Scientists on Elicitability, Propriety and Related Topics, Bern (Switzerland), 29 May 2015.

Davide Mercadante:
"A new and un-conventional ultrafast binding mechanism of intrinsically disordered proteins to structured partners", 49th Biophysical Society Meeting, Baltimore, USA, February 2015.

"Rescuing the overcollapse of Intrinsically Disordered Proteins by using a force field derived by a new paradigm", CECAM workshop on Intrinsically disordered Proteins, Zurich, Switzerland, June 2015.

Wolfgang Müller:
"de.NBI-SysBio: Systems Biology Standards and Data Management", 10th CeBiTec Symposium: Bioinformatics for Biotechnology and Biomedicine, Bielefeld, Germany, March 23 – 25, 2015.

"Data management for systems biology", 7th International Practical Course in Systems Biology, Gothenburg, Sweden, June 3, 2015.

"Data Management for NMTrypI: Progress and Plans", NMTrypI Consortium Meeting, Hamburg, Germany, September 9 – 11, 2015.

"Data and model management for systems biology", de.NBI Late Summer School 2015 in Microbial Bioinformatics, Justus-Liebig-University, Gießen, Germany, September 20 – 26, 2015.

"FAIRDOM Data management for ERASysAPP projects", 2nd ERA-SysAPP Exchange Networking and Info Day, Munich, Germany, December 8 – 9, 2015.

Prajwal P. Nandekar:
"Dynathor: Dynamics of the Complex of Cytochrome P450 and Cytochrome P450 Reductase in a Phospholipid Bilayer", 18th Results and Review Workshop, High Performance Computing Center (HLRS), Stuttgart, Germany, October 5 – 6, 2015.

Sebastian Ohlmann:
"Hydrodynamics of the common envelope phase", Conference STEPS 2015 (Stellar end products) at ESO Garching, Germany, July 6 – 10, 2015; Annual meeting of the German astronomical society, Kiel, Germany, September 14 – 18, 2015.

Workshop "Combining grid-based hydrodynamics with gravitational dynamics", in Leiden, Netherlands, September 30 – October 2, 2015; Invited seminar at the Argelander Institute for Astronomy in Bonn, Germany, October 15, 2015; Invited Seminar at the Department of Physics and Astronomy at Macquarie University, Sydney, Australia, November 6, 2015.

Rüdiger Pakmor

"Type Ia Supernovae form white dwarfs mergers", Astronum 2015, Avignon, France, June 6, 2015; Carnegie SN Ia Progenitor Workshop, Pasadena, USA, August. 5, 2015.

"Magnetic fields in the GigaGalaxy disks", Virgo Consortium Meeting, Durham University, Durham, UK, July 30, 2015.

"Cosmic ray driven winds in galaxies", Virgo Consortium Meeting, Leiden, Netherlands, December 17, 2015.

Christoph Pfrommer:

"Cosmic ray physics in Arepo", Virgo Consortium Meeting at Lorentz Center, Leiden, The Netherlands, December 14 – 18, 2015.

Martin Pippel:

"De novo assembly of highly repetitive genomes", Sequencing, Finishing, and Analysis in the Future (SFAF), Santa Fe, May 27 – 29, 2015.

Kai Polsterer:

"Astroinformatics / Machine Learning in Astronomy: lessons learned from learning machines", Colloquium, AIP, Potsdam, March 2015; Colloquium, Universitätssternwarte Wien, Austria, June 2015; Colloquium, eScience Institute, University of Washington, Seattle, USA, July 2015; AstroInformatics 2016, Dubrovnik, Croatia, October 2015.

"Analyzing Complex and Structured Data via Unsupervised Learning Techniques", Focus Meeting on Statistics and Exoplanets, XXIX general assembly of IAU, Hawaii, USA, August 2015.

"Astronomy in the Cloud: using cloud infrastructure to cooperate, scale, and democratise e-science", Splinter on E-Science & Virtual Observatory at the Annual Meeting of the Astronomische Gesellschaft, Kiel, September 2015.

"E-Science and E-Infrastructure Challenges in Astronomy", Perspectives of Astrophysics in Germany 2015-2030, AIP, Potsdam, December 2015.

Stefan Richter:

"Working at HITS / Arbeiten am HITS", Think beyond the limits – 20 Jahre Klaus Tschira Stiftung, Studio Villa Bosch, Heidelberg, January 23, 2015.

Lee Rosenthal:

"The Contribution of AGN to Galaxy IR Luminosity", American Astronomical Society, Seattle, WA, USA, January 6, 2015; Haverford College, Haverford, PA, USA, May 7, 2015.

Kevin Schaal:

"High-redshift accretion shocks in Illustris", Leiden University, Virgo Consortium Meeting, Leiden, Netherlands, December. 15, 2015.

"Hydrodynamic shocks in cosmological simulatons", Harvard-Smithsonian Center for Astrophysics, Cambridge (MA), USA, September. 24, 2015.

"TENET – Discontinuous Galerkin Hydrodynamics with adaptive mesh refinement", Massachusetts Institute of Technology, Cambridge (MA), USA, August. 9, 2015.

"Finding and interpreting shocks in cosmological simulations", Durham University, Virgo Consortium Meeting, Durham, UK, July 31, 2015.

Roman Schefzik:

"Physically coherent probabilistic weather forecasts via discrete copula-based ensemble postprocessing methods", Talk, European Geosciences Union General Assembly, Vienna (Austria), April 13 – 17, 2015; Talk, Workshop for Young Scientists on Elicitability, Propriety and Related Topics, Bern (Switzerland), May 29, 2015; Talk, German Statistical Week, Hamburg (Germany), September 15 – 18, 2015.

"Proper scoring rules in probabilistic weather forecasting: Assessment of ensemble copula coupling", Talk, Workshop for Young Scientistis on Elicitability, Propriety and Related Topics, Bern (Switzerland), May 28 – 29, 2015.

"Empirical copula-based ensemble postprocessing methods", Talk, Mini-Symposium on Statistical Postprocessing of Ensemble Forecasts, Heidelberg (Germany), Heidelberg (Germany), July 15, 2015; Talk, ECMWF Workshop on Calibration and Verification of Ensemble Forecasts, Reading (UK), August 19, 2015.

Michael Schick:

"A parallel stochastic multilevel method for the solution of the incompressible Navier-Stokes equations with uncertain parameters", CSC Seminar, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany, February 2, 2015.

Michael Schick, Vincent Heuveline:

"A Polynomial Chaos Method for Uncertainty Quantification in Blood Pump Simulation", 1st International Conference on Uncertainty Quantification in Computational Sciences and Engineering (UNCECOMP), Crete Island, Greece, May 25 – 27, 2015.

Patrick Schmidt:

"Creating predictive densities with a set of point forecasts by individual rationalization", Young Statisticians Workshop, German Statistical Society (DStatG), Hamburg (Germany), September 14, 2015.

Christine Simpson:

"A localized model for star-formation and stellar feedback in highly resolved simulations of galaxies", Harvard University, Boston, MA, USA, May 20, 2015.

"Supernova-driven winds simulated on a moving mesh", University of California at Santa Cruz, Santa Cruz, CA, USA, August. 21, 2015.

"Cosmic Ray Driven Outflows in a Realistic ISM", University of Leiden, Leiden, Netherlands, December. 17, 2015.

Chen Song:

"A Polynomial Chaos Method for Uncertainty Quantification in Blood Pump Simulation", International workshop on Uncertainty Quantification in Computational Fluid Dynamics, Paris, France, May 18 – 20, 2015.

"A Towards Blood Pump Simulation Using Polynomial Chaos and the Variational Multiscale Method", 1st international conference on Quantification of Uncertainty in Engineering, Sciences and Technology (QUEST), Beijing, China, October 19 – 21, 2015.

Volker Springel:

"Exascale Simulations of the Evolution of the Universe including Magnetic Fields", SPPEXA Annual Meeting, Garching, March 2015.

"The Auriga Project", Virgo Consortium Meeting, Durham University, Durham, UK, July 2015.

Antonia Stank:

"A combinatorial puzzle: Modelling disaggregase co-chaperone complexes", Molecular Graphics & Modelling Society (MGMS) Conference on "Exploring Mechanisms in Biology: Theory and Experiment", Singapore, Nov. 25 – 27, 2015.

Andreas Weidemann:

"Reaction kinetics database SABIO-RK", COMBINE 2015: 6th Computational Modeling in Biology Network Meeting, Salt Lake City, Utah (USA), October 12 – 16, 2015.

Rainer Weinberger:

"AGN Feedback in Cosmological Simulation", Virgo Consortium Meeting, Leiden, Netherlands, December 16, 2015.

Ulrike Wittig:

"Money for biocuration: strategies, ideas & funding", 8th International Biocuration Conference, Beijing, China, April 23 – 26, 2015. "Data Upload into SABIO-RK via SBML", Beilstein ESCEC Symposium, Rüdesheim, Germany, September 14 – 18, 2015.

9.3 Memberships

Tilmann Gneiting:

Fellow, European Centre for Medium-Range Weather Forecasts (ECMWF); Reading (UK; 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; Institute of Mathemati-

cal Statistics (IMS) Representative, Committee of Presidents of Statistical Societies (COPSS) Awards Committee.

Martin Golebiewski:
German delegate at the ISO technical committee 276 Biotechnology (ISO/TC 276), International Organization for Standardization (ISO); Convenor (chair) of ISO/TC 276 Biotechnology working group 5 "Data Processing and Integration", International Organization for Standardization (ISO); Chair of the national German working group "Data Processing and Integration in Biotechnology", German Institute for Standardization (DIN); Member of the national German standardization committee ("Nationaler Arbeitsausschuss") NA 057–06–02 AA Biotechnology, German Institute for Standardization (DIN); Member of the board of coordinators of the COMBINE network (Computational Modeling in Biology network); Member of the Richtlinienausschuss (German committee for engineering standards) VDI 6320 "Datenmanagement im Bereich Life Sciences", Association of German Engineers (VDI).

Frauke Gräter:
Member of BIOMS (Heidelberg Center for Modeling and Simulation in the Biosciences) Steering Committee; Faculty member, Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg; Associated faculty member, HGS MathComp Graduate School, University of Heidelberg; Faculty member, Hartmut Hoffmann-Berling International Graduate School of Molecular and Cellular Biology (HBIGS), University of Heidelberg.

Wolfgang Müller:
Member of the Scientific Advisory Board of the BioModels Database; Member of the Richtlinienausschuss (German committee for engineering standards) VDI 6320 "Datenmanagement im Bereich Life Sciences", Association of German Engineers (VDI).

Kai Polsterer:
Member of the IEEE Task Force on Mining Complex Astronomical Data; Member of the Standing Committee on Science Priorities of the International Virtual Observatory Alliance; Member of the Knowledge Discovery in Databases Interest Group of the International Virtual Observatory Alliance.

Andreas Reuter:
Member of the Board of Trustees of Max-Planck-Institute for Astronomy; 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 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, University of Heidelberg; Member of the search committee for a professorship on "Scientific Visualization" at University of Heidelberg; Member of the scientific committee of the 6th International Conference on "High Performance Scientific Computing".

Volker Springel:
Member of the Interdisciplinary Center for Scientific Computing (IWR), Heidelberg Heidelberg; External Scientific Member of the Max-Planck-Institute for Astronomy, Heidelberg; 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"; Member of the Scientific Advisory Board of the Gauss Centre for Supercomputing (GCS); Member of the International Advisory Board of the Institute for Computational Cosmology, Durham University, UK.

Alexandros Stamatakis:
Member of the steering committee of the Munich Supercomputing System HLRB at LRZ; Member of the scientific advisory board of the Computational Biology Institute in Montpellier, France.

Michael Strube:
Editorial Board: Dialogue & discourse Journal.

Rebecca Wade:
Member of Scientific Advisory Council of the Leibniz-Institut für Molekulare Pharmakologie (FMP), Berlin-Buch; Member of Scientific Advisory Board of the Max Planck Institute of Biophysics, Frankfurt; 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 Graduate School faculty, Interdisciplinary Center for Scientific Computing (IWR), DKFZ-ZMBH Alliance of the German Cancer Research Center and the Center for Molecular Biology at Heidelberg University; Mentor, BioMedX, Heidelberg, "Selective Kinase Inhibitors" Team.

Anna Wienhard:
Editorial Board, Geometriae Dedicata; Scientific Advisory Board, Springer Lecture Notes in Mathematics; Scientific Advisory Board (Kuratorium), Internationales Wissenschaftsforum Heidelberg; Scientific Committee, Heidelberg Laureate Forum; Fellow of the American Mathematical Society; Network Executive Committee, Research Network "Geometric Structures and Representation Varieties".

9.4 Contributions to the Scientific Community

Program Committee Memberships

Jonathan Fuller, Xiaofeng Yu:
Heidelberg Unseminars in Bioinformatics (HUB), Heidelberg, Germany, 2015.

Jonathan Fuller:
GTC Bio Protein-protein interaction advisory board, 2015.

Sebastian Martschat:
4th Joint Conference on Lexical and Computational Semantics, Denver, Col., June 4 – 5, 2015.

Andreas Reuter:
Deutsche Forschungsgemeinschaft; Fonds zur Förderung der wissenschaftlichen Forschung (Österreich); Member of the Scientific Committee of the 3rd Heidelberg Laureate Forum, August 2015.

Organization Committee Memberships (Chair)

Camilo Aponte-Santamaría:
"XIV Latin American Workshop on Nonlinear Phenomena (LAWNP)" co-chair of one of the contributed sessions, Cartagena de Indias, Colombia, September 2015; "59th Annual Meeting of the Biophysical Society", co-chair of the Mechanosensation session platform, Baltimore, USA, February 2015.

Martin Golebiewski:
HARMONY 2015: The Hackathon on Resources for Modeling in Biology, chair of the "Interstandard session", Wittenberg Germany, April 20 – 24, 2015; COMBINE 2015: 6th Computational Modeling in Biology Network Meeting, chair of the session "Community & Interoperability", Salt Lake City, Utah (USA), October 12 – 16, 2015; 16th International Conference on Systems Biology (ICSB), session chair "Big Data for Systems Biomedicine – Data and modeling standards", Singapore, November 23 – 26, 2015.

Ana Peon-Nieto:
Organizer of the "What is ... ?" Seminar, University of Heidelberg.

Ana Peon-Nieto, Gye-Seon Lee, Daniele Alessandrini:
Conference Higher Teichmüller theory and Higgs bundles-Interactions and new trends, IWH Heidelberg, November 2 – 6, 2015.

Christoph Pfrommer:
Scientific Organizing Committee for the 11th International Conference on High Energy Density Laboratory Astrophysics, SLAC National Accelerator Laboratory, Stanford, USA; Scientific Organizing Committee for Feedback over 44 orders of magnitude: from gamma-rays to the Universe, Perimeter Institute for Theoretical Physics, Waterloo, Canada; Member of the Scientific Organizing Committee of the "Heidelberg Joint Astronomical Colloquium".

Andreas Reuter:
Scientific Chair of the 3rd Heidelberg Laureate Forum, August 2015.

Michael Strube:
Program Co-Chair, 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, July 26 – 31, 2015.

Rebecca Wade, Wolfgang Müller, Vincent Heuveline:
Festsymposium in honor of Andreas Reuter, Studio Villa Bosch, Heidelberg, November 27, 2015.

Anna Wienhard:
Anna Wienhard with (Virginie Charette (University of Sherbrooke), LEAD Fanny Kassel (Université de Lille I (Sciences et Techniques de Lille Flandres Artois)), Karin Melnick (University of Maryland)) Connections for Women: Dynamics on Moduli Spaces of Geometric Structures, Mathematical Sciences Research Institut (MSRI), Berkeley, January 15 – 16, 2015; Anna Wienhard with (Marc Burger (Eidgenössische TH Zürich-Hönggerberg), David Dumas (University of Illinois at Chicago), Olivier Guichard (Université de Strasbourg I (Louis Pasteur)), François Labourie (Université Nice Sophia-Antipolis)) Workshop: Dynamics on Moduli Spaces, Mathematical Sciences Research Institut (MSRI), Berkeley, April 13 – 17, 2015; Anna Wienhard with Olivier Guichard (Université de Strasbourg I (Louis Pasteur)), Workshop: Recent Advances in Surface Group Representations, IRMA, University of Strasbourg, September, 28 – October, 2, 2015; 6th Heidelberg-Karlsruhe-Strasbourg Geometry Day, Heidelberg, October 30, 2015; Co-organizer of Upstream Mentoring Network, University of Heidelberg.

Workshop Organization

Tilmann Gneiting, Sebastian Lerch, Roman Schefzik:
Organizers, Mini Symposium on Statistical Postprocessing of Ensemble Forecasts, Heidelberg (Germany), July 15, 2015.

Martin Golebiewski:
COMBINE Tutorial "Modelling and Simulation Tools in Systems Biology", National University of Singapore (NUS), Singapore, November 25, 2015; Meetings of ISO/TC 276 Biotechnology working group "Data Processing and Integration", Berlin, Germany, July 15 – 16, 2015 and Tokyo, Japan, October 26 – 30, 2015.

Frauke Gräter:
Co-organizer of a mini-symposium on "Mechanics of Silk", within the 9th European Solids Mechanics Conference, July 6th – 10th, 2015.

Renate Kania:
Workshop "Money for biocuration: strategies, ideas & funding", 8th International Biocuration Conference, Beijing, China, April 23 – 26, 2015.

Wolfgang Müller:
ERASysAPP workshop "Reproducible and Citable Data and Models", Rostock-Warnemünde, Germany, September 14 – 16, 2015.

Rüdiger Pakmor, Markus Kromer, Nancy Elias de la Rosa, Stefan Taubenberger:
Special session "Hunting down the elusive progenitors and explosion mechanisms of Type Ia supernovae", European Week of Astronomy and Space Science, La Laguna, June 25, 2015.

Kai Polsterer:
Co-organizing the splinter on E-Science & Virtual Observatory at the Annual Meeting of the Astronomische Gesellschaft 2015.

Alexandros Stamatakis:
Co-organizer of 7th Advanced Course on Computational Molecular Evolution, Hinxton UK, April 2015.

Ulrike Wittig:
Workshop "Money for biocuration: strategies, ideas & funding", 8th International Biocuration Conference, Beijing, China, April 23 – 26, 2015.

Research Program Organization

Anna Wienhard:
Program on Dynamics on Moduli Spaces of Geometric Structures, MSRI, Berkeley (together with D. Canary, W. Gold- man, F. Labourie, H. Masur), January – May 2015.

Referee Work

Camilo Aponte-Santamaría:
Biophysical Journal; Journal of Molecular Graphics and Modeling.

Nikos Gianniotis:
IEEE Transactions on Cybernetics; IJCAI Conference; Pattern Analysis and Applications.

Tilmann Gneiting:
Senior Editor, Annals of Applied Statistics.

Frauke Gräter:
Biophysical Journal; Journal of the American Chemical Society; PLoS Journals; Nature Journals; Proceedings of the National Academy of Sciences; German Research Society (DFG); PRACE; HSFP.

Davide Mercadante:
PLoS One; Food Chemistry; Process Biochemistry.

Kai Polsterer:
Annals of Applied Statistics; Astronomy and Computing.

Siegfried Schloissnig:
Biotechnology and Biological Sciences Research Council (BBSRC); Bioinformatics; BMC Evolutionary Biology.

Rebecca Wade:
Associate Editor: Journal of Molecular Recognition, PLOS Computational Biology; Section Editor: BMC Biophysics; 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 Applications · Open Access Bioinformatics.

Visiting Positions

Anna Wienhard:
Simons Visiting Professor and Eisenbud Professor, Mathematical Science Research Institute (MSRI), Berkeley, January – March 2015; Moore Distinguished Scholar, California Institute of Technology, September 2014 – April 2015.

9.5 Awards

Tilmann Gneiting:
Distinguished Achievement Medal, American Statistical Association Section on Statistics and the Environment, 2015.

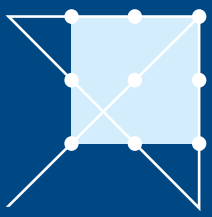
Svenja Jacob:
Otto-Haxel Award by the Department for Physics and Astronomy at Heidelberg University for the best theoretical master thesis in the winter semester 2015/16.

Davide Mercadante:
HITS award for the excellent research in the field of Intrinsically Disordered Proteins.

Prajwal Nandekar:
Member of winning and runner-up teams in Hackathon sessions on "Setting up simulation" and "Macromolecular simulation software", respectively, CECAM Extended Software Development Workshop", at Forschungszentrum Jülich, Germany, October12 – 17, 2015.

Christoph Pfrommer:
European Research Council Consolidator Grant (€ 2 million), 2015.

Alexandros Stamatakis, Alexey Kozlov, Tomas Flouri, Kassian Kobert, Andre Aberer, Mark Holder:
Teaching excellence certificate, based on student evaluation results, by the dean of the CS faculty at KIT for regular class on "Introduction to Bioinformatics for Computer Scientists", winter term 2014/15.



Edited by
HITS gGmbH
Schloss-Wolfsbrunnenweg 35
D-69118 Heidelberg

Editor
Dr. Peter Saueressig
Public Relations

Contact
Dr. Peter Saueressig
Phone: +49-6221-533 245
Fax: +49-6221-533 298
@PeterSaueressig

Our e-mail addresses have the following structure:
Firstname.lastname@h-its.org

Twitter: @HITStudies
Facebook: /HITStudies
Youtube: /TheHITSters
Instagram: /the_hitsters

Pictures
HITS gGmbH (unless otherwise indicated)

Layout and Design
FEUERWASSER | grafik . web . design
www.feuerwasser.de

ISSN 1438-4159